The Limits of Progress: Understanding the Performance of State-of-the-Art Language Models in Argument Mining

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Marc Feger

geboren in

Mönchengladbach

Düsseldorf, November 2025

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Stefan Dietze
2. Prof. Dr. Martin Mauve

Tag der mündlichen Prüfung:

Wahr ist alles an dem Tag, da es gedruckt wird
- Heinrich Heine -

Abstract

Identifying arguments is a foundational stage for many forms of manual and automated discourse analysis, including the study of political deliberation, online dialogue, and scientific reasoning. As such, the availability of rich open-access text corpora combined with powerful *Language Models* (LMs) has sparked overall interest in *Argument Mining* (AM) research.

Nevertheless, two significant shortcomings persist. Specifically, there is an incomplete account of the systemic factors that ostensibly propel progress and validate its acclaimed achievements and enduring constraints that this oversight imposes on both the reliable extraction of arguments across heterogeneous domains, particularly in the fast-paced context of social media.

To be precise, Twitter, recently rebranded as 𝕏, has become a global agora that shapes public opinion and offers a rich source of conversational data, but systematic investigation of argumentation on this platform remains scant. Second, despite steady methodological advances, most argument systems are validated only on the datasets for which they were designed, leaving their generalizability untested. This thesis tackles these issues with three interlinked studies.

At the data level, it introduces and fully documents the creation of the first conversation-based corpus for mining arguments on Twitter. Findings reveal that Twitter debates frequently involve meaningful information exchange and explicit reasoning. Furthermore, although human argument recognition is highly sensitive to conversational context, the *state-of-the-art* (SOTA) is driven by LMs that appear to attain strong performance when trained on such data. Closer inspection, however, indicates that this performance is largely superficial and does not reflect robust representational alignment of these models with the expected class semantics.

At the representation level, this work addresses the inherently entangled representations of arguments on Twitter and proposes a pre-training strategy for LMs that disentangles and refines them internally. By explicitly modeling argument structure at the component level and capturing distinctions between those components, this approach generates representations that are inherently more interpretable and more faithful to the intended semantics. The resulting model surpasses earlier baselines on Twitter and generalizes robustly across topics.

Finally, an extensive re-evaluation and cross-application analysis assesses the most task-relevant approaches and corpora, including those introduced in this thesis. Experiments reveal sharp performance drops caused by dataset-specific artifacts related to topic and content, which current SOTA LMs undesirably exploit when identifying arguments. Hence, this challenges the informative value of isolated baseline experiments and the transferability of findings in AM.

Taken together, the thesis offers both a novel resource and a refined pre-training approach for LMs that enhance the generalization of argument signals on Twitter, alongside a systematic comparative study that highlights the limitations of existing research and sets the stage for future advances in genuine argument identification.

Acknowledgments

I first want to express my sincere gratitude to Martin Mauve. Long before this dissertation began, you played a decisive role in shaping my academic journey. You welcomed me to your chair and sparked my interest in online argumentation. Your trust, along with your consistently open, supportive, and thoughtful leadership, has had a lasting impact on me.

Furthermore, my deepest thanks and highest respect go to Stefan Dietze. Your vision, expertise, honest feedback, and dedication have left a lasting mark on both this work and on me personally. You have been a formative influence long before this doctoral journey began. During my master's studies, your lectures and seminars shaped my motivation to pursue research, seek broader scientific connections, and approach knowledge with a critical mindset. From the very start of my work, you have provided continuous guidance and encouragement, for which I am truly grateful. Reading through this thesis, I am constantly reminded of how much I have learned from you and how far excellent supervision and constructive criticism can lead.

I am also sincerely thankful to Katarina Boland for her subject-matter expertise and for the enriching intellectual exchanges that ultimately led to our joint publication. Your warm, creative, and open spirit is a constant source of inspiration, and your passionate coffee breaks added both energy and insight to our discussions.

To everyone whose contributions I may not have explicitly mentioned, please accept my heartfelt thanks. Your support, whether visible or behind the scenes, was essential to the success of this work. In this sense, I am grateful to Sabine Freese and Lisa Lorenz for their reliable and efficient administrative assistance throughout my journey.

Moreover, I am particularly indebted to Jan Steimann, both professionally and personally. There are many qualities I could name, but few would fully capture the depth of my appreciation. Simply, thank you for being my true friend.

My roots and the truest source of support throughout this journey lie with my family. I thank my parents, Gabriela and Franz Josef Feger, who, despite every circumstance, made it possible for their children to learn, to strive, and to accomplish what had once been out of their reach, allowing me, as the first in our family's history, to pursue a doctorate.

Most of all, I thank my beloved wife, Aylin Feger. We met at the age of 13, have grown side by side ever since, and have always pushed each other forward. You have been by my side for more than half my life, and I can say with full conviction: you are my foundation.

Finally, I would like to answer a question often posed by my late Swiss aunt, Elisabeth Renggli:

«*Hesch den öppis gschafft?*»
Hopefully, I did!

Symbols

Although every effort has been made to keep this thesis as accessible as possible, it cannot be entirely free of mathematics. The mathematical components have been kept to a minimum, yet some notation is unavoidable. To provide clarity, a brief overview of the symbols and formulas used throughout the text is given here, and additional notation is introduced where necessary.

Learning

f, F_θ, θ : Abstract target function f, its approximation F_θ, the set of learnable parameters θ.

$C_{\hat{W}}, \hat{W}$: Classification model $C_{\hat{W}}$ with learnable parameters $\hat{W}$.

G_W, W : Feature representation model G_W with learnable parameters W.

$\mathcal{T}, t, t^{(i)}$: A text $\mathcal{T}$, $t = (t^{(1)}, \ldots, t^{(n)})$ a token sequence of a textual unit from $\mathcal{T}$.

$\mathcal{Y}, y, \hat{y}$: A set of possible labels $\mathcal{Y}$, y the true, and $\hat{y}$ the predicted label.

$h, h^{CLS}, h^{(i)}$: Summary h^{CLS} of token representations $h^{(i)}$ via $G_W(t)$, $h = h^{CLS}$ for simplicity.

$\mathcal{H}, \mathbb{R}^d, d$: Representation space $\mathcal{H}$ of G_W, a subset of the d-dimensional real vector space $\mathbb{R}^d$.

Measures

M, Md, SD : Mean, median, and standard deviation.

Δ : Difference or change between two values.

α : Krippendorff's coefficient measuring inter-annotator agreement.

ρ : Spearman's non-parametric correlation coefficient for monotonic relationships.

macro : Averaging strategy computing a metric per class and averaging equally across classes.

P, R, F1 : Precision, recall, and F1 as their harmonic mean.

$F(df_1, df_2)$: F-statistic, ratio of explained to residual variance, degrees of freedom df_1, df_2.

$t(df_1)$: t-statistic, standardized difference between means, degrees of freedom df_1.

p, p_{corr} : Probability (uncorrected or corrected) of observing a result under the null hypothesis.

d : Cohen's effect size quantifying the magnitude of a difference between two means.

η^2_G : Generalized effect size indicating the proportion of variance explained by a specific effect.

Operations

$S(n)$: The n-dimensional unit sphere on which vectors are normalized to unit length.

$p(y|t)$: Conditional probability distribution of labels y given input t.

Contents

Chapter 1

Introduction

Communication. The word *communication* comes from the Latin word *communis*, meaning *commonness*. To communicate is to create something in common with someone else, to share an idea, a thought, or a feeling, and is about getting a sender and a receiver in tune with each other (Schramm and Roberts, 1971).

What happens, however, when there is no common ground to start with? Then the task is not just to share but to build what can be shared. How can this common ground be found? Where are its limits, and how can these limits be identified and, in the best interest, overcome? This is where argumentation comes into play, for it does not assume commonness but actively works to create it (Eemeren and Grootendorst, 2003; Eemeren et al., 2014).

Arguments, in this sense, shape human reasoning, constitute a basis for inferring meaning in challenging situations, guide collective decisions, and slip almost effortlessly into everyday conversations (Cohen, 1984, 1987; Walton, 1996; Eemeren and Grootendorst, 2003; Toulmin, 2003; Mercier and Sperber, 2011; Eemeren et al., 2014), while standing as cornerstones of a deliberative democracy (Lawrence et al., 2017; Iñaki Goñi, 2025).

However, developing systems and methods that can mimic or deal with this human characteristic of applying language is a non-trivial problem in *Natural Language Processing* (NLP) (Dusmanu, Cabrio, and Villata, 2017; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Reimers and Gurevych, 2019; Slonim et al., 2021).

At the heart of this lies one of its problems, *Argument Mining* (AM), which can be described as *the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language* (Lawrence and Reed, 2019).

While numerous surveys (Daxenberger et al., 2017; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Schaefer and Stede, 2021; Vecchi et al., 2021; Ajjour et al., 2023; Feger, Boland, and Dietze, 2025; Romberg et al., 2025; Stahl et al., 2025) demonstrate that AM is an active research area applied across a wide range of genres, its particular relevance becomes evident in the context of social media. Thereby, these platforms are increasingly emerging as the primary arena of *Computer-Mediated Communication* (CMC) (Simpson, 2002), where messages are disseminated and received by many within the dynamics of mass communication (Perelman et al., 1969; Schramm and Roberts, 1971).

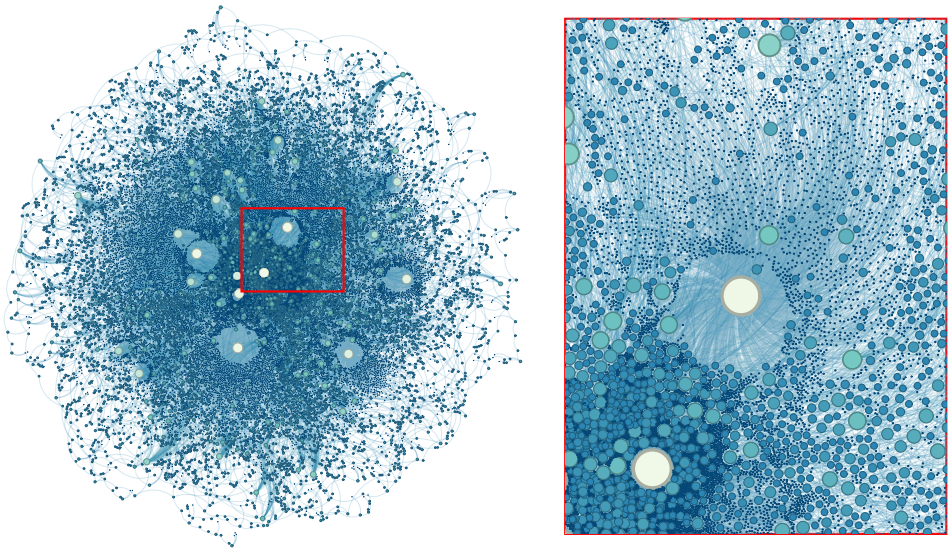


Figure 1.1: A Twitter conversation graph spanning ~30k tweets, created from three original tweets for each of six controversial topics. Node size and brightness indicate the in-degree, that is, the number of replies received by each tweet. The graph reveals conversations where individual tweets attracted up to 4k direct replies, as well as threads extending to eighty-one consecutive tweet–reply exchanges.

Considering, for example, a social media platform like Twitter (now 𝕏)[1] (Lippi and Torroni, 2016; Schaefer and Stede, 2021), even a brief examination of the platform reveals the broad diversity of tweets. As is typical of user-generated content, they can vary widely in purpose (Dusmanu, Cabrio, and Villata, 2017; Schaefer and Stede, 2021), tone (Agarwal et al., 2011), and even show inconsistency when written by the same author (Tan, Lee, and Pang, 2014).

Under this circumstance, a single tweet may serve as a provocation, a rallying cry, a joke, or nothing more than a mundane remark about lunch (Rogers, 2013). Although some tweets reflect a sincere desire to engage in ongoing debates, others arise from very different motivations, which is precisely what makes this type of user-generated data valuable for public interest research (Rogers, 2013; Schaefer and Stede, 2021; Independent Technology Research, 2023).

Once replies begin, a tweet can develop into a rapidly expanding conversation network involving thousands of participants (Nishi et al., 2016), as shown in Figure 1.1. Without any doubt, as these conversations grow and their context becomes more complex, manually tracking each tweet quickly becomes infeasible. Nevertheless, identifying argumentative excerpts within them remains crucial for meaningful debate analysis (Lopes Cardoso et al., 2023), underscoring once more the importance of automated solutions.

[1] Twitter is used hereafter to remain consistent with the publications.

Arguing, by whatever means, is not confined to Twitter. Rather, Twitter can be conceived as a microcosm that reflects certain fundamental challenges of AM research in general.

What unites all AM projects is that, alongside conceptual questions such as what constitutes an argument (Freeman, 1991, 2000; Toulmin, 2003; Walton, Reed, and Macagno, 2008; Lawrence and Reed, 2019), they also confront practical challenges related to the comparability and transferability of findings (Daxenberger et al., 2017; Geirhos et al., 2020; Thorn Jakobsen, Barrett, and Søgaard, 2021), as well as the assessment of the so-called *state-of-the-art* (SOTA), that is, identifying the field's current best performance or approach (Lippi and Torroni, 2016).

While these challenges persist, they unfold in parallel with the rapid progress of *Language Models* (LMs) (Rogers, Kovaleva, and Rumshisky, 2020; Saphra et al., 2024; Li et al., 2025; Zhao et al., 2025), whose advances in capturing language, context, and nuance shape how such issues can be addressed (Chen et al., 2024; Cabessa, Hernault, and Mushtaq, 2025).

This thesis arises from the premise that reliable and automated AM is and will become an even more important technological aid for social media, society, and NLP in general because the torrent of sources and the accelerating pace of debates will or already do exceed the fixed limits of human attention. Without reliable AM research, that struggle will only intensify.

1.1 Motivation and Problem Statement

Generally speaking, a straightforward way to understand the problem of AM is to recognize that *despite the lack of an exact definition, researchers within this field usually focus on analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and analyze textual data at hand* (Habernal and Gurevych, 2017). In this process, increasing attention has focused on automatically identifying arguments based on their constituent elements, a prerequisite for argument-centric downstream tasks (Cabrio and Villata, 2018; Lawrence and Reed, 2019; Schaefer and Stede, 2021; Vecchi et al., 2021; Ajjour et al., 2023).

In this context, the problem statement of this thesis is aimed at a bilateral gap of AM:

G1 - The Twitter Gap: Despite the relevance of social media and Twitter's role in public interest research (Rogers, 2013; Independent Technology Research, 2023), argument identification in full Twitter conversations is understudied (Schaefer and Stede, 2021).

G2 - The Generalization Gap: Although advances in argument identification are frequently reported across studies, concerns are growing over the real robustness and task alignment of current SOTA LMs in NLP and AM alike (Geirhos et al., 2020; Saphra et al., 2024). These issues and limitations are rarely addressed in AM research (Daxenberger et al., 2017; Thorn Jakobsen, Barrett, and Søgaard, 2021).

In short, without the ability of LMs to automatically and consistently recognize arguments, downstream tasks are prone to failure. Accordingly, claims of SOTA performance and findings must rely on model- and data-centric AM that generalizes robustly and is evaluated against the volatility of real-world debates, particularly those shaped by social media such as Twitter.

1.2 Research Questions

In this section, the research questions of this thesis are presented as they derive from the research gaps motivated in the previous Section 1.1. Each research question corresponds to one of the primary publications of this thesis, which are listed in the next Section 1.3. It is important to note that these questions represent a canonical progression from the previous ones and reflect an expansion of the overall research scope. Moreover, each research question is further divided into sub-questions, which individually contribute to the overarching answer.

While **Q1** and **Q2** are still closely tied to **G1** (i.e., Twitter-specific issues), **Q3** is formulated in a more general manner and addresses **G2**, namely the broader generalization gap across the literature. **Q3** thus extends the core ideas of **Q1-2** beyond the Twitter context to the more general task of argument identification in scholarly literature.

It should also be emphasized that **Q1-2** are not explicitly stated in their respective publications, whereas **Q3** is reflected in its corresponding publication, albeit in a paraphrased form. Nevertheless, each chapter concludes with a discussion of the corresponding research questions and its sub-questions, with Chapter 3 addressing **Q1**, Chapter 4 answering **Q2**, and Chapter 5 concerning **Q3**. The overarching answers are then synthesized in the concluding remarks of Chapter 6.

Q1: Can arguments be extracted within entire Twitter conversations?

- **Q1.1:** How can an annotation scheme for arguments in conversations be designed?
- **Q1.2:** What annotation quality can a resulting gold dataset achieve?
- **Q1.3:** What baseline results do SOTA models yield on this dataset?

Q2: Do SOTA models inherently predict arguments in Twitter conversations?

- **Q2.1:** Can the semantics of Twitter arguments be encoded and represented?
- **Q2.2:** How well do learned representations generalize across topics?

Q3: Does reported progress in argument identification reflect genuine advances?

- **Q3.1:** How comparable are existing AM benchmark datasets?
- **Q3.2:** Do SOTA models transfer their abilities across benchmarks?
- **Q3.3:** What do SOTA models actually learn when identifying arguments?

1.3 Publications and Contributions

The following types of publications delineate the scholarly output generated over the course of this thesis and their function within the overarching problem statement of AM.

Primary Publications constitute the backbone of the thesis. Within these, the central research questions **Q1-3** are addressed, the core methodology is articulated, and the key findings are presented. For each primary study, the specific contributions of all authors are outlined, and their scope and context in advancing the thesis argument are explained.

Secondary Publications are complementary to these and stem from the broader research contributions of this thesis or from the ideas, challenges, and solutions they present. These publications are not included in the main body of the thesis but are suggested for further reading as they provide additional context on the broader scope of the research. The main author's contributions are indicated, with each placed within the research line and explicitly linked to the relevant primary publication.

1.3.1 Primary Publications

The following peer-reviewed publications represent the core contributions of this thesis and reflect the central research questions, methodologies, and findings of this thesis:

Paper 1: Marc Feger and Stefan Dietze (May 2024a). "TACO – Twitter Arguments from COnversations". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, pp. 15522–15529. URL: https://aclanthology.org/2024.lrec-main.1349

> **Scope and Context:** For answering **Q1**, this paper presents a data-centric contribution, introducing the first dataset of *Twitter Arguments from COnversations* (TACO) that captures full conversations on six controversial topics, enabling the study of argument components within real-world social media interactions. It also highlighted both theoretical and practical challenges related to bias, usability, and the establishment of reliable ground truth data for AM. Baseline evaluation involved fine-tuned LMs, following best practices within AM literature. While achieving numerical SOTA performance, the error analysis indicated that these results stemmed from model confusion between tweet classes belonging to the same overarching category (argument vs. no-argument) and from reliance on superficial signals rather than true task alignment. These early limitations motivated efforts to improve model-sided capabilities in reliably handling and generalizing elements from different types of tweets toward their joint and pragmatic role in constituting arguments.
>
> **Contributions:** The contributions and achievements of this paper are the result of ongoing and close cooperation between both authors. Marc Feger conceived the research idea, obtained Twitter research access, and oversaw data collection and hosting.

He also developed the annotation framework and managed the recruitment, training, and compensation of annotators. Stefan Dietze contributed equally by providing domain expertise, actively shaping the planning, execution, and evaluation of the experiments and acquired data, and offering valuable feedback on editorial decisions and conceptual discussions throughout the process. In particular, it should be noted that he initiated the error analysis, which ultimately revealed the limitations that gave rise to the subsequent research questions and contributions of this thesis. Both authors were involved in every written iteration of this paper. Of special note is that the experience and supervision of Stefan Dietze proved essential in fruitfully implementing the reviewers' feedback, which ultimately led to the acceptance of this paper at the LREC-COLING conference.

Paper 2: Marc Feger and Stefan Dietze (June 2024b). "BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2256–2266. DOI: 10.18653/v1/2024.findings-naacl.146. URL: https://aclanthology.org/2024.findings-naacl.146

Scope and Context: In response to **Q2**, this paper presents a model-centric contribution by demonstrating that enhancing pre-training and incorporating theoretical insights on structural argument components into the baseline model of TACO through text augmentation and *Contrastive Learning* (CL) improves its performance at the representational level. *WRAPresentations* (WRAP), the resulting LM of this process, not only boosts performance across diverse classes of tweets but also enables effective cross-topic AM on TACO, whereas relying solely on standard cross-entropy loss fails to preserve critical argument-specific signals when applied to *out-of-distribution* (o.o.d) data. This work lays the groundwork for rethinking the foundations of AM by raising two broader questions. First, based on the applied text augmentation and cross-topic performance of WRAP, it examines whether detecting arguments on Twitter truly requires in-domain training or if models can instead be effectively trained using data from other domains. Second, it investigates whether the representational improvements proposed here generalize beyond Twitter and whether the limitations addressed by WRAP persist in other benchmarks as well.

Contributions: This work is the result of the close collaboration of both authors. Marc Feger originated the research idea and was responsible for the planning, execution, and evaluation of the experiments as well as the data analysis. Stefan Dietze contributed to these aspects and provided critical feedback on experimental and editorial decisions while also engaging in in-depth discussions on possible improvements, particularly in relation to the limitations identified in the first main contribution of this thesis and their connection to the comparative and generalization experiments. Both authors were involved in every written effort that ultimately shaped the work into its current form and led to its acceptance at the NAACL conference on the first attempt.

Paper 3: Marc Feger, Katarina Boland, and Stefan Dietze (July 2025). "Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, pp. 23900–23915. URL: https://aclanthology.org/2025.acl-long.1164/

Scope and Context: Addressing **Q3**, this work provides an evaluation-focused contribution and demonstrates that despite strong benchmark performance, SOTA LMs for argument identification tend to overfit to dataset-specific artifacts, which limits their cross-dataset generalization. Task-specific pre-training on structural signals of argument components, as introduced with WRAP, yields slightly improved representations and transferability across datasets with differing theoretical assumptions. However, overall generalization remains low. The same pattern can be observed when training on combined data sources, where both improved pre-training and vanilla LMs showed small improvements compared to single-dataset transfer experiments that involve training on one dataset and testing on another. Yet performance remained far below the SOTA on individual benchmarks. Strikingly, performance drops most for datasets that had previously been considered easy to learn. Perhaps most surprising are the manipulation experiments where only content words are retained. Under these conditions, models achieve nearly the same or even better results, which suggests that they may not capture much beyond lexical content. Altogether, these findings expose key limitations of current benchmark assumptions and emphasize the need to reconsider data, model, and evaluation practices that define the current notion of the SOTA within AM.

Contributions: This work is also the result of an utterly positive team effort by all authors. Marc Feger originated the research idea and was responsible for the planning, implementation, execution, and evaluation of the experiments as well as the data analysis. Both Katarina Boland and Stefan Dietze provided critical feedback on experimental decisions, contributed to revisions of the paper, and actively engaged in discussions regarding the central ideas. In particular, Katarina Boland's input on dataset collection, assessment, and evaluation, as well as her contributions to the experimental design concerning sufficient dataset sizes, were essential. Equally important was Stefan Dietze's critical engagement in formulating the research questions and interpreting the observed results, which directly guided the design of the manipulation experiments. He further emphasized the need to assess the empirical and practical relevance of the findings, a contribution that proved central to the overall success of this paper. All authors were involved in every written iteration of the paper. Especially Katarina Boland's and Stefan Dietze's handling of reviewer feedback, along with the additional experiments carried out in response, as well as Stefan Dietze's guidance on the structure of the paper, contributed decisively to its acceptance in the special track of the ACL conference.

1.3.2 Secondary Publications

In addition to the primary publications, the following works are related to or have emerged alongside the thesis research. These include collaborative efforts, secondary studies, and supporting investigations that complement the main line of inquiry.

Paper 4: Jan Steimann, Marc Feger, and Martin Mauve (2022). "Inspiring Heterogeneous Perspectives in News Media Comment Sections". In: *Human Interface and the Management of Information: Visual and Information Design*. Ed. by Sakae Yamamoto and Hirohiko Mori. Cham: Springer International Publishing, pp. 118–131. ISBN: 978-3-031-06424-1. DOI: https://doi.org/10.1007/978-3-031-06424-1_10

> **Scope and Context:** The development of this work was shaped by foundational insights from TACO, particularly in the creation and evaluation of data annotations. The collaboration that emerged from this context was vital, with WRAP to assess the relevance of online comments proving crucial to follow-up work (Steimann, 2025). This collaboration underscores the broader applicability and transferability of methods and findings from this thesis (Feger and Dietze, 2024a,b), highlighting their relevance and the value of AM for advancing research in related domains.
>
> **Contributions:** The idea of recommending comments based on other comments originated with Martin Mauve. Building on this foundation, Jan Steimann developed the concept into a comprehensive ecosystem of components for suggesting comments across diverse communities. He designed the recommendation model, which in follow-up publications incorporated the approach proposed in the second contribution of this thesis, implemented the two-step method for timely suggestions, conducted the evaluation, and authored the full paper. Marc Feger strengthened the work by implementing and hosting the annotation platform and by offering valuable guidance on conducting the annotations and assessing their quality, drawing on the first contribution of this thesis. The joint effort ultimately led to the paper being accepted in its current form at the HCII conference.

Paper 5: Yannick Zelle, Thibault Grison, and Marc Feger (2024). "SciTok - A Web Scraping Tool for Social Science Research". In: *HCI International 2023 – Late Breaking Posters*. Ed. by Constantine Stephanidis et al. Cham: Springer Nature Switzerland, pp. 103–109. ISBN: 978-3-031-49212-9. DOI: https://doi.org/10.1007/978-3-031-49212-9_14

> **Scope and Context:** This contribution builds on ideas from the data extraction tool used to gather TACO, which was developed for Twitter (Feger and Dietze, 2024a) and was guided by content moderation approaches specific to that platform (Grison et al., 2023), to enable data collection and content analysis for other platforms, notably TikTok.
>
> **Contributions:** Yannick Zelle initiated his bachelor thesis project in collaboration with Marc Feger, building on an idea by Thibault Grison to extract data on content moderation related to LGBT topics from TikTok. He was responsible for developing the technical solutions, conducting the experiments, and writing the bachelor thesis.

Thibault Grison and Marc Feger provided the conceptual input, shaped the research direction, and supported editorial decisions. In addition, Marc Feger contributed to the development and validation of the data extraction tool, drawing on his experience with data gathering from Twitter in the first contribution of this thesis. In addition, he carefully reviewed the work, supervised the resulting publication, and actively supported its further elaboration and development, which ultimately led to its acceptance as a conference paper on the first submission trial at the HCII conference.

Paper 6: Simon B. Weber, Marc Feger, and Michael Pilgermann (2024). "Don't Stop Believin': A Unified Evaluation Approach for LLM Honeypots". In: *IEEE Access* 12, pp. 144579–144587. DOI: 10.1109/ACCESS.2024.3472460

Scope and Context: This contribution illustrates that research into conversational dynamics, the creation of standardized annotation frameworks and evaluation methodologies as employed for TACO, the exploration of representation properties in LMs, as exemplified with WRAP, and the generalization of methods across datasets as explored in the third paper extend well beyond the boundaries of AM. The overarching aim was to identify an LMs capable of generalizing across different conversational turns (Feger and Dietze, 2024a,b; Feger, Boland, and Dietze, 2025) to distinguish between legitimate and malicious responses in an attacker and defender scenario and simulating a server under attack, framing a form of technical dialogue between the two sides (Weber, 2025).

Contributions: Simon Weber, bringing in his expertise in cybersecurity, initiated the project by proposing the evaluation of *Large Language Models* (LLMs) as honeypot backends, preparing the datasets, and generating the evaluation data. Together with Marc Feger, who contributed his expertise in NLP, he developed and tested the annotation framework. Building on insights from the first and second contributions of this thesis, Marc Feger led the annotation evaluation, analyzed conversational dynamics between honeypots and attackers, and devised the distance-based method for assessing response convincingness, drawing on prior experience with semantic similarity and LMs. Simon Weber conducted the dataset comparisons and analyzed command complexity as well as, at Marc Feger's request, the methodological generalizability in relation to the third contribution of this thesis. Simon Weber and Marc Feger co-authored the paper in equal parts, while Michael Pilgermann enriched the project through idea exchange and provided valuable feedback on the drafts, which ultimately led to its acceptance and publication in the renowned IEEE Access journal on the first submission attempt. The collaboration itself was initiated on the suggestion of Martin Mauve, whose role, though not named as co-author, was instrumental in bringing the involved researchers together.

Paper 7: Marie Braun and Marc Feger (2025). "TextLabel: A Web Application for Streamlining Text Annotation and Classification". In: *HCI International 2025 Posters*. Ed. by Constantine Stephanidis et al. Cham: Springer Nature Switzerland, pp. 243–250. ISBN: 978-3-031-94171-9. DOI: https://doi.org/10.1007/978-3-031-94171-9_21

Scope and Context: This paper addresses practical challenges encountered during the annotation process for several contributions of this thesis (Feger and Dietze, 2024a; Steimann, Feger, and Mauve, 2022; Weber, Feger, and Pilgermann, 2024) and particularly bottlenecks arising from repeated cycles of model training and evaluation required in real-world implementation scenarios (Feger and Dietze, 2024b; Feger, Boland, and Dietze, 2025). Central to this contribution is a web-based application, TextLabel, which provides a click-and-select pipeline for constructing and executing text classification workflows, including annotation, training, and evaluation, with minimal coding effort. By addressing these challenges through an integrated and user-friendly interface, the work offers improvements that benefit not only AM but also NLP projects more broadly.

Contributions: Marc Feger initiated the development of an automated tool for text annotation and model training, building on the efforts of the main contributions and the first secondary contribution of this thesis, thereby laying the foundation for the project. Under his supervision, Marie Braun significantly advanced this work through both her bachelor's and master's theses, in which she combined theoretical considerations with practical implementations for data annotation and the standardization of LM training, including their abstraction into a pipeline architecture. She further contributed substantially to the design and implementation of both the backend and frontend, along with other technical aspects related to the deployment of the tool. Her contributions strengthened the technical implementation and included conducting and evaluating the field experiments. Marc Feger guided the process, ensuring their coherence and scientific rigor, and oversaw their transformation into a research paper. The paper itself was written and revised in close collaboration between the two authors, which ultimately led to its immediate publication by the HCII conference.

1.4 Outline

This thesis follows the canonical order of its contributing publications along the corresponding research questions **Q1-3**, as these build on one another in such a way that the conclusion of each serves as the initial question for the next.

Chapter 2, at first, establishes the conceptual and methodological foundations of this thesis. It starts with a motivating introduction that uses practical examples and real-world use cases to show where AM is used. Drawing on theoretical literature, the chapter continues by describing the communicative role and structure of arguments, provides a formal task description of AM, and outlines its canonical pipeline. The focus then turns to applied research, reviewing prevailing practices, genre coverage, and computational approaches to argument identification. Collectively, these preliminaries provide a descriptive overview that contextualizes and motivates the empirical investigations presented in the subsequent chapters.

Chapter 3 is primarily data-centric and addresses **Q1** by introducing the TACO dataset, describing its construction, the modeling of argument components within coherent Twitter conversations, and the development of a corresponding annotation scheme. Furthermore, this chapter presents the final dataset and baseline results. It also identifies their limitations in preparation for **Q2**, focusing on ambiguous tweet representations and the baseline model's tendency to exploit artifacts, such as text length, that are not tied to the target class semantics.

Chapter 4 is model-centric and is concerned with **Q2** by introducing WRAP, an enhanced TACO baseline. It shows how class definitions can be explicitly turned into a learning task for enhancing the representations associated with their intended semantics. Best-practice approaches are then systematically compared, showing that text augmentation and CL are particularly effective for improving semantic representations in TACO. Beyond concrete gains in baseline performance and cross-topic generalization for AM on Twitter, the chapter also raises the broader question of how well AM baselines generalize at all.

Chapter 5, for the purpose of answering **Q3**, is evaluation-centric, surveying and re-evaluating the findings from Chapter 3 and 4, while also revisiting the literature and SOTA models and datasets presented in Chapter 2. In doing so, it systematically examines the pairwise and combined transferability of baseline methods and datasets. In contrast to the descriptive overview in the preliminaries, this survey chapter provides a more fine-grained and numerical analysis of model performance and dataset properties that defines the current literature. The analysis shows that while some improvements achieved in this thesis can be partially transferred to other research contributions, the generalizability between these remains fundamentally limited. Comparative experiments highlight that much of the reported SOTA performance is not the result of capturing the underlying class semantics of arguments but rather of exploiting dataset-specific artifacts related to topics and content.

The final Chapter 6 then summarizes the primary contributions and findings of this thesis (**Q1-3**) and outlines directions for future research in the field of AM.

Chapter 2

Preliminaries of Argument Mining

This section presents the fundamental theoretical and practical concepts of *Argument Mining* (AM), which form the basis for the discussions and approaches adopted throughout this thesis, with particular emphasis on argument identification, which constitutes the central focus of this work. In one way or another, the preliminary concepts presented here reflect the best practices available at the time this thesis and its respective contributions were developed.

The preliminaries begin with a brief overview of AM applications, highlighting the matter's relevance and diverse domains of use while providing context for the research problem that motivates the subsequent theoretical and practical explorations. Thereafter, the section delves into the key theoretical foundations that underpin AM, as well as the practical methodologies that guide its implementation. In particular, a detailed survey is provided of existing literature, including both AM frameworks and the LMs that are commonly utilized with them, offering insights into their history, areas of application, and trends.

2.1 Understanding the World Through Argument Mining

Undoubtedly, the range of applications in which arguments are formulated and employed is not only rooted in a long historical tradition but has also become nearly ubiquitous in contemporary contexts. Thereby, it is a common aspect of daily life to engage with information, develop personal interpretations, and, when appropriate, communicate these perspectives to others, which often shift quickly into the digital sphere. Thereby, as of October 2025, approximately 5.66 billion people worldwide use social media primarily to stay connected, follow news, and engage with trending topics (DataReportal, 2025). Regardless of the setting in which such communication occurs, it is an iterative process that involves the collection of information, the formation of beliefs, and the articulation of messages and feedback (Schramm and Roberts, 1971; Eveland and Cooper, 2013).

To provide an initial sense of context, Figure 2.1 illustrates several common settings in which arguments occur. These include social media platforms such as Twitter, political and legal discourse involving figures like Elon Musk in relation to his acquisition of Twitter and his connections to Donald Trump, as well as scientific writing and reviews, extending to user feedback on products and films.

Figure 2.1: A brief inspiration of possible application areas for AM.

While a comprehensive account of all AM applications and research exceeds the scope of this thesis, various notable scientific contributions should be mentioned to provide a general overview. Surveys covering 103 datasets on various argument quality criteria (Romberg et al., 2025) and 71 datasets related to computational argumentation (Stahl et al., 2025) are to be listed in this context. Likewise, 59 datasets focusing on topic ontologies, that is, the thematic coverage of datasets within AM research, are reported (Ajjour et al., 2023). Another key contribution is the third paper of this thesis (Feger, Boland, and Dietze, 2025), which reviews 52 datasets related to argument detection. Earlier studies also provide valuable insights, with 18 datasets surveying argument-centric tasks (Cabrio and Villata, 2018) and 6 datasets identifying specific argument components like claims (Daxenberger et al., 2017).

Given such comprehensive literature, it is unsurprising that research in the field of AM has yielded a diverse body of work, reflecting the wide-ranging contexts in which arguments occur. Therein, prominent areas of investigation encompass political discourse (Haddadan, Cabrio, and Villata, 2019), as well as user-generated content on social media platforms, including Twitter (Feger and Dietze, 2024a; Schaefer and Stede, 2021), Reddit (Hidey et al., 2017), and various other digital environments where online discussions take place (Walker et al., 2012; Habernal and Gurevych, 2015; Swanson, Ecker, and Walker, 2015; Stab et al., 2018; Cheng et al., 2022). Closely related to these are deliberative applications in the context of participation, including online civic engagement (Liebeck, Esau, and Conrad, 2016), e-rulemaking (Niculae, Park, and Cardie, 2017), and deliberation in political science (Falk and Lapesa, 2022).

Another important area of deliberation, in which the usage of arguments comes to mind, is e-commerce, where purchasing decisions and consumer opinions play a significant role. Such contributions include arguments found in Amazon product reviews (Chen et al., 2022), the comparison of arguments related to different brands or products (Panchenko et al., 2019), or their relevance (Wachsmuth, Stein, and Ajjour, 2017; Feger, Steimann, and Meter, 2020).

Further domains involve communication that is not directly connected to public participation platforms like social media or civic engagement forums. These include discussions of financial reports (Alhamzeh et al., 2022), presentations of essays (Stab and Gurevych, 2014, 2017), or research (Lauscher, Glavaš, and Ponzetto, 2018; Fergadis et al., 2021) including peer reviews (Hua et al., 2019; Fromm et al., 2021) and abstracts (Mayer, Cabrio, and Villata, 2020b).

In addition, arguments also appear in other contexts of knowledge dissemination, for example, on Wikipedia (Biran and Rambow, 2011; Aharoni et al., 2014; Levy et al., 2018), as well as in everyday news (Al-Khatib et al., 2016b; Hautli-Janisz et al., 2022) and law (Palau and Moens, 2009; Teruel et al., 2018; Poudyal et al., 2020; Grundler et al., 2022; Habernal et al., 2023).

At this stage, it would certainly be possible to identify numerous additional applications and studies related to AM. However, the central observation is that AM does not address a narrowly defined or isolated problem. Instead, it constitutes a broadly applicable problem that extends across various domains and applications of communication. In any context where communication and the exchange of ideas occur, AM either already serves a practical function or presents clear potential for meaningful implementation to make sense to the real or digital world.

However, to develop effective approaches for AM, it becomes crucial to consider the nature of the data involved. Specifically, one must distinguish whether arguments are already explicitly annotated or present within the data or whether they must first be scraped and automatically identified. This distinction has significant implications for the design of AM systems and the selection of appropriate techniques, as it directly influences the complexity and scope of the task.

2.1.1 Explicit Argument Data

Ideally, the data under consideration is already available in a structured format, where arguments do not need to be extracted from raw text but can instead be accessed directly. Such explicit argument data often originates from applications specifically designed to support the exchange and analysis of arguments. Examples include debating platforms such as *Dialog-Based Online Argumentation* (D-BAS) (Krauthoff et al., 2018) or *Social Linked Arguments* (SoLAr) (Feger, 2021), which guide users through a dialogue process and force the formulation of arguments via predefined input fields. Similar structures can also be found in platforms and formats developed to facilitate and organize real-world online deliberation (Kriplean et al., 2012; Schneider and Meter, 2019), as well as in web-based discussion environments like `kialo.com` (Mezza, Wobcke, and Blair, 2024), `args.me` (Ajjour et al., 2019), or `procon.org`, which allow users to systematically construct and organize pro and con arguments.

The concept of the *Argument Web* (AW) (Reed et al., 2017) builds on such platforms by seeking to integrate explicit argument data within a broader ecosystem of specialized tools, services, and systems for which the focus does not lie within the conventional boundaries of NLP, particularly in terms of LMs or automated pipelines to identify arguments. Instead, the focus lies on the structured collection of arguments explicitly indicated as such by users, supported by standardized frameworks and interfaces that guide users through debates and facilitate the articulation and storage of the resulting data, for instance, using the *Argument Interchange Format* (AIF) (Rahwan and Reed, 2009).

Consequently, from an NLP standpoint, the need for automatic argument identification is significantly reduced in these settings. This is because the platforms themselves, through a combination of user incentives and platform design, make arguments explicit and directly accessible. This supports tasks such as argument similarity (Misra, Ecker, and Walker, 2016), argument search (Boltužić and Šnajder, 2014; Ajjour et al., 2019), or the acquisition of research data (Walker et al., 2012; Al-Khatib et al., 2016a; Reed et al., 2017).

Nonetheless, it remains debatable whether such applications and data sources are widespread or merely academic idealizations of structured debate on social media, rather than reflecting platforms where users engage at scale and arguments must be distinguished from other content after a text is published.

2.1.2 Arguments within Noise

On the other side of explicitly structured argument data is a broad category of data, which may even represent the majority of cases, that exists in an unstructured form and requires the automated identification and extraction of arguments along with related methodologies to access and pre-process the data.

In such applications, arguments are not directly available, nor are the contents readily accessible to AM models. This includes, for example, spoken interactions such as transcribed debates (Haddadan, Cabrio, and Villata, 2019; Shnarch et al., 2020; Alhamzeh et al., 2022), images (Kiesel et al., 2021; Liu et al., 2023), and audio (Olshefski et al., 2020; Mestre et al., 2021). In these cases, AM efforts are closely linked to traditional data pre-processing tasks and the associated challenges, such as transcription errors and related issues.

While platforms such as Twitter (Schaefer and Stede, 2021), Reddit (Hidey et al., 2017), or news agencies like the BBC (Hautli-Janisz et al., 2022) provide access to their data, it is still necessary to, for example, extract conversational structures (Feger and Dietze, 2024a), approximate such (Bosc, Cabrio, and Villata, 2016), or develop appropriate scrapers to extract the content (Steimann, Feger, and Mauve, 2022; Zelle, Grison, and Feger, 2024).

Broadly speaking, data is most commonly unstructured and requires substantial effort to make it suitable for AM research, posing the challenge of distinguishing arguments from noisy or irrelevant content across sources that differ in complexity (Lopes Cardoso et al., 2023), but offering access to large volumes of unfiltered user-generated data.

2.1.3 Emerging Applications in Large Language Models

In this context, AM can also be considered gaining practical relevance for modern applications of LLMs such as ChatGPT (OpenAI, 2025). Models like *Large Language Model Meta AI* (LLaMA) (Grattafiori et al., 2024) or *Generative Pre-trained Transformer* (GPT) (Radford et al., 2019) not only set new standards for the use of NLP technologies but are increasingly moving into the public and scientific spotlight because of their apparent ability to engage in human-like communication. Thereby, such models do not merely process language but aim to emulate human reasoning and discourse, including the formulation, exchange, and evaluation of arguments (Bender et al., 2021; Saphra et al., 2024; Zhou et al., 2024a), and are increasingly being integrated into AM research, with applications ranging from argument identification and relation modeling (Cabessa, Hernault, and Mushtaq, 2025) to argument generation (Chen et al., 2024) and instruction tuning (Stahl et al., 2025).

A central point of current debate is whether such are genuine *Large Reasoning Models* (LRMs) capable of reasoning about factual relationships, referring to their ability to construct coherent arguments to solve problems and to communicate the underlying process effectively (Lawsen, 2025; Shojaee et al., 2025). Against this backdrop, the relevance of AM goes beyond the use of such models, as it also concerns the acquisition and structuring of high-quality data on which these data-hungry systems depend (Bender et al., 2021).

2.2 Argument Mining in Theory

In addition to exploring the potential applications of *Argument Mining* (AM), it is essential to address its foundations. While the mining aspect is reserved for later discussion, the term argument involves defining the concept and engaging with theoretical frameworks explaining what constitutes an argument. Although no exhaustive definition exists (Wagemans, 2016; Stab et al., 2018; Lopes Cardoso et al., 2023; Feger, Boland, and Dietze, 2025), the term argument can be approached from various theoretically grounded perspectives (O'Keefe, 1977; Reed and Walton, 2003; Toulmin, 2003; Lopes Cardoso et al., 2023).

Arguments as a Product: The product-oriented perspective, typically associated with a bottom-up approach, focuses on the structural properties of individual argument instances. This perspective aims to identify abstract, transferable patterns, often referred to as argument schemes or frameworks, capable of describing a wide range of argument types (Eemeren and Grootendorst, 2003; Eemeren et al., 2014) and to find those features of an argument that stay the same and independent of their use case or content (Toulmin, 2003).

Arguments as a Process: In contrast, the process-oriented perspective, similar to a top-down approach, views arguments as communicative acts embedded in discourse in which the focus shifts to their pragmatic and dialogical dimensions (Eemeren and Grootendorst, 2003; Eemeren et al., 2014), how arguments are used, negotiated, and responded to within communication (Schramm and Roberts, 1971; Toulmin, 2003).

While the distinction between arguments as products and as processes is not always clearly defined (Lopes Cardoso et al., 2023), the process-based perspective provides a valuable conceptual framework for understanding the purpose of arguments within communicative contexts. It contributes to a more profound understanding of what an argument is, along with the theoretical considerations that are relevant to its practical application. In contrast, the product-based perspective is more closely concerned with applied aspects such as structure, formalization, and identification, which will be addressed in subsequent Subsection 2.2.2 of this thesis.

2.2.1 The Communicative Role of Arguments

When considering arguments as a dynamic process, the notion of argumentation naturally comes to mind. This association is by no means inaccurate. In fact, the terms *Argument Mining* (AM) and *Argumentation Mining* (AM) are often used interchangeably to describe the computational analysis of argumentative discourse (Lawrence and Reed, 2019).

Nonetheless, it remains to be clarified which specific characteristics and purposes of communication, particularly those related to argumentation, arguments have within that process.

Communication: According to the basic Schramm model (Schramm and Roberts, 1971), communication is understood as the exchange of information, ideas, or attitudes between at least two participants. This exchange presupposes three essential components. These are a sender, a message, and a receiver, whose interaction forms the foundation of any communicative act and enables the transmission and interpretation of meaning.

Specifically, the sender encodes information or thoughts into a message using natural language, supported by verbal and non-verbal signals, all shaped by personal experience. Aiming to trigger similar thoughts and mental images in the receiver's mind, the message is then interpreted by the receiver, who decodes it based on their background and perspective. This interpretation, in turn, is translated into individual thoughts and returned to the sender as feedback, expressing the receiver's understanding of the original message. Communication is therefore not a one-way transmission but a dynamic and mutual exchange of meaning.

Communication and Problems: As introduced, the term communication originates from the Latin word communis and refers to the pursuit of commonness (Schramm and Roberts, 1971). However, this common ground can be called into question when different experiences, inaccurate information, misinterpretations, or poorly encoded messages from communicators impair mutual understanding and must first be actively established and questioned.

In this context, messages are shaped with the intention of conveying to the receiver that one's own views and assumptions are valid and of providing justification for them (Toulmin, 2003).

Ideally, this leads to the identification of common ground if the issue is tame, in the sense of having a self-evident solution upon which a consensus can be reached. However, many disputes involve abstract problems shaped by moral values, personal experiences, or political motivations. In such situations, there is no clear solution, no absolute right or wrong, but rather better or worse compromises that need to be established (Rittel and Webber, 1973).

This becomes especially apparent in CMC (Simpson, 2002), particularly in social media environments, which represent contemporary forms of classic mass communication (Schramm and Roberts, 1971) and share similar communicative challenges. These, among others, include messages being exposed to outer noise that interferes with the original message (Shannon, 1948), for example, through interpretation by many minds (Perelman et al., 1969) and their amendment in the process (Boyd, Golder, and Lotan, 2010).

Hence, it is within these problem-oriented forms of communication, which themselves are prone to errors, where attempts are made to establish commonness, where advantages and disadvantages of solutions are considered, that specific messages need to take on a particular role therein.

Argumentation: In a broader sense, one might argue that argumentation, at its most fundamental level, is the existence of a shared language and thus a technique that enables communication to take place (Perelman et al., 1969), and that it does not fundamentally differ from what is generally understood as the use of everyday language (Eemeren and Grootendorst, 2003; Eemeren et al., 2014).

More specifically, in the context of problems and communication, argumentation can be understood as a linguistic necessity, meaning a specific mode of communication concerned with what is defensible, plausible, or probable, particularly in cases where no self-evident, obvious, or exact solutions to a problem exist or can be decided (Perelman et al., 1969).

Combining these perspectives on communication, argumentation can be considered a verbal, social, and rational activity that involves the use of language directed toward others, grounded in individual assumptions, and aimed at persuading by presenting standpoints either supporting or in opposition to a particular issue (Eemeren and Grootendorst, 2003).

However, not every expression necessarily constitutes such a specific message, and the interpretation of such expressions depends on their pragmatics, meaning the inclusion of contextual information and background knowledge, which allows them to fulfill a context-specific function within the communication process (Eemeren and Grootendorst, 2003; Eemeren et al., 2014).

Against this background, an argument can be understood as a specific type of message in the communication process. It is a problem-oriented exchange that invites mutual feedback and adaptation of the message itself. This message serves a persuasive purpose by providing justification through feedback on the credibility of one's viewpoint, beliefs, and experiences, and it requires careful attention to the pragmatic-dialectical context, including relevant background information, the broader conversational setting, and the recipient's interpretation, all of which are also subject to change.

2.2.2 The Constitution of Arguments

Rather than viewing arguments as embedded within the interactive process of argumentation between at least two opponents, an alternative approach considers arguments as the outcome, the product, of this form of communication, emphasizing the identification of those components that are fundamental to their meaning, function, and representation in language (Freeman, 1991; Simosi, 2003; Katzav and Reed, 2004; Freeman, 2011; Lopes Cardoso et al., 2023).

In fact, there are commonly established theories (Perelman et al., 1969; Eemeren and Grootendorst, 2003; Reed and Walton, 2003; Toulmin, 2003; Katzav and Reed, 2004; Walton, Reed, and Macagno, 2008; Eemeren et al., 2014) that define stereotypical patterns of arguments, which lead to individualized lists of schemes that must be named on a case-by-case basis through the use of identification questions specific to each argument type therein (Wagemans, 2016).

For instance, there are schemes that encompass as many as 104 distinct stereotypes of arguments, such as the argument-from-expert-opinion, argument-from-position-to-know, argument-from-lack-of-knowledge, or the slippery-slope-argument (Walton, Reed, and Macagno, 2008). Actually, entire periodic tables of arguments have been developed, outlining how they can theoretically be described (Wagemans, 2016).

Due to the large number of different argument schemes, initial concerns have arisen regarding their practicability beyond the idealized conditions of argumentation theory, particularly about making these diverse stereotypical forms accessible and consistently usable for AM systems (Katzav and Reed, 2004; Rahwan and Reed, 2009). This problem is best exemplified by the widely used Toulmin model (Toulmin, 2003), which comprises six components that are not always present or used in real-world arguments (Freeman, 1991; Simosi, 2003; Habernal and Gurevych, 2015). Moreover, the underspecification or even contradiction of identification questions and their interpretation also merits critical consideration of such kinds of argument schemes (Wagemans, 2016).

A Standard Approach to Arguments: Despite the absence of a universally accepted nomenclature and definitional scope for arguments (Stab et al., 2018), a broadly standardized framework is employed within the field of AM (Thomas, 1986; Freeman, 1991, 2011; Lawrence and Reed, 2019), which serves as a conceptual foundation for articulating the theoretical principles underlying argument structure and function of the respective components. Therein, an argument is described as consisting of two kinds of statements where the premises provide reasons to support the main point, the conclusion, an argument is trying to establish.

In this context, deviations in terminology from the aforementioned components may arise, such as the colloquial use of a conclusion as a claim, which might be employed synonymously in the standard approach (Freeman, 1991). To maintain conceptual clarity, the terms premise and conclusion will be used consistently from this point until Subsection 2.3.2, which is devoted to explicating various frameworks that build upon the standard approach.

Accordingly, it is important to bear in mind that a variety of alternative descriptions exist, sometimes referred to as the flat or simple argument model (Aharoni et al., 2014; Stab et al., 2018; Fromm et al., 2021), all grounded in the notion of representing arguments in terms of their structural composition, or, put differently, their argument kernel (Fergadis et al., 2021).

In this sense, arguments can be understood as a canonical structure $A = \langle C, P \rangle$ with $P = \{P_1, \ldots, P_k\}$ where $k \geq 1$, C denotes the conclusion, and P the set of premises (Freeman, 1991, 2011; Wachsmuth, Stein, and Ajjour, 2017; Feger, Steimann, and Meter, 2020).

However, despite the widespread use of this representation, one has to decide whether an argument in its minimal form ($k = 0$) can consist solely of a conclusion without any explicitly stated premises (Palau and Moens, 2009; Habernal, Eckle-Kohler, and Gurevych, 2014). Enthymemes exemplify this issue (Hitchcock, 1985; Lawrence and Reed, 2019) because they rely on a premise presumed to be common knowledge or considered obvious to the audience.

In any case, the minimal condition for something to qualify as an argument is the explicit presence of the point it seeks to establish, that is, the conclusion, for without such a focal element the argument would be pointless (Freeman, 1991, 2011).

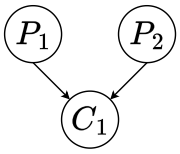
While the primary focus may lie on the formalized structure of arguments, their actual linguistic realization as textual structures may also be worth considering in this context. Thereby, textual structures and their interrelations are ideally made explicit through specific lexical cues (Cohen, 1984, 1987) like *because*, *therefore*, or *for this reason*, which indicate functional relationships between text segments by marking both the transition points and the scope of the connected elements (Mann and Thompson, 1987; Knott and Dale, 1994).

Although such cues are not always overtly realized, particularly when argument components remain implicit or when less attention is given to fair and deliberate language use (Lopes Cardoso et al., 2023), they also do not always convey unambiguous meaning, as illustrated by the word *since*, which can function as either a temporal or a causal marker (Pitler and Nenkova, 2009). Nonetheless, these cues constitute an integral part of language. Thereby, authors frequently employ them strategically to facilitate the reader's identification and interpretive processes (Cohen, 1984, 1987; Knott and Dale, 1994). For example, concluding elements can be emphasized by cue phrases such as *therefore* or *as a result*, sequences by *first* or *then*, contrasts by *however* or *but*, and justifications by *because*, *since*, or *as*, thus directly emphasizing special parts of the text and giving an indication about the role of these individual statements in an argument (Cohen, 1984; Knott and Dale, 1994).

Turning back to the structure of arguments, the canonical model of premises and conclusion can be expanded to include various stereotypical patterns and derived hybrid forms that arise from the interplay of these components (Thomas, 1986; Freeman, 1991, 2011; Lawrence and Reed, 2019; Lopes Cardoso et al., 2023).

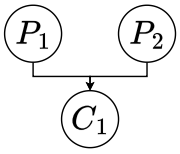
Typical structures of conclusion and premise include the following:

Convergent Argument Structures have multiple independent premises that each support the conclusion on their own.



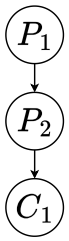
> *Example*: Because the road is icy $_{P_1}$ and the car has worn-out tires $_{P_2}$, it is reasonable to conclude that driving fast would be dangerous. $_{C_1}$

Linked Argument Structures have premises that work together to support the conclusion.



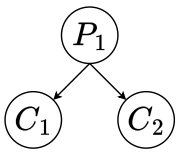
> *Example*: Because Altbier has a rich, malty flavor $_{P_1}$ and that flavor pairs perfectly with its top-fermented character $_{P_2}$, it follows that Altbier offers a uniquely balanced taste experience. $_{C_1}$

Sequential Argument Structures present premises in a logical chain, where each premise supports the next one, leading to the conclusion.



> *Example*: After the new medication was administered too late to effectively combat the infection $_{P_1}$, as a matter of fact, this delay allowed the T virus to spread unchecked $_{P_2}$, therefore humanity will turn into zombies $_{C_1}$.

Divergent Argument Structures have one premise that supports multiple independent conclusions and represent a special case in which separate arguments emerge.



> *Example*: This coffee is strong $_{P_1}$, so it is not only great for waking up in the morning $_{C_1}$ but also useful for pulling all-nighters. $_{C_2}$

2.2.3 The Argument Mining Pipeline

After outlining how arguments are constituted, AM will be defined more precisely as a structured sequence of standardized tasks, as is conventionally described in scholarly literature (Palau and Moens, 2009; Peldszus and Stede, 2013; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Vecchi et al., 2021; Cabessa, Hernault, and Mushtaq, 2025). Thus, the following pipeline outlines the objectives and challenges of the first three AM steps, focusing on defining the tasks rather than prescribing methods, while acknowledging, in line with the no-free-lunch theorem (Wolpert and Macready, 1997), that no single approach might address all of them equally well (Habernal, Eckle-Kohler, and Gurevych, 2014).

Argument Mining as a Formalized Learning Task: In fact, each step in AM can be regarded as a *Supervised Learning* (SL) task (Lippi and Torroni, 2015), in line with how such tasks are typically defined in the *Machine Learning* (ML) literature (Mitchell, 1997; Wolpert and Macready, 1997; Bishop, 2006).

To be precise, let $\mathcal{T}$ denote the set of textual units (e.g., sentences, sub-sentences, or words) extracted from a text source (e.g., a document), and let $\mathcal{Y}$ denote the set of possible labels associated with the respective step of the AM pipeline, specified later in this section.

It is assumed that there exists an (unknown) target function $f : \mathcal{T} \to \mathcal{Y}$, which maps each unit $t \in \mathcal{T}$ to its correct label $y \in \mathcal{Y}$.

Since f is inaccessible, the task of AM is to learn a parameterized function $F_\theta : \mathcal{T} \to \mathcal{Y}$, with θ denoting the learnable parameters, that serves as an approximation of f. In ML terminology, the learnable function F_θ is referred to as the model, whereas f designates the abstract target function. Throughout this thesis, the uppercase notation is used for the parameterized model, while the lowercase notation denotes the underlying, unknown function.

In practice, models often estimate conditional probabilities $p(y|t)$ over all labels in $\mathcal{Y}$, and the prediction $\hat{y}$ is taken as the label with the highest probability. Rule-based systems skip this probabilistic step and deterministically define which label might apply to a given input t. Both can thus be considered defining a conditional mapping from inputs to predictions $\hat{y}$, differing only in whether this mapping is probabilistic or determined by fixed rules.

The objective during training is to optimize F_θ by learning θ on a labeled dataset $(t_i, y_i)_{i=1}^n$ such that the predictions $\hat{y}_i = F_\theta(t_i)$ approximate the true labels $y_i = f(t_i)$. To this end, a loss function, such as the cross-entropy loss (Zhang and Sabuncu, 2018), is applied to quantify and reduce prediction error, thereby guiding the learning of θ toward alignment of F_θ with f.

This procedure is referred to as *Supervised Fine-Tuning* (SFT).

Argument Unit Detection: Building on this, the initial phase of AM focuses on identifying segments of text that are relevant to argumentation, referred to as *Argumentative Discourse Units* (ADUs), without yet assigning them specific roles such as premise or conclusion. This step operates on a predefined, non-overlapping segmentation of the input text, with each segment denoted by $t \in \mathcal{T}$. Segmentation is typically performed at the token, sub-sentence, or sentence level, depending on the desired granularity.

Although segmentation can be treated as a task in its own right, it is most often provided by the labeled dataset as a result of pre-processing during annotation. Nonetheless, when segmentation is modeled explicitly, it is commonly approached as a sequence labeling problem using token-level tagging schemes, which enable models to learn ADU boundaries directly from annotated data (Habernal and Gurevych, 2015; Habernal et al., 2023).

In one way or another, at this stage, the goal is to distinguish between segments that contribute to an argument and those that do not. The term argumentative is used here in a broad sense, encompassing any kind of argument component. Accordingly, the label space for this sub-task can be defined as $\mathcal{Y}_{ADU} = \{\text{ADU}, \neg\text{ADU}\}$.

Thereby, the objective is to detect ADUs by approximating the function $f_{AUD} : \mathcal{T} \rightarrow \mathcal{Y}_{ADU}$, which maps each text segment to its ADU status. The resulting ADUs, in turn, form the basis for the subsequent classification of argument components. Nonetheless, it can be debated over whether this step is necessary or can be replaced by directly labeling argument components, which by definition are ADUs.

For a better understanding, however, consider the following example in which the ADUs are emphasized.

Example (John F. Kennedy, 1962):

> President Pitzer, Mr. Vice President, Governor, Congressman Thomas, Senator Wiley, and Congressman Miller, Mr. Webb, Mr. Bell, scientists, distinguished guests, and ladies and gentlemen: [...]
>
> We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept [...]
>
> Thank you.

Argument Component Classification: Given the set of ADUs, denoted by $\mathcal{T}_{ADU}$, the next step is to assign a functional role within an argument to each unit. These roles are drawn from a predefined label set like $\mathcal{Y}_{Components} = \{\text{Conclusion, Premise, } \varnothing\}$ based on the definition and function of each component within the argument. Hence, this step transforms raw ADUs into argument components or excludes them as non-components, if necessary, by approximating $f_{ACC} : \mathcal{T}_{ADU} \rightarrow \mathcal{Y}_{Components}$, which maps each ADU to its corresponding component label.

In the following, the previously introduced ADUs are explicitly labeled as either conclusion or premise, reflecting their anticipated component types.

Example (John F. Kennedy, 1962):

> We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept [...]

Argument Structure Prediction: After the individual components of an argument have been classified and are represented as $\mathcal{T}_{Components}$, the final step is to determine the structural connections between them. These are drawn from labels like $\mathcal{Y}_{Structure} = \{\text{Is-Premise-Of, } \varnothing\}$, indicating the type of relation (or absence thereof) between pairs of components. Consequently, for all pairs of distinct component assignments $(t_i, t_j) \in \mathcal{T}_{Components} \times \mathcal{T}_{Components}$, with $i \neq j$, the goal is to predict the structural role that t_i has regarding t_j. This, in turn, can be modeled by $f_{ASP} : \mathcal{T}_{Components} \times \mathcal{T}_{Components} \rightarrow \mathcal{Y}_{Structure}$, which maps each ordered pair of argument components, or subset thereof, to its corresponding structural label. Based on this, the most plausible overall argument structure can be constructed, reflecting how the components interact within the argument.

To illustrate this, the identified conclusions and premises are mapped onto a hybrid structure based on the standard approach to arguments introduced earlier.

Example (John F. Kennedy, 1962):

Conclusion:

C_1 We choose to go to the moon in this decade and do the other things

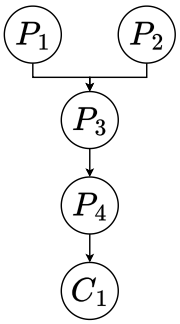
Premises:

P_1 they are (not) easy

P_2 they are hard

P_3 that goal will serve to organize and measure the best of our energies and skills

P_4 that challenge is one that we are willing to accept



2.3 Argument Mining in Practice

This section complements the theoretical perspective on AM by examining how existing studies engage with real-world data and apply theoretical frameworks in practice. It aims to provide a general overview of key research contributions, all centered on the common pursuit of identifying arguments in text. To this end, selected studies are presented and discussed, highlighting the most relevant methodological approaches and developments in the field.

The selection and presentation of related work on argument identification is primarily based on a central contribution of this thesis presented in Chapter 5, which systematically evaluated 52 relevant AM studies and benchmarks in English language (Feger, Boland, and Dietze, 2025).

Building on the aforementioned work, which compiled contributions on AM from the ACL Anthology, systematic surveys (Daxenberger et al., 2017; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Schaefer and Stede, 2021; Vecchi et al., 2021; Ajjour et al., 2023), Google Scholar, and Google Datasets, the here presented overview extends this foundation by incorporating additional characteristics not addressed but necessary for a more comprehensive overview of the field. In contrast, it is not assessed whether the datasets contain a sufficient number of instances, as this information is mainly relevant for application-oriented research questions. Apart from this difference, the same selection criteria, either directly or in an adapted form, were applied. The selected contributions were required to meet the following conditions:

Sentential: The data is available in sentence-based form or can be aggregated accordingly (e.g., from token-level annotations to sentence-level).

Binary: The annotations distinguish between arguments and no-arguments or can be simplified accordingly, for instance, by identifying a central element such as the conclusion that defines an argument.

Reproducible: The data is (largely) available and accessible as described in the paper.

Related: The datasets are available either in their original (primary) form, in a curated version, or in an updated form. Additionally, datasets are included in the overview if they are cited in other works as a reference for the theoretical framework used, regardless of whether they are directly employed or merely serve as a source of conceptual inspiration.

For the datasets identified through the selection process, additional attributes were systematically collected to enable a more comprehensive comparison across studies and to enrich the descriptive analysis. These features are not intended to provide scientific rigor through empirically standardized comparisons or methods, as such rigor is generally lacking across the reviewed studies, but rather to indicate overarching trends within the literature landscape.

Framework: The theoretical framework that defined arguments, informed the annotation guidelines, or served as a conceptual foundation (e.g., through reference to the Toulmin model) was recorded as explicitly described in the respective publication or its supplementary materials (e.g., repositories or appendix).

Best Model & F1 Score: The best-performing model was identified based on the highest baseline results for argument identification. These results were either explicitly stated or reported as part of the AM pipeline (excluding argument structure prediction) and, when not directly provided, inferred from supplementary materials such as confusion matrices. No distinction was made between F1 variants (macro, micro, weighted) or evaluation strategies (e.g., fixed splits vs. cross-validation), as no consistent practice exists across studies. When multiple test sets were available, the average performance was reported. If a mixed test set was used, the corresponding result was selected.

The full outcomes of this process are shown in Table 2.1, which is used as a guide for the sections that follow. These sections will outline quantitative trends in the data that are pertinent to surveying the field and motivating several methodological decisions underlying the main contributions of this thesis.

While a more in-depth, feature-level analysis and interpretation follows in Chapter 5, this section does not provide a full-scale numerical comparison. Though the presentation and discussion of fundamental dataset characteristics (genres, annotation frameworks, and applied models) remain essential for a comprehensive understanding of the current literature, which in turn helps identify key challenges and opportunities in AM research.

It must be acknowledged, however, that each dataset was developed under specific conditions, involving varying levels of expertise, theoretical foundations, and annotation methodologies. As each of the respective contributions shows, these differences naturally manifest across genres, definition scopes, and the linguistic and structural features of the texts involved. Nonetheless, as the following sections will demonstrate, this does not mean that these works exist in isolation. In fact, many of them inform, inspire, or are adopted by others, and at some point come to address the constitution of arguments and their identification to some extent.

Dataset	Paper	Genre	Definition	Framework	Model	Best F1
VG	Reed et al., 2008	Mixed	Claim-based	Walton, 1996		
OC	Biran and Rambow, 2011	Online Debate	Claim-based	Own	Handcrafted + HB	47.41
WTP	Biran and Rambow, 2011	Online Debate	Claim-based	Own	Handcrafted + HB	49.91
CDC	Aharoni et al., 2014	Encyclopedia	Claim-based	Stance-based (Own)		
MT	Peldszus and Stede, 2015	Microtext	Claim-based	Freeman, 1991, 2011	Handcrafted + EG	86.90
CE	Rinott et al., 2015	Encyclopedia	Claim-based	CDC		
AEC	Swanson, Ecker, and Walker, 2015	Online Debate	Implicit-Markup	Own		
WD	Habernal and Gurevych, 2015	Online Debate	Claim-based	Toulmin, 2003	Handcrafted + SVM^{hmm}	34.80
WEBIS	Al-Khatib et al., 2016a	Online Debate	Argumentative	Stance-based	Handcrafted + NB	92.20
AFS	Misra, Ecker, and Walker, 2016	Online Debate	Conclusion-based	AEC		
ASC	Wojatzki and Zesch, 2016	Twitter Debate	Claim-based	Stance-based	Handcrafted + SVM	66.00
PE	Stab and Gurevych, 2017	Academic	Claim-based	Toulmin, 2003	Handcrafted + SVM (ILP)	82.60
CMV	Hidey et al., 2017	Online Debate	Claim-based	Freeman, 2000		
CDCP$^{(*)}$	Niculae, Park, and Cardie, 2017	Online Debate	Argumentative	Own	Handcrafted + SVM	73.50
QMC	Levy et al., 2018	Encyclopedia	Argumentative	CDC	DNN$'_{suff,w2v}$	69.00
ARGUMINSCI	Lauscher, Glavaš, and Ponzetto, 2018	Academic	Claim-based	Toulmin, 2003		
UKP	Stab et al., 2018	Mixed	Evidence or Reasoning	Stance-based	dip2016 + biclstm + mtl	66.62
RCT	Mayer et al., 2018	Academic	Claim-based	PE	Handcrafted + SVM (SSTK)	78.00
USELEC	Haddadan, Cabrio, and Villata, 2019	Spoken Debate	Claim-based	CE	FastText + BiLSTM	84.30
ACQUA	Panchenko et al., 2019	Mixed	Argumentative	Own	InferSent + XGBoost	85.00
AMPERE	Hua et al., 2019	Academic	Argumentative	CDCP	BiLSTM + CRF	62.64
ASRD	Shnarch et al., 2020	Spoken Debate	Argumentative	Stance-based	BERT	57.00
ABSTRCT	Mayer, Cabrio, and Villata, 2020a	Academic	Claim-based	RCT	SciBERT + GRU + CRF	87.00
VACC	Morante et al., 2020	Online Debate	Claim-based	Stance-based		
ECHR	Poudyal et al., 2020	Legal	Conclusion-based	Own	RoBERTa	76.50
SCIARK	Fergadis et al., 2021	Academic	Claim-based	Toulmin, 2003	SciBERT + BiLSTM	74.20
AMSR	Fromm et al., 2021	Academic	Claim-based	UKP	PeerBERT	78.90
SDAT	Hansen and Hershcovich, 2022	Twitter Debate	Argumentative	CDC	RoBERTa + XGBoost	67.00
IAM	Cheng et al., 2022	Mixed	Claim-based	CDC	RoBERTa	72.36
FINARG	Alhamzeh et al., 2022	Spoken Debate	Claim-based	Cohen, 1987	DistilBERT	80.00
TACO	Feger and Dietze, 2024a	Twitter Debate	Inference-Information	Own	BERTweet	85.06

Table 2.1: Overview of the 31 datasets that meet the sentential, binary label, and reproducibility criteria, including associated theoretical frameworks, best-performing models, and reported F1 scores. (*) The binary criterion was not met, but the dataset served as a framework for another study.

2.3.1 Genre Coverage in Contemporary Research

While various domains in which AM is applied have already been touched upon in Section 2.1, the following sections provide a description of the collected datasets, presented in Table 2.1, in terms of their genres. These refer to the origin of a dataset and characterize the nature of the platforms from which the data were collected. It thus reflects the primary source characteristics and highlights the focus placed on specific data sources and application domains within the research. A more detailed analysis is provided in the contribution presented in Chapter 5.

Online Debate: Unsurprisingly, online debates, including those conducted on Twitter, constitute a significant portion of the collected studies. In fact, a total of 12 works (~39%) focus on this genre. On the one hand, there are platforms specifically designed to facilitate formal debates. On the other hand, discourse also arises on general comment-based platforms, where interaction develops more organically and is not shaped by predefined constraints of online argumentation systems.

The first group includes the *Argument Extraction Corpus* (AEC) (Swanson, Ecker, and Walker, 2015), derived from `4forums.com` and `createdebate.com`, and the *Web Discourse* (WD) dataset (Habernal and Gurevych, 2015), which contains data also collected from `createdebate.com` and extended by data from `debate.org`. In addition, the *Web Information Systems* (WEBIS) dataset (Al-Khatib et al., 2016a), named after the research group at Bauhaus-Universität Weimar, was compiled from `idebate.org`. The *Argument Facet Similarity* (AFS) dataset (Misra, Ecker, and Walker, 2016), on the other side, used the same source and also included material from `procon.org`, where facets are typically paraphrased propositions.

Comment-based platforms where argumentation takes place, though not as explicitly structured or guided as in dedicated argumentation systems, are represented in several datasets as well. These include the *Online Comments* (OC) corpus from `livejournal.com` and posts from the *Wikipedia Talk Pages* (WTP) dataset (Biran and Rambow, 2011), respectively. The *Consumer Debt Collection Practices* (CDCP) dataset (Niculae, Park, and Cardie, 2017), also referenced as the *Cornell eRulemaking Corpus* (CDCP), consists of user comments on rule proposals from `regulation.org`, while *Change My View* (CMV) (Hidey et al., 2017) contains discussions from the thread of the same name on `reddit.com`. To this end, the *Vaccination Corpus* (VACC) (Morante et al., 2020) draws on archived web content from `archive.org` related to vaccination.

Twitter also belongs to this category, representing a hybrid between argumentation and comment platforms. Within the context of this category, one might argue that Twitter functions as a general CMC platform that incorporates elements of online argumentation yet lacks predefined input structures and, unlike dedicated argumentation systems, is socially mainstream and attracts a broad and diverse audience. In this respect, such debates can range from scientific (Hafid et al., 2022), political (Marchetti-Bowick and Chambers, 2012), social (Bhatti, Ahmad, and Park, 2021), and sustainability (Hansen and Hershcovich, 2022) to mixed topics (Feger and Dietze, 2024a; Fafalios et al., 2018), while supporting multimodal forms of expression that go beyond text and also include arguments in images (Liu et al., 2022).

However, studies focusing on twitter.com are relatively scarce within the subset of contributions that meet the quality criteria applied in this analysis. In fact, among available resources, only the *Argument Stance Classification* (ASC) corpus (Wojatzki and Zesch, 2016), sometimes called *Tweet Stance Classification* (TSC) (Schaefer and Stede, 2019), the *Sustainable Diet Arguments on Twitter* (SDAT) corpus (Hansen and Hershcovich, 2022), and *Twitter Arguments from COnversations* (TACO) (Feger and Dietze, 2024a), a contribution of this thesis, provide data for mining arguments on this widely used yet underrepresented platform in the context of AM (Schaefer and Stede, 2021). These three datasets account for only one out of four addressing online debate and just one in ten of the total considered in this survey.

In practice, a major challenge in working with Twitter content concerns the criterion of reproducibility (Independent Technology Research, 2023), as datasets are often removed to comply with Twitter's data retention policies (Bosc, Cabrio, and Villata, 2016; Dusmanu, Cabrio, and Villata, 2017). In many other cases, datasets are not publicly released or referenced, which further complicates their reuse and validation (Procter, Vis, and Voss, 2013; Llewellyn et al., 2014; Addawood and Bashir, 2016; Addawood, Schneider, and Bashir, 2017; Bhatti, Ahmad, and Park, 2021). Moreover, such datasets relied on Twitter's 1% *Application Programming Interface* (API), raising concerns about representativity (Morstatter, Pfeffer, and Liu, 2014; Morstatter et al., 2021), which has since been discontinued and replaced by an advanced but costly API that hampers public-interest research (Independent Technology Research, 2023).

Academic: In addition, several sources were included that are situated within academic, scientific, or educational contexts. This category accounts for seven contributions (~22%) in total. A notable trend is the reliance on scientific publication platforms. For example, the *Randomized Clinical Trials* (RCT) and *Abstracts of Randomized Clinical Trials* (ABSTRCT) datasets (Mayer et al., 2018; Mayer, Cabrio, and Villata, 2020a) both focus on abstracts collected from pubmed.gov. Within the same context of scientific contributions, openreview.net, a platform for peer review in areas like *Artificial Intelligence* (AI), served as the source for the *Argument Mining for Peer Reviews* (AMPERE) (Hua et al., 2019) and the *Argument Mining in Scientific Reviews* (AMSR) (Fromm et al., 2021) dataset, respectively.

Similarly, the Dr. Inventor project (Fisas, Ronzano, and Saggion, 2016), which focuses on publications containing scientific discourse in the field of computer graphics, was incorporated into the *Argumentative Analysis of Scientific Publications* (ARGUMINSCI) corpus (Lauscher, Glavaš, and Eckert, 2018), named after the corresponding system but also known as *Scientific Arguments* (SCIARG) (Lauscher, Glavaš, and Ponzetto, 2018). In turn, undocs.org, which provides official United Nations documents and sustainability reports, was included as a source in building the *Scientific Argumentation Knowledge* (SCIARK) corpus (Fergadis et al., 2021). This corpus examined how sustainability goals are reflected in scientific abstracts collected from databases such as pubmed.gov and semanticscholar.org. The study further explored the generalizability of SCIARK in relation to ABSTRCT.

Moreover, in the context of educational communication, essayforum.com, a platform where users seek support in writing persuasive essays, was addressed in the *Persuasive Essays* (PE) dataset (Stab and Gurevych, 2017).

Mixed: Beyond clearly delineated genres, there are four datasets (~13%) that also blend content from a range of heterogeneous sources. For example, this includes a dataset on *Various Genres* (VG) (Reed et al., 2008) from AraucariaDB (Reed and Rowe, 2004), a database of arguments. Complementing this, diverse web-based materials from `archive.org` and `commoncrawl.org` were also incorporated into the *Ubiquitous Knowledge Processing* (UKP) corpus (Stab et al., 2018), developed by the lab of the same name at the Technical University of Darmstadt. Similarly, the *Argumentation in Comparative Question Answering* (ACQUA) corpus (Panchenko et al., 2019) stems from a project with the identical designation, funded by the Deutsche Forschungsgemeinschaft, and is also referred to as the *Comparative Sentences* (COMPSENT) corpus. In addition, user-generated content from various online forums was taken into account with the *Integrated Argument Mining* (IAM) dataset (Cheng et al., 2022).

Encyclopedia: Three sources (~10%) in the selection focus exclusively on the knowledge platform `wikipedia.org`, addressing *Context Dependent Claim* (CDC) (Aharoni et al., 2014), their supporting evidence in the *Context Dependent Evidence* (CE) corpus (Rinott et al., 2015), and retrieving them through *Querying of Main Concepts* (QMC) (Levy et al., 2018).

Spoken Debate: Moreover, another three sources (~10%) were included that focus on specific thematic domains, particularly in the areas of politics, public discourse, and economics, where debates are typically conducted in spoken form. This includes data from `debates.org`, which provides official transcripts of United States presidential debates from 1960 to 2016, forming part of the *U.S. Election Debate* (USELEC) dataset (Haddadan, Cabrio, and Villata, 2019), also referred to as *USElecDeb60To16* (USELEC). Furthermore, the *Automatic Speech Recognition of Debates* (ASRD) corpus (Shnarch et al., 2020) spoken debate is also transcribed. Within the economic domain, materials from `financialmodelingprep.com` were used for the *Finanical Arguments* (FINARG) corpus (Alhamzeh et al., 2022), with transcribed financial calls of Amazon, Apple, Microsoft, and Facebook.

Legal: A corpus containing legal documents was also included in the selection (~3%). Specifically, judgments and decisions, as made available through `hudoc.echr.coe.int`, were made available in the *European Court of Human Rights* (ECHR) corpus (Poudyal et al., 2020). It is important to note, however, that this source represents only one among several that address law (Teruel et al., 2018; Grundler et al., 2022; Habernal et al., 2023), though it is excluded based on the selection criteria mentioned earlier.

Microtext: Lastly, the selection includes a distinctive dataset (~3%) consisting of the *Microtext* (MT) dataset (Peldszus and Stede, 2015). Such texts are characterized by their concise form and manual composition, and one can argue at this point whether this renders them comparable to comments or tweets.

2.3.2 Conceptual Frameworks for Defining Arguments

In the following, the underlying theoretical frameworks concerning the conceptual structure of arguments are examined. Particular emphasis is placed on the central role of the so-called standard approach to arguments (Thomas, 1986; Freeman, 1991, 2011; Lawrence and Reed, 2019), as in Subsection 2.2.2, and how it is addressed in practical AM research.

For this sake, the primary labels of the underlying datasets are presented first. This is followed by an outline of the theoretical frameworks employed in generating these labels, with particular attention paid to their relation to the standard approach in argument construction.

Claim(Conclusion)-based: This category refers to definitions that identify either a claim or a conclusion as the minimal constituent of an argument. While both terms denote the minimal element of an argument, speaking of a claim emphasizes the language of the arguing process, whereas speaking of a conclusion emphasizes the argument as the end product of that process (Freeman, 1991).

Building on this, referring to a statement as a claim within argumentation theory situates it within the process of argumentation. It is a proposition advanced for the acceptance of others, carrying the obligation to provide backing if challenged, and one for which reasons are or can be provided, whether those have yet been examined (Freeman, 1991; Toulmin, 2003). By contrast, calling it a conclusion, a particular type of claim, places it within the product of argumentation as the point reached in that process for which reasons have been presented and deliberated (Freeman, 1991).

Datasets belonging to this category are characterized primarily by their reliance on claims. In numerical terms, the majority of the 31 selected datasets, 19 in total (~61%), are claim-based, whereas only two (~6%) are conclusion-based.

The conclusion-based datasets are AFS, which focuses on the facets, which refer to the essential aspects of arguments, and ECHR, which draws on conclusions due to the nature of legal texts. Such legal texts, as reflected in their pleadings, contain judgments, in other words, decisions that have been reached and evaluated through the argumentative process, and thus pertain more to conclusions than to claims.

In particular, debates in online forums, as in ASC and CMV, or in spoken form, as in USELEC, are found in two out of five claim-based datasets. A similar pattern can be observed in two of the six cases in which a claim-based dataset belongs to the academic genre, with examples including PE, RCT, and AMPERE. This breadth reflects, on the one hand, the versatility of claims and, on the other, their open nature, since claims need not be conclusively substantiated in the way that conclusions must be.

It should be noted, however, that the respective definitions are not necessarily uniform, where the same holds true for conclusions or other definitional approaches that may be employed. As will be elaborated later in this section, there are naturally different perspectives on which framework to adopt, to what extent such frameworks should be applied, and what elements a framework for defining arguments and their components ought to include.

For instance, in the context of academic datasets such as ABSTRCT, claims are described using the term conclusion to characterize the summarizing and persuasive function of scientific abstracts. By contrast, in ARGUMINSCI and SCIARK, they are presented, as explained in the next section, in an argumentatively synthesized form. While an in-depth analysis of the motivations behind this remains open to debate and likely reflects individual design choices, it would also encompass questions about the differences, similarities, and even contradictions that arise when such frameworks are applied mutually. Further details relevant to this point are provided in the contribution of this thesis, presented in Chapter 5.

Argumentative: This term serves as a generic designation, defining sentences as argumentative insofar as they satisfy the minimal requirements of an argument, such as the presence of a conclusion or another minimally defined component, without further specifying the precise role of each component. In this sense, argumentative texts are argument products, and those displaying these structures are to be regarded as presenting single arguments (Freeman, 1991). The term thus serves as a coarse-grained category, independent of detailed argument structure, and does not address whether a statement is more appropriately labeled as a claim or a conclusion.

In total, seven of the selected datasets (~23%) fall under this type of definition. Here, too, except for legal texts and microtexts, this category is represented across all genres. Notably, it is particularly prevalent in more dynamic environments such as online platforms, as seen in WEBIS, CDCP, and SDAT, or in spoken contexts such as ASRD, where this softening of the argument concept is applied.

Others: Lastly, this category functions as a catch-all for definitions that do not fit the previously specified types, bringing together conceptually diverse, mostly custom-made approaches. Thereby, this category comprises three (~10%) of the selected contributions.

At the outset, the AEC dataset should be noted, as it introduces the so-called markup hypothesis. This hypothesis is grounded in five lexical cues (*so*, *if*, *but*, *first*, *I agree that*) that signal arguments or the relations between their components, such as those between a conclusion and a premise. This approach, in turn, can be compared to the idea of discourse markers for signaling rhetorical or structural relations (Cohen, 1984, 1987; Mann and Thompson, 1987; Knott and Dale, 1994; Pitler and Nenkova, 2009), as discussed in Subsection 2.2.2.

Furthermore, in the UKP dataset, evidence or reasoning are considered as components of arguments. In this regard, it should be noted that these two components are primarily defined by the premise, that is, the element that provides the basis for a claim when required. Thereby, in the UKP dataset, a claim is understood as a stance in favor of or against a given topic, while evidence or reasoning, without further formal definition, are treated as the elements that support or refute the claim. In combination, they are intended to be interpreted as supporting or opposing arguments regarding a specific topic addressed. While the underlying labels of evidence or reasoning represent a distinctive feature of the UKP dataset, the specific way of defining arguments as favoring or attacking a given topic is elaborated later in this section in the context of stance-based frameworks.

Furthermore, the TACO dataset, which constitutes one of the main contributions of this thesis, also falls into this category. Anticipating the more detailed description of the definitions underlying the constitutive elements of inference and information in Chapter 3, it should be noted that these elements are guided by the standard definition of AM, which focuses on the automatic extraction of arguments based on the structures of inference and reasoning (Lawrence and Reed, 2019). Here, inference is defined broadly to include claims, conclusions, and other minimal argumentative elements, while information encompasses premises, evidence, reasoning, and contextual details, all of which can inform the inference component of an argument. On the one hand, these components align with argumentative definitions through their intentional inclusiveness, accommodating the variety of motivations behind tweets and avoiding the narrow focus found in, for example, conclusion-based definitions. On the other hand, they also parallel claim(conclusion)–based definitions, as they likewise divide arguments into distinct components to achieve greater granularity.

Turning back to the respective frameworks that underlie the definitions of the labels, they should first be grouped to provide a clearer understanding of their nature. In particular, attention shall be directed to the interrelations between the respective frameworks, the ways in which they serve as mutual sources of inspiration, and their connection to the standard approach to arguments outlined in Subsection 2.2.2. This discussion is presented solely in descriptive form, while experiments concerning transferability can be found in the respective contribution in Chapter 5.

Literature-based: To begin with, ten frameworks (~32%) draw on the traditional literature of argumentation theory, as discussed in Section 2.2, and integrate the conceptual foundations into their respective definitions, thereby grounding their approaches in established principles.

Foremost among these is Toulmin's model of arguments (Toulmin, 2003). Various sources referencing Toulmin serve as direct points of reference for the design of their labels in six of the selected datasets. Interestingly, only WD and ARGUMINSCI annotate all six components of the Toulmin model (claim, data, qualifier, rebuttal, warrant, backing) within a test iteration, yet both note that many of these components are often implicit or, as in the case of ARGUMINSCI, virtually absent. In ARGUMINSCI, it was observed that only the claim and data components were readily identifiable, whereas the remaining components were either difficult to detect or entirely missing. This observation is consistent with the broader criticism of the Toulmin model and its practical applicability (Freeman, 1991, 2000, 2011; Lopes Cardoso et al., 2023). Similarly, SCIARK, citing the complexity of the Toulmin model, adopts a simplified version consisting solely of claim and evidence. Furthermore, for reasons of practicality, the procedure employed in PE, which serves as the methodological basis for RCT and, in turn, as the methodological foundation for ABSTRCT, reduces the Toulmin model to only two components, namely claim and premise.

In this context, the Freeman model (Freeman, 1991, 2000, 2011), widely regarded as the standard approach to arguments (Lawrence and Reed, 2019; Lopes Cardoso et al., 2023), holds that an argument consists of only two components, the conclusion and the premises, whereby the conclusion is, in Toulmin's terms, an ordinary claim.

In this approach, there is no dispute over the equivalence of claims and conclusions, as an argument cannot exist without an attempt to establish at least one point (Freeman, 1991). By contrast, Toulmin's additional elements, like data, warrants, and backing, are more controversial, while in the standard approach they may all be subsumed under the category of premises (Freeman, 1991). This simplification of arguments originates in a critique of the Toulmin model and its practical applicability, a critique that is supported by the fact that even those who explicitly refer to Toulmin often reduce their frameworks to two components for the sake of simplicity.

With this background, it is striking that this model is referenced as a foundational framework in only two datasets, namely MT and CMV. Thereby, for CMV, it is noted that its framework is nonetheless comparable to that of PE, as both rely on the same two types of components. In the case of CMV, however, this decision is justified directly with reference to Freeman's standard approach and not indirectly via problems observed in the actual application of Toulmin's argument model.

Furthermore, the Walton model for arguments (Walton, 1996; Walton, Reed, and Macagno, 2008), which is also based on two main components and aims to reach a conclusion, albeit conceptualized primarily from a dialectical perspective between two opponents, has only been applied in the context of VG. It should be noted that although the Walton model is also conclusion-based, it is characterized in particular by a topological classification of clearly defined argument types and by associated guiding questions for understanding these arguments. The central difference from Freeman's standard approach is that Walton divides structures and forms of everyday arguments into concrete categories such as an argument-from-example, slippery-slope-argument, or argument-from-arbitrariness-of-a-verbal-classification, while Freeman primarily describes them based on their components and basic structures, as shown in Subsection 2.2.2. Although both approaches are comparable in terms of the constitution of arguments, they differ both in the level of detail and in the way in which the respective components are specified and further grouped. Whereas Freeman remains at a rather superficial level, Walton captures differentiated contexts between the components and names them.

Lastly, the Cohen model, which is similar to Freeman's model in its basic structure but focuses more on computational aspects for the actual identification of arguments, is mentioned in the literature primarily in connection with FINARG. Yet, a characteristic feature of the Cohen model is not only that the terms claim and conclusion are used synonymously, but also that particular emphasis is placed on so-called cues that indicate relationships between the individual argument components. Thereby, claims or conclusions function as central nodes to which supporting evidence is arranged in hierarchical form. Similar to the use of discourse markers, these structures are recognized on a rule-based basis by interpreting certain linguistic patterns or cues as indications of arguments.

Stance-based: The selection also includes frameworks that, while not directly rooted in classical argumentation theory, build on Toulmin or, following Freeman, adopt a two-component model. These are extended by incorporating a stance, a positive or negative attitude, toward a given topic and classifying arguments or their components in relation to that stance.

On the whole, 12 of the selected datasets (~39%) use stance-based frameworks. Of these, two are particularly noteworthy, as they are frequently cited and are characteristic of the other frameworks used.

In the selection in Table 2.1, this type of framework appears for the first time in the CDC dataset and was later cited by CE, which serves as a reference for USELEC, QMC, SDAT, and IAM as the basis for the definitions used there. Although it is reasonable to assume that CDC had a significant influence on this framework variant, it should be noted that this work was published as part of the IBM Project Debater, a project in which automated argumentation was gaining considerable popularity at the time (Slonim et al., 2021).

In the CDC framework, topics are defined as a short, usually controversial statement that defines the subject of interest, while a context-dependent claim, from which the name of the data set derives, is described as a general, concise statement that directly supports or disputes the topic. Context-dependent evidence, in turn, refers to a section of text that directly supports a CDC in the context of the given topic. Here, the term context is understood in the narrower sense of indicating the thematic relevance of the respective argument components. However, in terms of component selection, the CDC framework deliberately distances itself from Toulmin's model but relies on a claim-based approach that additionally integrates a topic-dependent contextualization of the arguments.

A comparable approach can be found in the UKP dataset framework, which in turn was used by AMSR as a basis. Here, a stance-based method is likewise used for reasons of simplification. Unlike the CDC approach, however, the topics are defined as some matter of controversy for which there is an obvious polarity to the possible outcomes, that is, a question of being either for or against the use or adoption of something, or the commitment to some course of action, etc. It is therefore characteristic of this approach that the topics themselves are understood as claims for which there is either supporting or contradictory evidence or reasoning.

In a broader sense, this type of framework draws on Freeman's model without explicitly naming it, distilling its core idea to the minimal assumption that arguments are products of discourses in which certain statements are put forward to support others (Freeman, 1991), and extending it to encompass the possibility of a negative stance together with its thematic context.

Own: Lastly, it is also important to identify those frameworks that introduce independent definitions for their frameworks. A total of nine (~29%) of the selected datasets report that they have developed their own definition, while others do not mention this aspect at all.

To begin with, the AEC dataset, which is referenced by AFS, has to be highlighted. Although AEC does not explicitly state that it is based on one of the previously described framework types, it relies on a self-developed markup hypothesis. This hypothesis specifies the argumentative function of the key phrases *so*, *if*, *but*, *first*, and *I agree that* within a sentence. The AEC framework therefore concentrates, although within a more limited scope, on the role of discourse markers and their connection to the explicit articulation of arguments and, for this reason, is also comparable to a very simplistic derivative of Cohen's model (Cohen, 1984).

A similar case can be observed in the ACQUA dataset, which focuses on arguments that establish relative evaluations, in other words, arguments that classify items as better or worse, which can be argued to be a special case of stance.

For the remaining datasets, it can be stated that they either introduce independent definitions for a two-component model, as in the case of inference and information in TACO, or, when several components are involved, specify a hierarchy of different manifestations that indicate how elements are to be categorized.

Taken together, these types of frameworks can therefore be assumed, similar to Freeman's standard model, to regard two components as constitutive of arguments. The differences, however, lie in the ways in which these components are labeled or identified. Some approaches rely on simplified indicators such as discourse markers, whereas others develop their own terminologies, either to contribute to a more nuanced understanding of the components or to adapt the framework to specific use cases, such as Twitter. Given their shared focus on identifying arguments and their overlapping considerations regarding their constitution, it can be assumed that the respective approaches, at least at first glance, are closely related. However, it is questionable to what extent these approaches are transferable to one another. This issue, therefore, must be considered from a practical perspective and is addressed in Chapter 5.

2.3.3 Computational Approaches to Argument Identification

Another central focus of this section is the presentation of various approaches and solution strategies about the models and features used and that have defined the SOTA at various points in the development of the field. This overview is intended to contextualize the methodological orientation and focus of the main contributions of this thesis (Feger and Dietze, 2024a,b; Feger, Boland, and Dietze, 2025) in relation to existing approaches reported in Table 2.1 and to their respective peak scores summarized in Figure 2.2.

Before proceeding, a brief digression on language modeling is in order, as this is a central approach to advancing linguistic intelligence in models for both NLP and automated AM.

The History of Language Models: In this regard, four successive and overlapping stages of *Language Models* (LMs) have emerged (Saphra et al., 2024; Zhao et al., 2025).

Statistical Language Models (SLMs) assume that language can be modeled through statistical properties of word sequences observed in their immediate context (Jelinek, 1998; Rosenfeld, 2000; Gao and Lin, 2004; Hastie, Tibshirani, and Friedman, 2009). A common approach uses n-gram models (e.g., bi- or trigrams) represented as discrete units, each forming a dimension in a feature vector. More generally, SLMs may also employ composite feature vectors integrating characteristics such as keyword frequencies, sequence length, or other statistics (Hastie, Tibshirani, and Friedman, 2009). While easy to compute, these SLMs remain limited in semantic richness, as they treat features as independent and without deeper contextualization (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov, Yih, and Zweig, 2013; Rodríguez et al., 2018; Svete and Cotterell, 2024).

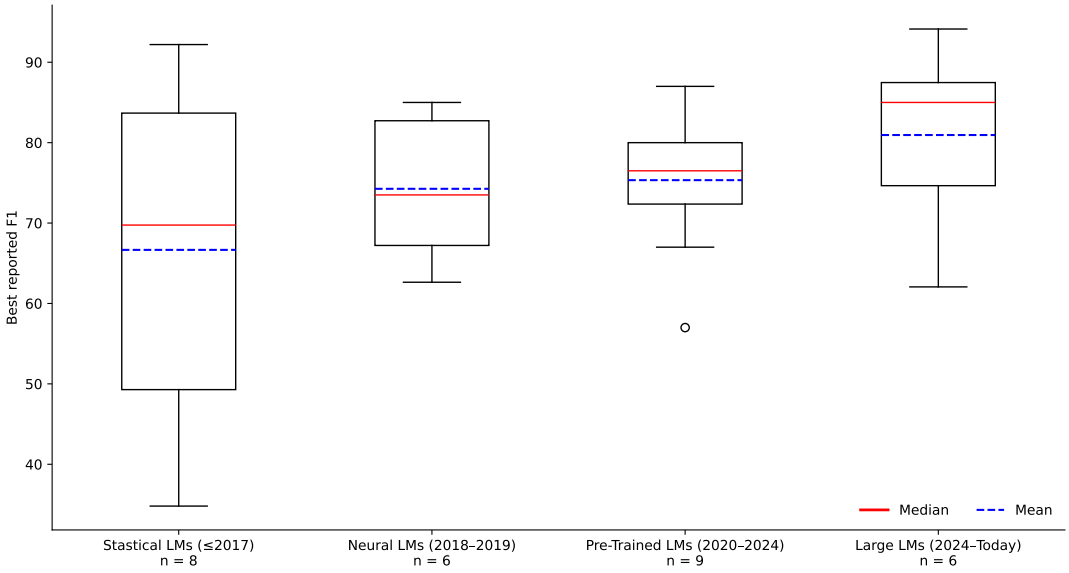


Figure 2.2: Best F1 scores reported in Table 2.1, without consistent differentiation into micro, macro, or weighted averages, etc. The values from 2024 onward (right-hand side) pertain to the pre-2024 datasets QMC, ASRD, IAM (Chen et al., 2024) and PE, ABSTRCT, CDCP (Cabessa, Hernault, and Mushtaq, 2025), but were obtained using LLMs. Across all phases of LMs, results converge within a narrower range, accompanied by a clear overall rise in reported F1 scores.

Neural Language Models (NLMs) take a step further by using *Neural Networks* (NNs) to meaningfully learn word sequence probabilities (Bengio et al., 2003). A key contribution was the introduction of distributed word representations, i.e., dense vectors in which the meaning of a word is spread across many dimensions (Mikolov et al., 2013b). Word meaning, to some extent, emerges from contextual features learned directly by the model from input data (Turian, Ratinov, and Bengio, 2010), thereby transcending SLMs and enabling first attempts towards task-agnostic *Representation Learning* (RL) (Bengio, Courville, and Vincent, 2013). With *Word to Vector* (Word2Vec) (Mikolov et al., 2013a,b), a characteristic model of this generation was introduced that efficiently learned static word representations and demonstrated their ability to capture semantic relations between words, as in *king* – *man* = *queen* - *woman*.

Pre-trained Language Models (PLMs) mark a new generation of LMs aimed at learning general-purpose, context-dependent representations from large-scale unlabeled corpora. These models are pre-trained not only to capture complex properties of word usage, such as syntax and semantics, but also to model how these properties vary across different linguistic contexts (Peters et al., 2018; Devlin et al., 2019; Ethayarajh, 2019). This allows them to represent phenomena like polysemy, where the meaning of a word depends on its context, for example, when deciding whether the discourse marker *since* is interpreted causally or temporally (Pitler and Nenkova, 2009). Moreover, they can be efficiently adapted to a wide range of NLP tasks (Rogers, Kovaleva, and Rumshisky, 2020; Sun et al., 2020), during which task-specific information is integrated into the learned representations with relatively little additional training.

An early approach in this line is the use of *Bidirectional Long Short-Term Memorys* (BiLSTMs), which process sequences in both directions, left to right and right to left, so that each word is represented in relation to its full surrounding context (Schuster and Paliwal, 1997; Graves and Schmidhuber, 2005). Unlike NLMs, this yields dynamic, context-sensitive representations that vary with the sentence in which a word occurs. A prominent PLM based on BiLSTMs is *Embeddings from Language Models* (ELMO) (Peters et al., 2018).

A further advancement in this direction was achieved with the transformer architecture, which introduced a modular encoder and decoder design built entirely on attention mechanisms and fully connected *Feedforward Neural Networks* (FFNs) (Vaswani et al., 2017). In this setup, the encoder stack maps an input sequence into contextualized hidden representations, and the decoder stack then auto-regressively generates an output sequence, consuming the previously generated tokens, basic units of text such as words, subwords, or characters, as additional input. In both stacks, the attention mechanisms make it possible to capture dependencies across arbitrary distances in the input sequence (Vig, 2019). Building on this foundation, the revolutionary *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) was introduced, triggering a surge of interest in transformers (Liu et al., 2019; Sanh et al., 2019; Rogers, Kovaleva, and Rumshisky, 2020) and their ability to generate powerful, context-sensitive language representations. BERT also popularized training strategies such as *Masked Language Modeling* (MLM) (predicting masked words in a sentence) and *Next Sentence Prediction* (NSP) (classifying whether two sentences follow each other) and has sparked a wave of contributions around PLMs in general.

Some of these focus on different architectures, such as *Bidirectional and AutoRegressive Transformers* (BART) (Lewis et al., 2020) that use both transformer stacks, or GPT-2 (Radford et al., 2019) that rely on the decoder-stack and *Next Token Prediction* (NTP). Others propose modified pre-training strategies, for example, the *Robustly Optimized BERT Pre-training Approach* (RoBERTa) (Liu et al., 2019), which removes NSP in favor of larger datasets, longer training, and focusing on the MLM objective.

Large Language Models (LLMs), in turn, have demonstrated that generative pre-training of the decoder-stack of PLMs (Radford and Narasimhan, 2018), combined with scaling by increasing parameter size, training data, or training duration (Henighan et al., 2020; Kaplan et al., 2020; Rae et al., 2021; Hoffmann et al., 2022; Wei et al., 2022b), systematically improves their performance on established downstream tasks (Saphra et al., 2024; Zhao et al., 2025). Furthermore, scaling has been shown to give rise to emergent abilities that are absent in smaller models but appear in larger ones (Wei et al., 2022b). An example is the prompting paradigm, where models such as GPT-3 can perform few-shot learning (Brown et al., 2020), solving tasks without additional SFT from only a few input–output examples provided in the input prompt. This capability emerges only beyond a certain scale and leads to performance well above random guessing (Wei et al., 2022b). Among others, this also includes abilities such as chain-of-thought prompting (Kojima et al., 2022; Wei et al., 2022c), which enables these models to solve complex problems by producing intermediate reasoning steps before giving a final answer, as well as instruction following (Hu et al., 2021; Wei et al., 2022a; Dettmers et al., 2023; Zhang et al., 2024) and even writing and executing code (Chen et al., 2021; Nye et al., 2021). Building on this insight, ever larger models have been developed, such as the LLaMA-family (7–65B) (Touvron et al., 2023), GPT-3 (175B) (Brown et al., 2020), and *Pathways Language Model* (PALM) (540B) (Chowdhery et al., 2022), which operate at parameter counts several orders of magnitude larger than those of earlier PLMs such as BERT (110M). To emphasize this qualitative leap, the research community coined the term LLMs.

The History of Language Models for Mining Arguments: The transition from hard-wired heuristics to the ability to solve complex tasks marks a decisive advance in the scientific understanding of LMs and is central to their historical development. Viewed through the lens of task-solving ability, four generations can be distinguished accordingly (Saphra et al., 2024; Zhao et al., 2025). Early SLMs supported specific applications by providing probability estimates that enhanced task-specific methods, while NLMs sought to learn task-agnostic representations and thereby reduced the need for manual feature engineering. With the advent of PLMs, contextualized representations could be generated and flexibly adapted to downstream tasks, and through scaling effects these models evolved into LLMs that act as general-purpose task solvers with a greatly expanded range of capabilities.

This development is also reflected in the application of the respective LMs to AM research, as illustrated in Figure 2.2 and detailed in Table 2.1.

In the early years of AM ($\leq$ 2017), SLMs were predominantly relied upon. These models typically made use of handcrafted features such as n-grams, stylistic indicators including sentiment and grammatical structures, and other statistical characteristics.

For decision-making, these features were often combined with classification techniques such as *Support Vector Machines* (SVMs) (Cortes and Vapnik, 1995), *Naive Bayes* (NB) (Lewis, 1998), or more tailored approaches such as *Hybrid Baseline* (HB) and *Evidence Graph* (EG).

In the years that followed, up until 2019, the field increasingly shifted towards NLMs.

On the one hand, improved feature representations such as Word2Vec or *FastText* (FastText) (Mikolov et al., 2018) at the word level, *Inferential Sentence Embeddings* (InferSent) (Conneau et al., 2017) at the sentence level, and BiLSTMs became central. In addition to these advanced approaches to feature representation, other classification models such as *Conditional Random Fields* (CRFs) (Lafferty, McCallum, and Pereira, 2001), FFNs, *Deep Neural Network* (DNNs) (Rosenblatt, 1958; LeCun, Bengio, and Hinton, 2015), and *eXtreme Gradient Boosting* (XGBoost) (Chen and Guestrin, 2016) were used to leverage these features.

On the other hand, the publication of BERT and its impact on NLP research can also be traced in AM, starting in 2020 and denoting the shift towards PLMs. Since then, models such as BERT, RoBERTa (Liu et al., 2019), and their derivatives, including *Scientific BERT* (SciBERT) (Beltagy, Lo, and Cohan, 2019) and smaller variants like *Distilled BERT* (DistilBERT) (Sanh et al., 2019), have been the most widely applied.

More recently, starting in 2024, LLMs have been investigated through the application of models from the *Fine-tuned LAnguage Net* (FLAN)-family (Chung et al., 2022; Wei et al., 2022a; Tay et al., 2023) and from the LLaMA-family (Touvron et al., 2023). These models have found their way into AM research (Chen et al., 2024; Cabessa, Hernault, and Mushtaq, 2025), where they have been applied to existing datasets such as QMC, ASRD, and IAM (Chen et al., 2024), as well as PE, ABSTRCT, and CDCP (Cabessa, Hernault, and Mushtaq, 2025).

Considering the best reported results in the papers, a clear trend emerges. In the early years up to 2017, results obtained with SLMs were scattered between 34.80 and 92.20 F1, with reasonably good mean and median values ($M = 66.66, Md = 69.75, n = 8$). During the early NLM era (2018–2019), results improved ($M = 74.26, Md = 73.50, n = 6$) and converged to best scores between 62.64 and 85.00 F1. In the subsequent PLM era from 2020 onwards, results increased slightly ($M = 75.34, Md = 76.50, n = 9$) while tightening further to a range between 57.00 and 87.00 F1. Finally, the results of the LLM phase starting in 2024 showed another numerical leap in best reported scores ($M = 80.95, Md = 85.00, n = 6$), with values ranging from 62.06 to 94.13 F1.

Taken together, and considering that the observed leaps may also reflect datasets becoming simpler for LMs as the models themselves grow more powerful, these figures, nonetheless, indicate a steady numerical increase over the years. At the same time, they also show that expectations for the corresponding baseline reports have risen, particularly given that PLMs, and especially LLMs, continue to produce increasingly strong results despite the lack of full understanding of their internal mechanisms (Geirhos et al., 2020). With this background, and given that the contributions of this thesis were made in the later post-BERT phase of PLM development, LMs of the BERT-family are considered as the SOTA.

Chapter 3

Building the First Conversation-Based Argument Mining Dataset for Twitter

This section introduces the initial contribution of this thesis, presenting the first dataset specifically designed for *Twitter Arguments from COnversations* (TACO), along with the associated challenges encountered before its creation that were not included in the respective paper (Feger and Dietze, 2024a).

3.1 Twitter Conversations

For the purpose of obtaining conversation-based tweets, the Twitter API v2 for academic projects[1] and its advanced functionality were used. This API enables a full archive search of Twitter from March 2006 onwards, with advanced query options for extracting graph-based data, namely conversations.[2]

From a theoretical perspective, the representation of a Twitter conversation is to be understood in analogy to a reply-tree (Nishi et al., 2016), whose structure is a rooted in-tree in which a designated root-tweet can be reached from every reply-tweet. Technically, the conversation-id opens a conversation-thread by assigning the identifier of the very first tweet to all, if any, of its subsequent reply-tweets, each linked via a reply-relation to the tweet they refer to. However, the conversation-id also preserves deleted tweets structurally and their actual replies within the conversation, but not to whom they replied.

For the purpose of wording, a tweet can be considered conversation-starting if it receives at least one reply-tweet, thereby initiating a conversation-thread beyond its root-tweet. A conversation, in turn, is defined as a set of conversation-threads initiated by conversation-starting tweets, which indicate their topical focus through the use of designated hashtags. While retweeting and quoting tweets can spark side-conversations in separate threads, they are excluded due to the possible broken-telephone effect, where the original message becomes distorted as it is passed on (Boyd, Golder, and Lotan, 2010). The structure of Twitter conversations, as described here, is schematically illustrated in Figure 3.1.

[1] Access was free (10 million tweets per month) and granted by Twitter after reviewing the research idea.
[2] The dedicated extraction tool is available at: Twitter-Conversation-Extraction.

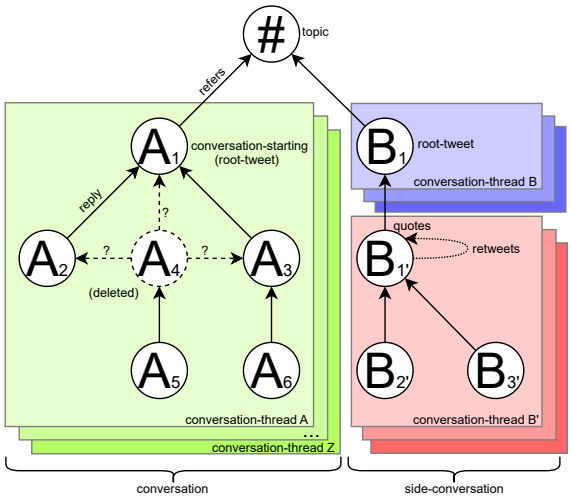


Figure 3.1: Simplified illustration of conversations (left) on Twitter. Retweets and quotes create copies of their original tweets for redistribution and thus cause side-conversations (right) in their own conversation-threads, which are excluded.

3.2 Modeling Argumentation on Twitter

At the outset of Chapter 2, the standard approach of argument structure, as well as its specific representation in the academic literature, was examined in detail.

Although there is widespread agreement (Feger, Boland, and Dietze, 2025) that arguments generally follow the standard approach (Thomas, 1986; Freeman, 1991, 2011) consisting of a minimal element that defines the argument and additional components that supplement it, the specific terminology used to describe these elements often varies across sources (Lopes Cardoso et al., 2023; Feger, Boland, and Dietze, 2025). For example, some authors refer to the components as premise and conclusion (Poudyal et al., 2020; Grundler et al., 2022), while others describe them as evidence and claim (Rinott et al., 2015; Mayer, Cabrio, and Villata, 2020a; Fergadis et al., 2021; Hansen and Hershcovich, 2022) or premise and claim (Hidey et al., 2017; Stab and Gurevych, 2017; Fromm et al., 2021).

Moreover, given Twitter's prominence as a widely used social media platform, it cannot be assumed in advance that annotators approach the content without pre-existing opinions about the topics being discussed. As a result, there is a risk that annotations may reflect individual beliefs about what is right or wrong, rather than being based on an objective analysis of the text segments that function as components of an argument. This potential bias could compromise the reliability of the annotation process by conflating personal judgment with analytical evaluation.

The Constitution of Arguments in Tweets: In order to address terminological inconsistencies, such as the use of different labels for components that served similar functions, or to account for genuine differences where applicable, and to avoid the need to assess quality criteria such as logical cogency (Wachsmuth et al., 2017; Romberg et al., 2025), the focus was shifted away from fixed terminology. Instead, attention was directed toward the purpose that individual components fulfilled within the broader context of a conversation at hand.

From this perspective, a more functional understanding of argument components along with their interrelations was abstracted to highlight their roles within the argument rather than focusing on specific terminological labels. This approach was shaped by guiding questions that a reader might reasonably ask when interpreting the data.

At the core of this approach is the idea that the canonical structure of an argument, commonly represented as $A = \langle C, P \rangle$, can be meaningfully generalized through two guiding questions. These questions are intended to capture the functional essence of the components without relying on fixed terminology. The minimal component C, which conveys the central message or takeaway of the argument, and the supplementary components P, which provide necessary support or context, are reinterpreted in terms of what a reader actively seeks to understand.

C then corresponds to the question: *What should I take away?*

P corresponds to the question: *What do I need to know?*

As outlined in the main text (Feger and Dietze, 2024a) and presented in Appendix A.1, the annotation guide supplemented these questions with definitions drawn from the Cambridge Dictionary[3], while also addressing the annotators' confidence in assigning the labels. Thereby, the following questions and definitions were provided to the annotators:

Inference: *A guess that you make or an opinion that you form based on the information that you have*[3], as captured in *What does the author of the tweet want you to believe?*

Information: *Facts or details about a person, company, product, etc.*[3], as reflected in *What does the author of the tweet want you to know?*

Difficulty: *How well did the components emerge?* as indicated by *Easy* (directly visible), *Normal* (repeated consideration), and *Hard* (strong concerns).

While inference and information were further elaborated upon in the paper, the difficulty was used solely as a control and support variable to evaluate and refine the annotation guidelines.

Indeed, the average difficulty rating per annotator shows that in at least 70% of all annotations, the task was generally perceived as easy. Instances that caused uncertainty were used as discussion points among annotators during the training phase, before the start of independent annotation. These discussions aimed not only to develop simplification rules but also to encourage critical reflection on the abstraction level of the annotated components. This process helped ensure greater consistency across annotators and promoted a shared understanding of how to handle ambiguous or borderline cases.

[3] dictionary.cambridge.org

For example, among the simplification rules developed, annotators agreed that rhetorical questions, as do hashtags, can carry inferential meaning. Quotes, dates, and personal stories were likewise interpreted as forms of information. To maintain consistency, annotators were also encouraged to avoid assigning components in cases where such assignments might appear forced or artificially constructed, such as interpreting an insult based solely on inferred intent.

Manifestations of Argument Components in Tweets: To facilitate a more nuanced understanding of the various forms that argument components may take, the observed instances of inference and information within the tweets were accordingly grouped into four distinct classes. These classes differ based on the specific realization of each component, which may be either present or absent in a given instance. In order to facilitate a clearer understanding of this classification, the defining characteristics of each class have been systematically outlined and documented in the annotation guidelines. The following outlines descriptions and examples of the four classes in which inference and information can manifest within a tweet.

Reason refers to a type of tweet in which the inference is directly derived from information explicitly provided within the tweet itself, such as citations, references to events, or factual statements. This type of tweet underscores the author's effort to anchor their argument in traceable and concrete information, which reflects the motivation *to try to understand and to make judgments based on practical facts*[3].

Example:

> Opinion: As the draconian (and then some) abortion law takes effect in #Texas, this is not an idle question for millions of Americans. A slippery slope towards more like-minded Republican state legislatures to try to follow suit. #abortion #F24 https://t.co/sMKUdhRF1q

Statement is a type of tweet in which only the inference is expressed, typically as *something that someone says or writes officially, or an action done to express an opinion*[3]. This type of tweet, in turn, presents an inference without explicitly providing the supplementing information, instead relying on context or prior knowledge assumed to be known by the reader.

Example:

> Men shouldn't be making laws about women's bodies #abortion #Texas

Notification refers to a type of tweet that is restricted to the straightforward dissemination of information, without engaging in inference. This class typically includes content such as media outlets sharing links to their articles or announcements that serve primarily to inform rather than to persuade or evaluate.

Example:

> #Mexico top court declares criminalizing #Abortion_06 unconstitutional
>
> - JURIST-News Mexico's Supreme Court of Justice of the Nation ruled Tuesday that total #criminalization of #abortion is unconstitutional. #AbortoLegalMexico #USSupremeCourt #SupremeCourt https://t.co/xLj5PZijOL

None is a type of tweet that contains neither inference nor informational content. It lacks reasoning, factual grounding, or contextual detail and may consist primarily of vague, emotionally charged, or displaced content that offers no clear persuasive or informative contribution to public discourse or debate in general.

Example:

> @sinnfeinireland Blah blah blah blah blah blah

Argumentation in Twitter Conversations: Building on the respective argument components and their resulting manifestations, which constitute the observable products of argumentation, it is imperative to also consider their integration into the broader communication process. In this regard, the contextual factors that are essential for interpreting the communicative function of arguments within argumentation in general (Eemeren and Grootendorst, 2003; Eemeren et al., 2014) cannot be overlooked for Twitter conversations.

Given that Twitter conversations are structured as reply trees, where tweets are interconnected and explicitly reference the tweet they respond to, it is essential to account for the contextual relationship between tweets in order to ensure accurate annotation and interpretation. For example, a tweet may be recognized as a Reason when viewed in relation to a preceding Notification tweet that provides the relevant contextual information, while the same tweet might be classified as Statement if treated in isolation.

Furthermore, entire conversation chains can be used to resolve issues of co-references, such as people being substituted by their corresponding pronouns used to refer to them in subsequent tweets (Bosc, Cabrio, and Villata, 2016). For example, if a person is mentioned in one tweet and later referred to as *he*, *she*, *his*, or *hers*, the broader conversational context can provide clarity (Andy, Callison-Burch, and Wijaya, 2020). Nonetheless, the primary focus lies on direct replies that are explicitly linked to a specific preceding tweet, thereby establishing an immediate contextual relation between the reply and the tweet it addresses.

In this context, conversations can be formally modeled as transition graphs of messages, where each state corresponds to a specific type of message and captures how different message types relate to and influence one another (Shannon, 1948). More precisely, a conversation can be abstracted as a probabilistic transition graph, which is a directed graph in which each node represents a message state and each edge represents a possible transition between states, labeled with a probability that indicates the likelihood of that transition occurring.

Formally, a probabilistic transition graph is defined as a tuple $G = (S, E)$, where S is a finite set of states and $E \subseteq S \times S$ is a set of directed edges. Each edge $(s, s') \in E$ carries a label $p(s'|s) \in [0, 1]$, representing the conditional probability of transitioning from state s to state s'. Moreover, for each fixed state $s \in S$, the probabilities of all outgoing edges sum to 1.

When applied to Twitter conversations, each state corresponds to one of the four possible manifestations of inference and information, that is, $S = \{\text{Reason}, \text{Statement}, \text{Notification}, \text{None}\}$.

Consequently, the transition probabilities represent the likelihood that a tweet of a given type will be followed by a reply assigned to any of the four classes, capturing how argumentation evolves across conversational turns via $p(\text{Reply}|\text{Tweet})$, compare Table 3.2.

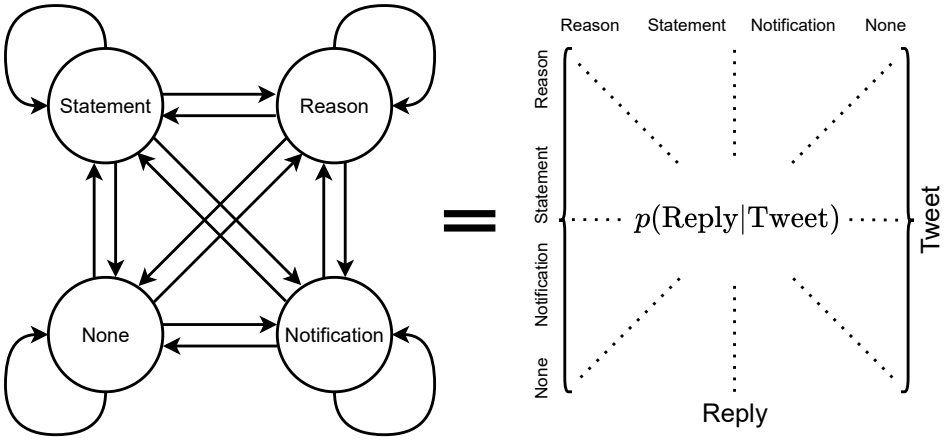


Figure 3.2: Twitter conversations as a transition graph, where nodes correspond to the distinct tweet classes: Reason Statement, Notification, and None. These types are distinguished by the manifestation of argument components, specifically inference and information. Directed edges in the graph represent transition probabilities $p(\text{Reply}|\text{Tweet})$, indicating the likelihood that a tweet of one type receives a reply of another type. The model captures the interaction patterns and structural dynamics of argumentation in Twitter conversations.

3.3 Baseline Evaluation with Pre-Trained Language Models

The following discussion addresses the baseline evaluation of TACO, with particular attention to the choice of the model-architecture family and the underlying mechanisms, as illustrated by the processes shown in Figure 3.3. Moreover, the following aspects are of central importance, as the same principles reappear in different versions throughout the thesis, particularly in the following Chapter 4.

For the baselines, PLMs, specifically BERT-based architectures, were chosen, since they represented the SOTA at the time the paper was written, compare Subsection 2.3.3. In particular, BERT, RoBERTa, DistilBERT, and *A pre-trained language model for English Tweets* (BERTweet), which was pre-trained specifically for Twitter language, were used. This work, however, concentrates on BERTweet as the main baseline. The comparison with the other models is taken up again in Chapter 5 during the re-evaluation of TACO, where it is shown once more that BERTweet and its improved version WRAP, presented in Chapter 4, outperform the others.

While all these models differ in their pre-training and size, they belong to the same BERT-based family of PLMs. The description of contextualized representations and the classification mechanism presented next, as illustrated in Figure 3.3, follows the principles introduced in the original BERT paper (Devlin et al., 2019).

In essence, these models were taught to generate word representations in context, which provide a rich foundation for downstream tasks and enable the distinction between different senses and nuances of meaning. Here, a word representation refers to a mathematical object associated with each word, typically a vector whose dimensions correspond to latent features, that is, properties not directly observable but inferred from the data itself, which capture syntactic and semantic aspects (Turian, Ratinov, and Bengio, 2010; Bengio, Courville, and Vincent, 2013). Beyond that, such a representation can also be understood as an embedding into a vector space so that these two terms can be read synonymously.

On top of these, a lightweight predictor can be added to perform classification or other specific objectives without the need to train the entire model from scratch.

Sequence Classification as a Formalized Learning Task: To be precise, a PLM of the BERT-family can be described as an embedding model, which, in turn, is a learned function G_W, parameterized by weights W (Sun et al., 2020). These weights are obtained through respective pre-training, which allows the model to be used directly for downstream tasks without training from scratch.

Given a textual sequence $t \in \mathcal{T}$ (e.g., a sentence in a document), it can be represented as a token sequence $t = (t^{(1)}, \ldots, t^{(n)})$, where each $t^{(i)}$ is a basic unit such as a word or sub-word.

For sequence classification, the BERT-family introduces a special token, namely the $[CLS]$ token, which is placed at the beginning of the input sequence and aggregates contextual information from all other tokens, thus serving as a pooled representation of the entire sequence.

The model then maps t to the corresponding sequence of contextualized token representations $(h^{\text{CLS}}, h^{(1)}, \ldots, h^{(n)}) = G_W(t)$, where each $h^{(i)} \in \mathcal{H} \subseteq \mathbb{R}^d$ is a d-dimensional vector (typically $d = 768$) in the representation space $\mathcal{H}$ of G_W. Therein, each $h^{(i)}$ encodes token $t^{(i)}$ in the context of the entire sequence t. By convention, however, the h^{CLS} representation, although itself also context-dependent like the others, is widely adopted as a proxy for the representation of the entire input sequence.

Apart from the ongoing debate about whether h^{CLS} is sufficiently informative or whether alternative pooling strategies such as straight mean pooling over each token representation $h^{(1)}, \ldots, h^{(n)}$ are more useful (Reimers and Gurevych, 2019), the use of the $[CLS]$ token remains the standard method for sequence classification (Devlin et al., 2019; Rogers, Kovaleva, and Rumshisky, 2020; Sun et al., 2020).

Furthermore, for sequence classification, h^{CLS} is often simply referred to as the classification token $h = G_W(t)$, a convention that will be followed throughout the rest of this thesis.

Considering this, h is passed to a classification head $C_{\hat{W}}$, parameterized by its weights $\hat{W}$. In the BERT-family, this head typically consists of one or two FFN layers on top of G_W, transforming its input into a vector of real-valued outputs, known as logits. Formally, if the classification token is denoted as $h = G_W(t)$, then the logits can be expressed as $z = C_{\hat{W}}(h)$.

The probability distribution across classes is then determined by applying the softmax function $p(y \mid t) = \text{softmax}(z)$. Thereby, z are the raw class scores, which are normalized into a class probability distribution by the softmax function. As formalized in Subsection 2.2.3, the highest probability is then selected for the final class prediction $\hat{y}$.

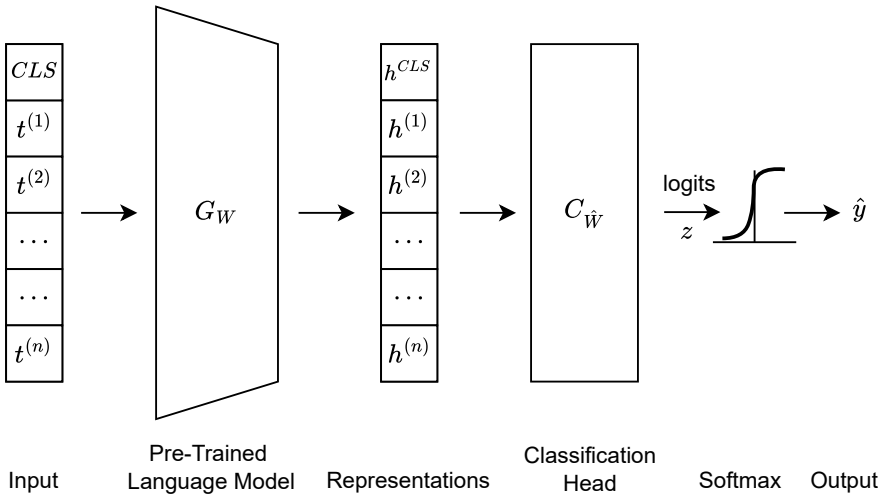


Figure 3.3: Sequence classification with BERT-like PLMs. The input sequence t of tokens $t^{(i)}$ (e.g., words or sub-words) is prefixed with a special $[CLS]$ token. A representation model G_W (e.g., BERT) maps t to contextualized token representations $(h^{\text{CLS}}, h^{(1)}, \ldots, h^{(n)}) = G_W(t)$. For classification, the pooled vector h^{CLS} (hereafter referred to as $h = G_W(t)$ for simplicity), which integrates contextual information from the entire sequence, is passed to a classification head $C_{\hat{W}}$. Its logits (raw outputs) are transformed into class probabilities via softmax, and the final prediction is obtained according to Subsection 2.2.3.

The overall model can thus be described as a composite function $F_\theta = C_{\hat{W}} \circ G_W$, where the parameter set $\theta = \{W, \hat{W}\}$ combines both the weights of the PLM G_W, which are used to derive meaningful representations of the input, and the separate weights of the classification head $C_{\hat{W}}$ for making predictions based on them.

At the bottom line, for the BERT-family, the sequence classification task boils down to several intermediate steps $t \xrightarrow{G_W} h \xrightarrow{C_{\hat{W}}} y$, where the entire set of parameters $\theta = \{W, \hat{W}\}$ can be fine-tuned together to achieve better classification results while also improving the representations to a certain extent. The latter will be examined at greater depth in Chapter 4.

Marc Feger and Stefan Dietze.

"TACO - Twitter Arguments from COnversations"

In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 15522–15529, Torino, Italia. ELRA and ICCL.
Acceptance Rate: ∼44.8%

Now turning to the specific paper (Feger and Dietze, 2024a), the following provides an overview of its content and outlines the key findings regarding research question **Q1** and its sub-questions **Q1.1-3** introduced in Section 1.2. After a brief statement on the general implications and significance of this work for the dissertation, the respective paper will be presented.

3.4.1 Summary

Twitter Arguments from COnversations (TACO) is the first publicly available dataset for *Argument Mining* (AM) based on full Twitter conversations, containing 1,814 annotated tweets from 200 discussions on six peak-performing hashtags, #Abortion, #Brexit, #TwitterTakeover, #GOT, #LOTRROP, and #SquidGame, sampled from around 600,000 tweets collected between 2020 and 2021. Specifically, each collected conversation is represented as a complete reply tree, starting with an initial tweet that sparked reactions and branching out into threads that trace the full progression of the corresponding sub-discourse down to the very last reply.

First, to address **Q1.1** (*How can an annotation scheme for arguments in conversations be designed?*), this paper presents the theoretical simplification of arguments and introduces a custom annotation framework grounded in Cambridge Dictionary[3] definitions. Given the lack of a universal definition of arguments and as the first attempt to annotate arguments within coherent Twitter conversations, it defines inference (opinions or guesses based on information) and information (facts or details) as the core components of arguments and discusses the design choices behind this modeling approach in light of established literature and best practices.

Reflecting the diverse communicative intents on Twitter, tweets are classified as Reasons (inference and information), Statements (inference only), Notifications (information only), and None (no elements). By its design, this differentiation supports a hierarchy of both a binary categorization between argument tweets, which contain inference, and no-argument ones, as well as a more detailed analysis of the specific classes each tweet falls into.

Based on this hierarchy, an annotation guide was developed and refined through iterative testing and optimization. In their final iteration, the annotations were designed with respect to the conversational context of tweets, that is, the preceding and subsequent replies.

Second, to answer **Q1.2** (*What annotation quality can a resulting gold dataset achieve?*), the process achieved a high inter-annotator agreement among six experts, with a Krippendorff's α of 0.718, and a strong retention rate, with 95.6% of all annotations passing a strict majority vote of 50% agreement regarding one of the four tweet classes defined.

Although the distribution varies across individual classes, the proportion of argument and no-argument tweets is balanced within the final dataset.

Based on the final annotations, an analysis of the conversations revealed that users on Twitter frequently engage in informed discourse, primarily involving reasoning or additional information, where reply patterns showed that such context-rich, argumentatively informed tweets were most frequently responded to in kind.

To address **Q1.3** (*What baseline results do SOTA models yield on this dataset?*), the evaluation involved the SFT of several SOTA PLMs in a sequence classification setup with cross-validation. The best performance was achieved with BERTweet, reaching 85.06% macro F1 for argument detection and 72.49% for tweet-level classification.

In its conclusion, considering **Q1.1**–**3**, the paper answers the main research question **Q1** (*Can arguments be extracted within entire Twitter conversations?*) by showing that arguments, defined through inference and information, can be identified across conversational threads and diverse topics with substantial agreement ($\alpha = 0.718$). Following best practices, BERTweet further achieved strong results, with 85.06% macro F1 for argument detection and 72.49% for tweet-level classification. Nonetheless, fine-tuning alone did not ensure strong performance across all classes, as superficial signals led to underperformance in some cases. In fact, the length of tweets and the presence of discourse markers show a strong alignment with class distinctions. Moreover, misclassifications at the tweet-level, for instance, labeling a Statement as a Reason, still contribute to ~43% of true positives in argument detection.

These results highlight the need for further work on refining model representations to better capture and generalize the diverse and pragmatic functions of different tweet classes in constructing arguments. They also suggest that strong performance in argument detection may not necessarily reflect a model's ability to reliably distinguish between the various classes that emerge of inference and information.

3.4.2 Importance and Impact on this Thesis

Assessing practical AM on Twitter under real-world conditions requires ground truth data that captures the structural components of arguments and reflects their dynamics across diverse topics of full conversations, for which this study introduces the first foundational resources.

Following best practices, findings showed that SFT did not consistently deliver strong performance across all tweets. Error analysis revealed that much of the baseline performance was not solely due to the model's ability to distinguish arguments from no-arguments based on the annotated data intent. Instead, superficial correlations, such as length, discourse marker usage, and tweet-level misclassifications that still matched the broader category of a tweet, raised concerns about the model's genuine alignment in representing argument components. In turn, the findings of this paper (Feger and Dietze, 2024a) provide the foundation for the second main contribution of this thesis (Feger and Dietze, 2024b), which addresses **Q2** (*Do SOTA models inherently predict arguments in Twitter conversations?*) in particular.

TACO – Twitter Arguments from COnversations

Marc Feger[1], **Stefan Dietze**[2]
[1]HeiCAD - Heine Center for Artificial Intelligence and Data Science,
[2]GESIS - Leibniz Institute for the Social Sciences
[1]Düsseldorf Germany, [2]Cologne Germany
marc.feger@hhu.de, stefan.dietze@gesis.org

Abstract

Twitter has emerged as a global hub for engaging in online conversations and as a research corpus for various disciplines that have recognized the significance of its user-generated content. Argument mining is an important analytical task for processing and understanding online discourse. Specifically, it aims to identify the structural elements of arguments, denoted as information and inference. These elements, however, are not static and may require context within the conversation they are in, yet there is a lack of data and annotation frameworks addressing this dynamic aspect on Twitter. We contribute *TACO*, the first dataset of *T*witter *A*rguments utilizing 1,814 tweets covering 200 entire *CO*nversations spanning six heterogeneous topics annotated with an agreement of 0.718 Krippendorff's α among six experts. Second, we provide our annotation framework, incorporating definitions from the Cambridge Dictionary, to define and identify argument components on Twitter. Our transformer-based classifier achieves an 85.06% macro F1 baseline score in detecting arguments. Moreover, our data reveals that Twitter users tend to engage in discussions involving informed inferences and information. TACO serves multiple purposes, such as training tweet classifiers to manage tweets based on inference and information elements, while also providing valuable insights into the conversational reply patterns of tweets.

Keywords: Argument Mining, Twitter Conversations, Resource, Inference and Information Extraction

1. Introduction

Social media has created an open network of voices, connecting people globally and allowing them to exchange ideas and engage in discussions on any topic of interest. Despite these benefits, maintaining healthy and substantial online deliberation and promoting transparent information exchange remain key challenges in this area (Chadwick and Howard, 2008; Ruiz et al., 2011). Twitter, now X, serves as a global hub for opinions, news, and information, recognized for its research value and user-generated content prior to its rebranding (Kwak et al., 2010; Boyd et al., 2010).

In this context, argument mining has emerged as a valuable technique to identify the structure of inference and reasoning presented as arguments in natural language and is closely related to information extraction, fact checking, citation and opinion mining (Lawrence and Reed, 2019). This involves the automatic identification and extraction of arguments expressed in text, thus enabling researchers to analyze and understand the nature and structural elements of online discussions. Over the past years, the field of argument mining has undergone significant development in various domains, such as legal texts (Moens et al., 2007; Wyner et al., 2010), newspapers (Reed et al., 2008; Mochales and Moens, 2011), essays (Stab and Gurevych, 2014; Persing and Ng, 2016; Wachsmuth et al., 2016), Wikipedia articles (Levy et al., 2014, 2017), and sources of conflicting content, such as user comments (Park and Cardie, 2014), dialogues (Swanson et al., 2015), and web discourses (Habernal and Gurevych, 2017). While these works made initial contributions to the field, more recent research has focused on the detection of arguments from heterogeneous sources of arbitrary web text (Daxenberger et al., 2017; Levy et al., 2018; Stab et al., 2018). Despite addressing different aspects of argument mining, all studies involve the detection of inference and information (Palau and Moens, 2009; Daxenberger et al., 2017) as part of online discourse.[1]

(1) Men shouldn't be making laws about women's bodies #abortion #Texas
(2) 'Bitter truth': EU chief pours cold water on idea of Brits keeping EU citizenship after #Brexit https://t.co/j3DteyWcMg via @TheLocalEurope
(3) Opinion: As the draconian (and then some) abortion law takes effect in #Texas, this is not an idle question for millions of Americans. A slippery slope towards more like-minded Republican state legislatures to try to follow suit. #abortion #F24 https://t.co/sMKUdhRF1q
(4) @sinnfeinireland Blah blah blah blah blah blah

Table 1: Example tweets that contain inference (1), information (2), a combination of both (3), or none of either (4).

For argument mining on Twitter, research has expanded from specific topics like football (Llewellyn et al., 2014) and encryption (Addawood and Bashir, 2016) to encompass various subjects, including Brexit and Grexit (Dusmanu et al., 2017). Recent studies have focused on structuring tweets into debates via semantic similarity on the topics Iran, Grexit, Apple Watch and Game of Thrones (Bosc et al., 2016), and with isolated tweet-reply

[1]These components, inference and information, are defined along the annotation framework in Section 2.1.

pairs on climate (Schaefer and Stede, 2020), neither capturing entire conversations.

This diversification has led to specialized tasks, such as identifying pro and con arguments in Planned Parenthood tweets (Bhatti et al., 2021) and evaluating scientific support in Covid-19 and climate-related tweets (Hafid et al., 2022).

Despite the progress in argument mining, the scope of related research on Twitter is restricted to a micro-level perspective, solely examining individual tweets and neglecting the interrelated reply tweets that make up the wider context of Twitter discussions. So far, no ground truth data for assessing arguments in entire Twitter conversations exists (Schaefer and Stede, 2021). With our work, we contribute the following to advance the field of argument mining on Twitter:

1 **Annotation Framework.** Our specialized argument mining framework for Twitter conversations evolved from an extensive analysis of the elements defining arguments in relevant literature and iterative deliberations among our experts. It builds on Cambridge Dictionary's[2] definitions to define and identify inference and information within tweets.

2 **Conversation-Based Ground Truth Data.** Our TACO[3] dataset comprises 1,814 tweets, covering 200 entire conversations from six widely-discussed Twitter events. It was annotated by six experts with a high agreement score of 0.718 Krippendorff's α. TACO is available to the public in compliance with Twitter's data policy.

3 **Baseline Classification Model.** Our published transformer-based classifier[4], underlines TACO's significance in argument mining by achieving an 85.06% macro F1 for detecting arguments in tweets and 72.49% macro F1 for identifying combinations of inference and information. This classifier can be employed in both cases to aid in building new datasets and tweet curation.

2. Constructing the TACO dataset

Given the brevity of tweets, which were originally limited to 140 respectively 280 characters, structural elements of arguments such as inference or information tend to be scattered across distinct messages (Kwak et al., 2010; Boyd et al., 2010; Addawood and Bashir, 2016; Dusmanu et al., 2017). At the same time, tweets tend to be rather diverse in nature. Some tweets indicate a genuine interest in contributing to ongoing debates, while others may express different motivations, such as a what-i-had-for-lunch-like tweet (Rogers, 2013).

2.1. Annotation Framework

With no one-size-fits-all definition of what an argument is (Palau and Moens, 2009; Habernal et al., 2014; Stab et al., 2018), the crucial challenge is how to identify arguments on Twitter. However, one potential strategy for simplifying this task is to differentiate between tweets that contain an inference as a key component of an argument and those that do not (Palau and Moens, 2009; Stab and Gurevych, 2014; Bosc et al., 2016; Daxenberger et al., 2017; Lawrence and Reed, 2019). Our aim here is not to create a new formalism for arguments, but rather to integrate established theories and provide a reusable set of definitions that can be applied to Twitter.

To define this critical component of an argument, we refer to the Cambridge Dictionary, which defines *inference as a guess that you make or an opinion that you form based on the information that you have*. We also utilize their description of *information* as *facts or details about a person, company, product, etc.*.

Taken together, argument mining on Twitter involves determining if a tweet contains an inference (**Argument**) or not (**No-Argument**) by also considering its combination with information, as illustrated in Figure 1.

Tweets categorized as an **Argument** can be:

Statement, a tweet where only inference is presented like *something that someone says or writes officially, or an action done to express an opinion* (see tweet 1 in Table 1).

Reason, a tweet where the inference is based on information mentioned in the tweet, such as references, and thus reveals the author's motivation *to try to understand and to make judgments based on practical facts* (see tweet 3 in Table 1).

In contrast, tweets that are categorized as **No-Argument** are defined by the absence of inference and can be described as:

Notification, a tweet that is limited to only providing information, such as media channels promoting their articles (see tweet 2 in Table 1).

None, a tweet that provides neither inference nor information (see tweet 4 in Table 1).

2.2. Data Sampling and Annotation

Twitter conversations are shown to have a strong focus on various topics, often driven by hashtags (Hughes and Palen, 2009; Rogers, 2013; Zhou and Chen, 2014). We utilized Twitter's API v2 to collect a corpus of tweets and their reply-relations, enabling the extraction of entire conversations.

[2] dictionary.cambridge.org
[3] github.com/TomatenMarc/TACO
[4] huggingface.co/TomatenMarc/TACO

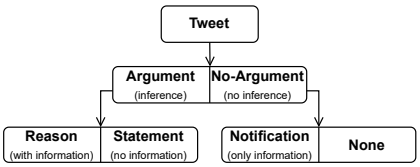


Figure 1: Hierarchy of inference and information.

While earlier studies of argument mining on Twitter mostly focused on one or two topics (Llewellyn et al., 2014; Addawood and Bashir, 2016; Dusmanu et al., 2017; Schaefer and Stede, 2020; Bhatti et al., 2021; Hafid et al., 2022), we aimed to create a more comprehensive corpus by collecting tweets from six controversial topics: #Abortion, #Brexit, #GOT, #TwitterTakeover, #SquidGame, and #LOTRROP. Over a period of seven months, we collected ~600k tweets around key-dates[5] related to these hashtags, resulting in a significant increase in tweet volume. The allocation of the ~600k tweets and the profiles of the hashtag topics are as follows:[5]

#Abortion (5%) pertaining to the 'Row v. Wade' lawsuit, which challenges the Texas abortion ban (S.B.8) after six weeks, discussed and obtained between August 15 and October 16, 2021.

#Brexit (70.9%) relates to the global discussion of Great Britain's departure from the European Union on February 1, 2020. This historic event was queried from January 1 to March 1, 2020.

#GOT (10.2%) was gathered from April 1 to May 1, 2019, for expressing criticism about the final season of 'Game of Thrones'.

#TwitterTakeover (3.1%) was queried from April 1 to May 1, 2022, in association with Elon Musk's position as Twitter's largest shareholder, raising concerns about freedom of speech and initiating its transition to X.

#SquidGame (8.5%) explores economic inequality's impact on moral choices while also advertising the corresponding Netflix series, collected from September 10 to October 10, 2021.

#LOTRROP (2.3%) discussed the release of the trailer for Amazon's first season of 'The Lord of the Rings: The Rings of Power' and the related debate on representation and diversity in media, which we tracked from February 1 to March 1, 2022.

Text annotation is a multifaceted task, encompassing reading, comparing, memorization, and developing consensus about the data (Guetterman et al., 2018; Geiger et al., 2020). Under the author's guidance, two annotation rounds were performed with distinct expert groups to classify tweets based on our proposed annotation framework for Twitter conversations.[5]

[5]For the meta-data or examples, see: README.md

In step 1, three experts and the paper's first author refined the framework's guidelines and assessed their generalizability across topics. We sampled 300 conversation-starting tweets for #Abortion and #Brexit, given their significance in argument mining (Wachsmuth et al., 2017a; Dusmanu et al., 2017; Stab et al., 2018; Levy et al., 2018; Bhatti et al., 2021). Only when complete agreement on a tweet's classification was reached did it proceed as a candidate for step 2 annotation.

In step 2, the first author and two additional experts annotated 200 full conversations, tracing the sequences of reply-relations from the starting tweets to the final replies in a conversation, thereby considering the conversational context.[5] This included 50 conversations for both #Abortion and #Brexit, and 25 conversations for each of the remaining four topics. In total, 236 #Abortion, 285 #Brexit, 192 #GOT, 166 #TwitterTakeover, 226 #SquidGame, and 209 #LOTRROP tweets were considered, averaging 219 tweets per topic.

In these steps, 1,814 annotated tweets were generated, including 1,314 conversation-based tweets of all topics and 500 distinct conversation-starting tweets for #Abortion and #Brexit, with a strong 0.718 Krippendorff's α agreement, given the author's involvement in both phases.

2.3. TACO Dataset

The final TACO ground truth (see Table 2) involved a strict majority vote approach, discarding tweets with less than 50% agreement among assigned votes for a specific class, resulting in 1,734 tweets, accounting for 95.6% of the annotated tweets.

Category	Argument		No-Argument	
	865 (49.88%)		869 (50.12%)	
Class	Reason	Statement	Notification	None
	581 (33.50%)	284 (16.38%)	500 (28.84%)	369 (21.28%)

Table 2: Class distribution in the TACO dataset.

3. Argument Mining on Twitter

We trained a soft-max classifier on top of different transformer models to utilize TACO in two tasks: (1) detecting inference in tweets (**Argument** vs. **No-Argument**), and (2) classifying tweets based on combinations of inference and information.

To obtain a first baseline for the usability of TACO, we fine-tuned the classifiers within an ordinary sequence classification approach integrating TACO's tweets, whose labels, created using our conversation-based annotation framework, encode implicit conversational context without detailing the conversation structures.[6]

[6]This approach is pragmatic, avoiding the complexity of modeling conversation hierarchies during fine-tuning.

For task (2), we fine-tuned the classifiers using 10-fold cross-validation, whereby the experimental results showed that BERTweet (Nguyen et al., 2020) provides superior classification performance. Further it is worth noting that the results for task (1) are aggregations of task (2), with a focus on presence or absence of inference.

The benefits of the BERTweet classifier for TACO extend beyond theory and are supported by cross-validation demonstrating strong performance for argument mining across the conversation-based tweets. Our results demonstrate this effectiveness, with the baseline model achieving an 85.06% macro F1 for inference detection of task (1) and 72.49% for classification of task (2), as seen in Table 3.

In terms of TACO's data, the classification model had access to the following text features indicating tweet classes. The length of tweets differs among classes, with Reason being the longest on average (213), None the shortest (63), and Notification (156) and Statement (122) falling in the middle character range. URL usage varies, with Reason (34.6%) and Notification (71.6%) having the highest, while None and Statement use them less than 8.11%. The usage of discourse marker[7] aligns with the argumentativeness of tweets: Reason (32.9%) is highest, followed by Statement (19%), Notification (11.4%), and None (8.7%).

Although there are misclassifications in task (2), ~43% of them are counted as true positives in task (1). The misclassification of Reason as Statement (or vice versa) still falls into the category of **Argument**, the same being true for the mutual misclassification of Notification and None, which still contributes as **No-Argument**.

Besides inference detection, the model also faces the challenge of correctly classifying tweets by identifying the informative parts, which is distinct from detecting inference and may not be entirely satisfactory due to the varied forms in which information can be presented, like URLs or fragmented text passages. Different tweets may employ language for specific intents involving rhetorical devices or visual elements, adding complexity to identifying information's formal attributes.

Error analysis revealed that Statements, although typically lacking information, can include URLs when misclassified as Reason or Notification, as URLs might seem like apparent markers of information, resulting in an increased average length of 172. In fact, 22.09% of these misclassified Statements contained URLs, predominantly from internal sources like memes, GIFs, and videos. This circumstance might be influenced by Twitter's recommendation system, which rewards tweets that attract a large audience and have entertaining contents attached.[8]

Task	Instance	Precision (%)	Recall (%)	F1 (%)	macro F1 (%)
Category	*Argument*	83.59	**87.17**	**85.34**	**85.06**
	No-Argument	**86.66**	82.97	84.77	
Class	*Reason*	73.69	75.22	74.45	**72.49**
	Statement	54.37	59.15	56.66	
	Notification	79.02	**77.60**	78.30	
	None	**83.87**	77.51	**80.56**	

Table 3: Baseline argument mining performance.

Further investigation found that 24.42% of these incorrect assignments in the Statement class contained discourse markers, which complicated classification because these markers often organized relations between information and inference, leading to multiple stacked inferences that amplified the message's tone rather than being perceived as information in the first instance.

To provide context, the model achieved a F1 score of 80.56% in detecting tweets lacking inference and information (None), which often led to conversation halts in about one-third (33.15%) of cases where a tweet received no further replies. Reason was the second most common in ending conversations at a rate of 29.73%, potentially indicating knockout arguments of further interest. Additionally, a Statement received no replies in 19.04% of cases, while Notification had no replies in 18.08%.

3.1. Conversational Reply Patterns

In our final analysis of the conversation-based ground truth data, we explored state transitions between connected tweet-reply pairs (t, r) to reveal TACO's value in understanding reply patterns and providing insights into conversation progression, as shown in Table 4.

$P(r\|t)$	Reason	Statement	Notification	None
Reason	**0.51**	0.12	0.31	0.06
Statement	**0.38**	0.21	0.33	0.08
Notification	0.26	0.08	**0.57**	0.09
None	0.26	0.08	**0.44**	0.22

Table 4: Transition probability of a tweet with class t (row) having a reply with class r (column).

Our findings reveal users often reply with informed inferences (Reason) or additional information (Notification), with less common conversations solely based on inference (Statement) or lacking both elements (None). Additionally, **Argument** relies on informed inference, while **No-Argument** depends on information usage in replies, reflecting a preference for informed debates.

[7] dictionary.cambridge.org/us/grammar/british-grammar/discourse-markers

[8] github.com/twitter/the-algorithm

4. Conclusion

This paper presents the first ground truth dataset, TACO, for conversation-based argument mining on Twitter, efficiently annotated by six experts using our purpose-built annotation framework that incorporates definitions of argument constituting elements from the Cambridge Dictionary. Unlike previous datasets that often rely on isolated tweets without the contextual framework of conversations, TACO offers fully annotated and coherent Twitter conversations across six topics used for training a transformer-based classification model, providing valuable resources for future research in this domain. Furthermore, the provided classifier effectively differentiates tweets that make arguments from those that do not, based on the presence of inference. Additionally, our multi-class approach sufficiently identifies tweet classes, especially those that lack information and inference. Our findings suggest the need for further research to enhance the semantic features of our proposed tweet classes, possibly by fine-tuning BERTweet for more generalized representations according to our framework.

5. Ethics Statement

In the context of this study, which uses Twitter data, we have adhered to ethical practices and privacy principles and ensured data protection by limiting the publication of TACO to tweet IDs in accordance with Twitter's terms of service. Our annotation process, which involved volunteer experts, has been carefully designed to limit data access to what was strictly necessary and to ensure ethical standards, fair compensation and data integrity. Access to the original dataset is restricted to non-harmful research, subject to appropriate data protection agreements with the authors. It should also be noted that the TACO dataset covers sensitive topics that may contain language and images that some may find offensive.

6. Acknowledgments

We express our gratitude to Aylin Feger, Tillmann Junk, Andreas Burbach, Talha Caliskan, and Aaron Schneider for their valuable contributions as experts to the annotation process. Additionally, we extend our sincere appreciation to the anonymous reviewers, whose fair and insightful assessments have significantly enhanced the quality of our work. Your willingness to share your expertise and provide constructive feedback is deeply appreciated.

7. Bibliographical References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. 2021. Argument mining on Twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Andrew Chadwick and Philip N Howard. 2008. *Routledge Handbook of Internet Politics*, 1st edition. Routledge.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA. Association for Computing Machinery.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of twitter users regarding the brexit referendum. *Computational Social Networks*, 4(1):6.

Timothy C Guetterman, Tammy Chang, Melissa DeJonckheere, Tannya Basu, Emily Scruggs, and V G Vinod Vydiswaran. 2018. Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of medical Internet research*, 20(6):e231.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. 2022. Scitweets - a dataset and annotation framework for detecting scientific online discourse. In *Proceedings of the 31st ACM International Conference on Information; Knowledge Management*, CIKM '22, page 3988–3992, New York, NY, USA. Association for Computing Machinery.

Amanda Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6:248–260.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 591–600, New York, NY, USA. Association for Computing Machinery.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 462–468, Reykjavik, Iceland. European Language Resources Association (ELRA).

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 225–230, New York, NY, USA. Association for Computing Machinery.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda. 2016. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining*, 6(1):26.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Damien Pfister. 2014. "a short burst of inconsequential information:" networked rhetorics, avian consciousness, and bioegalitarianism. *Environmental Communication: A Journal of Nature and Culture*, 9:1–19.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Richard Rogers. 2013. Debanalizing twitter: The transformation of an object of study. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 356–365, New York, NY, USA. Association for Computing Machinery.

Carlos Ruiz, David Domingo, Josep Lluís Micó, Javier Díaz-Noci, Koldo Meso, and Pere Masip. 2011. Public sphere 2.0? the democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics*, 16(4):463–487.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments

from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017a. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. *Approaches to Text Mining Arguments from Legal Cases*, pages 60–79. Springer Berlin Heidelberg, Berlin, Heidelberg.

Xiangmin Zhou and Lei Chen. 2014. Event detection over twitter social media streams. *The VLDB Journal*, 23(3):381–400.

Chapter 4

Learning Generalizable Argument Representations for Twitter

This section presents the second contribution (Feger and Dietze, 2024b) of this thesis, which addresses the representation-related limitations of the TACO baseline model identified in Chapter 3 from a modeling perspective. The focus lies on the role of representations and possible strategies for improving them, considerations that were not part of the published paper but are essential for understanding the improved representation model, WRAP, introduced here.

4.1 Representation Learning

At the heart of ML lies the challenge of transforming raw observations of the real world into descriptions, i.e., representations of how this world unfolds, which make the information contained therein accessible to models and enable them to perform downstream tasks such as classifications (Bengio, 2013; Bengio, Courville, and Vincent, 2013).

Generally speaking, ML, and so is NLP, is bound by the overarching problem that the performance of a model depends heavily on the quality of the representations of the data to which it is applied (Bengio, Courville, and Vincent, 2013; Le-Khac, Healy, and Smeaton, 2020).

Traditionally, and as noted in Subsection 2.3.3 for AM, this challenge has been addressed through manual feature engineering (Bengio, Courville, and Vincent, 2013; Zhao et al., 2025). In this paradigm, domain experts design and select features by defining descriptors that capture relevant aspects of the data, guided by prior knowledge, heuristics, and intuition.

However, these engineered features and their creation are often time-consuming, domain-specific, and difficult to scale. In addition, manually designed features remain fixed once they are specified. This means that, due to their design, they cannot adapt to new data distributions or tasks and may also fail to capture factors that are relevant but unknown to the designer or cannot be easily represented within handcrafted descriptors.

Representation Learning (RL), on the other side, offers a fundamentally different approach. Rather than relying on human intuition, models learn directly from data to construct internal representations that ideally capture task-relevant patterns and abstractions.

This shift from engineering to learning reduces dependence on handcrafted design, allows representations to adapt to specific problems, and enables the discovery of features that may not be accessible to human designers or which were not thought of in the first place.

With this in mind, an entire chapter, Chapter 5, is dedicated to a more detailed discussion of the issues arising from this. But before turning to these issues, it is first necessary to clarify what is meant by RL, how the learning process is generally structured, and how meaningful representations can be obtained.

At its core, the idea behind RL involves *learning representations of data that make it easier to extract useful information when creating classifiers or other predictors* (Bengio, Courville, and Vincent, 2013). More specifically, *it refers to the process of learning a parametric mapping from the raw data space to a feature vector or tensor, in the hope of capturing and extracting more abstract and useful concepts that can improve performance on a range of downstream tasks* (Le-Khac, Healy, and Smeaton, 2020).

Representation Learning as a Formalized Learning Task: In practice, RL describes the process of learning a mapping $g : \mathcal{T} \to \mathcal{H}$ from the input space $\mathcal{T}$ to an abstract representation space $\mathcal{H}$ with desirable properties (Bengio, 2013; Bengio, Courville, and Vincent, 2013; Le-Khac, Healy, and Smeaton, 2020; Torralba, Isola, and Freeman, 2024).

This mapping is typically referred to as a feature encoder (Le-Khac, Healy, and Smeaton, 2020) and, in the case of NLP tasks, is implemented as a learnable parametric function $G_W(t)$ that transforms an input sequence $t = (t^{(1)}, \ldots, t^{(n)})$ of tokens $t^{(i)}$ into a corresponding sequence of token representations $(h^{(1)}, \ldots, h^{(n)})$. In this vein, discriminative tasks such as sequence classification require a single representation of the entire sequence that aggregates the information contained in $h^{(1)}, \ldots, h^{(n)}$, which, for simplicity, can be denoted as $h = G_W(t)$.

Although the mechanisms described above are consistent with the BERT architecture presented in Section 3.3, the notion of RL highlights that a model G_W learns its parameters W to encode meaningful representations of the input, typically starting from scratch.

In the case of *Pre-trained Language Models* (PLMs), however, as considered in this thesis, the weights W are not learned anew but are to be further refined and adapted under task-specific conditions that go further than the shallow adaptation alongside fine-tuning the classification head for a downstream task (Sun et al., 2020).

Hence, in such discriminative approaches to RL, the objective is to optimize the conditional probability $p(y \mid t)$ in the classification pipeline $t \xrightarrow{G_W} h \xrightarrow{C_{\hat{W}}} y$ by specifically aligning the parameters W of G_W with the general requirements of the task, i.e., to the semantics that disambiguate each $y \in \mathcal{Y}$ to be what it has to be (Bengio, 2013) before fine-tuning $C_{\hat{W}}$.

While in classical SFT the information formally flows back through the entire model, most adjustment signals originate from the lightweight FFN classification head $C_{\hat{W}}$ and thus provide rather unspecific guidance for fundamentally updating W (Merchant et al., 2020; Zhou and Srikumar, 2022). In contrast, the focus here lies on deliberately improving W, so that the adapted representations themselves become better suited for later downstream decisions.

Principles of a Desirable Representation: Unquestionably, it is difficult to define universal criteria for what makes a representation good. The answer always depends on the specific goal or application in mind. Still, key contributions (Bengio, 2013; Bengio, Courville, and Vincent, 2013; Le-Khac, Healy, and Smeaton, 2020; Torralba, Isola, and Freeman, 2024) in the field emphasize a set of general assumptions that guide how to think about such representations.

Before turning to these principles, it is important to emphasize that the properties of RL outlined here are based on the aforementioned key contributions, yet paraphrased and reorganized in a didactic manner to highlight their interrelations.

First, there are assumptions about the explanatory factors underlying data (factor-based assumptions). These assumptions concern the nature and organization of the latent explanatory factors that give rise to observable data. These are:

Multiple Explanatory Factors: Data is usually generated by a combination of latent, explanatory factors rather than a single source of variation. Effective representations should disentangle these factors, allowing the model to capture them in a structured way.

Sparsity: Building on the idea of multiple factors, sparsity ensures that only a few factors are active for any given example. In this way, redundancy is reduced, the most relevant variables are highlighted, and interpretability as well as efficiency of the learned representations are improved.

Depth and Hierarchy: Extending these assumptions, a central principle for building representations is their hierarchical organization. The guiding idea is that concepts in the real world are themselves often organized hierarchically, where more abstract notions stand above and capture less abstract ones.

Second, there are assumptions about the spatial and geometric organization of data in the representation space (space-related assumptions). In contrast, these assumptions focus on the geometric and spatial organization of data in the input and representation space. They describe how data points are distributed and what kinds of neighborhood relations should be preserved. These can be described as:

Manifold: The manifold assumption states that real-world data, although embedded in a high-dimensional input space, do not occupy this space uniformly. Instead, data points tend to concentrate on regions of much lower dimensionality that form a smooth, underlying manifold. A good representation should uncover and align with this low-dimensional structure, as it often reflects the true organization and explaining factors of the data.

Smoothness: Building on the manifold view, another assumption is that nearby inputs should also correspond to nearby representations. This smoothness principle ensures that the geometry of the learned space respects meaningful similarities in the input space.

Temporal and Spatial Coherence: In the same sense, underlying concepts are expected to vary only gradually across time or space, even if individual raw observations fluctuate more abruptly. This assumption reflects the idea that data trajectories evolve smoothly along the underlying manifold. Enforcing coherence therefore encourages representations to capture stable and slowly changing (robust) explanatory factors of variation.

Natural Clustering: Taken together, these assumptions suggest that similar data points should form natural clusters in the representation space. A good representation makes such clusters explicit, separating distinct groups while maintaining similarity within them.

Beyond these fundamental assumptions, there are additional properties that increase the usefulness of representations in practice. They highlight the potential for transfer across tasks and unlabeled data.

Shared Factors across Tasks: Building on these principles, many underlying factors are not confined to a single task but generalize across related problems. Representations that capture such shared structure are especially valuable, as they can be reused and facilitate transfer across tasks, topics, or even domains.

Semi-Supervised Utility: In the same spirit, a strong representation should already capture useful structures for unlabeled data and only require limited supervision to adapt to downstream tasks. This property is crucial whenever labeled data are scarce, but unlabeled data is abundant.

Taken together, these considerations lead to three overarching properties identified as the hallmarks of good representations (Le-Khac, Healy, and Smeaton, 2020). They summarize the outcome of the assumptions and desirable properties introduced above as:

Distributed: Drawing all these aspects together, good representations should use distributed encodings, where information is spread across many dimensions. This allows for greater expressive power compared to local or handcrafted representations.

Abstraction and Invariance: At the same time, they should capture abstract concepts that remain robust and invariant to irrelevant perturbations, such as noise or smaller changes.

Disentangled: Finally, they should separate independent factors of variation, making them more interpretable, reusable across tasks, and better aligned with human-understandable structures and concepts.

4.2 Turning TACO Definitions into Representations

Again, there are many approaches to learning representations, as best demonstrated by how examples like Word2Vec or the BERT-family learn to encode word meaning. In this contribution (Feger and Dietze, 2024b), however, the aim is different.

Here, the focus is on improving the representations of the TACO baseline model, which is a PLM already capable of encoding language and achieving good classification results but whose representations remain entangled in terms of the TACO classes. Yet, as has been shown in Chapter 3, about 43% of instances were misclassified at the instance level, although they still counted as true positives at the category level. For example, a Reason could be labeled as a Statement, and despite the misclassification, the outcome would still be correct in terms of assigning the broader category as an argument. This indicates that while both the annotations and the classification performance can be considered of high quality, the underlying representations themselves are not yet sufficiently disentangled.

Hence, the aim is to disentangle these representations so that they reflect the desired TACO hierarchy in a way that corresponds to human interpretation given the constituting elements of arguments and can subsequently also be used as supervision for models. In other words, the representations should first and foremost align with how humans interpret the argument constitution while at the same time being structured in a way that allows computational models to leverage them systematically. For this purpose, the principles of learning desirable representations outlined in Section 4.1 are to be considered.

In particular, this involves transferring the TACO hierarchy into a disentangled representation space of the baseline model while addressing abstraction and invariance, thereby fostering robust and generalizable properties that capture the essence of tweet classes, delineate their boundaries, and, respectively, reveal what distinguishes arguments from no-arguments.

It should be noted in advance that these representations are distributed due to their design, as they are derived from PLMs in general and from the TACO baseline in particular. Moreover, abstraction is partly a matter of modeling the relevant factors and properties of the desired representation space, while invariance and the remaining parts of abstraction are more closely tied to the training procedure itself and will be discussed later.

Consequently, this section first discusses how the desirable properties of the TACO hierarchy can be modeled within a corresponding theoretical representation space. Building on this, several practical considerations are then presented, including fundamental aspects that are not explicitly addressed in the implementation section of the contribution itself. In both cases, it is important to stress that both the modeling of the theoretical space and its practical realization are integral parts of this paper's contribution. The respective assumptions regarding, for example, the desired structure of this space and the actual disentanglement of the baseline's representations are ultimately subject to experimental validation, which, as will be shown in the paper, confirms the assumptions made in the following parts.

Deriving a Disentangled Representation Space from the TACO Hierarchy: Regarding the TACO dataset, this section focuses on mapping its tweet hierarchy into a theoretical embedding space so that properties and relations between classes and categories become specifiable according to the hierarchy's semantics.

To achieve this, the hierarchy, which already reflects a human-understandable structure, must be transformed into an equivalent spatial structure that is both measurable and metrically interpretable for computational models (Xing et al., 2002; Chopra, Hadsell, and LeCun, 2005; Hadsell, Chopra, and LeCun, 2006), thereby following the smoothness and cluster assumption. By design, this corresponds to an d-dimensional metric space $\mathcal{H} \subseteq \mathbb{R}^d$ as typically assumed in PLMs (Devlin et al., 2019) and in the context of RL (Bengio, Courville, and Vincent, 2013).

This transformation, shown in Figure 4.1, builds directly on the main characteristic of the TACO hierarchy, namely that a tweet can be annotated and interpreted according to its two constitutive elements, inference and information. By that, the decisive factor in determining whether an argument arises is the presence or absence of the inference component, while the categories are further subdivided into classes according to the presence or absence of the information component.

Formally, the category of arguments consists of the two subclasses Reason (presence of inference and information) and Statement (presence of inference but absence of information). In contrast, no-arguments are divided into Notification (absence of inference but presence of information) and the residual class None (absence of both inference and information).

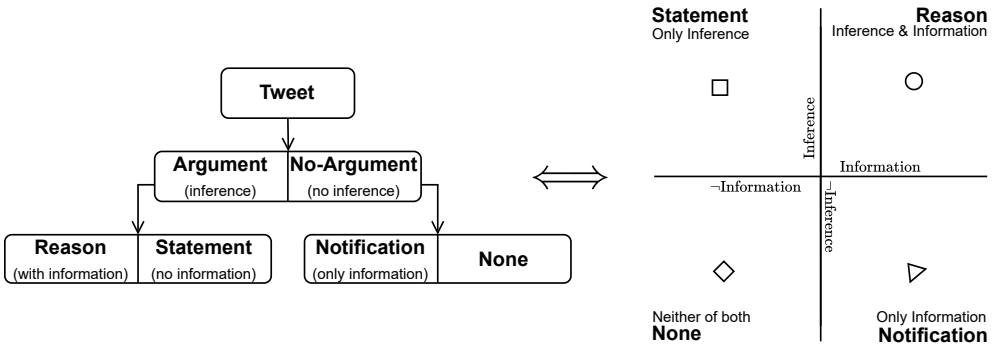


Figure 4.1: The TACO tweet hierarchy is mapped onto a representation space ($\mathcal{H} \subseteq \mathbb{R}^2$) defined by two axes: inference and information. Positive values indicate presence, negative values absence. The four quadrants correspond to classes (shown by symbols), with opposite classes (e.g., Reason vs. None) in orthogonal quadrants. Related classes (e.g., Reason and Statement) are adjacent along the information axis, while argument vs. no-argument categories are separated along the inference axis (upper vs. lower half).

Applied to the canonical structure of arguments from Subsection 2.2.2, a tweet within the TACO hierarchy can thus be represented as $h = \langle information, inference \rangle$.

For practical reasons, the values of this formal representation are treated as general variables, not as actual text or sets of text as in the examples presented in Subsection 2.2.2. Instead, following the principle of Section 4.1, they are expressed as numerical values to establish a bridge between the descriptive examples and a numerical vector in a metrical space.

Accordingly, a tweet can be placed along the information and inference axes, together forming a two-dimensional Cartesian space.

Considering these assumptions, four quadrants emerge, each corresponding to one of the four classes defined for TACO. The first quadrant, consisting of positive axis values, can be understood as the manifold where embeddings of the Reason class reside. The orthogonal third quadrant, containing only negative axis values, corresponds to the None class. The second quadrant represents embeddings of the Statement class, since it combines positive inference values with negative information values. Opposite to this, the Notification class is located, defined by positive information values but the absence of positive inference values.

Thus, it can further be assumed that the proposed space describes the location of a representation, i.e., in which class-quadrant the corresponding embedding resides.

More precisely, each dimension corresponds to one component, expressed in terms of positive or negative values (presence vs. absence) in the $\langle information, inference \rangle$-space. To illustrate the structure of the space, the ones-vector $\mathbb{1}_2 = \langle 1, 1 \rangle$ can be taken as a reference. Formally, any vector representation $h = \langle information, inference \rangle$ in this space can then be obtained by dimension-wise multiplication $h = \mathbb{1}_2 \odot s$ with a scaling vector $s \in \mathbb{R}^2 \setminus \{0\}$. In this, the zero vector is explicitly ignored, as otherwise a trivial solution or representational collapse would occur (Torralba, Isola, and Freeman, 2024). In turn, the magnitude of s determines the distance from the origin (zero vector), while the sign of its components specifies the quadrant.

Hence, the class membership of h is determined solely by its sign pattern. For example, $h = \langle 1, 1 \rangle$ lies in the Reason quadrant, $h = \langle -1, 1 \rangle$ in the Statement quadrant, $h = \langle 1, -1 \rangle$ in the Notification quadrant, and $h = \langle -1, -1 \rangle$ in the None quadrant.

Learning Desirable TACO Representations from Similarities and Contrasts: Following the modeling of the theoretical representation space, and with it the conception of how corresponding representations should be spatially arranged, this section turns to more practical aspects of the implementation. The focus is not on mathematical properties in depth but on the underlying ideas, as the formal details are addressed in the subsequent paper.

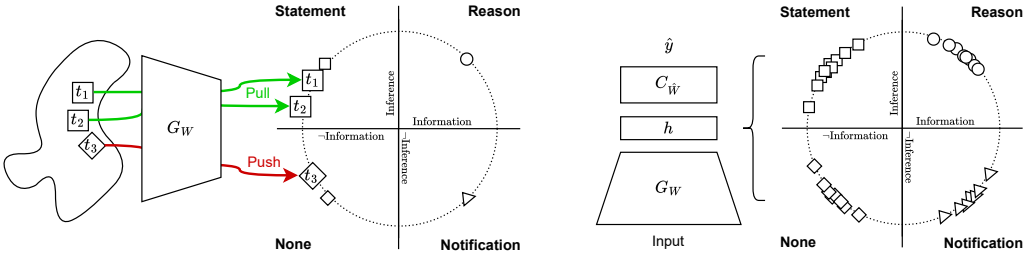
Taking a step back, it is important to note that the focus here no longer lies on the original task of instance-based classification into discrete labels but on comparing them regarding whether they should cluster according to their class semantics or be separated from one another, along with the corresponding implementation that adapts the representations accordingly (Le-Khac, Healy, and Smeaton, 2020).

In extension of Section 3.3, which states that classic SFT assumes pre-trained language representations are largely well structured and require only minor adjustments for better class alignment, the present objective instead calls for a deeper restructuring of the representation space to enable more advanced classifiers (Bengio, Courville, and Vincent, 2013; Le-Khac, Healy, and Smeaton, 2020). By that means, the goal is to disentangle the explanatory factors underlying the representation and to align them with the semantics that define each class so that not only within-class similarity but also contrasts between them become explicitly encoded.

This captures the core idea of *Contrastive Learning* (CL), where representations are learned such that similar instances are pulled closer together and dissimilar instances are pushed apart (Chopra, Hadsell, and LeCun, 2005; Chen et al., 2020; Le-Khac, Healy, and Smeaton, 2020).

The mechanism underlying this process is based on pairwise comparisons of input instances. Pairs that share the same label (e.g., two Reason tweets) are treated as positive pairs, and their embeddings are to be aligned (pulled together), while pairs with different labels (e.g., Reason vs. Statement) are treated as negative pairs, and their embeddings are repelled (pushed apart).

More specifically, CL is defined through alignment, denoting the pulling effects, and uniformity, denoting the separation effects (Wang and Isola, 2020). The latter is usually modeled on the unit hypersphere, also referred to as the n-sphere $S(n) = \{ h \in \mathbb{R}^{n+1} : \|h\| = 1 \}$, which constrains embeddings to move along its surface rather than expanding arbitrarily, since the unit length binds all vectors. In two dimensions, as in TACO, this reduces to stable trajectories along the circle $S(1)$. Distances are then understood not in terms of Euclidean magnitude but as angular separation. The hypersphere thus stabilizes representations and enforces both within-class similarity and between-class separation by assigning regions of angular proximity to clusters and leaving angular gaps between them.



(a) Contrastive Learning

(b) Optimized Representation Model

Figure 4.2: Comparison of (a) CL of the TACO representation space and (b) the optimized representation models. In (a), instances of the same class are pulled together and different ones pushed apart on the unit sphere, yielding a disentangled representation space whose quadrants abstract and align with the TACO hierarchy and which serves as the basis for further optimization during classification in (b). Augmented data is used during training to enforce abstraction and transformation invariance in the embeddings.

Taken to the intended space for TACO, these clusters ideally converge toward the outermost points on the unit circle $S(1)$ that best represent their respective classes, provided that attraction and repulsion effects are sufficiently strong. These outermost (central) points can be compared to the example embeddings from the previous section, where the space was theoretically constructed. For instance, $h = \langle 1, 1 \rangle$ for Reason or $h = \langle -1, 1 \rangle$ for Statement when projected onto $S(1)$. All these aspects can be comprehended as part of Figure 4.2a.

Put simply, CL enables further optimization of the TACO baseline PLM, aligning the representations more closely with class semantics before classical SFT. The optimized embedding space then forms clusters that capture these semantics, organizing data along stable angular trajectories toward maximally separated positions in the class quadrants of the intended space. More generally, even without directly relying on the original classification objective, a PLM with such an extended pre-training is already familiar with the elementary semantics of the classes and categories it will encounter. This is summarized in Figure 4.2b.

Learning Generalizable Representations Through Abstraction and Invariance: Besides CL, the principles of abstraction and invariance presented earlier in Section 4.1 have to be considered for learning representations as well.

Abstraction in general means moving away from details and instead focusing on the underlying structures or principles. It is about identifying what is essential while ignoring incidental aspects such as superficial or spurious associations (for example, arguments being tied to a specific topic) that can mislead the model (Clark et al., 2019; Geirhos et al., 2020; Thorn Jakobsen, Barrett, and Søgaard, 2021). In terms of RL and CL for TACO, this means that the baseline PLM should capture the core semantics of how the classes are constituted, the information they convey, and the inferences they support while avoiding a focus on surface-level features such as topics, names, or numbers.

For instance, take the rather simplistic argument, *We should study because it helps our future.* In another case one might say, *We should exercise because it improves our health.* Although the topics differ, both sentences share the same constitution, a beneficial outcome justifies a recommended action. What changes on the surface is the subject, but what remains constant is the relation between and presence of the inference and its information.

The same can be stated in more elaborate reformulations. Again, taking the simplistic argument, it can also appear as *People ought to work out since it benefits their well-being* or *In reason of its advantages for the body, exercising is important for us.* Here the pronouns, discourse markers, and stylistic choices vary, and even the order of the argument components may change, yet the underlying composition and idea of the argument remain intact. Even more extremely, one can also change the topic and the overall idea entirely, as in *We should smoke because it calms our nerves*, without altering the fact that it still represents an argument. This persistence across different forms is what abstraction aims to capture and can also be understood in terms of generalization more broadly (Bengio, Courville, and Vincent, 2013).

Building on this idea, invariance means that a disentangled representation remains stable under changes in external semantics (Wang et al., 2024), which implies coherence despite variations in irrelevant signals that co-occur in the data (Torralba, Isola, and Freeman, 2024).

Transferred to TACO, this means that instances of the same class are expected to share essential properties that remain stable even when certain features like topics, markers, or other surface features vary. This connects directly to the assumption of shared factors, which holds that different instances are governed by common latent factors that define their class membership. In line with this assumption, as well as with the principle underlying the semi-supervised utility assumption, good representations capture these fundamental properties so that they can transfer to multiple, potentially unseen, instances and their variations without having been explicitly included during training. Without such mechanisms, models risk overfitting to surface details of the training data and fail to learn the deeper regularities required for genuine task alignment (Geirhos et al., 2020; Thorn Jakobsen, Barrett, and Søgaard, 2021).

To address both these aspects of RL, training is performed on augmented rather than original data (Le-Khac, Healy, and Smeaton, 2020; Wang and Isola, 2020). Formally, let $T(t)$ denote a transformation for a given input $t \in \mathcal{T}$ that yields a correlated and semantically similar reformulation $t' = T(t)$. According to the smoothness assumption, if t and t' are similar ($t \approx t'$) and preserve the underlying semantics of the original label ($y = y'$), then the corresponding representations should remain close in the embedding space, i.e., $G_W(t) \approx G_W(t')$.

By enriching the training distribution in this way, the model is encouraged to focus on invariant properties of the data manifold rather than memorizing spurious details, thereby fostering robust generalization across different manifestations of the same class. Building on this perspective, the CL approach in this paper is extended more comprehensively by systematically using such augmented variations from the original data. Moreover, this reduces the dependency on the unaltered training samples and enables more realistic downstream experiments, since one can explicitly test whether models have genuinely captured the desirable properties (Zhuang et al., 2021) rather than simply memorizing the original data (Geirhos et al., 2020).

Marc Feger and Stefan Dietze.

"BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter"

In: Findings of the Association for Computational Linguistics: NAACL 2024, pages 2256–2266, Mexico City, Mexico. Association for Computational Linguistics.
Acceptance Rate: ∼12.5%

The following section turns to the specific paper (Feger and Dietze, 2024b), providing an overview of its content and summarizing the key findings regarding research question **Q2** and its sub-questions **Q2.1**–**2** as introduced in Section 1.2. After a brief discussion of the broader implications and significance of this work for the dissertation, the paper itself is presented.

4.3.1 Summary

This paper explores cross-topic generalization and task-specific RL for TACO, with an emphasis on enhancing the identification of the argument constituting elements of inference and information. This work is motivated by misclassifications between classes that belong to the same overarching category, for example, distinguishing Statement, which conveys only inference, from Reason, which combines inference with information but still functions as an argument even when confused. To address the pragmatic variability inherent in tweet language and to overcome the limitations of blurred representations that affect both performance and reliance in the TACO benchmark, the best-performing baseline model BERTweet is refined.

To push the predictive capabilities and foster cross-topic generalization, the study aims at more reliably encoding the presence or absence of inference and information within the embedding space of BERTweet, prioritizing their functional role in argument construction over superficial features like the number of words it takes to express them or topic-specific wording.

To address **Q2.1** (*Can the semantics of Twitter arguments be encoded and represented?*), a theoretical embedding space is constructed based on the hierarchy defined in the TACO benchmark, illustrating how ideal embeddings would be expected to cluster. This space is two-dimensional, with each axis representing either inference or information. The direction along each axis indicates the presence or absence of that component, resulting in four distinct sectors that correspond to the four classes defined in TACO.

Building on the theoretical embedding space as a target, this paper presents the training and parameterization process using a Siamese network with contrastive loss. The objective is to guide the learned embeddings along fixed trajectories on the unit sphere, encouraging the formation of clusters that tend toward their ideal, outermost points within each quadrant. The model is then trained on pairs of augmented tweets that incorporate randomized topic terms, subtle sentence variations, and the removal of Twitter-specific elements such as hashtags, while ensuring that the label semantics remain intact.

Within this architecture, different expressions of inference and information are compared based on semantic similarity via cosine distance as a measure, with similar instances brought closer together and dissimilar ones pushed apart in the embedding space. The final representation model is called WRAP.

Subsequent cluster analysis of the optimized embedding space of WRAP reveals that the tweet representations can, to a certain extent, be aligned with the theoretically expected ideal representations, reflecting the intended role of each class within TACO's hierarchy.

Guided by **Q2.2** (*How well do learned representations generalize across topics?*), downstream evaluations involved cross-validation on shuffled topics (closed-topic), as defined by the TACO baseline, and a leave-one-out setup where each topic was used once for testing while training on the others (cross-topic). In addition, ablated TACO baseline models with varying pre-training strategies, including standard cross-entropy and pre-training on the same augmented tweets used for WRAP, were evaluated under identical conditions. For interpretability and to evaluate the capabilities of PLMs, the same experiments were also conducted using SLMs, incorporating tweet length as the sole feature, along with a *Random Forest* (RF) that employs *Term Frequency-Inverse Document Frequency* (TF-IDF) features to assess word importance.

Thereby, the CL strategy of WRAP proves most effective, enabling strong generalization and robust representation, with macro F1 scores of 86.62% for inference, 86.30% for information, and 75.29% for the respective classes of TACO in the closed-topic setting, and 86.27%, 84.90%, and 73.54%, respectively, in cross-topic evaluation.

Regarding the sub-questions **Q2.1**–**2**, and in light of the modified variant of BERTweet, namely WRAP, two distinct insights emerge in relation to the overarching question **Q2** (*Do SOTA models inherently predict arguments in Twitter conversations?*). First, learning both the contrasts and similarities of semantics along the TACO hierarchy produces a sufficiently analogous embedding space, which improves the disentanglement of the $[CLS]$ representation in BERTweet and promotes invariance to irrelevant signals, thereby enhancing generalization across topics. Second, however, the results indicate that these properties of desirable representations do not hold for the classical SFT approach, where pre-trained embeddings are superficially adapted in any way necessary to fit the task.

Consequently, the results of this paper raise two open questions. The first concerns whether WRAP is transferable beyond Twitter and whether AM data from other domains can be applied to Twitter. The second addresses the extent to which the shortcomings of the classical sequence classification approach observed in this narrow study can also be found more generally in the broader literature on argument identification.

4.3.2 Importance and Impact on this Thesis

Developing robust AM models for Twitter's dynamic and diverse topics involves addressing those challenges identified for TACO, which point to representational model limitations that hinder predictive performance and undermine trust in the models' learned capabilities.

The findings suggest that task-specific pre-training, as implemented in WRAP, guided by argument theory, combined with corresponding methodological approaches, enables more robust and generalizable AM models for Twitter than those obtained with current best practices. As such, this work provides a foundation for critically evaluating the current SOTA in AM, especially in terms of the reliability of benchmark results and generalization capacity of PLM models beyond the Twitter context. Hence, this paper (Feger and Dietze, 2024b) provides the foundational insights that motivate a more rigorous evaluation of argument identification in general. This, in turn, gives rise to **Q3** (*Does reported progress in argument identification reflect genuine advances?*), which is addressed in the third main contribution of this thesis (Feger, Boland, and Dietze, 2025).

BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter

Marc Feger
HeiCAD - Heine Center for Artificial Intelligence and Data Science
Düsseldorf Germany
marc.feger@hhu.de

Stefan Dietze
GESIS - Leibniz Institute for the Social Sciences
Cologne Germany
stefan.dietze@gesis.org

Abstract

Argument mining, dealing with the classification of text based on inference and information, denotes a challenging analytical task in the rich context of Twitter (now 𝕏), a key platform for online discourse and exchange. Thereby, Twitter offers a diverse repository of short messages bearing on both of these elements. For text classification, transformer approaches, particularly BERT, offer state-of-the-art solutions. Our study delves into optimizing the embeddings of the understudied BERTweet transformer for argument mining on Twitter and broader generalization across topics. We explore the impact of pre-classification fine-tuning by aligning similar manifestations of inference and information while contrasting dissimilar instances. Using the TACO dataset, our approach augments tweets for optimizing BERTweet in a Siamese network, strongly improving classification and cross-topic generalization compared to standard methods. Overall, we contribute the transformer WRAPresentations and classifier WRAP, scoring 86.62% F1 for inference detection, 86.30% for information recognition, and 75.29% across four combinations of these elements, to enhance inference and information-driven argument mining on Twitter.

1 Introduction

Twitter (now 𝕏) is a global hub for opinions, news, and information and serves as a primary data source for research, which had already recognized the value of its user-generated content prior to its transition to 𝕏 (Kwak et al., 2010; Boyd et al., 2010).

Argument mining describes the process of classifying texts by assessing their written content in terms of information and inference elements to identify arguments (Palau and Moens, 2009; Peldszus and Stede, 2013; Lawrence and Reed, 2019).

In the intersection of traditional machine learning and natural language processing, pre-trained transformers like BERT (Devlin et al., 2019) and its specialized variants, such as BERTweet (Nguyen et al., 2020), have set state-of-the-art classification standards (Houlsby et al., 2019; Sun et al., 2019). During fine-tuning, transformers create universal text representations providing contextual features for a soft-max classifier, meaning additional layers on top of the pre-trained model that are jointly optimized for downstream tasks (Devlin et al., 2019).

Thereby, the field of argument mining has also witnessed the benefits of transformer models like BERT for cross-topic classification (Bhatti et al., 2021; Thorn Jakobsen et al., 2021) and argument similarity (Reimers and Gurevych, 2019; Reimers et al., 2019; Thakur et al., 2021).

Besides the common methods of adjusting the in-task performance through parameter tweaks (Lan et al., 2019; You et al., 2019) or incorporating augmentations (Feng et al., 2021; Thakur et al., 2021), multi-task learning is recommended as an additional fine-tuning strategy (Sun et al., 2019; Stab et al., 2018). Thereby, multi-task learning denotes a prior phase of fine-tuning representations on auxiliary tasks such as clustering or semantic similarity before proceeding to the actual classification step and is argued to effectively reduce a model's sensitivity to spurious correlations (Liu et al., 2019; Tu et al., 2020), which in turn is key to cross-topic argument mining (Thorn Jakobsen et al., 2021).

We believe that acquiring robust and meaningful representations, in the sense of perceiving the constituent elements of arguments, prior to classification is particularly useful for the nuanced task of argument mining when applied to diverse topics.

Generalizability in terms of cross-topic classification is crucial for practical argument mining in realistic scenarios, both in general research (Daxenberger et al., 2017; Stab et al., 2018) and specifically on Twitter (Schaefer and Stede, 2021), necessitating models to focus on argument components while avoiding reliance on spurious correlations like topic words (Thorn Jakobsen et al., 2021).

In this paper, we pioneer the optimization of the understudied transformer BERTweet for argument mining on Twitter. Thereby, we refine its representations of tweets within the embedding space by specializing BERTweet to better encode inference and information across diverse topics.

Utilizing the TACO dataset (Feger and Dietze, 2024), offering the first strong baseline evaluations of BERTweet for argument mining on Twitter, we optimize the model's representation layers in a multi-task approach by accentuating the contrast between inference and information while centering similar manifestations before the actual classification step, for which we assume proximity to imply shared class signals (van Engelen and Hoos, 2020).

We achieve this by configuring a Siamese BERTweet network using SBERT (Reimers and Gurevych, 2019). Applying contrastive loss (Hadsell et al., 2006) and text augmentation techniques (Wei and Zou, 2019), this network teaches BERTweet to cluster tweet embeddings according to their respective roles in argument mining, that is, to generally encode the presence or absence of both inference and information in those representations used for classification. Hence, we aim for classifications driven by the argument constituting elements, steering clear of spurious correlations.

Utilizing BERTweet's enhanced embeddings, it excels in both closed and cross-topic argument mining on Twitter, outperforming several standard methods (Schaefer and Stede, 2021) in this domain.

Towards inference and information-driven argument mining on Twitter, we contribute:[1]

- A pre-classification fine-tuning approach for BERTweet, enhancing its capacity to represent information and inference for closed and cross-topic argument mining on Twitter.
- An augmentation strategy to reduce spurious entity and topic signals while increasing sentence variability in tweets.
- WRAPresentations[2], an enhanced BERTweet embedding model driven by inference and information, obtained through contrastive optimization on augmented TACO tweets.
- WRAP[3], our tweet argument classifier leveraging WRAPresentations for argument mining across diverse topics on Twitter.

[1] github.com/TomatenMarc/TACO-Fiesta
[2] huggingface.co/TomatenMarc/WRAPresentations
[3] huggingface.co/TomatenMarc/WRAP

2 Twitter Arguments from Conversations

Our primary dataset[4], TACO (Feger and Dietze, 2024), encompasses 1,734 tweets from 200 entire conversations spanning six topics: #Abortion (25.9%), #Brexit (29.0%), #GOT (11.0%), #LOTR-ROP (12.1%), #SquidGame (12.7%), and #TwitterTakeover (9.3%). So far, it stands as the sole publicly available labeled tweet dataset tailored for inference and information extraction, strategically addressing reply-patterns inherent to their emerging conversational contexts during annotation.

Annotations were conducted by six experts according to the Cambridge Dictionary definitions, differentiating *inference* as *a guess that you make or an opinion that you form based on the information that you have* and *information* as *facts or details about a person, company, product, etc*. With a robust agreement of 0.718 Krippendorff's α, four classes emerged of these elements: *Reason* (inference and information), *Statement* (inference without information), *Notification* (information without inference), and *None* (neither element).

Table 1 details the class distribution of TACO.

Reason	Statement	Notification	None
581 (33.50%)	284 (16.38%)	500 (28.84%)	369 (21.28%)

Table 1: The class distribution of tweets in TACO.

On TACO, Vanilla BERTweet serves as the best performing baseline, excelling with 74.45% F1 for Reason, 56.66% F1 for Statement, 78.30% F1 for Notification, and 80.56% F1 for None after fine-tuning on these classes (Feger and Dietze, 2024).

3 Inference and Information-Driven Representations for Mining Arguments

In text classification, transformers like BERTweet use the final hidden state of the first token $[CLS]$ as the sequence representation. Classification involves a soft-max classifier added as an extension after the final representation layer, determining the label assignment for a tweet t by evaluating the probability of each possible label y as:

$$p(y|h) = softmax(\hat{W}h) \quad (1)$$

where, $\hat{W}$ signifies the task-specific weights of the classification head, and h represents the final representation of t obtained with the transformer. Achieved through pooling an entire sequence representation via $[CLS]$, h is expressed as

[4] github.com/TomatenMarc/TACO

$G_W(t) = h$, where the transformer is considered an independent function $G_W(t)$ with its distinct weights W, taking t as input. For the specific classification task, both $\hat{W}$ and W are jointly fine-tuned by maximizing the log-probability of the correct label, where h implicitly undergoes optimization.

For optimizing class assignments on TACO, we emphasize the impact of specializing h for encoding inference and information before classification.

Hence, we consider the pre-classification specialization of an embedding h as a contrastive problem of semantic similarity, where tweets with similar expressions of the text dimensions inference and information are brought closer together, while those lacking in similarity are positioned farther apart.

3.1 Embedding Inference and Information

We measure the semantic similarity between two tweet representations, denoted as h_1 and h_2, using cosine distance:

$$D(h_1, h_2) = 1 - cos(h_1, h_2) \in [0, 2] \quad (2)$$

a standard metric (Mikolov et al., 2013; Kim, 2014; Tai et al., 2015; Chen and He, 2020) for assessing text vector similarity. $D(h_1, h_2)$ reflects complete equivalence at 0, orthogonality at 1, and absolute dissimilarity at 2. Mainly defined by the cosine similarity $cos(h_1, h_2) \in [-1, 1]$, where -1 represents complete dissimilarity, 1 indicates equivalence, and values closer to 0 suggest orthogonality, this distance is length-independent and primarily influenced by the angle between two embeddings.

Building on this circumstance, we assume that the actual representation h of a tweet can be normalized and lies on the n-sphere:

$$S(n) = \{h \in \mathbb{R}^{n+1} : \|h\| = 1\} \quad (3)$$

Transferred to the Cartesian nature of arguments $h = \langle information, inference \rangle$, we consider their representations to live on the unit sphere $h \in S(1)$ (Wang and Isola, 2020; Khosla et al., 2020; Chen and He, 2020). In h, 1 signifies full presence, and -1 implies total absence of a component. Consequently, an ideal class center on the unit sphere heads towards the pole $\langle 1, 1 \rangle$ for Reason, $\langle -1, 1 \rangle$ for Statement, $\langle 1, -1 \rangle$ for Notification, and $\langle -1, -1 \rangle$ for None. A breakdown of this is shown in the upper part of Figure 1, acknowledging the realistic expectation that the actual embeddings may differ from the ideals while the objective is to get them closer to them.

3.2 Contrastive Siamese Network

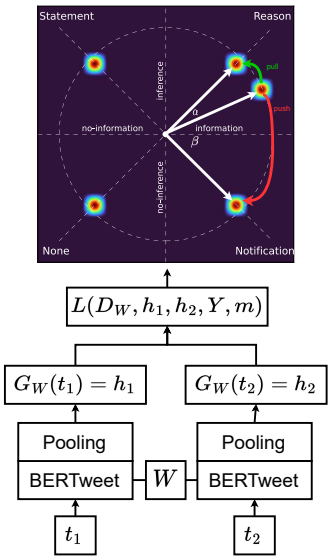


Figure 1: Visualization of the employed Siamese BERTweet architecture, with parameterized cosine distance $D_W(h_1, h_2)$ and contrastive loss $L(D_W, h_1, h_2, Y, m)$. Atop this architecture, the Cartesian embedding space for an argument representation $h = \langle information, inference \rangle$ is presented as target.

To address semantic similarity, a prevalent strategy involves enhancing representations through learning a metric (Chopra et al., 2005; Xing et al., 2002; Hadsell et al., 2006). Precisely, metric learning entails the implicit acquisition of a metric $D_W(h_1, h_2)$ parameterized by the weights W of the representation model G_W (Chopra et al., 2005).

We seek to find W such that the target metric:

$$D_W(t_1, t_2) = 1 - cos(G_W(t_1), G_W(t_2)) \quad (4)$$

is smaller if t_1, t_2 are semantically similar, and higher if not.

By utilizing the identical embedding function $G_W(t)$ (BERTweet) with shared weights W to learn the metric, our architecture is referred to as a Siamese network (Bromley et al., 1993; Chopra et al., 2005). Similar and dissimilar tweet pairs are provided as input to this network. To update the weights and optimize the network's performance, a loss function is applied on top of this architecture.

To attain the goal of increasing the differentiation between similar and dissimilar pairs, it is suggested to employ the contrastive loss (Chopra et al., 2005; Hadsell et al., 2006):

$$L(D_W, h_1, h_2, Y, m) = (Y)\frac{1}{2}D_W(h_1, h_2)^2 + (1-Y)\frac{1}{2}\{max(0, m - D_W(h_1, h_2))\}^2 \quad (5)$$

where, h_1, h_2 are two representations ($G_W(t_i) = h_i$) of different tweets t_1, t_2 to be optimized given $D_W(h_1, h_2)$ as metric. Y denotes the binary label indicating if t_1, t_2 are similar ($Y = 1$) or contrasting ($Y = 0$). Furthermore, a margin value $m > 0$ is introduced as the minimal distance between two contrasting tweets.

When establishing m, our objective was to set $D_W(h_1, h_2)$ in a way that maximizes contrast between dissimilar pairs while avoiding overestimation of their true distance. Focusing on $D_W(h_1, h_2) \in [0, 1]$, representing positive similarity, we selected $m = 0.5$. This choice intuitively represents the minimum threshold for high similarity, yielding optimal results in our study.

With $m = 0.5$ we ensure that even if a representation closely matches an ideal center but is labeled as dissimilar, the optimized representation pushes 60° away and into an adjacent quadrant.

3.3 Augmentation of TACO

In the initial phase of processing TACO data, we generated a unique copy for each tweet through augmentation, denoted as A-TACO. Employing EDA (Easy Data Augmentation) techniques (Wei and Zou, 2019) of (1) synonym replacement, random (2) insertion, (3) swap, and (4) deletion, this procedure segregates our total ground truth into A-TACO, for optimization the embedding space of BERTweet prior to classification, and TACO, designated for fine-tuning and evaluating classifiers.

Maintaining independence between optimization and evaluation data is crucial to avoid further spurious correlations (Thorn Jakobsen et al., 2021) and ensure that the data includes essential class signals, thus enabling broad generalization across varying sentence structures and cross-topic evaluations.

Following technique (1), we utilized spaCy[5] to automatically identify as many entities and preselected keywords related to the six topics in the TACO dataset as possible. Subsequently, we replaced these words with the $[MASK]$ token, a placeholder commonly used by BERT-like models, including BERTweet, for predicting missing words.

Particularly, we utilized BERTweet as a fill-mask model to generate new tokens for those masked in the input sequence (Kumar et al., 2020).

In order to increase the variability of word choice and sentence structure while minimizing semantic changes, the techniques (2-4) were applied to 10-90% of all words. Optimal coherence, with an average cosine distance of ~ 0.08 between the $[CLS]$ tokens of tweets and augmentations, is seen at a replacement rate of 10%, maintaining semantic consistency with entity and topic words being almost entirely changed or removed. Again, step (1) was applied to avoid reintroducing topic words. Refer to Table 2 for an augmentation example.[6]

TACO	**Elon Musk** ready with 'Plan B' **if Twitter** rejects his offer **Read @USER Story | HTTPURL #ElonMusk #ElonMuskTwitter #TwitterTakeover HTTPURL**
A-TACO	**Wenger** ready with 'Plan B' **as Wenger** rejects his offer **- HTTPURL via @USER**

Table 2: An augmented Notification reminiscent of a general blog comment after replacing entities (Elon Musk and Twitter are changed to Wenger), deleting topic or entity references, including hashtags, and rewording the tweet while retaining its original substance.

4 Experimental Setup

This section outlines the protocols used for evaluating and optimizing BERTweet's embedding space with A-TACO and follow-up classification on TACO. We select macro F1 scores[7] for evaluation in response to the imbalanced distribution across TACO's four classes, guaranteeing an equitable analysis and underscoring a model's adeptness at managing heterogeneous data distributions. In our subsequent classification analysis, we also present the micro F1 scores[7] for each tweet class. Beyond this, we consider Recall to account for the generalizability of a model to unknown topics after fine-tuning in the pre-classification phase.

4.1 Models

In our approach, it is important to differentiate between the pre-classification fine-tuning for specializing embeddings and their subsequent fine-tuning tailored for mining arguments on TACO. In this context, we compare different ablations of our fine-tuning pipeline for embeddings before and upon classification, comparing their prediction strength with various common baseline models.

[5] spacy.io

[6] For more examples, see: README.md

[7] Precision and Recall for experiments are in the repository.

For both tasks, we utilize the Vanilla BERTweet model[8], with 12 transformer blocks and 12 self-attention heads processing sequences of up to 128 tokens, consistent with the best performing model reported for TACO (Feger and Dietze, 2024).

The first embedding model derived from Vanilla BERTweet, enhanced as described in Section 3 by applying contrastive loss within the Siamese network utilizing A-TACO to improve the cosine distance $D_W(t_1, t_2)$ for similar or dissimilar tweets, is referred to as WRAPresentations. For comparison, we introduce a second derivative, Augmented BERTweet, which undergoes pre-classification fine-tuning using the same tweets of A-TACO as WRAPresentations but directly optimizes $p(y|h)$ with standard cross-entropy loss.

Both these strategies aim to improve the representation $G_W(t) = h$ of any tweet t used for subsequent classification $p(y|h)$ on TACO by incorporating augmented tweets of A-TACO and adjusting the internal weights W in different ways to better encode argument components for each model G_W.

For classification on TACO, we utilize TF-IDF representations, where word frequency is widely recognized as a feature in strong baselines for argument mining on Twitter, which are Support Vector Machine (SVM) (Addawood and Bashir, 2016), Logistic Regression (LR) (Bosc et al., 2016; Dusmanu et al., 2017), and Random Forest (RF) (Dusmanu et al., 2017). These models go beyond considering individual words by incorporating tweet-related features like emoji, URL, and hashtag frequencies. Despite this, their potential for cross-topic generalizability remains unexplored.

For each classifier, we evaluate the average class length for classification to examine linguistic feature acquisition.

4.2 Pre-Classification Fine-Tuning

To enhance BERTweet's embeddings, we chose TACO's golden tweets with flawless annotation agreement, accounting for 70.3% of all tweets, with class distribution remaining largely consistent.

For the final evaluation, we employed the original golden tweets for #Abortion but augmentations of golden tweets for the remaining five topics during fine-tuning. #Abortion was deemed as holdout topic due to its highest dissimilarity when compared to the remaining topics, posing a greater classification challenge (Thorn Jakobsen et al., 2021). This provided initial insights into cross-topic generalization and the efficacy of fine-tuning with augmentations and predicting given real tweets. Pairs were formed for all tweet combinations, denoting tweets of the same class as similar $Y = 1$ and those of different classes as dissimilar $Y = 0$, yielding more dissimilar than similar pairs.

For the final validation set, 86,142 pairs were generated. The optimization data, divided into fine-tuning and test sets with a stratified $60/40$ ratio, yielded 307,470 and 136,530 candidate pairs, respectively. To ensure a balance between similar and dissimilar pairs, we chose the largest possible set such that both similar and dissimilar pairs are equally represented (Bromley et al., 1993; Chopra et al., 2005) while maintaining all tweets of the respective splits.

In total, 162,064 pairs were obtained for fine-tuning, 71,812 for testing, and 53,560 for final validation of the enhanced BERTweet representations prior to classification.

For all transformer models, we performed fine-tuning over 5 epochs using an A100 GPU with 40 GB of memory, a batch size of 32, and a learning rate of $4e^{-5}$, which proved to be optimal for all models. The Siamese BERTweet network is implemented using SBERT (Reimers and Gurevych, 2019) as depicted in the lower part of Figure 1.

Additionally, we applied different fine-tuning strategies for WRAPresentations using both $[CLS]$ pooling, later used for classification, and $[MEAN]$ pooling, recommended for better sentence embeddings (Reimers and Gurevych, 2019).

4.3 Argument Mining on TACO

We evaluate the practicality of BERTweet's specialized embeddings on TACO, given the three argument mining tasks of (1) inference detection, (2) information recognition, and (3) classification of all four tweet classes, with a concurrent aim for cross-topic generalization.

For task (3), we trained a feed-forward neural network with two linear layers on top of each embedding model, undergoing 5 additional fine-tuning epochs with the best performing parameters having a learning rate of $4e^{-5}$ and batch size of 8, corresponding to the best model and parameters reported for TACO (Feger and Dietze, 2024). Again, we used a single A100 GPU with 40 GB of memory. Thereby, the results for tasks (1) and (2) are aggregations specific to class elements of task (3) predictions, focusing on inference or information.

[8] huggingface.co/vinai/bertweet-base

Extending our ablation strategy, classifiers were evaluated in two different setups to investigate the general effects of fine-tuning embeddings prior to classification and their subsequent adaptability to actual class signals (Peters et al., 2019).

In the first setup (Frozen), freezing embeddings allowed us to assess the benefits attributable to pre-classification fine-tuning. In the second setup (Dynamic), embeddings underwent further fine-tuning during classification head optimization, where we assessed their adaptability to task-specific learning. Success in both setups signifies a model's ability to represent argument components prior to classification and to adapt these fine-tuned representations to the specific classes of inferences and information.

We employed a 6-fold shuffled cross-validation, maintaining consistent splits for all classifiers across the six topics of TACO, to establish an upper-bound (Thorn Jakobsen et al., 2021). This closed-topic validation was then compared with cross-topic validation, where each of the six topics served as a unique testing set, and the remaining five topics were utilized for fine-tuning (Bosc et al., 2016; Daxenberger et al., 2017; Stab et al., 2018). Lower performance is expected in cross-topic validation, as classifiers are exposed to unseen topics.

5 Results

In this section, each model is investigated with respect to the actual tweets of TACO. First, we assess the embeddings of each transformation model in terms of their baseline notion of argument components and in terms of the four tweet classes, focusing on the structural differences of their representations. Second, we evaluate the different models in both closed and cross-topic classifications to determine their applicability to, and generalizability across, topics.

5.1 Results: Pre-Classification Fine-Tuning

Model	P	R	F1
Vanilla BERTweet-$[CLS]$	50.00	100.00	66.67
Augmented BERTweet-$[CLS]$	65.69	86.66	**74.73**
WRAPresentations-$[CLS]$	**66.00**	84.32	74.04
WRAPresentations-$[MEAN]$	63.05	**88.91**	73.78

Table 3: Evaluation of within-class similarity and between-class separability of all transformer models using $[CLS]$ tokens as used during classification. These models were fine-tuned with A-TACO and evaluated on the TACO holdout topic #Abortion. Suffixes indicate pooling methods for optimizing the embedding spaces.

After pre-classification fine-tuning to enhance semantic similarity, we evaluate the optimized embedding models for classifying tweet pairs as similar or dissimilar given $D_W(t_1, t_2)$.

All fine-tuning strategies outperformed Vanilla BERTweet in terms of F1, compare Table 3.

We excluded WRAPresentations with $[CLS]$ pooling for follow-up classification due to the absence of discernible benefits in F1 compared to Augmented BERTweet and WRAPresentations using $[MEAN]$ pooling for pre-classification fine-tuning, each showing higher Recall scores.

Hence, we will refer to WRAPresentations-$[MEAN]$ as WRAPresentations.

In comparing Augmented BERTweet and WRAPresentations, both models show similar overall performance in terms of F1, but diverge in their emphasis on Precision and Recall. The results suggest that contrastive fine-tuning of representations is not inherently superior to directly optimizing $p(y|h)$ with augmented tweets. However, this strategy enhances Recall, with further distinctions expected in downstream task evaluations.

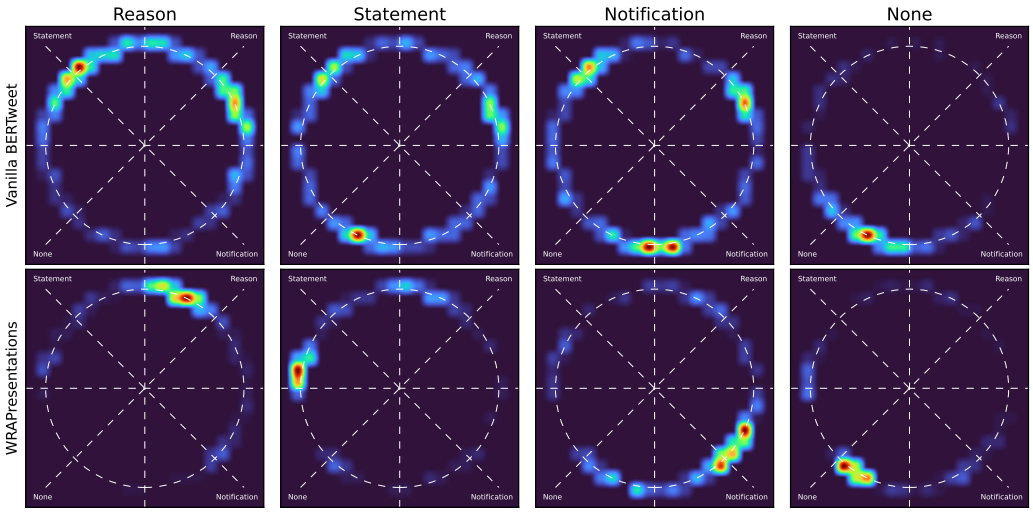
Nonetheless, we assume that the enhanced Recall at this stage is already a first indicator for later generalizations of classifications across topics. Additionally, we confirmed the effectiveness of pre-classification fine-tuning with A-TACO when applied to real tweets from an unseen topic.

Furthermore, we visually explored BERTweet's embedding space before and after fine-tuning, utilizing $[CLS]$ representations of all original tweets in TACO, as depicted in Figure 2(a).

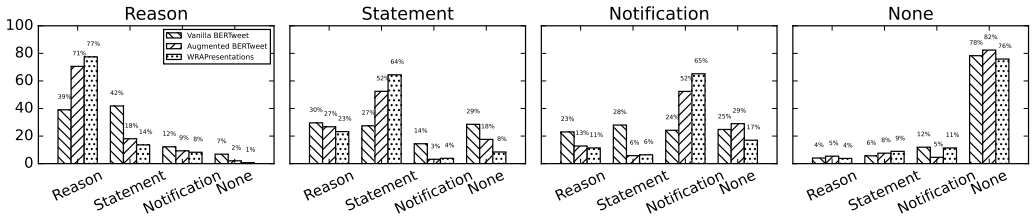
Applying t-SNE for dimensional reduction (van der Maaten and Hinton, 2008; Jawahar et al., 2019), comparing Vanilla BERTweet with WRAPresentations showed enhanced class quadrant density, compare Figure 2(a), suggesting an improvement of class semantics given inference and information for a majority of tweets. Similar patterns, albeit at smaller numbers, are observed for Augmented BERTweet, see Figure 2(b).

Numerically, WRAPresentations improved tweet order by 38% for Reason, 37% for Statement, and 41% for Notification over Vanilla BERTweet. Despite a -2% decrease in the None class quadrant, None remains predominant, refer to Figure 2(b).

Augmented BERTweet closely matches WRAPresentations in representing tweets, excelling by 6% for None but lagging behind by -6% for Reason, -12% for Statement and -13% for Notification.



(a) t-SNE embeddings of tweet class $[CLS]$ tokens before and after fine-tuning given inference and information.



(b) Distribution of classes within the projected quadrants of the expected $\langle information, inference \rangle$ space.

Figure 2: Investigation on the impact of BERTweet's fine-tuning for the transfer of class semantics onto the expected $\langle information, inference \rangle$ space in terms of the $[CLS]$ tokens for tweet classification. Considering the classes, (a) highlights the tightening of tweet embeddings towards their respective ideal class poles. Considering the distribution of tweets, (b) emphasizes that each expected quadrant corresponds to the anticipated majority class.

Model	Inference		Information		Multi-Class	
	Frozen	Dynamic	Frozen	Dynamic	Frozen	Dynamic
Closed-Topic (6-fold) Validation						
Length	62.34		71.47		38.26	
RF + TF-IDF	76.12		80.56		55.65	
Vanilla BERTweet	73.12	84.54	66.49	83.55	42.87	71.05
Augmented BERTweet	84.49	**86.68**	79.22	84.57	67.07	73.80
WRAPresentations	**86.88**	86.62	**81.54**	**86.30**	**71.07**	**75.29**
Cross-Topic (6-fold) Validation						
Length	61.99		71.55		38.17	
RF + TF-IDF	73.93		80.16		53.29	
Vanilla BERTweet	70.28	83.15	66.15	82.22	39.00	68.12
Augmented BERTweet	84.20	84.25	79.38	83.31	66.41	69.99
WRAPresentations	**86.83**	**86.27**	**81.54**	**84.90**	**70.93**	**73.54**

Table 4: Macro F1 scores of each classifier for inference and information detection, and all four classes.

Model	Reason		Statement		Notification		None	
	Frozen	Dynamic	Frozen	Dynamic	Frozen	Dynamic	Frozen	Dynamic
Closed-Topic (6-fold) Validation								
Length	61.68		20.19		14.47		56.72	
RF + TF-IDF	69.35		17.30		63.35		72.62	
Vanilla BERTweet	66.05	74.98	*00.00*	53.99	43.80	77.62	61.63	77.62
Augmented BERTweet	74.50	76.82	49.53	58.37	70.95	**80.28**	73.29	79.71
WRAPresentations	**77.34**	**78.14**	**58.66**	**60.96**	**72.61**	79.36	**75.67**	**82.72**
Cross-Topic (6-fold) Validation								
Length	61.78		19.32		14.49		57.09	
RF + TF-IDF	68.61		13.33		62.75		68.46	
Vanilla BERTweet	63.57	73.15	*00.00*	47.40	35.79	74.92	56.64	77.01
Augmented BERTweet	75.18	75.10	46.34	51.74	71.61	75.71	72.50	77.42
WRAPresentations	**77.13**	**77.05**	**57.62**	**58.33**	**73.05**	**78.45**	**75.91**	**80.33**

Table 5: Micro F1 scores for classifiers identifying the four classes in inference and information detection.

5.2 Results: Classification and Generalization

For simplicity, we present the outcomes of the RF classifier as best performing baseline and the average class length as minimal-performance indicator.

When turning to the closed-topic validation, WRAPresentations outperforms all classifiers except task (1), where dynamic embeddings in Augmented BERTweet exhibit performance nearly equivalent, as demonstrated in the upper half of Table 4. Quantitatively, WRAPresentations yields 86.88% F1 for task (1), 81.54% F1 for task (2), and 71.07% F1 for task (3) when frozen. Dynamically optimizing embeddings, WRAPresentations achieves 86.62% F1 for task (1), 86.30% F1 for task (2), and 75.29% F1 for task (3).

Shifting our attention to the more demanding task of cross-topic validation, assessing a classifier's ability to generalize to unseen topics, WRAPresentations demonstrates superior performance over all evaluations, thereby achieving 86.83% F1 for task (1), 81.54% F1 for task (2), and 70.93% F1 for task (3) when frozen. With dynamically adjusted embeddings, it achieves 86.27% F1 for task (1), 84.90% F1 for task (2), and 73.54% F1 for task (3), compare lower half of Table 4.

Further, WRAPresentations clearly improved performance for Statement, the least common and most difficult class to predict when comparing the remaining classifiers. Thereby, all other classifiers perform below or slightly above chance agreement for closed-topic validation and generalization across topics for this class, where Vanilla BERTweet even achieved 00.00% F1 when frozen, showcasing the necessity for adjusting classifiers and embeddings to specific classes, see Table 5.

6 Discussion

WRAPresentations consistently outperforms all models, except for a marginal -0.06% F1 decrease compared to Augmented BERTweet with dynamic representations for task (1) of closed-topic evaluation, while totally excelling across topics.

Augmented BERTweet performs stronger in detecting instances without inference, as demonstrated by the substantial 9.33% F1 increase for the Notification class with dynamic embeddings, see upper half of Table 5. Considering that tasks (1) and (2) are aggregations derived from the results of task (3), WRAPresentations enhances the overall performance of task (3) for achieving the best results, prioritizing an improvement in task (2) while incurring a slight decrease in task (1).

This effect emerges as further refinements for additional classification improvements can partially overwrite the enriched representations of inference and information in tweets, exposing unconsidered class signals during optimization of the head.

However, examining WRAPresentations' frozen states, superior in closed and cross-topic validation, underscores the advantages of our pre-classification fine-tuning focused on semantic similarity in tweets for enhanced classification strength, see Table 4, 5.

Supported by these cross-validated results, it appears that WRAPresentations can establish robust inference and information-driven representations for tweet classification, owing to our multi-task approach for systematically contrasting the argument-constituting elements in its embedding space, demonstrating adaptability and generalizability for all three argument mining tasks on Twitter, including the difficult Statement identification.

7 Conclusion and Ongoing Work

Our pre-classification multi-task fine-tuning approach considerably improves the specification of embeddings of BERTweet to encode diverse manifestations of inference and information, especially supporting the classification of tweets in TACO.

Enhanced by contrastive learning of semantic similarity, BERTweet's optimized embeddings excel a diverse range of argument mining approaches for Twitter, showcasing superior adaptability to class signals and cross-topic generalization.

In this regard, we can successfully contribute WRAPresentations, a contrastively optimized embedding model, and the advanced classification model WRAP for inference and information-driven argument mining across diverse topics on Twitter.

We also provide grounds for assuming that the augmentation of tweets constitutes a valuable asset within this domain of research.

Given our successful pre-classification fine-tuning with augmented tweets showing strong impact towards original tweets, we pose the two broader questions for argument mining regarding: (1) the necessity of using tweets for detecting arguments on Twitter, requiring investigation of whether tweet-like instances from other domains alone are sufficient, and (2) whether WRAPresentations or our contrastive learning approach can be transferred to build strong classifiers for domains other than Twitter.

Limitations

For our work, we report the following limitations:

The field of argument mining on Twitter is subject to Twitter's strict data regulations, which allow only the publication of tweet identifiers but not their text. The costly Twitter API, offering only 1,500 free queries per month, complicates research reproducibility and risks data loss from deleted tweets when fetched by their identifiers. For this study, we used the TACO dataset from our previous study, which gave us full access to the data. Access to the source dataset can be granted on request for non-harmful research purposes, subject to appropriate and mandatory data protection agreements.

Acknowledgments

We are very grateful for the attentive and constructive feedback from our anonymous reviewers. Your willingness to share your expertise and provide constructive feedback is deeply appreciated.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. 2021. Argument mining on Twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024. TACO – Twitter Arguments from COnversations. *Preprint*, arXiv:2404.00406.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR*, abs/2004.11362.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, WWW '10, page 591–600, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7:1–31.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume abs/2005.10242.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Reducing BERT pre-training time from 3 days to 76 minutes. *CoRR*, abs/1904.00962.

Chapter 5

Evaluating the Generalizability of Argument Mining Datasets and Models

In this chapter, the third and final paper (Feger, Boland, and Dietze, 2025) is discussed, focusing on the generalizability of the TACO annotations, the improved representation model WRAP, and other existing baseline models that define the SOTA of AM more broadly. The chapter begins with an introduction to aspects of generalizability, in particular the problem of models learning properties that are not truly related to the task itself but rather to hidden shortcuts that nonetheless lead to seemingly good results. Following the discussion of the paper itself, a dedicated section addresses the fundamental issue of limited generalizability, highlighted as an open challenge in the study, and outlines a resulting shared task.

5.1 Shortcut Learning

Can horses do math? What may sound like an odd question in the context of an academic thesis attracted considerable public attention in the early 20th century with the case of Clever Hans (Pfungst, 1907). In fact, Hans was the horse of Mr. von Osten, a German mathematics teacher. By tapping his hoof, Hans appeared able to add, subtract, multiply, and divide, leading Mr. von Osten to insist that Hans was genuinely intelligent.

However, systematic investigation discovered that Hans failed whenever he could not see the questioner or when the questioner did not know the correct answer (Pfungst, 1907; Rosenthal, 1966). Yet, Hans consistently succeeded when the questioner was present and aware. It was thus concluded that Hans was not actually solving problems but responding to subtle, unconscious cues such as slight changes in posture, facial expressions of excitement when he reached the correct answer, or bodily tension that signaled when to stop tapping.

This story of Clever Hans is therefore a classic example of unconscious cues (Clever-Hans-cues) or experimenter effects (Rosenthal, 1966) and demonstrates a cognitive bias in which successful performance in tasks associated with human intelligence is easily mistaken for evidence of comparable cognitive abilities (Marr, 1982). Beyond the above, it demonstrates that it is sometimes difficult for the observer to imagine that a complex challenge could be solved in a fundamentally different way than the human way (Geirhos et al., 2020).

A similar assumption persists in ML as well. Although artificial neurons, as in a DNN, differ substantially from biological neurons while being inspired by their architecture (Rosenblatt, 1958), their success that drives the hype of modern ML, especially around LLMs like ChatGPT, naturally invites the belief that they rely on information in the same way humans do (Geirhos et al., 2020).

In recognition of this, certain research, for example, indicates that LLMs and the human brain utilize similar prediction and context mechanisms, yet these are mere correlations, not shared causal mechanisms (Caucheteux, Gramfort, and King, 2022; Goldstein et al., 2022).

However, while showing that clear differences exist at the implementation level, there is often a tacit assumption at the algorithmic level that human-like performance implies a human-like strategy for solving a problem (Marr, 1982; Funke et al., 2020; Geirhos et al., 2020).

This phenomenon can be thought of as a symbol for problems of dataset biases (Torralba and Efros, 2011), or for learning in an anti-causal direction (Schölkopf et al., 2012), all of which can be summarized under the notion of *Shortcut Learning* (SCL) (Geirhos et al., 2020).

Put simply, SCL does not offer a mathematical formulation but describes the general problem that models, like Hans, often perform well not by solving the intended task but by exploiting hidden cues or spurious correlations in the data. These cues act as shortcuts that fail once they are absent or altered, exposing a lack of genuine task alignment and generalization, understood as having learned an ability that is truly of interest and that is invariant to mere memorization of a specific dataset or reliance on its particular quirks (Geirhos et al., 2020).

5.2 Constraints of Evaluating the State-of-the-Art

In the wider sense, SCL can be understood as a fundamental limitation of the standard approach to training and testing models. Standard practice assumes that datasets can be divided into subsets that serve different functions. This partitioning is typically done at random, splitting the ground truth into training, validation, and test subsets (Kohavi, 1995; Hastie, Tibshirani, and Friedman, 2009; Geirhos et al., 2020).

A common split, for example, is 60/20/20, where 60% percent of the data are used for training, 20% for validation during training, and 20% are held out for final testing. Alternative approaches such as k-fold cross-validation, in which the data are divided into k-folds, are essentially based on the same principle. Each fold is used once for testing, while the remaining $k-1$ folds are used for training. In practice, these samples are drawn randomly from the ground truth and assumed *independent and identically distributed* (i.i.d). This means, although being disjoint samples of the data, they still share similar characteristics. Such evaluations therefore measure performance within the closed scope of a benchmark rather than under real-world *out-of-distribution* (o.o.d) shifts (Torralba and Efros, 2011; Rendle, Zhang, and Koren, 2019; Geirhos et al., 2020).

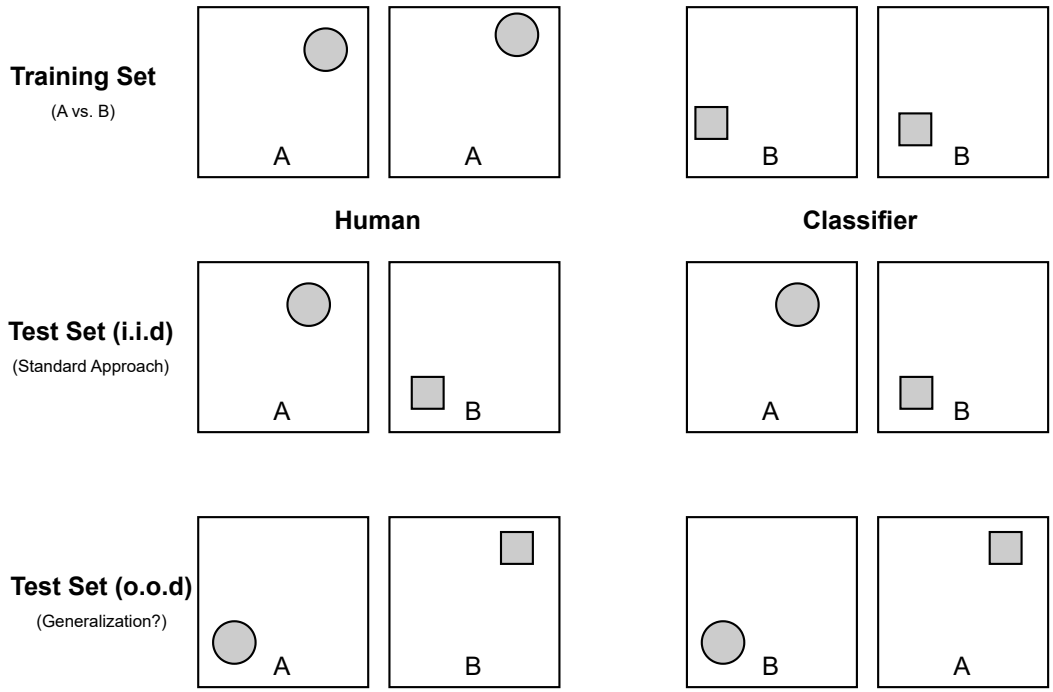


Figure 5.1: Adapted example from the original SCL paper (Geirhos et al., 2020). The task is to distinguish circles (A) from squares (B). While color and size are identical, the training data differ in shape and position. Circles are located at the top right, while squares are located at the lower left. On an *independent and identically distributed* (i.i.d) test set, this leads to seemingly good results, but on an *out-of-distribution* (o.o.d) test set with positions swapped, it becomes clear that the model relies on position rather than shape. These are so-called shortcut opportunities. While this toy example involves only a few factors, with more factors it becomes increasingly unclear which features a classifier, or a model in general, truly relies on.

An adapted example in Figure 5.1, taken from the original SCL paper (Geirhos et al., 2020), illustrates this using the seemingly simple task of distinguishing circles from squares. For simplicity, both shapes share the same background, area, and color. Furthermore, the performance of a classifier is compared with that of a human observer. No matter how, the dataset is arranged with the circles always in the top right corner and the squares always in the bottom left corner. In turn, on i.i.d test data, both humans and classifiers achieve comparable performance, seemingly indicating a shared understanding of object characteristics. However, when the test set is changed to include o.o.d conditions where the positions of the shapes are swapped, the classifier fails. This reveals that the model, apart from humans with real-world experience and intuition, had not learned the concept of shape at all but had instead relied on the positional cue as a shortcut. In fact, this problem would not have been exposed under standard i.i.d testing.

While this example involves only two dimensions of variation, similar issues become much more complex in the real world (Ilyas et al., 2019), for instance in domains such as natural language, where shortcut opportunities are less obvious to the human eye (Gururangan et al., 2018; Thorn Jakobsen, Barrett, and Søgaard, 2021). This connects to the principle of least effort in linguistics (Zipf, 1949), which describes the tendency of speakers to reduce effort by favoring simpler forms, often without consciously noticing it. Nonetheless, these reduced forms are still understood, as in replacing *we are* with *we're* or shortening *television* to *TV*.

In ML and NLP, such tendencies of favoring simple solutions are further reinforced by loss functions like cross-entropy, which are not designed to evaluate why a model assigns high probability to an answer but only that it does so. Consequently, it can be argued if models are rewarded for adopting easier solutions rather than developing robust, semantically grounded solutions. Humans, by contrast, have an intuitive understanding and real-world experience of what constitutes certain things, whereas a model is only exposed to a form of artificially generated experience shaped by its architecture, created representations of the world it is employed in, training data, and the way it is guided (Marr, 1982; Geirhos et al., 2020).

Taking this into consideration, it is important to recognize that while models can achieve high performance on specific datasets, this should not be mistaken for evidence that they have acquired the underlying ability of interest. Thus, SCL emphasizes that each model is only an approximation, an expectation of how the real world unfolds, and that fixed i.i.d performance indicators that only consider numbers can be misleading in this regard.

In this sense, SCL calls for greater emphasis on o.o.d testing not only through larger datasets (Halevy, Norvig, and Pereira, 2009) or bigger models (Saphra et al., 2024; Zhou et al., 2024b; Zhao et al., 2025) but through systematic evaluation of certain data properties, for example, by explicitly removing potential shortcut opportunities (Geirhos et al., 2020).

Hence, SCL denotes testing whether a model sustains performance in the presence of shortcut opportunities or collapses once these are removed and thus questions the genuine assignment of terms like intelligence or SOTA for certain models and their results. Within the scope of this thesis and of AM in general, such shortcut opportunities are examined in more detail in the corresponding publication (Feger, Boland, and Dietze, 2025) presented here.

Marc Feger, Katarina Boland, and Stefan Dietze.

"Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments"

In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.
Acceptance Rate: ∼20.3%

5.3.1 Summary

This study, at first, reassesses *Argument Mining* (AM) benchmarks and their cross-applicability, questioning whether reported results reflect genuine progress or overfitting to data artifacts.

From an initial pool of 52 AM resources, 17 sentence-level datasets, including TACO, were mainly selected based on reproducibility, size, and relevance. The collection is organized by primary labels such as argumentative sentences, claims, conclusions, and custom-defined elements like implicit markups, inference, or others.

Concerning **Q3.1** (*How comparable are existing AM benchmark datasets?*), the analysis focused on characterizing the selected benchmarks based on the definitions used to annotate arguments therein. It also involved comparing the benchmarks using broader statistical properties of sentence-level features, such as length, readability, entropy, and part-of-speech distributions, alongside word-level features derived from their vocabularies.

Despite differences, it was found that the definitions often inform each other and reflect a shared understanding of what constitutes an argument. Additionally, the analysis of sentence-level features revealed a strong correlation between arguments and no-arguments, with Spearman's $\rho \geq 0.97$ within datasets and $\rho \geq 0.68$ across datasets as such. In contrast, at the word-level, both datasets and labels differed mainly in semantic content. While stop and function words, punctuation, and discourse markers overlapped by over 73%, remaining words showed only 19% overlap on average, driving lexical separation in content and vulnerability to SCL.

For assassin **Q3.2** (*Do SOTA models transfer their abilities across benchmarks?*), practical experiments included BERT, RoBERTa, and DistilBERT, which are frequently reported as SOTA, as well as WRAP. Accordingly, these models were evaluated in two setups using stratified samples. In the first, a model was trained on one dataset and tested on another. In the second, the model was trained on all but one dataset, which was used for testing.

Experimental results indicate that although the best-performing models often achieve around 0.79 macro F1 on benchmarks, they struggle in o.o.d test settings. In these generalization tests, 97% perform worse, with 62% falling below 0.65 and 8% dropping under 0.5 macro F1.

Regarding **Q3.3** (*What do SOTA models actually learn when identifying arguments?*), controlled input manipulation was performed for each experiment investigating whether performance remains stable after removing punctuation, stop words, function words, or discourse markers. If true, then a model is likely to rely on shallow artifacts related to content.

In fact, the ablation experiments suggest that performance remained largely stable as long as topical terms are retained, even when omitting nearly every second word in a sentence.

While none of the models fully generalize across benchmarks, WRAP showed slightly better robustness to varying data and achieved performance gains in generalization, particularly towards TACO, averaging 0.75 macro F1 in cross-dataset evaluations.

Moreover, training on the combination of all datasets while holding out one dataset at a time for evaluation consistently improved the performance of the PLMs. However, none of them outperformed established benchmarks in absolute terms. Still, they showed more robust and stable performance compared to training on one dataset and testing on another.

Error analysis suggests that generalization issues are not solely attributable to SCL alone. Across all models, arguments are correctly classified in ~28% of cases, while no-arguments are correctly recognized in ~37% of cases. In addition, the misclassification rate for no-arguments (~13%) is lower than for arguments (~22%). This suggests that arguments are not only more difficult to recognize but that contextual differences (e.g., different pragmatic function in discourse) and annotation practices might also contribute to differences in classification.

Based on **Q3.1–3** and with respect to the overarching research question **Q3** (*Does reported progress in argument identification reflect genuine advances?*), it is concluded that the datasets share similar properties and labels of the same task but differ primarily in their content. As a result, those PLMs assumed to define the SOTA for AM, even when following best practices, often exploit such shortcuts and learn dataset-specific features rather than the intended task. This undermines both generalization and contextualization capabilities and holds true even when such models are applied to seemingly simple datasets in which discourse markers are explicitly used to signal arguments. These findings confirm what has long been suspected in AM, namely that common baselines do not reflect genuine progress but rather the exploitation of discriminative shortcuts, raising questions about how benchmarks should be expanded to more accurately measure progress and alignment with the intended task of argument identification.

5.3.2 Importance and Impact on this Thesis

This study is the first to systematically and empirically demonstrate the limited generalization of current approaches to argument identification.

Although task-specific pre-training provides some robustness and transferability (e.g., to Twitter data), the reported SOTA results do not demonstrate inherent model capabilities, as they collapse under shortcut-controlled o.o.d evaluations across benchmarks. The recognition of these limitations, subsequently, highlights the need for AM models, benchmarks, and evaluations that emphasize task alignment and robustness over dataset-specific optimization, a direction to be further outlined in the shared task[1] presented after this paper.

Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments

Marc Feger
Heinrich-Heine-University
Düsseldorf, Germany
marc.feger@hhu.de

Katarina Boland
Heinrich-Heine-University
Düsseldorf, Germany
katarina.boland@hhu.de

Stefan Dietze
GESIS - Leibniz Institute for the
Social Sciences & Heinrich-Heine-University
Düsseldorf, Germany
stefan.dietze@gesis.org

Abstract

Identifying arguments is a necessary prerequisite for various tasks in automated discourse analysis, particularly within contexts such as political debates, online discussions, and scientific reasoning. In addition to theoretical advances in understanding the constitution of arguments, a significant body of research has emerged around practical argument mining, supported by a growing number of publicly available datasets. On these benchmarks, BERT-like transformers have consistently performed best, reinforcing the belief that such models are broadly applicable across diverse contexts of debate. This study offers the first large-scale re-evaluation of such state-of-the-art models, with a specific focus on their ability to generalize in identifying arguments. We evaluate four transformers, three standard and one enhanced with contrastive pre-training for better generalization, on 17 English sentence-level datasets as most relevant to the task. Our findings show that, to varying degrees, these models tend to rely on lexical shortcuts tied to content words, suggesting that apparent progress may often be driven by dataset-specific cues rather than true task alignment. While the models achieve strong results on familiar benchmarks, their performance drops markedly when applied to unseen datasets. Nonetheless, incorporating both task-specific pre-training and joint benchmark training proves effective in enhancing both robustness and generalization.

1 Introduction

Undeniably, discourse gives people the opportunity to express and discuss their beliefs on any topic.

Argument mining, in this sense, is the automatic identification of the structure of inference and reasoning expressed as arguments presented in natural language (Lawrence and Reed, 2019).

Although there is no one-size-fits-all answer to *What is an argument?* (Stab et al., 2018), the idea suggests itself that arguments are latent yet observable and revolve around *how* they are constituted in terms of their logical scaffolding of argument discourse units, rather than *what* specific subject they address. In practice, these elements, whether sentences or sub-sentence segments, are pragmatically assigned functional roles, most commonly claims and premises, and form the fundamental building blocks of an argument (Stab and Gurevych, 2014; Daxenberger et al., 2017; Lawrence and Reed, 2019; Lopes Cardoso et al., 2023).

Consider the example *X should Y, because Z*, such as *Students should study, because it improves grades* or *We should reduce plastic use, because it minimizes ocean pollution*, which illustrates that the manifestation of an argument should ideally rely on structural components conveyed through functional patterns, while remaining agnostic of certain topics or other content-specific elements.

For this reason, one might assert that argument mining, in theory, is applicable across different corpora if the structural signals defining arguments are reliably identifiable from appropriately labeled data. Conversely, in practice, any inability to apply these signals to diverse datasets may expose systematic biases in the field, an issue that has long been informally discussed over coffee breaks.

Generalizability, in this regard, takes high priority, especially at leading NLP conferences such as ACL 2025, as it allows models to make reliable and reasonable predictions on data that does not correspond to their training data. This is especially true for real-world models, which should mimic human-like generalization abilities, where emerging evidence indicates that such models are often

fine-tuned to the specifics of established benchmark datasets, leading to unfounded optimism about their improvements (Saphra et al., 2024).

Consequently, concerns about vulnerability to shortcut learning (Geirhos et al., 2020) highlight the broader challenge of evaluating baselines beyond isolated benchmarks (Rendle et al., 2019).

Argument mining is one such area of natural language processing applications in which the ability to generalize is key. Hence, we ask for:

Q1: How comparable are the existing benchmark datasets for argument mining?

Q2: Do state-of-the-art argument mining models generalize to out-of-distribution data from other benchmarks?

Q3: Do these models acquire a generalizable concept of arguments?

In this context, there has been speculation that BERT (Devlin et al., 2019), known to pay great attention to basic syntax, nouns, and coreferences (Clark et al., 2019), is prone to learning shortcuts when mining arguments (Geirhos et al., 2020), where its generalization is limited to within-topic signals in datasets sharing similar argument and topic structures (Thorn Jakobsen et al., 2021).

Our aim is not to propose a new formalism for arguments or to pinpoint the best-performing argument mining model, but to use data from previous work in which different theories have been applied to see whether individual efforts and perspectives converge in terms of identifying arguments.

With this being said, we perform the first large-scale experimental assessment of benchmarks, systematically evaluating generalization across diverse argument mining datasets following a comprehensive review of datasets spanning 2008 to 2024.

For our study, we selected BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) as exemplary BERT-like models, widely recognized as standard baselines in various areas of natural language processing (Rogers et al., 2020), including recent research on argument mining (Shnarch et al., 2020; Mayer et al., 2020a; Fromm et al., 2021a; Alhamzeh et al., 2022; Feger and Dietze, 2024b). We also examine WRAP (Feger and Dietze, 2024a), the only transformer whose language representation pre-training is extended by leveraging contrasts of inference and information signals to generalize argument components. Although originally designed for cross-topic generalization on Twitter (𝕏), WRAP does not rely on tweet- or topic-specific features to enhance its generalizability, distinguishing it from the others and making it particularly interesting for research.

In this study, we start by detailing our process of finding argument mining benchmark datasets and explain the selection criteria and justifications in Section 2. The core characteristics of these datasets, addressing research question **Q1**, are then examined in Section 3. Next, we describe our experimental setup in Section 4, covering both result generation and the implementation of best practices for significance testing, which form the basis for answering **Q2 - Q3** in Section 5. The results of this paper are then discussed in Section 6 and concluded in Section 7.

In order not only to elucidate the process but also to foster discussion that may inspire new approaches for novel datasets and broader generalization of argument mining methods, we contribute:

1. A survey of argument mining datasets between 2008 and 2024, primarily from the ACL Anthology, that identified 52 relevant papers with datasets from leading NLP conferences.

2. The first large-scale re-assessment that combines benchmark evaluations for 17 selected argument mining datasets, including controlled manipulation experiments to determine whether the reported state-of-the-art models (BERT, RoBERTa, DistilBERT, WRAP) actually learn generalizable argument concepts.

3. Statistical evidence that shortcut learning undermines generalization in argument mining. Although each of the examined transformers delivers strong results on benchmarks, all struggle to varying degrees when applied to other datasets, with WRAP generally performing slightly better. These challenges are compounded by divergent argument definitions and inconsistent annotations across datasets.

2 Argument Mining Benchmark Datasets

This section outlines the dataset collection and selection process, emphasizing the rationale behind our choice of benchmark datasets for argument mining. The decisions for all 52 datasets reviewed are present in Appendix A.1. Additionally, the code and data are available in our repository[1].

[1] Limited-Generalizability

Dataset	Paper	Genre	Definition	Arguments	No-Arguments
ACQUA	(Panchenko et al., 2019)	Mixed	Argumentative	1,949	5,236
WEBIS	(Al-Khatib et al., 2016a)	Online Debate	Argumentative	10,804	5,543
ABSTRCT	(Mayer et al., 2020b)	Academic	Claim-based	1,308	7,323
ARGUMINSCI	(Lauscher et al., 2018)	Academic	Claim-based	6,554	9,548
CE	(Rinott et al., 2015)	Encyclopedia	Claim-based	1,546	85,417
CMV	(Hidey et al., 2017)	Online Debate	Claim-based	979	1,593
FINARG	(Alhamzeh et al., 2022)	Spoken Debate	Claim-based	4,607	8,310
IAM	(Cheng et al., 2022)	Mixed	Claim-based	4,808	61,715
PE	(Stab and Gurevych, 2017)	Academic	Claim-based	2,093	4,958
SCIARK	(Fergadis et al., 2021)	Academic	Claim-based	1,191	10,503
USELEC	(Haddadan et al., 2019)	Spoken Debate	Claim-based	13,905	15,188
VACC	(Morante et al., 2020)	Online Debate	Claim-based	4,394	17,825
WTP	(Biran and Rambow, 2011)	Online Debate	Claim-based	1,135	7,274
AFS	(Misra et al., 2016)	Online Debate	Conclusion-based	5,150	1,036
UKP	(Stab et al., 2018)	Mixed	Evidence or Reasoning	11,126	13,978
AEC	(Swanson et al., 2015)	Online Debate	Implicit-Markup	4,001	1,374
TACO	(Feger and Dietze, 2024b)	Twitter Debate	Inference-Information	864	868

Table 1: The final 17 datasets that meet the sentential, binary label, and reproducibility criteria, each yielding at least 1,700 instances (850 per label) under a stratified 60/20/20 split, ensuring adequate size for the experiments.

2.1 Collection Process

As part of our data collection process, we examined the most recent and relevant survey papers on argument mining, primarily from the ACL Anthology (Daxenberger et al., 2017; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Vecchi et al., 2021; Schaefer and Stede, 2021; Ajjour et al., 2023), all of which catalog datasets addressing various subtasks within the field, where argument identification is a fundamental prerequisite for each.

To expand and back up our dataset collection, we searched Google Scholar and Google Dataset Search for the keyword *argument mining* to find contributions beyond survey papers.

Based on our assessment, we found 52 such papers with datasets, mostly from top NLP conferences like ACL, NAACL, LREC, or EMNLP.

2.2 Selection Criteria

The dataset selection process for this paper was conducted in two stages. In the primary inclusion phase, we evaluated all 52 datasets based on:

- **Sentential**: The data and labels are at the sentence-level or aggregatable to this level (e.g., from sub-sentence or token annotations). Tweets were excluded from classical sentence conventions due to their unique structure.
- **Binary**: The dataset assigns binary labels to distinguish argument from no-argument sentences (e.g., based on the presence or absence of claims or other argument components).
- **Reproducible**: The dataset is largely replicable, with minor discrepancies from the publication (e.g., updates or duplicate removal affecting size). To ensure reproducibility, we reviewed documentation, labels, guidelines, and tools, and attempted to resolve access issues (e.g., client-sided or coding errors).

We applied these criteria sequentially, excluding datasets immediately upon failing any condition, eliminating 24 of the initial 52. In the refined inclusion step, we assessed relationships and data sufficiency to ensure adequate evaluation and generalization sizes, leading us to consider:

- **Related**: Connections between datasets such as updated versions, additional non-task-related features (e.g., stance added to a claim), and curated subsets derived from repositories that serve as data sources rather than datasets.
- **Sufficiency**: For a stratified 60/20/20 split, each dataset must have at least 500 training instances and 150 evaluation instances per label. An initial analysis revealed that two in five datasets fell short of this threshold, and alternative splits (e.g., 70/15/15 or 80/10/10) would further reduce evaluation sizes, worsening the small-data issue.

In total, this process resulted in 17 datasets encompassing ~345k labeled sentences, each meeting the aforementioned criteria. The final selection of datasets included in this study is listed in Table 1.

3 Characterizing Argument Mining Benchmark Datasets and Definitions

Before addressing **Q1**, we briefly introduce the individual datasets, organizing them by their primary labels. We then give the answer to **Q1** in terms of comparing definitions in Section 3.1 and textual characteristics in Section 3.2.

Argumentative serves as an umbrella term, identifying arguments with markers or patterns that suggest structural components, without necessarily specifying their roles (e.g., as claim or inference). In this sense, ACQUA (Panchenko et al., 2019) contains 7,185 argumentative sentences from Common Crawl (Panchenko et al., 2018), covering topics like computer science and brands, categorizing comparisons (e.g., Matlab vs. Python) as argumentative or not. Similarly, WEBIS (Al-Khatib et al., 2016a) comprises 16,347 segments across 14 topics (e.g., culture, health) from iDebate, with user-assigned labels (introduction, for, against) mapped to argumentative and non-argumentative labels.

Claim-based approaches explicitly annotate for the presence of claims as the core of an argument. Thereby, ABSTRCT (Mayer et al., 2020b), sourced from PubMed, comprises 8,631 sentences extracted from abstracts related to five diseases (e.g., neoplasm, glaucoma). ARGUMINSCI (Lauscher et al., 2018) provides annotations for the Dr. Inventor dataset (Fisas et al., 2016) for computer graphics publications, totaling 16,102 sentences. CE (Rinott et al., 2015) contains 86,963 sentences from Wikipedia across 58 topics (e.g., one-child policy, physical education). CMV (Hidey et al., 2017) consists of 2,572 sentences from the *Change My View* subreddit, spanning a diverse range of topics. FINARG (Alhamzeh et al., 2022) comprises 12,917 sentences sourced from transcribed earnings calls of Amazon, Apple, Microsoft, and Facebook. Moreover, IAM (Cheng et al., 2022) contains 66,523 sentences from various online platforms across 123 topics (e.g., vaccination, multiculturalism), while PE (Stab and Gurevych, 2017) includes 7,051 annotated sentences from persuasive essays (e.g., about cloning). SCIARK (Fergadis et al., 2021) contains 11,694 annotated sentences from scientific literature (e.g., PubMed, Semantic Scholar) on sustainable development goals (e.g., well-being, gender equality), also considering generalization to ABSTRCT. On the other hand, USELEC (Haddadan et al., 2019) offers 29,093 sentences from transcripts of U.S. presidential debates from 1960 (Kennedy vs. Nixon) to 2016 (Clinton vs. Trump), transcribed from the Commission on Presidential Debates. VACC (Morante et al., 2020) offers 22,219 sentences from a mixed collection of online debates about vaccination, while WTP (Biran and Rambow, 2011) includes 8,409 sentences from Wikipedia Talk Pages on various topics (e.g., Darwinism, the Catholic Church).

Others represents a residual category encompassing a variety of distinct definitions. AFS (Misra et al., 2016) comprises 6,186 annotated sentences drawn from online debate platforms such as iDebate and ProCon for three topics (e.g., gay marriage, death penalty). Sentences are labeled based on whether they explicitly convey a specific argument facet, with conclusions serving as the core component of the argument. UKP (Stab et al., 2018) contains 25,104 sentences across eight topics (e.g., nuclear energy, minimum wage) for cross-topic argument mining from heterogeneous sources, where arguments provide evidence or reasoning to support or oppose a topic. On the other hand, AEC (Swanson et al., 2015) contains 5,375 sentences on four topics (e.g., evolution, gun control) from CreateDebate, highlighting simple argument signals with labels based on the implicit markups: so, if, but, first, I agree that. Finally, TACO (Feger and Dietze, 2024b) comprises 1,734 tweets spanning six topics (e.g., abortion, Squid Game). It is designed for cross-topic argument mining on Twitter, focusing on inference to shape arguments.

3.1 Comparing Argument Definitions

(Q1) *Argument definitions vary, reflecting a spectrum of perspectives that contribute to a shared understanding of arguments.* Central to this is the observation that definitions mutually inform each other in their concepts (Lopes Cardoso et al., 2023). For example, in Table 1 most papers are claim-based, but when comparing the definitions, some view a claim as argumentative (Lauscher et al., 2018; Fergadis et al., 2021), others as conclusive (Mayer et al., 2020b), as stances (Rinott et al., 2015; Hidey et al., 2017; Cheng et al., 2022; Stab and Gurevych, 2017), or as a hybrid concept of all these (Haddadan et al., 2019; Morante et al., 2020).

Hence, further clarification is needed, especially concerning their generalization as part of **Q2 - Q3**. Thereby, Table 2, with examples from different definitions, illustrates whether their efforts nevertheless converge in the identification of arguments despite different perspectives.

Label	Dataset	Example
ARG	ACQUA	We chose MySQL over PostgreSQL primarily because it scales better and has embedded replication.
	SCIARK	In this case, if symptomatic, the treatment should be surgery, clinical follow-up, and counseling.
	AEC	So it would seem that if there is a scientific theory of [. . .], it has been tested [. . .] and therefore [. . .].
¬ARG	WEBIS	The Mo Ibrahim Prize was first established in 2007, and the prize represents [. . .] African leadership.
	FINARG	For those unable to attend in person, these events will be webcast and you can follow [...] at URL.
	TACO	'Bitter truth': EU chief [...] on idea of Brits keeping EU citizenship after #Brexit URL via USER

Table 2: Examples of argument (ARG) and no-argument (¬ARG) sentences from various datasets. Despite differences in definitions and topics, the similarities within and distinctions between label groups underscore the shared endeavor of argument mining approaches in identifying arguments, though each emerged differently.

3.2 Comparing Dataset Dimensions

First, the two text dimensions used to analyze the selected datasets are presented. For dataset-wise correlations of these, please refer to Appendix A.2.

Sentence-Level: To capture a broad, macro-level view without delving into individual word details, we used spaCy[2] to extract key textual attributes. These features reveal the overall structural and statistical properties of sentences, enabling sentence-level characterization of each dataset by:

- *Length*: Measured by the number of words per sentence, which serves as an indicator of linguistic complexity and verbosity.
- *Stop/Function Word Ratio*: The ratio of stop (e.g., it, is, are) and function words (e.g., against, because, therefore), including discourse markers, to the other words in a sentence to show their relative frequency of use.
- *Type-Token Ratio*: The ratio of unique words to total words in a sentence, assessing lexical diversity.
- *Readability*: The Flesch Reading Ease score quantifies text clarity, with lower values ($0 \leq$) indicating complex academic language and higher values (≤ 100) denoting easy readability, understandable by an 11-year-old.
- *Entropy*: Quantifies lexical unpredictability and the amount of information in a sentence, with values ranging from 0 (fully predictable text) to 1 (maximal unpredictability).
- *Sentiment*: Defined by polarity, ranging from -1 (extremely negative) to 1 (extremely positive), and subjectivity, ranging from 0 (objective) to 1 (subjective), possibly revealing persuasive strategies through emotions.
- *Part-of-Speech Tags*: The distribution of the 17 universal POS tags reflects basic syntax, lexical composition, and stylistic variation.

[2] spacy.io

Word-Level: To compare datasets at the word level, we analyze the vocabulary of unique words used in each dataset. We extend this to words that convey the central semantic content of a sentence (e.g., government, abortion, freedom), that is, all words except stop and function words, discourse markers, and punctuation. Their relatedness or uniqueness is described using Jaccard similarity, a measure of similarity between two sets based on the ratio of their intersection to their union.

(Q1) *The sentence structures are strongly correlated across all datasets and labels.* On average, a sentence contains 21 words, with nearly every second word (48%) being a stop or function word. Sentences are lexically diverse (91% type-token ratio) yet highly readable (63% readability). The high predictability (22% entropy) and objective tone (43% subjectivity) suggest clear, structured writing with a slightly positive inclination (8% polarity). This is reinforced by the POS patterns, where sentences typically include five nouns, three punctuation marks, and two verbs, adpositions, and determiners, with other tags averaging below two.

Moreover, an average sentence closely aligns with both argument and no-argument sentences across these 24 sentence-level features (Spearman's $\rho \geq 0.97$), with a strong correlation ($\rho \geq 0.68$) across datasets. Slight differences exist in length, with an argument sentence averaging 24 words compared to 20 for a no-argument sentence, with readability scores of 60% and 64%, respectively.

(Q1) *Datasets and labels mainly differ in their semantic content.* Looking at the vocabularies, the datasets remain largely distinct, with 7–36% Jaccard similarity, a trend also observed for the semantic content words, reflecting their open-class.

In contrast, stop, function, and discourse words show over 73% overlap due to their closed nature.

Interestingly, while comparing sentences across labels shows similar patterns, words describing the core semantic content remain largely distinct, overlapping below 48% and 19% on average, reinforcing lexical separation. Undeniably, the datasets share overlapping content, e.g., when discussing the one-child policy (PE) and abortion (IAM, TACO, UKP) or, figuratively speaking, the death penalty (AEC). Similarly, when discussing vaccination (VACC) overlaps might occur with medical (ABSTRCT) or sustainability (SCIARK) topics.

However, we found that these similarities are not very pronounced and that the datasets and labels are largely disjointed in terms of their core semantic content. This could provide the models with a shortcut opportunity, not based on how the labels are constructed, but rather on what they are about.

4 Experimental Setup

In this section, we outline the experimental setup and the best practices used for statistical testing to generate the data needed to answer **Q2 - Q3**.

Sampling: To create fixed training, development, and test sets, we used a 60/20/20 stratified split for each of the 17 datasets in Table 1, selecting 850 instances per label, corresponding to 1,700 samples per dataset and 28,900 in total.

Transformers: We selected BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) as widely accepted standard baselines for NLP (Rogers et al., 2020), including argument mining (Shnarch et al., 2020; Mayer et al., 2020a; Fromm et al., 2021a; Alhamzeh et al., 2022; Feger and Dietze, 2024b). Further, we examined WRAP (Feger and Dietze, 2024a), the only transformer that is specifically pretrained for argument generalization. This applies contrastive learning to cluster similar manifestations of inference and information, separate dissimilar ones, and produce generalized embeddings robustly adaptable to downstream classification. However, our goal is to assess the generalizability of these state-of-the-art argument mining models, not to find the best. For these, we use the standard hyperparameter grid for GLUE (Wang et al., 2018), as accepted in the BERT and RoBERTa papers, balancing performance and time with a batch size of 32, 3 epochs, and a learning rate between 2e-5 and 5e-5, each trained on an A100 GPU.

Benchmarking and Generalization: The experiments presented here are the core investigations related to **Q2**. For each, we report the test results after tuning the hyperparameters to a target's development dataset, optimizing the macro F1 score to ensure equal importance of both labels.

We begin with an initial assessment using pairwise comparisons, following the transfer learning framework (Pan and Yang, 2010; Houlsby et al., 2019; Zhuang et al., 2019), where models are trained on one dataset and evaluated on others, including benchmarks on individual datasets. This yields a 17×17 matrix per model, with rows as training and columns as test data, see Figure 1.

Secondly, we conducted a supplementary experiment by training on all but one dataset and testing on the reserved one, forcing the models to generalize from joint benchmark data (Hays et al., 2023; Feger and Dietze, 2024a). Thereby, we will report the performance per model and evaluate each against the excluded dataset's state-of-the-art benchmark, compare Table 4 and Figure 1.

Disrupting Argument Signals: To build on the experiments addressing **Q2** and provide insight for **Q3**, we apply controlled input manipulation to both experiments described above. Specifically, we assess transformer performance after systematically removing stop and functional words (e.g., a, the, against, because), discourse markers, and punctuation using spaCy[2]. This process results in the elimination of around half the words in each sentence. It is therefore assumed that the removal of these lexical and syntactic elements, which also function as scaffolding for rhetorical and logical devices (Knott and Dale, 1994), suppresses the linguistic cues that, in theory, enable the distinction between the elements that constitute an argument and those that do not (Daxenberger et al., 2017; Opitz and Frank, 2019; Thorn Jakobsen et al., 2021). What remains is a lexical skeleton that primarily reflects topical and subject-related content while omitting functional and discursive elements, calling into question the model's ability to discern argued excerpts from mainly descriptive content (Lopes Cardoso et al., 2023), see Table 3.

Evaluation: We perform the experiments for **Q2 - Q3** and repeat them three times, each with varied samples and training initializations. To test significance, we use a two-way ANOVA with repeated measures for experimental robustness and one-tailed Student's t-tests for pairwise comparisons of models, see Appendix B for full details.

Label	Form	Example
ARG	Original	They should increase more routes to make people transport more easily.
	Manipulated	increase routes people transport easily
¬ARG	Original	Should governments spend more money on improving roads and highways?
	Manipulated	governments spend money improving roads highways

Table 3: Example from PE showing an argument (ARG) and no-argument (¬ARG) sentence in the original and manipulated form.

5 Results

In this section, we will address and answer questions **Q2 - Q3**. To this end, we will mainly focus on Figure 1, which compares the pairwise experiments to show which state-of-the-art argument mining model performs best, thus reflecting the current benchmark and generalization landscape. Tying in with this, we will then turn on Table 4 contrasting the state-of-the-art performance against those obtained by the models if trained on heterogeneous data. In addition, we elaborate on the insights gained from the controlled manipulations applied to these experiments. After that, we will discuss the significance of our results. However, for a better understanding, it can already be assumed that the results for each model and experiment follow a normal distribution, as confirmed with D'Agostino and Pearson's K^2 test ($p \geq .05$).

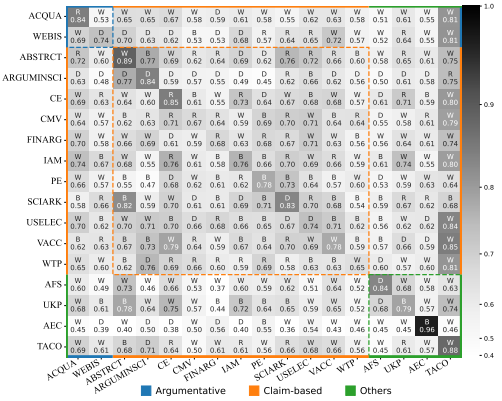


Figure 1: The best macro F1 scores from the benchmarking and pairwise generalization experiments, comparing WRAP (W), BERT (B), RoBERTa (R), and DistilBERT (D), indicate that strong performance is primarily achieved in the benchmark settings, as reflected along the main diagonal. Furthermore, WRAP excels in generalizing to TACO, as seen on the right.

(Q2) *Strong argument mining baselines do not necessarily imply strong argument generalization.* A notable observation in Figure 1 is the contrast between baselines on individual datasets and generalization across multiple datasets and definitions. Strikingly, 97% of generalization experiments fall below the mean benchmark result ($M = 0.79$), with 62% scoring under 0.65, while in 8% of cases generalization drops below 0.5 macro F1, highlighting the challenge of maintaining strong benchmark performances when tested on out-of-distribution datasets. We will further break down our answer:

Generalizability seems to be the exception rather than the norm. Given these circumstances, Table 1 shows several notable exceptions of good (≥ 0.75) to strong (≥ 0.8) generalizability across and within both definitional categories and genres, particularly for claim-based datasets. For instance, strong performance emerges within the academic domain, where SCIARK reaches 0.82 on ABSTRCT with BERT, and both ABSTRCT and ARGUMINSCI achieve 0.77 using BERT and DistilBERT. Evidence of cross-genre generalization also appears in cases such as IAM (mixed genre) and VACC (online debate), which achieve 0.76 and 0.79 on CE (encyclopedia) using RoBERTa and WRAP.

Broader generalization across definitions and genres is especially evident in UKP (evidence or reasoning, mixed), which surpasses 0.75 on both ABSTRCT (claim-based, academic) and CE (claim-based, encyclopedia) with BERT and WRAP. Similarly, TACO (inference-information, Twitter debate) consistently exceeds 0.8 across a vast range of definitions and genres with WRAP.

Still, both cross-definition and cross-genre generalization remain limited and exceptional.

Task-related pre-training appears to have a positive effect on overall performance and generalization. Numerically, WRAP ($M = 0.61, SD = 0.1$) shows the best overall performance in terms of macro F1. Notably, WRAP is the only model that attains a mean above 0.6 macro F1, while BERT ($M = 0.58, SD = 0.11$), RoBERTa ($M = 0.57, SD = 0.12$), and DistilBERT ($M = 0.56, SD = 0.11$) all perform worse. This performance advantage is particularly evident in cases where WRAP achieves the highest scores compared to the other models. In fact, WRAP demonstrates superior performance in 133 out of 289 experiments (46%), whereas BERT does so in 58 experiments (20%), RoBERTa in 50 experiments (17%), and DistilBERT in 48 experiments (17%).

	WRAP	BERT	RoBERTa	DistilBERT	SOTA	$\Delta_{max/min}$
ACQUA	**0.66**	0.6	*0.59*	*0.59*	**0.84**	0.18 / 0.25
WEBIS	0.63	**0.66**	*0.62*	0.65	**0.74**	0.08 / 0.12
ABSTRCT	**0.74**	**0.74**	**0.74**	*0.71*	**0.89**	0.15 / 0.18
ARGUMINSCI	**0.59**	*0.47*	0.55	0.5	**0.84**	0.25 / 0.37
CE	**0.77**	*0.72*	0.76	*0.72*	**0.85**	0.08 / 0.13
CMV	**0.63**	0.62	0.62	*0.58*	**0.67**	0.04 / 0.09
FINARG	*0.61*	0.62	**0.66**	0.65	**0.68**	0.02 / 0.07
IAM	**0.73**	*0.71*	**0.73**	**0.73**	**0.76**	0.03 / 0.05
PE	*0.65*	*0.65*	**0.69**	*0.65*	**0.78**	0.09 / 0.13
SCIARK	**0.75**	*0.73*	0.74	*0.73*	**0.83**	0.08 / 0.1
USELEC	**0.7**	0.66	0.68	*0.59*	**0.74**	0.04 / 0.15
VACC	*0.68*	**0.7**	*0.68*	0.69	**0.78**	0.08 / 0.1
WTP	**0.59**	0.55	0.55	*0.54*	**0.65**	0.06 / 0.11
AFS	*0.57*	0.58	0.59	**0.6**	**0.84**	0.24 / 0.27
UKP	**0.7**	*0.67*	**0.7**	0.68	**0.79**	0.09 / 0.12
AEC	0.52	**0.57**	*0.51*	0.56	**0.96**	0.39 / 0.45
TACO	**0.76**	0.61	0.65	*0.55*	**0.88**	0.12 / 0.33

Table 4: Transformers trained on all but the target benchmark are evaluated against their state-of-the-art baseline (SOTA), compare diagonal of Figure 1. *Minimum* and **Maximum** values indicate deviation from SOTA ($\Delta_{max/min}$). While all models fall short relative to SOTA, WRAP yields the best results in most cases.

Joint benchmark data for training may also help bootstrap reliable and improved generalization. Furthermore, the results of the supplementary experiment presented in Table 4 indicate that overall performance tends to improve when models are trained on joint benchmark data. Thereby, WRAP ($M = 0.66, SD = 0.07$), RoBERTa ($M = 0.65, SD = 0.07$), BERT ($M = 0.64, SD = 0.07$), and DistilBERT ($M = 0.63, SD = 0.07$) all achieve average macro F1 scores above 0.6, with values that are numerically higher than those observed in the pairwise setup. Again, WRAP shows the most consistent advantage, ranking first in 11 out of 17 experiments (65%).

(Q3) *State-of-the-art argument mining models are not solely defined by argument signals.* Following the controlled manipulation in the pairwise setup, all models dropped to similar levels, WRAP and BERT ($M = 0.56$, $SD = 0.09$), DistilBERT ($M = 0.55$, $SD = 0.1$), and RoBERTa ($M = 0.57$, $SD = 0.1$). Similar trends appear post-manipulation in the supplementary experiment for WRAP, RoBERTa, and DistilBERT ($M = 0.62$, $SD = 0.06$), and BERT ($M = 0.61$, $SD = 0.06$). With careful attention to detail:

Shortcut learning influences generalization of arguments, but task-related pre-training weakens the impact. For the pairwise experiments, BERT and DistilBERT showed almost no changes after manipulating inputs ($\Delta \leq 0.02$), while RoBERTa maintained its performance completely, suggesting that the overall performance of these models is not based on learning how arguments are constituted. In contrast, WRAP, which relies on its task-related pre-training to embed structural argument components across topics, showed the largest drop in macro F1 with $\Delta = 0.05$.

Jointly integrating benchmark data for training improves generalization and reduces shortcut reliance. The impact of WRAP towards robustness of generalization is also true for the supplementary experiment, where WRAP exhibited the largest performance drop ($\Delta = 0.04$) post-manipulation. Nonetheless, RoBERTa and BERT showed similar trends ($\Delta = 0.03$), while DistilBERT showed mostly no changes ($\Delta = 0.01$). Whereas the results in Table 4 show that each model underperformed relative to the state-of-the-art baselines, a notable pattern still emerged. This is, training on jointly integrated benchmark data raises the average macro F1 score to at least 0.64 for three out of four transformers and 0.63 for the lowest-performing model, compared to a maximum of 0.61 in pairwise transfer, achieved by WRAP. While only WRAP generalizes better in the pairwise setting and is less affected by lexical shortcuts, this advantage persists when trained on joined datasets. However, in this merged setting, RoBERTa and BERT also show improved robustness, despite their stronger reliance on shortcuts in the pairwise setup. Furthermore, average differences remain moderate with $\bar{\Delta}_{max} = 0.12$ and $\bar{\Delta}_{min} = 0.18$ while the models learn from heterogeneous data sources.

Differences in definitions of arguments reinforce the limitations of generalization. However, while signs of shortcut learning are found, it is undeniably not the sole limiting factor. Averaged across all models, misclassification patterns show that arguments are correctly classified 28% of the time and no-arguments 37%, suggesting that identifying no-arguments is easier. This is further supported by the lower misclassification rate for no-arguments (13%) compared to arguments (22%), highlighting practical differences in argument definitions that affect both generalization and benchmarks (e.g., due to conflicting annotations). This can also be observed when analyzing the misclassifications of individual models. Here, all models misclassify no-arguments as arguments in fewer than 16% of cases. In contrast, BERT, RoBERTa, and DistilBERT exhibit higher misclassification rates, ranging from 21% to 26%, while WRAP misclassifies arguments as no-arguments in 18% of cases, highlighting its superior generalization ability for arguments.

(Q2 - Q3) *The experiments demonstrate both statistical significance and practical relevance.* Repeated experiments support the robustness of these results. Regarding the pairwise experiments, a two-way repeated measures ANOVA for **Q2** showed a significant effect only when comparing model performances ($F(3, 864) = 69.47, \epsilon = 0.56, p_{\text{corr}} < .05, \eta^2_G = 0.03$), with negligible resampling or interaction effects. For **Q2**, paired one-tailed t-tests also showed that only model comparisons involving WRAP were significant ($p_{\text{corr}} < .05, 8.12 \leq t(288) \leq 10.14$), with moderate effect sizes ($0.39 \leq d \leq 0.49$). Similarly, repeating **Q3** revealed no significant effects, confirming that once ablated, the models perform comparably overall. Also, for **Q3**, when comparing pre- and post-manipulation results per model, only WRAP showed a relevant decrease ($p < .05, t(288) = -8.91, d = -0.49$). In terms of the supplementary experiments, repetition yielded no significant effects pre- and post-manipulation. However, regarding **Q3**, one-sided paired t-tests revealed significant post-manipulation decreases for WRAP, RoBERTa, and BERT ($p < .05, -5.52 \leq t(16) \leq -2.67, -0.58 \leq d \leq -0.41$), with WRAP showing the strongest effect.

6 Discussion

To summarize the limited generalization in argument mining addressed, Table 5 compares the best baseline results pre- and post-manipulation. On average, macro F1 differences remain close, within $\bar{\Delta}_{max} = 0.07$ and $\bar{\Delta}_{min} = 0.12$ per model, and in the best cases even exceed benchmark levels.

In the single case of AEC, which relies on only five keywords for arguments, overemphasis on these signals also appears to impair generalization. Although AEC attains the highest score (0.96) and experiences the largest post-manipulation drop (≤ 0.45, Table 5), its generalization is limited to 0.63 or even below 0.5, compare Figure 1. Given the low performance and minimal differences between pre- and post-manipulation results, BERT, RoBERTa, and DistilBERT do not clearly demonstrate an inherent ability to generalize arguments.

Although these challenges may be widespread, positive examples highlight the potential for future progress. This is particularly evident in cases involving diverse sources and topics (VACC, CE, TACO, UKP, IAM), where UKP, IAM, and TACO already aim for generalizable annotations.

	WRAP	BERT	RoBERTa	DistilBERT	SOTA	$\Delta_{max/min}$
ACQUA	*0.73*	0.77	0.76	**0.78**	**0.84**	0.06 / 0.11
WEBIS	*0.61*	0.66	0.66	**0.67**	**0.74**	0.07 / 0.13
ABSTRCT	*0.83*	**0.87**	0.84	**0.87**	**0.89**	0.02 / 0.06
ARGUMINSCI	0.78	**0.79**	*0.77*	*0.77*	**0.84**	0.05 / 0.07
CE	*0.75*	0.79	0.77	**0.81**	**0.85**	0.04 / 0.1
CMV	*0.57*	0.64	0.64	**0.65**	**0.67**	0.02 / 0.1
FINARG	0.62	*0.61*	0.66	**0.69**	**0.68**	-0.01 / 0.07
IAM	*0.66*	0.69	**0.71**	0.7	**0.76**	0.05 / 0.1
PE	*0.66*	0.67	0.71	**0.73**	**0.78**	0.05 / 0.12
SCIARK	*0.71*	**0.8**	0.77	0.79	**0.83**	0.03 / 0.12
USELEC	0.65	**0.66**	*0.62*	**0.66**	**0.74**	0.08 / 0.12
VACC	*0.67*	0.68	**0.69**	**0.69**	**0.78**	0.09 / 0.11
WTP	**0.58**	*0.54*	0.57	0.56	**0.65**	0.07 / 0.11
AFS	*0.78*	**0.81**	0.8	0.79	**0.84**	0.03 / 0.06
UKP	*0.74*	0.76	**0.78**	*0.74*	**0.79**	0.01 / 0.05
AEC	*0.51*	0.55	0.58	**0.59**	**0.96**	0.37 / 0.45
TACO	**0.77**	*0.76*	*0.76*	**0.77**	**0.88**	0.11 / 0.12

Table 5: Post-manipulation performance of each transformer compared to state-of-the-art (SOTA) results for baseline experiments per dataset. *Minimum* and **Maximum** values are highlighted, with $\Delta_{\text{max/min}}$ indicating their deviation from SOTA.

Despite limitations, the need for a unified structural approach to argument analysis becomes apparent. This is reinforced by the effectiveness of methodologies tailored to argument mining, as seen in WRAP's strong performance, averaging 0.75 when generalizing to TACO from all other datasets (Figure 1). Training on joint benchmark data further strengthens these abilities also for the standard transformers, even if numerical results fall short of the rarely doubted state-of-the-art (Table 4). Benchmarking should therefore build on combined datasets that capture the task's general demands, as in GLUE (Wang et al., 2018) and instruction-tuning benchmarks (Ouyang et al., 2022; Zhang et al., 2024), for which decoder-based argument mining (Cabessa et al., 2025) may be of interest.

7 Conclusion

We present the first large-scale re-evaluation of argument mining benchmarks through a generalization lens and evaluate whether the reported performance marks true progress. While structural patterns hold, thematic and content differences between labels and datasets favor shortcut learning. BERT, RoBERTa, and DistilBERT often rely on this to inflate benchmarks, while WRAP shows more resilience, likely due to its pre-training for argument generalization. Training on shared benchmark data further reduces shortcut reliance and improves generalization, notably in combination with WRAP. Our results stress the need to integrate different task demands and suggest re-framing argument mining as a joint generalizability task.

Limitations

This study did not separate direct from implicit arguments lacking clear structural and lexical cues, including discourse markers, and based on data analysis, assumed such cases are rare. However, this may affect interpretation, as implicit arguments are likely to depend on topical and content cues.

While we mostly used publicly available datasets, some require granted access.

Additionally, when extraction scripts were unavailable, we derived our procedures from both the available documentation and our understanding of the original process. This was particularly relevant for datasets where `.ann` files only provided annotated sequence boundaries for larger documents stored in `.txt` or `.json` formats. In such cases, we used spaCy[2] for sentence boundary extraction, which may produce boundaries that differ from the original assumptions. Nevertheless, we confirmed that over 95% of the extracted sentences ended with proper punctuation and began with a capital letter. We provide an extraction script[1] that automatically retrieves and processes all datasets considered.

The reproducibility of the experiments may be constrained by factors such as data size, runtime, and associated costs, with all experiments in this study running ~126 hours on a costly A100 GPU.

Acknowledgments

We sincerely thank the anonymous reviewers for their attentive and constructive feedback, which greatly contributed to improving the paper. Cheers!

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Yamen Ajjour, Johannes Kiesel, Benno Stein, and Martin Potthast. 2023. Topic ontologies for arguments. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1411–1427, Dubrovnik, Croatia. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence*, pages 48–59, Cham. Springer International Publishing.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016a. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogues by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5427–5433. AAAI Press.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive

and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024a. BERTweet's TACO fiesta: Contrasting flavors on the path of inference and information-driven argument mining on Twitter. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2256–2266, Mexico City, Mexico. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024b. TACO – Twitter arguments from COnversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15522–15529, Torino, Italia. ELRA and ICCL.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021a. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4758–4766.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021b. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4758–4766.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Nancy Green. 2018. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110, Brussels, Belgium. Association for Computational Linguistics.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artif. Intell. Law*, 32(3):1–38.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Marcus Hansen and Daniel Hershcovich. 2022. A dataset of sustainable diet arguments on Twitter. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40–58, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022.

QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3660–3669, New York, NY, USA. Association for Computing Machinery.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Hospice Houngbo and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, Maryland. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. Aifdb: Infrastructure for the argument web. In *Computational Models of Argument*, Frontiers in Artificial Intelligence and Applications.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins. 2023. Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *Computational Models of Argument*, Frontiers in Artificial Intelligence and Applications, pages 137–148.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020a. Transformer-based Argument Mining for Healthcare Applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, Santiago de Compostela / Online, Spain.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020b. Transformer-based argument mining for healthcare applications. In *European Conference on Artificial Intelligence*.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The discussion tracker corpus of collaborative argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1033–1043, Marseille, France. European Language Resources Association.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining*, pages 136–145, Florence, Italy. Association for Computational Linguistics.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P. Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the difficulty of evaluating baselines: A study on recommender systems. *ArXiv*, abs/1905.01395.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2310–2326, Mexico City, Mexico. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin,

Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based argument mining - modeling implicit argumentation using stance. In *Conference on Natural Language Processing*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685.

A Extended Descriptive and Experimental Details

This appendix provides additional data and details omitted from Sections 2 and 3.

A.1 Section 2

For Section 2 we present the entire decision-making process for the selection of the benchmark datasets used in this work, which is in Table 6.

A.2 Section 3

Figure 2 extends the analysis in Section 3.2 by showing pairwise Spearman's ρ correlations for all reproducible datasets, including those omitted from experiments due to their small size.

Figure 3 extends the vocabulary analysis from Section 3.2 by displaying word overlaps across all datasets with available data.

B Statistical Design Protocol

In this appendix we also explain our protocol for the best-practices of statistical testing as described in Section 4 and applied in Section 5.

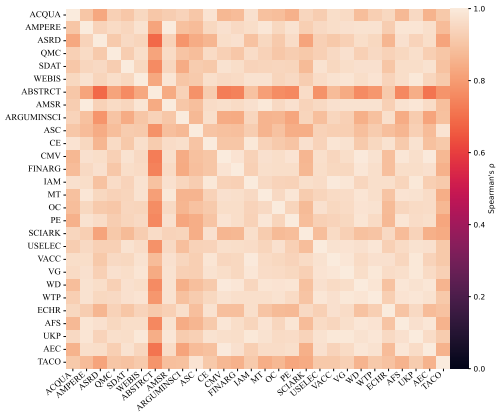


Figure 2: The correlations of the individual datasets (as well as the labels) in relation to the sentence-related features show a strong overall correlation ($\rho \geq 0.68$). Most strikingly, the ABSTRCT dataset stands out as medical texts exhibit different sentence structures from conventional ones, characterized by technical language, methodological details, and numerical values.

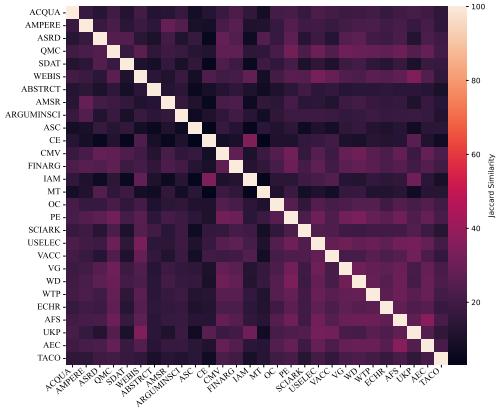


Figure 3: The word overlaps, measured by the Jaccard similarity between the vocabularies of two datasets, show that the datasets (as well as the labels) are generally distinct from each other. The overlaps range between 3–36%, with an average of 19%.

B.1 Two-Way Repeated Measures ANOVA

We employ a two-way repeated measures ANOVA to evaluate the effects of sampling (factor 1) and model choice (factor 2) on the macro F1 (dependent variable), with each dataset pair treated as a subject.

For valid inference, the following assumptions must be met:

- **Continuous Dependent Variable**: By definition, the macro F1 score is a continuous measure.
- **Within-Subject Design**: Each subject experiences every variation of both factors.
- **Normality**: The dependent variable is approximately normally distributed for each repeated measure (D'Agostino and Pearson's K^2 test).
- **Sphericity**: The variances of the differences between every pair of repeated measures are equal. If the Greenhouse-Geisser ϵ is below 0.75 (with values near 1 indicating compliance), we adjust the p-values (p_{corr}).

We can specifically evaluate for:

- **Sampling Effect**: Whether variations in data sampling (via different random seeds) influence model performance.
- **Model Choice Effect**: The performance differences among transformer models trained and evaluated on fixed samples. Each model is reinitialized in each trial using distinct random seeds to prevent carry-over effects.
- **Interaction Effect**: Whether the effect of sampling varies across the different models, offering insights into model stability under varying data conditions.

We evaluate the practical relevance of statistical significance using the effect size:

- **Generalized Eta Squared** (η^2_G): Proportion of the explained variance, interpreted as: ~0.01 (small), ~0.06 (moderate), ~0.14+ (strong).

B.2 One-Tailed Paired Student's t-Tests

Further, we conduct one-tailed paired t-tests as post-hoc analysis to identify directional differences (e.g., one model consistently outperforming another). These tests use the same assumptions as the prior ANOVA, except for sphericity. We apply the Bonferroni correction (p_{corr}) for multiple comparisons.

For these tests, we evaluate their practical relevance using the effect size:

- **Cohen's d**: The mean difference between paired conditions relative to the standard deviation of the differences, interpreted as: ~0.2 (small), ~0.5 (moderate), ~0.8+ (strong).

Dataset	Paper	Definition	Genre	Sent.	Binary	Reprod.	Related	Arg.	N-Arg.	Used
ACQUA	(Panchenko et al., 2019)	Argumentative	Mixed	Yes	Yes	Yes		1,949	5,236	Yes
AMPERE	(Hua et al., 2019)	Argumentative	Academic	Yes	Yes	Yes		6,729	242	No
ASRD	(Shnarch et al., 2020)	Argumentative	Spoken Debate	Yes	Yes	Yes		260	440	No
CDCP	(Niculae et al., 2017)	Argumentative	Online Debate	Yes	No					No
COMARG	(Boltužić and Šnajder, 2014)	Argumentative	Online Debate	No						No
EDIT	(Al-Khatib et al., 2016b)	Argumentative	Online Debate	Yes	No					No
IAC	(Walker et al., 2012)	Argumentative	Online Debate	No						No
MARG	(Mestre et al., 2021)	Argumentative	Spoken Debate	Yes	No					No
QMC	(Levy et al., 2018)	Argumentative	Encyclopedia	Yes	Yes	Yes		733	1,766	No
SDAT	(Hansen and Hershcovich, 2022)	Argumentative	Twitter Debate	Yes	Yes	Yes		387	210	No
WEBIS	(Al-Khatib et al., 2016a)	Argumentative	Online Debate	Yes	Yes	Yes		10,804	5,543	Yes
AAE	(Stab and Gurevych, 2014)	Claim-based	Academic	Yes	Yes	Yes	PE			No
ABSTRCT	(Mayer et al., 2020b)	Claim-based	Academic	Yes	Yes	Yes		1,308	7,323	Yes
AMECHR	(Teruel et al., 2018)	Claim-based	Legal	Yes	Yes	No				No
AMSR	(Fromm et al., 2021b)	Claim-based	Academic	Yes	Yes	Yes		839	561	No
ARGUMINSCI	(Lauscher et al., 2018)	Claim-based	Academic	Yes	Yes	Yes		6,554	9,548	Yes
ASC	(Wojatzki and Zesch, 2016)	Claim-based	Twitter Debate	Yes	Yes	Yes		147	568	No
CDC	(Aharoni et al., 2014)	Claim-based	Encyclopedia	Yes	Yes	Yes	CE			No
CE	(Rinott et al., 2015)	Claim-based	Encyclopedia	Yes	Yes	Yes		1,546	85,417	Yes
CMV	(Hidey et al., 2017)	Claim-based	Online Debate	Yes	Yes	Yes		979	1,593	Yes
CS	(Bar-Haim et al., 2017)	Claim-based	Encyclopedia	Yes	Yes	Yes	CE			No
DT	(Olshefski et al., 2020)	Claim-based	Spoken Debate	No						No
FINARG	(Alhamzeh et al., 2022)	Claim-based	Spoken Debate	Yes	Yes	Yes		4,607	8,310	Yes
IAM	(Cheng et al., 2022)	Claim-based	Mixed	Yes	Yes	Yes		4,808	61,715	Yes
MT	(Peldszus and Stede, 2015)	Claim-based	Microtext	Yes	Yes	Yes		112	337	No
OC	(Biran and Rambow, 2011)	Claim-based	Online Debate	Yes	Yes	Yes		702	7,824	No
PE	(Stab and Gurevych, 2017)	Claim-based	Academic	Yes	Yes	Yes		2,093	4,958	Yes
QT	(Hautli-Janisz et al., 2022)	Claim-based	Spoken Debate	Yes	No		AIFDB			No
RCT	(Mayer et al., 2018)	Claim-based	Academic	Yes	Yes	Yes	ABSTRCT			No
SCIARK	(Fergadis et al., 2021)	Claim-based	Academic	Yes	Yes	Yes		1,191	10,503	Yes
UGWD	(Habernal and Gurevych, 2017)	Claim-based	Online Debate	Yes	Yes	Yes	WD			No
USELEC	(Haddadan et al., 2019)	Claim-based	Spoken Debate	Yes	Yes	Yes		13,905	15,188	Yes
VACC	(Morante et al., 2020)	Claim-based	Online Debate	Yes	Yes	Yes		4,394	17,825	Yes
VG	(Reed et al., 2008)	Claim-based	Mixed	Yes	Yes	Yes	AIFDB	547	2,029	No
WD	(Habernal and Gurevych, 2015)	Claim-based	Online Debate	Yes	Yes	Yes		211	3,661	No
WTP	(Biran and Rambow, 2011)	Claim-based	Online Debate	Yes	Yes	Yes		1,135	7,274	Yes
ECHR	(Poudyal et al., 2020)	Conclusion-based	Legal	Yes	Yes	Yes		414	10,264	No
AFS	(Misra et al., 2016)	Conclusion-based	Online Debate	Yes	Yes	Yes	IAC	5,150	1,036	Yes
ARGSME	(Ajjour et al., 2019)	Conclusion-based	Online Debate	Yes	No					No
BASN	(Kondo et al., 2021)	Conclusion-based	Mixed	Yes	No					No
BIOARG	(Green, 2018)	Conclusion-based	Academic	Yes	No					No
DEMOSTHENES	(Grundler et al., 2022)	Conclusion-based	Legal	Yes	Yes	No				No
RSA	(Houngbo and Mercer, 2014)	Conclusion-based	Academic	Yes	No					No
AIFDB	(Lawrence et al., 2012)	AIF	Mixed	Yes	No					No
LAMECHR	(Habernal et al., 2023)	Custom Framework	Legal	Yes	No					No
ABAM	(Trautmann, 2020)	Evidence or Reasoning	Mixed	Yes	No		AURC			No
ASPECT	(Reimers et al., 2019)	Evidence or Reasoning	Mixed	Yes	No		UKP			No
AURC	(Trautmann et al., 2020)	Evidence or Reasoning	Mixed	Yes	Yes	No				No
BWS	(Thakur et al., 2021)	Evidence or Reasoning	Mixed	Yes	No		UKP			No
UKP	(Stab et al., 2018)	Evidence or Reasoning	Mixed	Yes	Yes	Yes		11,126	13,978	Yes
AEC	(Swanson et al., 2015)	Implicit-Markup	Online Debate	Yes	Yes	Yes	IAC	4,001	1,374	Yes
TACO	(Feger and Dietze, 2024b)	Inference-Information	Twitter Debate	Yes	Yes	Yes		864	868	Yes

Table 6: Summary of the 52 datasets from the reviewed papers, sorted by their applied definitions. Data collection followed the methodology described in Section 2.1, and selection criteria are detailed in Section 2.2. Empty entries indicate that the corresponding criteria were not further evaluated because a preceding criterion had already been rejected. The *Related* column indicates connections between datasets, like updates (e.g., AAE to PE, CDC to CE, RCT to ABSTRCT), additions of non-task-related features (e.g., CS adds stances to the claims from CE, ABAM adds aspects to the claims of AURC), or subsets from larger repositories (e.g., VG and QT from AIFDB, AEC and AFS from IAC).

5.4 Shared Task: Generalizability of Argument Identification in Context

In this short section, a shared task is described that will be published as one of four tasks of the Touché Lab on Argumentation Systems at CLEF 2026[1]. The description is based on the already accepted and peer-reviewed proposal and highlights the specific task derived from the final and previously presented paper (Feger, Boland, and Dietze, 2025) of this thesis.

5.4.1 Motivation

At the end of the previous paper (Feger, Boland, and Dietze, 2025), the problem is raised that the current assumptions about the SOTA performance of PLMs for AM cannot be sufficiently explained by an inherent capability of these models. Instead, the evidence suggests that these so-called SOTA models rely on topic-based shortcuts, which isolate datasets from one another and prevent genuine generalization across them.

Before turning to the main body of this shared task, Figure 5.2 should be considered to illustrate the problem of limited generalization more clearly and to provide a visual motivation for the underlying problem. For this purpose, a *Mini Language Model* (MiniLM)[2] (Wang et al., 2020, 2021), which is a smaller, distilled PLM from the BERT-family, was employed as a de-facto standard for obtaining PLM-based sentence embeddings (Reimers and Gurevych, 2019).

The *Uniform Manifold Approximation and Projection* (UMAP) visualization shows that the embeddings form dataset-specific clusters. For example, datasets such as UKP, which is intended to cover AM data across heterogeneous sources, or IAM, which is presented as the largest AM dataset integrating multiple sources, nonetheless form distinct, separate clusters. The same phenomenon can be observed for many other datasets as well.

Although some overlap and mixing occur, the datasets largely form isolated clusters that can be distinguished from each other. Hence, this visually supports the findings that the previously discussed PLMs encode dataset-specific characteristics in their pre-trained representation, which the standard SFT approach cannot overcome.

In this context, the contrast is particularly striking when comparing almost diametrically opposed datasets such as CE and IAM, or AFS and SCIARK, as well as academic datasets like SCIARK, AMPERE, ABSTRCT, and ARGUMINSCI.

Even within the same genre, and, as the paper (Feger, Boland, and Dietze, 2025) shows, with comparable sentence structures and identical target tasks, the representations are more strongly divided by dataset boundaries than unified by shared factors for identifying arguments.

[1] Advertised at CLEF 2026.
[2] Specifically: all-MiniLM-L6-v2.

Nevertheless, the visualization also shows that all datasets occupy a broader common space in which gaps remain but certain overlaps emerge, visible in the central area or at the fringes where clusters partially blend into one another.

Motivated by this, the shared task starts from the observation that reported SOTA in AM is inflated by thematic and dataset-specific shortcuts, while genuine generalization, although suggested by strongly correlated shared factors, is hindered because current SOTA models fail to remain invariant to these shortcuts. It is therefore important to recognize these limitations and to work toward genuine generalization, or at least to examine the problem in more depth. Taking on this challenge is a task for the AM research community.

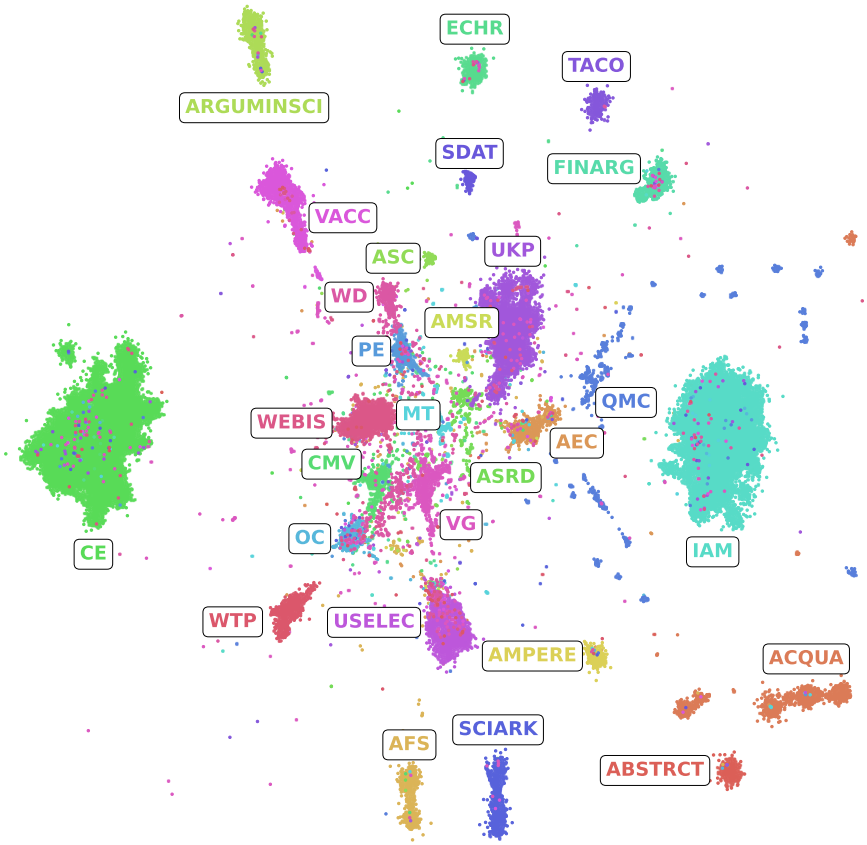


Figure 5.2: Visualization of the 28 reproducible datasets referenced in this paper (Feger, Boland, and Dietze, 2025). Representations were generated with MiniLM as a proxy for other PLMs of the BERT-family and visualized using UMAP. Despite some overlaps, clear clusters separate the datasets, showing that such models encode dataset-specific properties in their pre-trained representations for AM data. As shown in the paper, these properties cannot be overcome by superficial SFT in the classical approach.

Marc Feger, Julia Romberg, Katarina Boland, and Stefan Dietze.

"Generalizability of Argument Identification in Context"

In: Working Notes of the Conference and Labs of the Evaluation Forum: CLEF 2026, Touché Lab on Argumentation Systems, Jena, Germany, September 21-24, 2026.

Overview: Argument identification is a fundamental prerequisite for discourse analysis across domains such as political debate, online discussion, and scientific reasoning.

PLMs such as BERT (Devlin et al., 2019), designed for contextualized language representation, have demonstrated SOTA performance on established benchmarks. However, recent research suggests that SOTA performance often stems from exploiting spurious correlations (Thorn Jakobsen, Barrett, and Søgaard, 2021) and SCL (Geirhos et al., 2020), as benchmarks rely on specialized datasets that encourage models to capture dataset-specific patterns shaped by topic bias, argument definitions, and labeling schemes rather than abstractions that generalize across contexts (Feger, Boland, and Dietze, 2025).

Yet arguments are defined not only by form or content but also by their pragmatic function and contextualized use (Eemeren et al., 2014). Just as humans rely on context to identify and interpret arguments in discourse, so must machines. This task therefore examines how contextual cues can support automated argument identification, focusing on the impact of different types and amounts of context on building more generalizable and task-aligned systems.

Task: Given a sentence from a dataset along with metadata about its provenance, such as the source text and the dataset's annotation guidelines, predict whether the sentence can be annotated as an argument or not. In particular, the participants are encouraged to develop robust systems that generalize beyond lexical shortcuts to unseen datasets and investigate ways to exploit rich context information for this purpose.

Dataset: For the task, training data will be provided as a subset of the 17 benchmark datasets totaling about 345k labeled sentences and identified as most relevant for argument identification (Feger, Boland, and Dietze, 2025). This subset includes sentences each labeled as argument or no-argument, according to the respective dataset annotations, along with accompanying metadata such as IDs, generated training and development splits, links to original data sources and annotation guidelines, as well as the scripts used for data preparation.

Evaluation Setup: The systems will be evaluated on test data that differs from the development data. This includes partially or fully held-out portions of the datasets used for sampling, as well as newly created data reflecting diverse domains and annotation guidelines. This setup addresses the risk of data contamination in LLMs and participants' potential use of additional datasets during training. Generalizability will be measured using the macro F1 score. To evaluate the systems, the macro F1 score will be specified for each test dataset, along with the overall average of all these values.

Chapter 6

Conclusion

This thesis has demonstrated that *Argument Mining* (AM) is not a marginal curiosity but a cornerstone of modern automatic discourse analysis and, by extension, of *Natural Language Processing* (NLP) itself. In particular, two critical blind spots were addressed. The first is **G1**, which is referred to as **The Twitter Gap** and denotes the inherent neglect of conversational structure in social media discourse data, especially on Twitter. The second is **G2**, which is referred to as **The Generalization Gap** and describes the limited mutual validation of existing datasets and models that claim to define the *state-of-the-art* (SOTA) in AM.

For these gaps, this work has delivered three tightly interwoven primary studies (Feger and Dietze, 2024a,b; Feger, Boland, and Dietze, 2025). Each study was peer-reviewed and presented at flagship NLP venues (LREC-COLING, NAACL, and ACL). Together they provide new resources, methods, and empirical evidence that address the identified gaps while also opening new questions to guide future research. At the time of this thesis, these efforts have culminated in a shared task accepted at CLEF[1].

These papers, in turn, not only provided the foundation for several bachelor's and master's theses but also influenced four peer-reviewed secondary studies published in venues such as HCII and IEEE Access. These studies extend the findings to research domains as diverse as comment recommendation (Steimann, Feger, and Mauve, 2022), social science (Zelle, Grison, and Feger, 2024), cybersecurity with LLMs (Weber, Feger, and Pilgermann, 2024), and annotation tooling (Braun and Feger, 2025).

Building on the points stated here and the background knowledge gained throughout the thesis, this chapter concludes by revisiting the key findings regarding the research questions introduced in Section 1.2, with each main contribution discussed in relation to these. The future prospects arising from these results are then outlined before concluding with reflective thoughts that offer a meta-level interpretation of the findings and an optimistic outlook on AM research as a whole.

6.1 Key Findings

Turning to the key findings of this thesis, it first has to be noted that each main contribution pushes a different research question of this thesis forward, which will be discussed next.

Q1: ***Can arguments be extracted within entire Twitter conversations?***

First, *Twitter Arguments from COnversations* (TACO), the first AM corpus spanning entire Twitter conversations, has been compiled and annotated in the opening contribution of this thesis (Feger and Dietze, 2024a). By shifting the scope of analysis from isolated tweets to full discourse threads, TACO supplies a realistic ground truth for AM on Twitter and, therein, foregrounds the decisive roles of context dependence, pragmatic function, and tweet-level ambiguity in argument annotation and certain limitations regarding their automated recognition.

Based on the respective paper (Feger and Dietze, 2024a), the corresponding research question **Q1** can essentially be approached in two ways:

The first concerns the annotation itself. Across six different topics, 200 coherent conversations were annotated by six experts, resulting in a high inter-annotator agreement of $\alpha = 0.718$. Moreover, a definite majority decision regarding class assignment was reached in 95.6% of all 1,814 annotated tweets, which, in ordinary terms, can be regarded as strong. This reliability is supported by the iterative development of the guidelines, based on established definitions of argument components, the inference–information distinction from the Cambridge Dictionary[3], and the hierarchical categorization into four classes. For arguments, these are Reason (both elements) and Statement (inference only), while for no-arguments they are Notification (information only) and None (neither element). In addition, annotators' difficulty ratings indicate that at least 70% of all cases were generally perceived as easy.

On the other hand, it could thus be argued that not only the strong annotation results but also the classification outcomes speak for themselves. Following best practices, which included the use of different PLMs, hyperparameter tuning, and 10-fold cross-validation, BERTweet achieved a macro F1 of 85.06 for categorizing arguments vs. no-arguments and a macro F1 of 72.49 for classification, which can both be regarded as solid baseline performances.

Nevertheless, the other side of the answer to **Q1** emerges at the level of detail. Although arguments and no-arguments show similar classification performance with F1 scores of 85.34 and 84.77, the individual classes reveal a different picture. Most lie between 74.45 and 80.56, yet the Statement class falls sharply to 56.66 F1. Considering the macro and individual F1 scores, it is thus striking to see that BERTweet handles aggregated categories more reliably than the underlying classes. Error analysis shows that class-specific discrepancies cause this issue, with about 43% of misclassifications still correctly categorized.

Overall, the findings of this paper reveal a tension. On the one hand, the annotations were consistently perceived as easy by human annotators, and this perception was confirmed by the agreement measures. On the other hand, the best baseline results obtained with BERTweet reflect this to some extent, but a closer examination shows outcomes that point to confusion and a lack of clear separability rather than genuine task alignment.

Q2: ***Do SOTA models inherently predict arguments in Twitter conversations?***

Secondly, *WRAPresentations* (WRAP) was developed as an extension of BERTweet and addresses the representational challenges identified in TACO in a second contribution (Feger and Dietze, 2024b). To this end, a tailored CL framework with text augmentation promotes representation disentanglement, enhancing classification performance and enabling robust cross-topic transfer while maintaining invariance to paraphrasing. Building on this foundation, WRAP aligns with the structural semantics of arguments in the TACO hierarchy of tweets by modeling similarities and differences within different manifestations of inference and informational signals before task-specific SFT.

Regarding research question **Q2**, it is necessary to draw a further distinction based on the respective findings (Feger and Dietze, 2024b):

From this perspective, as shown by experimental results such as embedding space comparisons and cross-topic evaluations, the proposed CL method underlying WRAP captures class semantics more effectively. Quantitatively, WRAP raises the TACO baseline to 86.62 macro F1 for categories and 75.29 macro F1 for classes, with cross-topic experiments also yielding solid scores of 86.27 and 73.54 macro F1, respectively. Moreover, WRAP produces embeddings that approximate the ideal space derived from the TACO hierarchy and yields interpretable representations, with Reason aligned in 77% of cases, Statement in 64%, Notification in 65%, and None in 76%, whereas the vanilla BERTweet baseline achieves only 39%, 27%, 24%, and 78%. Thus it can be argued if WRAP constitutes an inherent SOTA model for TACO.

However, it is important to note that WRAP was explicitly designed and guided toward this capability through an extended pre-training phase aimed at abstracting the class semantics intended for TACO. This stands in contrast to standard best practices for sequence classification, in which PLMs depend on the downstream class information acquired only during conventional SFT following their general language pre-training.

Consequently, the vanilla approach to TACO does not exhibit this behavior beforehand. As demonstrated by the frozen parameter setup of BERTweet, where the representations cannot be adapted to downstream class information during SFT, the model yields 0.00 macro F1 for Statement. Moreover, Statement representations are misplaced in the Reason space in 42% of cases, while Reason itself accounts for only 39% there. Similarly, 27% of Statement embeddings occur in regions dominated by Reason with 30% and Notification with 29%. The Notification space, in turn, is almost evenly fragmented across Reason with 23%, Statement with 28%, Notification itself with 24%, and None with 25%. Considered as a whole, this indicates that certain classes cannot be classified at all before SFT or appear only in a heavily distorted manner, overlapping with the semantics of other classes.

At the same time, experiments suggest that general pre-training on augmented data can improve performance and yield results and representations somewhat comparable to those achieved with WRAP. Nevertheless, neither the vanilla approach nor augmented pre-training of BERTweet reaches the same level of robustness and consistency as the specifically designed CL framework inherent to WRAP.

In this sense, the findings for TACO demonstrate that established best practices for sequence classification using conventional SFT do not necessarily allow PLMs to rely on an inherent ability to identify tweets according to the intended class semantics.

Instead, these results raise a broader question regarding the applicability of WRAP to other datasets and, more fundamentally, the extent to which existing SOTA baselines suffer from similar limitations and therefore fail to reflect inherent task alignment for identifying arguments according to the intended semantics.

Q3: Does reported progress in argument identification reflect genuine advances?

Third, an extensive meta-analysis of the cross-applicability of the most relevant AM literature was conducted as the final contribution of this thesis (Feger, Boland, and Dietze, 2025). This included a comprehensive review of 52 AM studies and their datasets. From these, the 17 most relevant works were identified. These provide sentence-level annotations of arguments and no-arguments, represent reproducible and up-to-date sources, and contain a sufficient size with at least 850 instances per label.

About research question **Q3**, the following conclusions can be drawn from the contribution (Feger, Boland, and Dietze, 2025):

Survey-based comparison of the applied definitions and annotation guidelines reveals that the datasets often build on one another in their conception of argument, and, although they approach the same problem from different perspectives, they are not fundamentally distinct. This is supported by a more detailed structural investigation of direct comparability revealing that the datasets are strongly correlated across 24 surface-level sentence properties such as length, entropy, or part-of-speech tag distributions ($\rho \geq 0.68$). At the word level, however, the respective vocabularies of the datasets show two trends in terms of Jaccard similarity. It was shown that argument cues unite with at least 73% overlap, whereas content words divide with overlaps between 7–36%.

What stands out most is that, while sentence-level comparisons across labels show similar patterns, the words expressing core semantic content remain largely distinct, with overlaps below 48% and averaging 19%. This highlights lexical separation in specific words and suggests potential shortcuts by focusing on content rather than argument signals.

Concerning the re-evaluation of the baseline results and the transferability between datasets, two experimental approaches were pursued. In the first approach, models were trained pairwise on one dataset and tested on another, producing a 17×17 experiment matrix. This setup provided both i.i.d benchmarks on the main diagonal and o.o.d evaluations across datasets within a single report. In the second approach, one dataset was tested successively while training was conducted on the combination of all remaining datasets, thereby addressing both o.o.d training and testing. For all these experiments, the PLMs BERT, RoBERTa, DistilBERT, and WRAP were employed.

While the i.i.d results yielded strong baseline scores averaging 0.79 macro F1, the outcomes of the respective models across all experiments remained only between 0.56 and 0.61 macro F1, with WRAP achieving the numerically best results. The same pattern emerged in the joint benchmark experiments, where the models reached slightly higher average scores of 0.63 to 0.66 macro F1, with WRAP performing best. Yet they still fell far short of the baseline performance. A prime example of the contrast between strong i.i.d baseline results and weak o.o.d generalization is the AEC dataset, which appears relatively simple as it relies on exactly five keywords to distinguish arguments from no-arguments. While benchmark results of 0.96 macro F1 are achieved, transfer performance drops and is strongly located around the chance level of 0.5 macro F1.

Regarding the divergent i.i.d and o.o.d results and the identified shortcut opportunities, the experiments were repeated as controlled manipulation ablations in which only content-related words were retained. The findings showed that all models, except for WRAP, largely maintained their performance, and in the case of FINARG, even improved by 0.01 macro F1.

A threefold repetition using different random initializations and samples of all these experiments likewise showed no significant inconsistencies and overall significance ($p < .05$) of the experiments, as measured by a repeated-measures *Analysis of Variance* (ANOVA).

As for the research questions, this paper has shown that contemporary benchmarks capture dataset idiosyncrasies in the form of dataset-specific shortcuts rather than genuine argument signals. This finding reframes recent benchmark improvements and underscores the need for a methodological shift toward data, models, and evaluation setups that can generalize beyond narrowly optimized scenarios.

Taken more broadly, these limitations can also be read as a constraint of standard evaluation practices, which often impose SOTA expectations on models that exceed what they can reasonably achieve. These PLMs are inherently limited to exploiting superficial textual properties, since they receive only isolated input sequences. Expecting them to replicate human reasoning, which relies on broader context such as surrounding sentences and annotation guidelines, is therefore misguided. In this light, the present findings lay the groundwork for a shared task specifically designed to confront these limitations.

6.2 Future Work

The findings presented in this study should not be regarded as final, but rather as a foundation for future research in the field of *Argument Mining* (AM).

Future work could explore the following areas, either individually or in ways that complement and strengthen one another:

Comprehensive and Standardized Benchmarks: Building on the insights from this thesis, there is a clear need for greater standardization of benchmark datasets and a more consistent definition of fundamental task requirements in AM. Rather than focusing on isolated or narrowly scoped benchmarks, future efforts should emphasize the development of comprehensive evaluations, generalizable methodologies, and unified benchmark data.

A primary objective should be the robust detection of arguments. One possible direction would be the creation of integrated benchmark collections, akin to *General Language Understanding Evaluation* (GLUE) (Wang et al., 2018), which encompass various NLP sub-tasks. For example, the GLUE benchmark comprises a variety of tasks, including, among others, *Multi-Genre Natural Language Inference* (MNLI) (Williams, Nangia, and Bowman, 2018), *Stanford Sentiment Treebank* (SST) (Socher et al., 2013) for sentiment analysis, the *Corpus of Linguistic Acceptability* (CoLA) (Warstadt, Singh, and Bowman, 2019), and *Recognizing Textual Entailment* (RTE) (Dagan, Glickman, and Magnini, 2006), which originates from a series of annual textual entailment challenges (Wang et al., 2018).

In this sense, instruction-tuning datasets (Ouyang et al., 2022; Zhang et al., 2024; Stahl et al., 2025) may also serve as inspiration, as they combine a diverse range of tasks paired with varying natural language instructions and contextual information to support the generalized training of LLMs (Radford and Narasimhan, 2018; Mishra et al., 2022).

For future work, following the approach already adopted in the shared task[1] derived from this thesis, an important step would be the integration of existing benchmarks (Feger, Boland, and Dietze, 2025). Beyond this, it may also be valuable to not only group the data but also store task-related information and requirements, such as definitions, guidelines, and contextualized materials like the source documents analyzed alongside the annotated data.

Contextualization and Large Language Models: Preliminary research suggests that prompting LLMs with natural language instructions yields performance on par with baseline models across the QMC, ASRD, and IAM benchmarks (Chen et al., 2024). In this regard, it has already been reported for PE, CDCP, and ABSTRCT that instruction tuning can also increase the performance of LLMs even further (Cabessa, Hernault, and Mushtaq, 2025).

In this sense, one especially interesting direction for AM involves re-formulating traditional classification tasks, typically handled as sequence classification by PLMs from the BERT-family, into generation-based tasks for LLMs (Cabessa, Hernault, and Mushtaq, 2025) like those models from the LLaMA- or GPT-family.

This opens new avenues for examining how task performance relates to model contextualization and the injection of task information. Instead of relying on PLM classifiers that implicitly learn context from labeled data (Radford and Narasimhan, 2018; Feger, Boland, and Dietze, 2025), future work might embed full task information such as documents, definitions, and annotation guidelines directly into LLMs, creating training conditions that more closely mirror those of human annotators (Pangakis and Wolken, 2024).

However, the limited task alignment observed for PLMs (Feger, Boland, and Dietze, 2025) may also extend to LLMs, an issue that has already been noted for early GPT models (Geirhos et al., 2020). In this context, it remains debated whether these models develop genuine semantic understanding or mainly reproduce output patterns shaped by dataset-specific artifacts (Zhao et al., 2021; Kung and Peng, 2023), by factors related to scale, or by memorization effects due to exposure to existing benchmarks during training (Saphra et al., 2024).

Interpretability and Error Analysis: Building on this, it is essential to understand how such models operate in the execution of tasks. In the classic PLM setup, they are trained on annotated datasets with data and labels, and their predictions and internal mechanisms are subsequently reconstructed through methods such as *SHapley Additive exPlanations* (SHAP) and *Local Interpretable Model-agnostic Explanations* (LIME) (Ribeiro, Singh, and Guestrin, 2016; Lundberg and Lee, 2017; Bhatti, Ahmad, and Park, 2021), feature attribution techniques (Thorn Jakobsen, Barrett, and Søgaard, 2021), or manipulation experiments (Geirhos et al., 2020; Feger, Boland, and Dietze, 2025). Although these approaches can yield valuable insights, they illuminate only limited aspects of model behavior and remain inherently bounded by the adequacy of the explanatory framework itself (Liu, Yin, and Wang, 2019; Molnar, Casalicchio, and Bischl, 2020).

Beyond that, there is also the question of how far experimental design, corresponding flaws, and inherent model limitations are interdependent (Lawsen, 2025; Shojaee et al., 2025).

Nonetheless, it might be interesting to investigate whether directly prompting an LLM to generate explanatory reasoning for its decisions (Liu, Yin, and Wang, 2019; Gu, Tafjord, and Clark, 2024) improves interpretability and truly mirrors the model's decision process. This can involve techniques such as chain-of-thought prompting (Wei et al., 2022c), where the model generates an answer step by step, or least-to-most prompting (Zhou et al., 2023), which guides the model through explicitly decomposed subproblems toward a final decision.

Future research in AM could therefore investigate the interpretability of model decisions and how far these align with the justifications underlying human annotations in specific AM tasks.

Accounting for Human Abilities: Beyond models and datasets, attention should also be directed toward the humans whose argumentative abilities constitute the very task to be learned. These abilities often rely on world knowledge or reasoning over information not explicitly contained in the input. This highlights the need for benchmarks that move beyond hard class labels to also include the explanations annotators provide for their labeling decisions.

Inspiration for AM can be drawn from resources such as the *Evaluating Rationales And Simple English Reasoning* (ERASER) benchmark (DeYoung et al., 2020), which integrates multiple datasets and tasks like *Fact Extraction and VERification* (FEVER) (Thorne et al., 2018), where human annotations of rationales (supporting evidence) have been collected. In this setting it is not sufficient to predict a True or False label because models must also identify the relevant text passages that justify their predictions and thus align their rationales with human evidence annotations. Additional examples are provided by datasets such as *explanation-augmented Stanford Natural Language Inference* (e-SNLI) (Camburu et al., 2018) or *Common Sense Explanations* (CoS-E) (Rajani et al., 2019), where language models are trained to generate explanations in natural language that can be employed both during training and inference.

Incorporating such explanatory information may also strengthen the assessment of generalizability and reproducibility when annotation guidelines are reapplied by diverse annotators across differently distributed datasets. In this way, future research can explore the extent to which benchmarks and perspectives on arguments converge, how much generalization can realistically be expected from models and reported SOTA scores, and to what degree humans themselves can generalize in debates or reach consistent explanations for distinguishing arguments from no-arguments, if such is possible at all.

6.3 Closing Thoughts

The end of this thesis marks the beginning of new questions and challenges.

While the identification of limitations in current *Argument Mining* (AM) research represents one of the central contributions of this thesis, it also reveals that existing approaches, including those proposed herein, are often designed in isolation. Thereby, it cannot be excluded that they rely on features that are not intrinsically important for the task, even though they achieve *state-of-the-art* (SOTA) performance. This, in turn, raises doubts about whether later steps can be considered reliable when built on potentially flawed argument identification.

Undoubtedly, the findings presented in this thesis are not entirely exempt from these concerns, although they demonstrate comparatively better generalizability and overall numerical performance. Nonetheless, this thesis highlights the often-neglected need to critically examine how such results are achieved and whether they genuinely support the transfer of anticipated findings beyond the specific conditions in which they were obtained.

One possible explanation for this narrow perspective in the field is that AM is inherently tied to specific pragmatic-dialectical contexts in which arguments are articulated and interpreted. The recognition and interpretation of arguments can vary significantly from case to case, influenced not only by the subjective perceptions and prior knowledge of the recipient but also by a range of external factors such as temporal context, cultural background, discourse setting, and communicative intent. Such contextual factors influence how arguments are interpreted and annotated, making universal approaches difficult to develop. Unquestionably, this may suggest a need for locally tailored AM solutions that are closely aligned with the specific requirements of individual applications.

While this perspective has undoubtedly influenced the current AM landscape, it does not address the fundamental challenge of how interpretations and justifications are constructed within individual benchmark settings that are often presented as defining the SOTA. Moreover, it falls short of establishing a universally applicable paradigm for discourse analysis by isolating the research efforts.

At the same time, this thesis demonstrates that generalization across different benchmarks and models is indeed desirable and fertile for scholars, provided that appropriate attention is and will be given to the model perception, design of training strategies, the selection and preparation of data, and the development of robust evaluation frameworks.

It can thus be speculated if current limitations of AM research are symptoms of a fundamental research dilemma in which isolating novelty and the incremental pursuit of better benchmark scores often replace the search for more profound insights that actually live up to their promise (Lipton and Steinhardt, 2019; Bender and Koller, 2020). Relevant but not-so-good results are rejected from scholarly discourse and disappear down the drawer (Rosenthal, 1979), while standardized metrics are transfigured into targets (Strathern, 1997) whose numerical optimization says little about the actual quality of how they are satisfied (Goodhart, 1984).

Cheers!

Appendix

A.1 Annotation Guide for the TACO Dataset

Annotation

Dear annotators,

Arguments may be written in a wide variety of forms and may not always be directly apparent. Likewise, an **Argument** may address one topic or more, be rhetorically and linguistically heavy, or the inferences stated and reasons given may not be immediately obvious. Nevertheless, the individual text passages in a tweet can form an **Argument** in their effect. It is not important for you whether the **Argument** is right or wrong or whether it corresponds to your personal world view and opinion. To simplify this task, we differentiate between tweets that contain an inference as a key component of an **Argument** and those that do not.

To identify **Argument** constituents, we use the Cambridge Dictionary:

- **Inference**: *a guess that you make or an opinion that you form based on the information that you have*.
- **Information**: *facts or details about a person, company, product, etc.*.

Tweets that make an **Argument** can be classified as either:

- **Statement**: a tweet where only **Inference** is presented like *something that someone says or writes officially, or an action done to express an opinion*.
- **Reason**: a tweet where the **Inference** is based on **Information** mentioned in the tweet such as a source-reference or quotation, and thus reveals the author's motivation *to try to understand and to make judgments based on practical facts*.

Tweets that make **No-Argument** can be classified as either:

- **Notification**: a tweet that limits itself to only provide information like media channels promoting their latest articles.
- **None**: a tweet having neither inference or information, including hate-speech or spam.

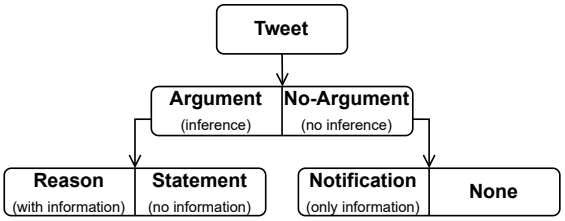


Figure 1: Hierarchy of Argument Mining on Twitter.

Before answering the questions ask yourself:

1 What does the author of the tweet want you to believe (**Inference**)?
2 What does the author of the tweet want you to know (**Information**)?
3 Does (1) emerge from (2)?

In the case of uncertainty consider the following:

1 Look the tweet up on Twitter
2 Follow the conversation down to the tweet.
3 *Concrete dates* are **Information**
4 *Quotes* and *headlines* are **Information**
5 *Experiences* are **Information**
6 *Considerations* are **Information**
7 *Standpoints* are **Inference**
8 *Rhetorical questions* are **Inference**
9 *Hashtags can* provide **Inference**
10 *Insults are no* **Inference**
11 *Exceptions might indicate* **Inference**
12 Tend to **0** in case of absolute doubt

Task:

1 Detect **Information** and **Inference**.
Mark each identified component with **1** else use **0**

2 For self-monitoring, specify if you have strong concerns about interpreting the tweet too strongly and selecting the component too artificially? Therefore, how well did the components emerge?
Mark the **Difficulty** with:

1: Easy (The component(s) is/are directly visible)
2: Normal (Repeated consideration reveals the component(s))
3: Hard (Strong concerns about any component(s) present)

Examples:

Tweet 1 (Reason):

The formula: Not everyone who voted Leave is racist. But everyone who's racist voted Leave. Not everyone who voted Leave is thick. But everyone who's thick voted Leave. The thick racists therefore called the shots, whatever the thoughts of the minority of others. #thick #Brexit

Topics: Brexit, Racism, Minority, Leave, Thick
Inference: 1
Information: 1
Difficulty: 1

In this example, **Information** and **Inference** are chosen because, on the one hand, a consideration is made, and on the other hand, a concrete image of Brexit is drawn. Specifically, the consideration *"everyone who's racist voted Leave"* and *"everyone who's thick voted Leave"* leads to the **Inference** (*"The thick racists [...] called the shots"*) relative to the exceptions *"Not everyone who voted Leave is racist"* and *"Not everyone who voted Leave is thick"*. Likewise, it alludes to the lack of hearing minorities, which is intended to emphasize the perceived backward step in Brexit (*#thick #Brexit*). Obviously, this tweet hints at the right-winged aspects of Brexit and reminds of the negative (*"thick"*) characteristics of marginalization as they are often present in racism and right-winged movements. For the entirety of the tweet, it is not relevant here whether the **Inference** and **Information** are true or false but contribute to the **Argument**.

Tweet 2 (Notification):

#Mexico top court declares criminalizing #Abortion_06 unconstitutional-JURIST-News Mexico's Supreme Court of Justice of the Nation ruled Tuesday that total #criminalization of #abortion is unconstitutional. #AbortoLegalMexico #USSupremeCourt #SupremeCourt `https://t.co/xLj5PZijOL`

Topics: Mexico, Abortion, Supreme Court, USA
Inference: 0
Information: 1
Difficulty: 1

In this example, only a summary of the article in the link below is presented in the tweet. To see this, the URL to the article must be followed. Likewise, the tweet summarizes and recites the main message of the article, as it also occurs in the headline.

Tweet 3 (Statement):

Possible tangible benefit of #Brexit: a fairer immigration system.

Topics: Immigration, Brexit, Benefit, Possibility
Inference: 1
Information: 0
Difficulty: 1

In this Tweet, only an **Inference** in the form of an assertion is made from anticipated and implicit **Information** (*the immigration system is unfair*). Any further elements are missing; the assertion is assigned to the topic in a substantial but challenging role.

Tweet 4 (None):

@sinnfeinireland Blah blah blah blah blah blah

Topics: Ireland $\rightarrow$ Brexit (?)
Inference: 0
Information: 0
Difficulty: 1

After a brief consideration, the individual components for an **Argument** are to be strongly doubted.

Tweet 5 (Reason):

Love my new favorite abortion shirt! Abortions are as good as, if not better than, all medical procedures! Get yourself this badass shirt and donate to @TEAFund: `https://t.co/vSiwJaMZjh` #abortion #AbortionBan #AbortionIsHealthcare #txlege #texas `https://t.co/5iTXpODVIv`

Topics: Abortion, Texas, Healthcare, Donation, Abortion-Shirt
Inference: 1
Information: 1
Difficulty: 2

This tweet primarily looks like an advertisement and call for donations. However, on closer reading, both main components are present as well. Here, the **Information** is provided by following the first URL to an information page or by the consideration *"Abortions are as good as, if not better than, all medical procedures"*. Where the consideration appears to have priority over the information page. Likewise, the **Inference** is in the hashtag #*AbortionIsHealthcare*. In both cases, the components of an **Argument** become apparent after short consideration (**Difficulty: 2**).

A.2 Statement on the Use of Artificial Intelligence (AI) Tools

This thesis was written by the author with the support of conventional academic tools and resources. ChatGPT[1] and LanguageTool[2] were employed in limited capacities to assist with clarifying technical phrasing, grammar checking, and paraphrasing. At no point did these tools generate original research content, interpret data, or substitute for the author's intellectual contributions and critical analysis. All AI-assisted input was reviewed and edited by the author to ensure alignment with the academic standards and integrity of the work.

[1] chatgpt.com
[2] languagetool.org

Glossary

ABSTRCT Abstracts of Randomized Clinical Trials. 29, 32, 33, 37, 40, 105, 114, 167

ACQUA Argumentation in Comparative Question Answering. 30, 36

ADU Argumentative Discourse Unit. 23, 24

AEC Argument Extraction Corpus. 28, 32, 35, 113

AFS Argument Facet Similarity. 28, 31, 35, 105

AI Artificial Intelligence. 29

AIF Argument Interchange Format. 16

AM Argument Mining. vii, 1, 3–11, 13–18, 20, 22, 23, 25, 26, 28, 29, 31, 33, 36, 39, 40, 49, 50, 59, 70, 71, 83, 86–88, 105, 106, 109, 110, 112, 114–117, 167, 168

AMPERE Argument Mining for Peer Reviews. 29, 31, 105

AMSR Argument Mining in Scientific Reviews. 29, 35

ANOVA Analysis of Variance. 113

API Application Programming Interface. 29, 41

ARGUMINSCI Argumentative Analysis of Scientific Publications. 29, 32, 33, 105

ASC Argument Stance Classification. 29, 31

ASRD Automatic Speech Recognition of Debates. 30, 32, 37, 40, 114, 167

AW Argument Web. 16

BART Bidirectional and AutoRegressive Transformers. 39

BERT Bidirectional Encoder Representations from Transformers. 38–40, 46–48, 60, 63, 87, 105–107, 112, 114, 167, 168

BERTweet A pre-trained language model for English Tweets. 46, 50, 69, 70, 110, 111

BiLSTM Bidirectional Long Short-Term Memory. 38, 40

CDC Context Dependent Claim. 30, 35

CDCP Consumer Debt Collection Practices. 28, 32, 37, 40, 114, 167

CE Context Dependent Evidence. 30, 35, 105

CL Contrastive Learning. 6, 11, 66–68, 70, 111, 168

CMC Computer-Mediated Communication. 1, 19, 28

CMV Change My View. 28, 31, 34

CoLA Corpus of Linguistic Acceptability. 114

COMPSENT Comparative Sentences. 30

CoS-E Common Sense Explanations. 116

CRF Conditional Random Field. 40

D-BAS Dialog-Based Online Argumentation. 15

DistilBERT Distilled BERT. 40, 46, 87, 112

DNN Deep Neural Network. 40, 84

ECHR European Court of Human Rights. 30, 31

EG Evidence Graph. 40

ELMO Embeddings from Language Models. 38

ERASER Evaluating Rationales And Simple English Reasoning. 116

e-SNLI explanation-augmented Stanford Natural Language Inference. 116

FastText FastText. 40

FEVER Fact Extraction and VERification. 116

FFN Feedforward Neural Network. 38, 40, 48, 60

FINARG Finanical Arguments. 30, 34, 113

FLAN Fine-tuned LAnguage Net. 40

GLUE General Language Understanding Evaluation. 114

GPT Generative Pre-trained Transformer. 17, 39, 114, 115

HB Hybrid Baseline. 40

IAM Integrated Argument Mining. 30, 35, 37, 40, 105, 114, 167

i.i.d independent and identically distributed. 84–86, 112, 113, 168

InferSent Inferential Sentence Embeddings. 40

LIME Local Interpretable Model-agnostic Explanations. 115

LLaMA Large Language Model Meta AI. 17, 39, 40, 114

LLM Large Language Model. 9, 17, 37, 39, 40, 84, 107, 109, 114, 115, 167

LM Language Model. vii, 3, 5–7, 9, 10, 13, 16, 36–40, 167

LRM Large Reasoning Model. 17

MiniLM Mini Language Model. 105, 106, 168

ML Machine Learning. 23, 59, 84, 86

MLM Masked Language Modeling. 38, 39

MNLI Multi-Genre Natural Language Inference. 114

MT Microtext. 30, 34

NB Naive Bayes. 40

NLM Neural Language Model. 38–40

NLP Natural Language Processing. 1, 3, 9, 10, 16, 17, 36, 38, 40, 59, 60, 86, 109, 114

NN Neural Network. 38

NSP Next Sentence Prediction. 38, 39

NTP Next Token Prediction. 39

OC Online Comments. 28

o.o.d out-of-distribution. 6, 84–88, 112, 113, 168

PALM Pathways Language Model. 39

PE Persuasive Essays. 29, 31, 33, 34, 37, 40, 114, 167

PLM Pre-trained Language Model. 38–40, 46–48, 50, 60, 63, 64, 67, 70, 71, 88, 105–107, 110–115, 167, 168

QMC Querying of Main Concepts. 30, 35, 37, 40, 114, 167

RCT Randomized Clinical Trials. 29, 31, 33

RF Random Forest. 70

RL Representation Learning. 38, 59–61, 64, 67–69

RoBERTa Robustly Optimized BERT Pre-training Approach. 39, 40, 46, 87, 112

RTE Recognizing Textual Entailment. 114

SCIARG Scientific Arguments. 29

SCIARK Scientific Argumentation Knowledge. 29, 32, 33, 105

SciBERT Scientific BERT. 40

SCL Shortcut Learning. 84–88, 107, 168

SDAT Sustainable Diet Arguments on Twitter. 29, 32, 35

SFT Supervised Fine-Tuning. 23, 39, 50, 60, 66, 67, 70, 105, 106, 111, 112, 168

SHAP SHapley Additive exPlanations. 115

SL Supervised Learning. 23

SLM Statistical Language Model. 36, 38–40, 70

SoLAr Social Linked Arguments. 15

SOTA state-of-the-art. vii, 3–5, 7, 11, 36, 40, 46, 50, 70, 71, 83, 86–88, 105–107, 109, 111–113, 116, 117

SST Stanford Sentiment Treebank. 114

SVM Support Vector Machine. 40

TACO Twitter Arguments from COnversations. 5, 6, 8, 9, 11, 29, 33, 36, 41, 46, 49, 59, 63–71, 83, 87, 88, 110–112, 168

TF-IDF Term Frequency-Inverse Document Frequency. 70

TSC Tweet Stance Classification. 29

UKP Ubiquitous Knowledge Processing. 30, 32, 35, 105

UMAP Uniform Manifold Approximation and Projection. 105, 106, 168

USELEC U.S. Election Debate. 30, 31, 35

VACC Vaccination Corpus. 28

VG Various Genres. 30, 34

WD Web Discourse. 28, 33

WEBIS Web Information Systems. 28, 32

Word2Vec Word to Vector. 38, 40, 63

WRAP WRAPresentations. 6–9, 11, 46, 59, 70, 71, 83, 87, 88, 111–113

WTP Wikipedia Talk Pages. 28

XGBoost eXtreme Gradient Boosting. 40

Bibliography

Feger, Marc and Stefan Dietze (May 2024a). "TACO – Twitter Arguments from COnversations". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 15522–15529. URL: https://aclanthology.org/2024.lrec-main.1349 (cit. on pp. 5, 8–10, 14, 16, 27–29, 36, 41, 43, 49, 50, 109, 110).

– (June 2024b). "BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2256–2266. DOI: 10.18653/v1/2024.findings-naacl.146. URL: https://aclanthology.org/2024.findings-naacl.146 (cit. on pp. 6, 8–10, 36, 50, 59, 63, 69, 71, 109, 111).

Pfungst, Oskar (1907). *Das Pferd des Herrn von Osten (der Kluge Hans): Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*. German. Digitized by Google from the library of the University of Michigan. WorldCat OCLC: 14803679. Leipzig: J. A. Barth, p. 198. URL: http://books.google.com/books?id=wLMxb2_9obYC&oe=UTF-8 (cit. on p. 83).

Shannon, C. E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on pp. 19, 45).

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort. An introduction to human ecology*. PsycINFO Database Record (c) 2016 APA, all rights reserved. Oxford, England: Addison-Wesley Press, pp. xi+573 (cit. on p. 86).

Rosenblatt, Frank (1958). "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519 (cit. on pp. 40, 84).

Rosenthal, Robert (1966). *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Crofts (cit. on p. 83).

Perelman, Chaïm, L. Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver (1969). *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press. ISBN: 9780268004460. URL: http://www.jstor.org/stable/j.ctvpj74xx (visited on 06/29/2025) (cit. on pp. 1, 19, 20).

Schramm, W. and D.F. Roberts (1971). *The Process and Effects of Mass Communication*. University of Illinois Press. ISBN: 9780252001970. URL: https://books.google.de/books?id=pUNQAQAAIAAJ (cit. on pp. 1, 13, 17–19).

Rittel, Horst W. J. and Melvin M. Webber (June 1973). "Dilemmas in a General Theory of Planning". In: *Policy Sciences* 4.2, pp. 155–169. ISSN: 1573-0891. DOI: 10.1007/BF01405730. URL: https://doi.org/10.1007/BF01405730 (cit. on p. 18).

O'Keefe, D. J. (1977). "Two concepts of argument". English. In: *Journal of the American Forensic Association* 13, pp. 121–128 (cit. on p. 17).

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results". In: *Psychological Bulletin* 86.3, pp. 638–641. DOI: 10.1037/0033-2909.86.3.638 (cit. on p. 117).

Marr, David (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. USA: Henry Holt and Co., Inc. ISBN: 0716715678 (cit. on pp. 83, 84, 86).

Cohen, Robin (July 1984). "A Computational Theory of the Function of Clue Words in Argument Understanding". In: *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. Stanford, California, USA: Association for Computational Linguistics, pp. 251–258. DOI: 10.3115/980491.980546. URL: https://aclanthology.org/P84-1055/ (cit. on pp. 1, 21, 32, 35).

Goodhart, C. A. E. (1984). "Problems of Monetary Management: The UK Experience". In: *Monetary Theory and Practice: The UK Experience*. London: Macmillan Education UK, pp. 91–121. ISBN: 978-1-349-17295-5. DOI: 10.1007/978-1-349-17295-5_4. URL: https://doi.org/10.1007/978-1-349-17295-5_4 (cit. on p. 117).

Hitchcock, David (1985). "Enthymematic Arguments". In: *Informal Logic* 7.2. DOI: 10.22329/il.v7i2.2707 (cit. on p. 21).

Thomas, Stephen N. (1986). *Practical Reasoning in Natural Language*. 3rd. Englewood Cliffs, NJ: Prentice Hall (cit. on pp. 20, 21, 31, 42).

Cohen, Robin (1987). "Analyzing the Structure of Argumentative Discourse". In: *Computational Linguistics* 13, pp. 11–24. URL: https://aclanthology.org/J87-1002/ (cit. on pp. 1, 21, 27, 32).

Mann, William C. and Sandra A. Thompson (1987). "Rhetorical Structure Theory: Description and Construction of Text Structures". In: *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*. Ed. by Gerard Kempen. Dordrecht: Springer Netherlands, pp. 85–95. ISBN: 978-94-009-3645-4. DOI: 10.1007/978-94-009-3645-4_7. URL: https://doi.org/10.1007/978-94-009-3645-4_7 (cit. on pp. 21, 32).

Freeman, James B. (1991). *A Theory of Argument Structure*. Berlin, Boston: De Gruyter Mouton. ISBN: 9783110875843. DOI: doi:10.1515/9783110875843. URL: https://doi.org/10.1515/9783110875843 (cit. on pp. 3, 20, 21, 27, 31–35, 42).

Knott, Alistair and Robert Dale (1994). "Using linguistic phenomena to motivate a set of coherence relations". In: *Discourse Processes* 18.1, pp. 35–62. DOI: 10.1080/01638539409544883. eprint: https://doi.org/10.1080/01638539409544883. URL: https://doi.org/10.1080/01638539409544883 (cit. on pp. 21, 32).

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). "Support-Vector Networks". In: *Mach. Learn.* 20.3, pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. URL: https://doi.org/10.1023/A:1022627411411 (cit. on p. 40).

Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 1137–1143. ISBN: 1558603638 (cit. on p. 84).

Walton, Douglas (1996). *Argumentation Schemes for Presumptive Reasoning*. 1st. Routledge. DOI: 10.4324/9780203811160. URL: https://doi.org/10.4324/9780203811160 (cit. on pp. 1, 27, 34).

Mitchell, Thomas M. (1997). *Machine Learning*. 1st ed. USA: McGraw-Hill, Inc. ISBN: 0070428077 (cit. on p. 23).

Schuster, M. and K.K. Paliwal (1997). "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2681. DOI: 10.1109/78.650093 (cit. on p. 38).

Strathern, Marilyn (1997). ""Improving ratings": audit in the British University system". In: *European Review* 5.3, pp. 305–321. DOI: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4 (cit. on p. 117).

Wolpert, D.H. and W.G. Macready (1997). "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82. DOI: 10.1109/4235.585893 (cit. on pp. 22, 23).

Jelinek, Frederick (1998). *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press. ISBN: 0262100665 (cit. on p. 36).

Lewis, David D. (1998). "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval". In: *Proceedings of the 10th European Conference on Machine Learning*. ECML '98. Berlin, Heidelberg: Springer-Verlag, pp. 4–15. ISBN: 3540644172 (cit. on p. 40).

Freeman, James B. (May 2000). "What Types of Statements are There?" In: *Argumentation* 14.2, pp. 135–157. ISSN: 1572-8374. DOI: 10.1023/A:1007846431353. URL: https://doi.org/10.1023/A:1007846431353 (cit. on pp. 3, 27, 33).

Rosenfeld, R. (2000). "Two decades of statistical language modeling: where do we go from here?" In: *Proceedings of the IEEE* 88.8, pp. 1270–1278. DOI: 10.1109/5.880083 (cit. on p. 36).

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781 (cit. on p. 40).

Simpson, James (Oct. 2002). "Computer-Mediated Communication". In: *ELT Journal* 56. DOI: 10.1093/elt/56.4.414 (cit. on pp. 1, 19).

Xing, Eric, Michael Jordan, Stuart J Russell, and Andrew Ng (2002). "Distance Metric Learning with Application to Clustering with Side-Information". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/c3e4035af2a1cde9f21e1ae1951ac80b-Paper.pdf (cit. on p. 64).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (Mar. 2003). "A neural probabilistic language model". In: *J. Mach. Learn. Res.* 3.null, pp. 1137–1155. ISSN: 1532-4435 (cit. on pp. 36, 38).

Eemeren, Frans H. van and Rob Grootendorst (2003). *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge University Press (cit. on pp. 1, 17, 19, 20, 45).

Reed, Chris and Douglas Walton (2003). "Argumentation Schemes in Argument-as-Process and Argument-as-Product". In: URL: https://api.semanticscholar.org/CorpusID:14536431 (cit. on pp. 17, 20).

Simosi, Maria (2003). "Using Toulmin's Framework for the Analysis of Everyday Argumentation: Some Methodological Considerations". In: *Argumentation* 17.2, pp. 185–202. ISSN: 1572-8374. DOI: 10.1023/A:1024059024337. URL: https://doi.org/10.1023/A:1024059024337 (cit. on p. 20).

Toulmin, Stephen E. (2003). *The Uses of Argument*. 2nd ed. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511840005 (cit. on pp. 1, 3, 17, 18, 20, 27, 31, 33).

Gao, Jianfeng and Chin-Yew Lin (June 2004). "Introduction to the special issue on statistical language modeling". In: *ACM Transactions on Asian Language Information Processing* 3.2, pp. 87–93. ISSN: 1530-0226. DOI: 10.1145/1034780.1034781. URL: https://doi.org/10.1145/1034780.1034781 (cit. on p. 36).

Katzav, Joel and Chris A. Reed (2004). "On Argumentation Schemes and the Natural Classification of Arguments". In: *Argumentation* 18.2, pp. 239–259. ISSN: 1572-8374. DOI: 10.1023/

B:ARGU.0000024044.34360.82. URL: https://doi.org/10.1023/B:ARGU.0000024044.34360.82 (cit. on p. 20).

Reed, Chris and Glenn Rowe (2004). "Araucaria: software for argument analysis, diagramming and representation". English. In: *International Journal on Artificial Intelligence Tools* 13.4. dc.publisher: World Scientific Publishing, pp. 961–979. ISSN: 0218-2130. DOI: 10.1142/S0218213004001922 (cit. on p. 30).

Chopra, S., R. Hadsell, and Y. LeCun (2005). "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1, 539–546 vol. 1. DOI: 10.1109/CVPR.2005.202 (cit. on pp. 64, 66).

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18.5. IJCNN 2005, pp. 602–610. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2005.06.042. URL: https://www.sciencedirect.com/science/article/pii/S0893608005001206 (cit. on p. 38).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. 1st. Information Science and Statistics. New York, NY: Springer. ISBN: 978-0-387-31073-2. DOI: 10.1007/978-0387310732 (cit. on p. 23).

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2006). "The PASCAL Recognising Textual Entailment Challenge". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Ed. by Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 177–190. ISBN: 978-3-540-33428-6 (cit. on p. 114).

Hadsell, R., S. Chopra, and Y. LeCun (2006). "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100 (cit. on p. 64).

Collobert, Ronan and Jason Weston (2008). "A unified architecture for natural language processing: deep neural networks with multitask learning". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, pp. 160–167. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177 (cit. on p. 36).

Reed, Chris, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens (May 2008). "Language Resources for Studying Argument". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias. Marrakech, Morocco: European Language Resources Association (ELRA). URL: https://aclanthology.org/L08-1553/ (cit. on pp. 27, 30).

Walton, Douglas, Christopher Reed, and Fabrizio Macagno (2008). *Argumentation Schemes*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511802034 (cit. on pp. 3, 20, 34).

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24.2, pp. 8–12. DOI: 10.1109/MIS.2009.36 (cit. on p. 86).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7 (cit. on pp. 36, 84).

Palau, Raquel Mochales and Marie-Francine Moens (2009). "Argumentation mining: the detection, classification and structure of arguments in text". In: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. ICAIL '09. Barcelona, Spain: Association for Computing Machinery, pp. 98–107. ISBN: 9781605585970. DOI: 10.1145/1568234.1568246. URL: https://doi.org/10.1145/1568234.1568246 (cit. on pp. 15, 21, 22).

Pitler, Emily and Ani Nenkova (Aug. 2009). "Using Syntax to Disambiguate Explicit Discourse Connectives in Text". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Ed. by Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li. Suntec, Singapore: Association for Computational Linguistics, pp. 13–16. URL: https://aclanthology.org/P09-2004/ (cit. on pp. 21, 32, 38).

Rahwan, Iyad and Chris Reed (May 2009). "The Argument Interchange Format". In: pp. 383–402. ISBN: 978-0-387-98196-3. DOI: 10.1007/978-0-387-98197-0_19 (cit. on pp. 16, 20).

Boyd, Danah, Scott Golder, and Gilad Lotan (2010). "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter". In: *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*. HICSS '10. USA: IEEE Computer Society, pp. 1–10. ISBN: 9780769538693. DOI: 10.1109/HICSS.2010.412. URL: https://doi.org/10.1109/HICSS.2010.412 (cit. on pp. 19, 41).

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (July 2010). "Word Representations: A Simple and General Method for Semi-Supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394. URL: https://aclanthology.org/P10-1040/ (cit. on pp. 38, 47).

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau (June 2011). "Sentiment Analysis of Twitter Data". In: *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Ed. by Meenakshi Nagarajan and Michael Gamon. Portland, Oregon: Association for Computational Linguistics, pp. 30–38. URL: https://aclanthology.org/W11-0705/ (cit. on p. 2).

Biran, Or and Owen Rambow (2011). "Identifying Justifications in Written Dialogues by Classifying Text as Argumentative". In: *International Journal of Semantic Computing* 05.04, pp. 363–381. DOI: 10.1142/S1793351X11001328. eprint: https://doi.org/10.1142/S1793351X11001328. URL: https://doi.org/10.1142/S1793351X11001328 (cit. on pp. 15, 27, 28).

Freeman, James B. (2011). *Argument Structure: Representation and Theory*. Springer Dordrecht. DOI: 10.1007/978-94-007-0357-5 (cit. on pp. 20, 21, 27, 31, 33, 42).

Mercier, Hugo and Dan Sperber (2011). "Why do humans reason? Arguments for an argumentative theory". In: *Behavioral and Brain Sciences* 34.2, pp. 57–74. DOI: 10.1017/S0140525X10000968 (cit. on p. 1).

Torralba, A. and A. A. Efros (2011). "Unbiased look at dataset bias". In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '11. USA: IEEE Computer Society, pp. 1521–1528. ISBN: 9781457703942. DOI: 10.1109/CVPR.2011.5995347. URL: https://doi.org/10.1109/CVPR.2011.5995347 (cit. on p. 84).

Kriplean, Travis, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett (2012). "Supporting reflective public thought with considerit". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. Seattle, Washington, USA: Association for Computing Machinery, pp. 265–274. ISBN: 9781450310864. DOI: 10.1145/2145204.2145249. URL: https://doi.org/10.1145/2145204.2145249 (cit. on p. 15).

Marchetti-Bowick, Micol and Nathanael Chambers (Apr. 2012). "Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Walter Daelemans. Avignon, France: Association for Computational Linguistics, pp. 603–612. URL: https://aclanthology.org/E12-1062/ (cit. on p. 28).

Schölkopf, Bernhard, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij (2012). "On causal and anticausal learning". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, pp. 459–466. ISBN: 9781450312851 (cit. on p. 84).

Walker, Marilyn, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King (May 2012). "A Corpus for Research on Deliberation and Debate". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 812–817. URL: https://aclanthology.org/L12-1643/ (cit. on pp. 14, 16).

Bengio, Yoshua (2013). "Deep Learning of Representations: Looking Forward". In: *Statistical Language and Speech Processing*. Ed. by Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan

Mitkov, and Bianca Truthe. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–37. ISBN: 978-3-642-39593-2 (cit. on pp. 59–61).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (Aug. 2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8, pp. 1798–1828. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50. URL: https://doi.org/10.1109/TPAMI.2013.50 (cit. on pp. 38, 47, 59–61, 64, 66, 67).

Eveland, William P. and Kathryn E. Cooper (2013). "An integrated model of communication influence on beliefs". In: *Proceedings of the National Academy of Sciences* 110.supplement_3, pp. 14088–14095. DOI: 10.1073/pnas.1212742110. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1212742110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1212742110 (cit. on p. 13).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781 (cit. on p. 38).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013b). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (cit. on p. 38).

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090/ (cit. on p. 36).

Peldszus, Andreas and Manfred Stede (Jan. 2013). "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *Int. J. Cogn. Inform. Nat. Intell.* 7.1, pp. 1–31. ISSN: 1557-3958. DOI: 10.4018/jcini.2013010101. URL: https://doi.org/10.4018/jcini.2013010101 (cit. on p. 22).

Procter, Rob, Farida Vis, and Alex Voss (2013). "Reading the riots on Twitter: methodological innovation for the analysis of big data". In: *International Journal of Social Research Methodology* 16.3, pp. 197–214. DOI: 10.1080/13645579.2013.774172. eprint: https://doi.org/10.1080/13645579.2013.774172. URL: https://doi.org/10.1080/13645579.2013.774172 (cit. on p. 29).

Rogers, Richard (2013). "Debanalizing Twitter: the transformation of an object of study". In: WebSci '13. Paris, France: Association for Computing Machinery, pp. 356–365. ISBN:

9781450318891. DOI: 10.1145/2464464.2464511. URL: https://doi.org/10.1145/2464464.2464511 (cit. on pp. 2, 3).

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts (Oct. 2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1631–1642. URL: https://aclanthology.org/D13-1170/ (cit. on p. 114).

Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim (June 2014). "A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics". In: *Proceedings of the First Workshop on Argumentation Mining*. Ed. by Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker. Baltimore, Maryland: Association for Computational Linguistics, pp. 64–68. DOI: 10.3115/v1/W14-2109. URL: https://aclanthology.org/W14-2109/ (cit. on pp. 15, 21, 27, 30).

Boltužić, Filip and Jan Šnajder (June 2014). "Back up your Stance: Recognizing Arguments in Online Discussions". In: *Proceedings of the First Workshop on Argumentation Mining*. Ed. by Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker. Baltimore, Maryland: Association for Computational Linguistics, pp. 49–58. DOI: 10.3115/v1/W14-2107. URL: https://aclanthology.org/W14-2107/ (cit. on p. 16).

Eemeren, Frans H. van, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans (July 2014). *Handbook of Argumentation Theory*. 1st ed. 61 b/w illustrations, 18 colour illustrations. Dordrecht: Springer, pp. XII+988. ISBN: 978-90-481-9472-8. DOI: 10.1007/978-90-481-9473-5. URL: https://doi.org/10.1007/978-90-481-9473-5 (cit. on pp. 1, 17, 19, 20, 45, 107).

Habernal, I., Judith Eckle-Kohler, and Iryna Gurevych (Jan. 2014). "Argumentation mining on the web from information seeking perspective". In: *CEUR Workshop Proceedings* 1341 (cit. on pp. 21, 22).

Llewellyn, Clare, Claire Grover, Jon Oberlander, and Ewan Klein (May 2014). "Re-using an Argument Corpus to Aid in the Curation of Social Media Collections". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 462–468. URL: https://aclanthology.org/L14-1651/ (cit. on p. 29).

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu (2014). "When is it biased? assessing the representativeness of twitter's streaming API". In: *Proceedings of the 23rd International Con-*

ference on World Wide Web. WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, pp. 555–556. ISBN: 9781450327459. DOI: 10.1145/2567948.2576952. URL: https://doi.org/10.1145/2567948.2576952 (cit. on p. 29).

Stab, Christian and Iryna Gurevych (Aug. 2014). "Annotating Argument Components and Relations in Persuasive Essays". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Ed. by Junichi Tsujii and Jan Hajic. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1501–1510. URL: https://aclanthology.org/C14-1142/ (cit. on p. 15).

Tan, Chenhao, Lillian Lee, and Bo Pang (June 2014). "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, pp. 175–185. DOI: 10.3115/v1/P14-1017. URL: https://aclanthology.org/P14-1017/ (cit. on p. 2).

Habernal, Ivan and Iryna Gurevych (Sept. 2015). "Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 2127–2137. DOI: 10.18653/v1/D15-1255. URL: https://aclanthology.org/D15-1255/ (cit. on pp. 14, 20, 23, 27, 28).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://doi.org/10.1038/nature14539 (cit. on p. 40).

Lippi, Marco and Paolo Torroni (2015). "Argument Mining: A Machine Learning Perspective". In: *Theory and Applications of Formal Argumentation*. Ed. by Elizabeth Black, Sanjay Modgil, and Nir Oren. Cham: Springer International Publishing, pp. 163–176. ISBN: 978-3-319-28460-6. DOI: 10.1007/978-3-319-28460-6_10 (cit. on p. 23).

Peldszus, Andreas and Manfred Stede (Sept. 2015). "Joint prediction in MST-style discourse parsing for argumentation mining". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 938–948. DOI: 10.18653/v1/D15-1110. URL: https://aclanthology.org/D15-1110/ (cit. on pp. 27, 30).

Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim (Sept. 2015). "Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 440–450. DOI:

10.18653/v1/D15-1050. URL: https://aclanthology.org/D15-1050/ (cit. on pp. 27, 30, 42).

Swanson, Reid, Brian Ecker, and Marilyn Walker (Sept. 2015). "Argument Mining: Extracting Arguments from Online Dialogue". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Alexander Koller, Gabriel Skantze, Filip Jurcicek, Masahiro Araki, and Carolyn Penstein Rose. Prague, Czech Republic: Association for Computational Linguistics, pp. 217–226. DOI: 10.18653/v1/W15-4631. URL: https://aclanthology.org/W15-4631/ (cit. on pp. 14, 27, 28).

Addawood, Aseel and Masooda Bashir (Aug. 2016). ""What Is Your Evidence?" A Study of Controversial Topics on Social Media". In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Ed. by Chris Reed. Berlin, Germany: Association for Computational Linguistics, pp. 1–11. DOI: 10.18653/v1/W16-2801. URL: https://aclanthology.org/W16-2801/ (cit. on p. 29).

Bosc, Tom, Elena Cabrio, and Serena Villata (May 2016). "DART: a Dataset of Arguments and their Relations on Twitter". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 1258–1263. URL: https://aclanthology.org/L16-1200/ (cit. on pp. 16, 29, 45).

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785 (cit. on p. 40).

Fisas, Beatriz, Francesco Ronzano, and Horacio Saggion (May 2016). "A Multi-Layered Annotated Corpus of Scientific Papers". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3081–3088. URL: https://aclanthology.org/L16-1492/ (cit. on p. 29).

Al-Khatib, Khalid, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein (June 2016a). "Cross-Domain Mining of Argumentative Text through Distant Supervision". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, pp. 1395–1404. DOI: 10.18653/v1/N16-1165. URL: https://aclanthology.org/N16-1165/ (cit. on pp. 16, 27, 28).

Al-Khatib, Khalid, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein (Dec. 2016b). "A News Editorial Corpus for Mining Argumentation Strategies". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Ed. by Yuji Matsumoto and Rashmi Prasad. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3433–3443. URL: https://aclanthology.org/C16-1324/ (cit. on p. 15).

Liebeck, Matthias, Katharina Esau, and Stefan Conrad (Aug. 2016). "What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld". In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Ed. by Chris Reed. Berlin, Germany: Association for Computational Linguistics, pp. 144–153. DOI: 10.18653/v1/W16-2817. URL: https://aclanthology.org/W16-2817/ (cit. on p. 14).

Lippi, Marco and Paolo Torroni (Mar. 2016). "Argumentation Mining: State of the Art and Emerging Trends". In: *ACM Trans. Internet Technol.* 16.2. ISSN: 1533-5399. DOI: 10.1145/2850417. URL: https://doi.org/10.1145/2850417 (cit. on pp. 2, 3).

Misra, Amita, Brian Ecker, and Marilyn Walker (Sept. 2016). "Measuring the Similarity of Sentential Arguments in Dialogue". In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Raquel Fernandez, Wolfgang Minker, Giuseppe Carenini, Ryuichiro Higashinaka, Ron Artstein, and Alesia Gainer. Los Angeles: Association for Computational Linguistics, pp. 276–287. DOI: 10.18653/v1/W16-3636. URL: https://aclanthology.org/W16-3636/ (cit. on pp. 16, 27, 28).

Nishi, Ryosuke, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda (May 2016). "Reply trees in Twitter: data analysis and branching process models". In: *Social Network Analysis and Mining* 6. DOI: 10.1007/s13278-016-0334-0 (cit. on pp. 2, 41).

Ribeiro, Marco, Sameer Singh, and Carlos Guestrin (June 2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Ed. by John DeNero, Mark Finlayson, and Sravana Reddy. San Diego, California: Association for Computational Linguistics, pp. 97–101. DOI: 10.18653/v1/N16-3020. URL: https://aclanthology.org/N16-3020/ (cit. on p. 115).

Wagemans, Jean (Jan. 2016). "Constructing a Periodic Table of Arguments". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2769833 (cit. on pp. 17, 20).

Wojatzki, Michael and Torsten Zesch (2016). "Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance". In: *Conference on Natural Language Processing*. URL: https://api.semanticscholar.org/CorpusID:85555944 (cit. on pp. 27, 29).

Addawood, Aseel, Jodi Schneider, and Masooda Bashir (2017). "Stance Classification of Twitter Debates: The Encryption Debate as A Use Case". In: *Proceedings of the 8th International Conference on Social Media & Society*. SMSociety17. Toronto, ON, Canada: Association

for Computing Machinery. ISBN: 9781450348478. DOI: 10.1145/3097286.3097288. URL: https://doi.org/10.1145/3097286.3097288 (cit. on p. 29).

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes (Sept. 2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 670–680. DOI: 10.18653/v1/D17-1070. URL: https://aclanthology.org/D17-1070/ (cit. on p. 40).

Daxenberger, Johannes, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych (Sept. 2017). "What is the Essence of a Claim? Cross-Domain Claim Identification". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2055–2066. DOI: 10.18653/v1/D17-1218. URL: https://aclanthology.org/D17-1218/ (cit. on pp. 1, 3, 14, 25).

Dusmanu, Mihai, Elena Cabrio, and Serena Villata (Sept. 2017). "Argument Mining on Twitter: Arguments, Facts and Sources". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2317–2322. DOI: 10.18653/v1/D17-1245. URL: https://aclanthology.org/D17-1245/ (cit. on pp. 1, 2, 29).

Habernal, Ivan and Iryna Gurevych (Apr. 2017). "Argumentation Mining in User-Generated Web Discourse". In: *Computational Linguistics* 43.1, pp. 125–179. DOI: 10.1162/COLI_a_00276. URL: https://aclanthology.org/J17-1004/ (cit. on p. 3).

Hidey, Christopher, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown (Sept. 2017). "Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum". In: *Proceedings of the 4th Workshop on Argument Mining*. Ed. by Ivan Habernal, Iryna Gurevych, Kevin Ashley, Claire Cardie, Nancy Green, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker. Copenhagen, Denmark: Association for Computational Linguistics, pp. 11–21. DOI: 10.18653/v1/W17-5102. URL: https://aclanthology.org/W17-5102/ (cit. on pp. 14, 16, 27, 28, 42).

Lawrence, John, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed (July 2017). "Using Argumentative Structure to Interpret Debates in Online Deliberative Democracy and eRulemaking". In: *ACM Trans. Internet Technol.* 17.3. ISSN: 1533-5399. DOI: 10.1145/3032989. URL: https://doi.org/10.1145/3032989 (cit. on p. 1).

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964 (cit. on p. 115).

Niculae, Vlad, Joonsuk Park, and Claire Cardie (July 2017). "Argument Mining with Structured SVMs and RNNs". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 985–995. DOI: 10.18653/v1/P17-1091. URL: https://aclanthology.org/P17-1091/ (cit. on pp. 14, 27, 28).

Reed, Chris, Katarzyna Budzynska, Rory Duthie, Mathilde Janier, Barbara Konat, John Lawrence, Alison Pease, and Mark Snaith (2017). "The Argument Web: an Online Ecosystem of Tools, Systems and Services for Argumentation". In: *Philosophy & Technology* 30.2, pp. 137–160. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0260-8. URL: https://doi.org/10.1007/s13347-017-0260-8 (cit. on p. 16).

Stab, Christian and Iryna Gurevych (Sept. 2017). "Parsing Argumentation Structures in Persuasive Essays". In: *Computational Linguistics* 43.3, pp. 619–659. DOI: 10.1162/COLI_a_00295. URL: https://aclanthology.org/J17-3005/ (cit. on pp. 15, 27, 29, 42).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 38).

Wachsmuth, Henning, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein (Apr. 2017). "Computational Argumentation Quality Assessment in Natural Language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 176–187. URL: https://aclanthology.org/E17-1017/ (cit. on p. 43).

Wachsmuth, Henning, Benno Stein, and Yamen Ajjour (Apr. 2017). ""PageRank" for Argument Relevance". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 1117–1127. URL: https://aclanthology.org/E17-1105/ (cit. on pp. 15, 21).

Cabrio, Elena and Serena Villata (2018). "Five years of argument mining: a data-driven analysis". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, pp. 5427–5433. ISBN: 9780999241127 (cit. on pp. 1, 3, 14, 22, 25).

Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom (2018). "e-SNLI: natural language inference with natural language explanations". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 9560–9572 (cit. on p. 116).

Fafalios, Pavlos, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze (2018). "TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets". In: *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*. Heraklion, Greece: Springer-Verlag, pp. 177–190. ISBN: 978-3-319-93416-7. DOI: 10.1007/978-3-319-93417-4_12. URL: https://doi.org/10.1007/978-3-319-93417-4_12 (cit. on p. 28).

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (June 2018). "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://aclanthology.org/N18-2017/ (cit. on p. 86).

Krauthoff, Tobias, Christian Meter, Michael Baurmann, Gregor Betz, and Martin Mauve (Sept. 2018). "D-BAS - A Dialog-Based Online Argumentation System". In: DOI: 10.3233/978-1-61499-906-5-325 (cit. on p. 15).

Lauscher, Anne, Goran Glavaš, and Kai Eckert (Nov. 2018). "ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing". In: *Proceedings of the 5th Workshop on Argument Mining*. Ed. by Noam Slonim and Ranit Aharonov. Brussels, Belgium: Association for Computational Linguistics, pp. 22–28. DOI: 10.18653/v1/W18-5203. URL: https://aclanthology.org/W18-5203/ (cit. on p. 29).

Lauscher, Anne, Goran Glavaš, and Simone Paolo Ponzetto (Nov. 2018). "An Argument-Annotated Corpus of Scientific Publications". In: *Proceedings of the 5th Workshop on Argument Mining*. Ed. by Noam Slonim and Ranit Aharonov. Brussels, Belgium: Association for Computational Linguistics, pp. 40–46. DOI: 10.18653/v1/W18-5206. URL: https://aclanthology.org/W18-5206/ (cit. on pp. 15, 27, 29).

Levy, Ran, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim (Aug. 2018). "Towards an argumentative content search engine using weak supervision". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2066–2081. URL: https://aclanthology.org/C18-1176/ (cit. on pp. 15, 27, 30).

Mayer, Tobias, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata (2018). "Argument Mining on Clinical Trials". In: *Computational Models of Argument*. Frontiers in

Artificial Intelligence and Applications, pp. 137–148. DOI: 10.3233/978-1-61499-906-5-137. eprint: https://doi.org/10.3233/978-1-61499-906-5-137. URL: https://doi.org/10.3233/978-1-61499-906-5-137 (cit. on pp. 27, 29).

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (May 2018). "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1008/ (cit. on p. 40).

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202/ (cit. on p. 38).

Radford, Alec and Karthik Narasimhan (2018). "Improving Language Understanding by Generative Pre-Training". In: URL: https://api.semanticscholar.org/CorpusID:49313245 (cit. on pp. 39, 114, 115).

Rodríguez, Pau, Miguel Ángel Bautista, Jordi Gonzàlez, and Sergio Escalera (2018). "Beyond One-hot Encoding: lower dimensional target embedding". In: *CoRR* abs/1806.10805. arXiv: 1806.10805. URL: http://arxiv.org/abs/1806.10805 (cit. on p. 36).

Stab, Christian, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych (Oct. 2018). "Cross-topic Argument Mining from Heterogeneous Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 3664–3674. DOI: 10.18653/v1/D18-1402. URL: https://aclanthology.org/D18-1402/ (cit. on pp. 14, 17, 20, 21, 27, 30).

Teruel, Milagro, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata (May 2018). "Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1640/ (cit. on pp. 15, 30).

Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal (June 2018). "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074/ (cit. on p. 116).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446/ (cit. on p. 114).

Williams, Adina, Nikita Nangia, and Samuel Bowman (June 2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://aclanthology.org/N18-1101/ (cit. on p. 114).

Zhang, Zhilu and Mert Sabuncu (2018). "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf (cit. on p. 23).

Ajjour, Yamen, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein (2019). "Data Acquisition for Argument Search: The args.me Corpus". In: *KI 2019: Advances in Artificial Intelligence*. Ed. by Christoph Benzmüller and Heiner Stuckenschmidt. Cham: Springer International Publishing, pp. 48–59. ISBN: 978-3-030-30179-8 (cit. on pp. 15, 16).

Beltagy, Iz, Kyle Lo, and Arman Cohan (Nov. 2019). "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: https://aclanthology.org/D19-1371/ (cit. on p. 40).

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning (Aug. 2019). "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes. Florence, Italy: Association for Computational Linguistics, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828/ (cit. on p. 67).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (cit. on pp. 38, 47, 64, 107).

Ethayarajh, Kawin (Nov. 2019). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 55–65. DOI: 10.18653/v1/D19-1006. URL: https://aclanthology.org/D19-1006/ (cit. on p. 38).

Haddadan, Shohreh, Elena Cabrio, and Serena Villata (July 2019). "Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4684–4690. DOI: 10.18653/v1/P19-1463. URL: https://aclanthology.org/P19-1463/ (cit. on pp. 14, 16, 27, 30).

Hua, Xinyu, Mitko Nikolov, Nikhil Badugu, and Lu Wang (June 2019). "Argument Mining for Understanding Peer Reviews". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2131–2137. DOI: 10.18653/v1/N19-1219. URL: https://aclanthology.org/N19-1219/ (cit. on pp. 15, 27, 29).

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry (2019). "Adversarial Examples Are Not Bugs, They Are Features". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf (cit. on p. 86).

Lawrence, John and Chris Reed (Dec. 2019). "Argument Mining: A Survey". In: *Computational Linguistics* 45.4, pp. 765–818. DOI: 10.1162/coli_a_00364. URL: https://aclanthology.org/J19-4006/ (cit. on pp. 1, 3, 18, 20–22, 25, 31, 33).

Lipton, Zachary C. and Jacob Steinhardt (Feb. 2019). "Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research." In: *Queue* 17.1, pp. 45–77. ISSN: 1542-7730. DOI: 10.1145/3317287.3328534. URL: https://doi.org/10.1145/3317287.3328534 (cit. on p. 117).

Liu, Hui, Qingyu Yin, and William Yang Wang (July 2019). "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 5570–5581. DOI: 10.18653/v1/P19-1560. URL: https://aclanthology.org/P19-1560/ (cit. on p. 115).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692 (cit. on pp. 38–40).

Panchenko, Alexander, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann (Aug. 2019). "Categorizing Comparative Sentences". In: *Proceedings of the 6th Workshop on Argument Mining*. Ed. by Benno Stein and Henning Wachsmuth. Florence, Italy: Association for Computational Linguistics, pp. 136–145. DOI: 10.18653/v1/W19-4516. URL: https://aclanthology.org/W19-4516/ (cit. on pp. 15, 27, 30).

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://api.semanticscholar.org/CorpusID:160025533 (cit. on pp. 17, 39).

Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher (July 2019). "Explain Yourself! Leveraging Language Models for Commonsense Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4932–4942. DOI: 10.18653/v1/P19-1487. URL: https://aclanthology.org/P19-1487/ (cit. on p. 116).

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410/ (cit. on pp. 1, 47, 105).

Rendle, Steffen, Li Zhang, and Yehuda Koren (2019). "On the Difficulty of Evaluating Baselines: A Study on Recommender Systems". In: *CoRR* abs/1905.01395. arXiv: 1905.01395. URL: http://arxiv.org/abs/1905.01395 (cit. on p. 84).

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *CoRR* abs/1910.01108. arXiv: 1910.01108. URL: http://arxiv.org/abs/1910.01108 (cit. on pp. 38, 40).

Schaefer, Robin and Manfred Stede (2019). "Improving Implicit Stance Classification in Tweets Using Word and Sentence Embeddings". In: *KI 2019: Advances in Artificial Intelligence*. Ed. by Christoph Benzmüller and Heiner Stuckenschmidt. Cham: Springer International Publishing, pp. 299–307. ISBN: 978-3-030-30179-8 (cit. on p. 29).

Schneider, Alexander and Christian Meter (June 2019). "Various Efforts of Enhancing Real World Online Discussions". In: *ECA 2019: Proceedings of the 3rd European Conference on Argumentation*. Groningen, Netherlands (cit. on p. 15).

Vig, Jesse (July 2019). "A Multiscale Visualization of Attention in the Transformer Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Marta R. Costa-jussà and Enrique Alfonseca. Florence, Italy: Association for Computational Linguistics, pp. 37–42. DOI: 10.18653/v1/P19-3007. URL: https://aclanthology.org/P19-3007/ (cit. on p. 38).

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman (2019). "Neural Network Acceptability Judgments". In: *Transactions of the Association for Computational Linguistics* 7. Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova, pp. 625–641. DOI: 10.1162/tacl_a_00290. URL: https://aclanthology.org/Q19-1040/ (cit. on p. 114).

Andy, Anietie, Chris Callison-Burch, and Derry Tanti Wijaya (Dec. 2020). "Resolving Pronouns in Twitter Streams: Context can Help!" In: *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*. Ed. by Maciej Ogrodniczuk, Vincent Ng, Yulia Grishina, and Sameer Pradhan. Barcelona, Spain (online): Association for Computational Linguistics, pp. 133–138. URL: https://aclanthology.org/2020.crac-1.14/ (cit. on p. 45).

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463/ (cit. on p. 117).

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ran-

zato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (cit. on p. 39).

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org (cit. on p. 66).

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408/ (cit. on p. 116).

Feger, Marc, Jan Steimann, and Christian Meter (Sept. 2020). "Structure or Content? Towards Assessing Argument Relevance". In: *Computational Models of Argument. Proceedings of COMMA 2020*. Ed. by Henry Prakken, Stefano Bistarelli, Santini Francesco, and Carlo Taticchi. Vol. 326. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 203–214. ISBN: 978-1-64368-107-8. DOI: 10.3233/FAIA200505. URL: http://doi.org/10.3233/FAIA200505 (cit. on pp. 15, 21).

Funke, Christina M., Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas S. A. Wallis, and Matthias Bethge (2020). "The Notorious Difficulty of Comparing Human and Machine Perception". In: *CoRR* abs/2004.09406. arXiv: 2004.09406. URL: https://arxiv.org/abs/2004.09406 (cit. on p. 84).

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann (Nov. 2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11, pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: https://doi.org/10.1038/s42256-020-00257-z (cit. on pp. 3, 40, 67, 68, 83–86, 107, 115, 168).

Henighan, Tom, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish (2020). "Scaling Laws for Autoregressive Generative Modeling". In: *CoRR* abs/2010.14701. arXiv: 2010.14701. URL: https://arxiv.org/abs/2010.14701 (cit. on p. 39).

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling Laws for Neural Language Models". In: *CoRR* abs/2001.08361. arXiv: 2001.08361. URL: https://arxiv.org/abs/2001.08361 (cit. on p. 39).

Le-Khac, Phuc H., Graham Healy, and Alan F. Smeaton (2020). "Contrastive Representation Learning: A Framework and Review". In: *IEEE Access* 8, pp. 193907–193934. DOI: 10.1109/ACCESS.2020.3031549 (cit. on pp. 59–62, 65, 66, 68).

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703/ (cit. on p. 39).

Mayer, Tobias, Elena Cabrio, and Serena Villata (2020a). "Transformer-Based Argument Mining for Healthcare Applications". In: *European Conference on Artificial Intelligence*. URL: https://ecai2020.eu/papers/1470_paper (cit. on pp. 27, 29, 42).

– (Aug. 2020b). "Transformer-based Argument Mining for Healthcare Applications". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence*. Santiago de Compostela / Online, Spain. URL: https://hal.science/hal-02879293 (cit. on p. 15).

Merchant, Amil, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney (Nov. 2020). "What Happens To BERT Embeddings During Fine-tuning?" In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. Online: Association for Computational Linguistics, pp. 33–44. DOI: 10.18653/v1/2020.blackboxnlp-1.4. URL: https://aclanthology.org/2020.blackboxnlp-1.4/ (cit. on p. 60).

Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl (2020). "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges". In: *ECML PKDD 2020 Workshops*. Ed. by Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, Ricard Gavaldà, João Gama, Linara Adilova, Yamuna Krishnamurthy, Pedro M. Ferreira, Donato Malerba, Ibéria Medeiros, Michelangelo Ceci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras, Peter Christen, Eirini Ntoutsi, Erich Schubert, Arthur Zimek, Anna Monreale, Przemyslaw Biecek, Salvatore Rinzivillo, Benjamin Kille, Andreas Lommatzsch, and Jon Atle Gulla. Cham: Springer International Publishing, pp. 417–431. ISBN: 978-3-030-65965-3 (cit. on p. 115).

Morante, Roser, Chantal van Son, Isa Maks, and Piek Vossen (May 2020). "Annotating Perspectives on Vaccination". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille,

France: European Language Resources Association, pp. 4964–4973. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.611/ (cit. on pp. 27, 28).

Olshefski, Christopher, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley (May 2020). "The Discussion Tracker Corpus of Collaborative Argumentation". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 1033–1043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.130/ (cit. on p. 16).

Poudyal, Prakash, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma (Dec. 2020). "ECHR: Legal Corpus for Argument Mining". In: *Proceedings of the 7th Workshop on Argument Mining*. Ed. by Elena Cabrio and Serena Villata. Online: Association for Computational Linguistics, pp. 67–75. URL: https://aclanthology.org/2020.argmining-1.8/ (cit. on pp. 15, 27, 30, 42).

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54/ (cit. on pp. 3, 38, 47).

Shnarch, Eyal, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim (Nov. 2020). "Unsupervised Expressive Rules Provide Explainability and Assist Human Experts Grasping New Domains". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2678–2697. DOI: 10.18653/v1/2020.findings-emnlp.243. URL: https://aclanthology.org/2020.findings-emnlp.243/ (cit. on pp. 16, 27, 30).

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang (2020). *How to Fine-Tune BERT for Text Classification?* arXiv: 1905.05583 [cs.CL]. URL: https://arxiv.org/abs/1905.05583 (cit. on pp. 38, 47, 60).

Wang, Tongzhou and Phillip Isola (2020). "Understanding contrastive representation learning through alignment and uniformity on the hypersphere". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org (cit. on pp. 66, 68).

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou (2020). "MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546 (cit. on p. 105).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (cit. on p. 17).

Bhatti, Muhammad Mahad Afzal, Ahsan Suheer Ahmad, and Joonsuk Park (Nov. 2021). "Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate". In: *Proceedings of the 8th Workshop on Argument Mining*. Ed. by Khalid Al-Khatib, Yufang Hou, and Manfred Stede. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1–11. DOI: 10.18653/v1/2021.argmining-1.1. URL: https://aclanthology.org/2021.argmining-1.1/ (cit. on pp. 28, 29, 115).

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba (2021). "Evaluating Large Language Models Trained on Code". In: *CoRR* abs/2107.03374. arXiv: 2107.03374. URL: https://arxiv.org/abs/2107.03374 (cit. on p. 39).

Feger, Marc (May 2021). "Online Argumentation and Social Media: What They Can Learn from Each Other". Master's thesis. Düsseldorf, Germany: Heinrich–Heine University Düsseldorf. DOI: 10.13140/RG.2.2.26976.81926. URL: https://www.researchgate.net/publication/352441941_Online_argumentation_and_social_media_What_they_can_learn_from_each_other (cit. on p. 15).

Fergadis, Aris, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou (Nov. 2021). "Argumentation Mining in Scientific Literature for Sustainable Development". In: *Proceedings of the 8th Workshop on Argument Mining*. Ed. by Khalid Al-Khatib, Yufang Hou, and Manfred Stede. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 100–111. DOI: 10.18653/v1/2021.argmining-1.10. URL: https://aclanthology.org/2021.argmining-1.10/ (cit. on pp. 15, 21, 27, 29, 42).

Fromm, Michael, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl (May 2021). "Argument Mining Driven Analysis of Peer-Reviews". In: *Proceedings of the AAAI Conference on Ar-*

tificial Intelligence 35.6, pp. 4758–4766. DOI: 10.1609/aaai.v35i6.16607. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16607 (cit. on pp. 15, 21, 27, 29, 42).

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685. arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685 (cit. on p. 39).

Kiesel, Johannes, Nico Reichenbach, Benno Stein, and Martin Potthast (Nov. 2021). "Image Retrieval for Arguments Using Stance-Aware Query Expansion". In: *Proceedings of the 8th Workshop on Argument Mining*. Ed. by Khalid Al-Khatib, Yufang Hou, and Manfred Stede. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 36–45. DOI: 10.18653/v1/2021.argmining-1.4. URL: https://aclanthology.org/2021.argmining-1.4/ (cit. on p. 16).

Mestre, Rafael, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman (Nov. 2021). "M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts". In: *Proceedings of the 8th Workshop on Argument Mining*. Ed. by Khalid Al-Khatib, Yufang Hou, and Manfred Stede. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 78–88. DOI: 10.18653/v1/2021.argmining-1.8. URL: https://aclanthology.org/2021.argmining-1.8/ (cit. on p. 16).

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen Carley (Aug. 2021). "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose". In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1, pp. 400–408. DOI: 10.1609/icwsm.v7i1.14401. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14401 (cit. on p. 29).

Nye, Maxwell I., Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena (2021). "Show Your Work: Scratchpads for Intermediate Computation with Language Models". In: *CoRR* abs/2112.00114. arXiv: 2112.00114. URL: https://arxiv.org/abs/2112.00114 (cit. on p. 39).

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir

Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving (2021). "Scaling Language Models: Methods, Analysis & Insights from Training Gopher". In: *CoRR* abs/2112.11446. arXiv: 2112.11446. URL: https://arxiv.org/abs/2112.11446 (cit. on p. 39).

Schaefer, Robin and Manfred Stede (2021). "Argument Mining on Twitter: A survey". In: *it - Information Technology* 63.1, pp. 45–58. DOI: doi:10.1515/itit-2020-0053. URL: https://doi.org/10.1515/itit-2020-0053 (cit. on pp. 1–3, 14, 16, 25, 29).

Slonim, Noam, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov (Mar. 2021). "An autonomous debating system". In: *Nature* 591.7850, pp. 379–384. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03215-w. URL: https://doi.org/10.1038/s41586-021-03215-w (cit. on pp. 1, 35).

Thorn Jakobsen, Terne Sasha, Maria Barrett, and Anders Søgaard (Aug. 2021). "Spurious Correlations in Cross-Topic Argument Mining". In: *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Ed. by Lun-Wei Ku, Vivi Nastase, and Ivan Vulić. Online: Association for Computational Linguistics, pp. 263–277. DOI: 10.18653/v1/2021.starsem-1.25. URL: https://aclanthology.org/2021.starsem-1.25/ (cit. on pp. 3, 67, 68, 86, 107, 115).

Vecchi, Eva Maria, Neele Falk, Iman Jundi, and Gabriella Lapesa (Aug. 2021). "Towards Argument Mining for Social Good: A Survey". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1338–1352. DOI: 10.18653/v1/2021.acl-long.107. URL: https://aclanthology.org/2021.acl-long.107/ (cit. on pp. 1, 3, 22, 25).

Wang, Wenhui, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei (Aug. 2021). "MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Compu-

tational Linguistics, pp. 2140–2151. DOI: 10.18653/v1/2021.findings-acl.188. URL: https://aclanthology.org/2021.findings-acl.188/ (cit. on p. 105).

Zhao, Tony Z., Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh (2021). *Calibrate Before Use: Improving Few-Shot Performance of Language Models*. arXiv: 2102.09690 [cs.CL]. URL: https://arxiv.org/abs/2102.09690 (cit. on p. 115).

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He (2021). "A Comprehensive Survey on Transfer Learning". In: *Proceedings of the IEEE* 109.1, pp. 43–76. DOI: 10.1109/JPROC.2020.3004555 (cit. on p. 68).

Alhamzeh, Alaa, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie (Dec. 2022). "It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset". In: *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Ed. by Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 163–169. DOI: 10.18653/v1/2022.finnlp-1.22. URL: https://aclanthology.org/2022.finnlp-1.22/ (cit. on pp. 15, 16, 27, 30).

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Rémi King (2022). "Deep language algorithms predict semantic comprehension from brain activity". In: *Scientific Reports* 12.1, p. 16327. ISSN: 2045-2322. DOI: 10.1038/s41598-022-20460-9. URL: https://doi.org/10.1038/s41598-022-20460-9 (cit. on p. 84).

Chen, Zaiqian, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park (Dec. 2022). "Argument Mining for Review Helpfulness Prediction". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 8914–8922. DOI: 10.18653/v1/2022.emnlp-main.609. URL: https://aclanthology.org/2022.emnlp-main.609/ (cit. on p. 15).

Cheng, Liying, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si (May 2022). "IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 2277–2287. DOI: 10.18653/v1/2022.acl-long.162. URL: https://aclanthology.org/2022.acl-long.162/ (cit. on pp. 14, 27, 30).

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin,

Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel (2022). *PaLM: Scaling Language Modeling with Pathways*. arXiv: 2204.02311 [cs.CL]. URL: https://arxiv.org/abs/2204.02311 (cit. on p. 39).

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei (2022). *Scaling Instruction-Finetuned Language Models*. arXiv: 2210.11416 [cs.LG]. URL: https://arxiv.org/abs/2210.11416 (cit. on p. 40).

Falk, Neele and Gabriella Lapesa (June 2022). "Scaling up Discourse Quality Annotation for Political Science". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 3301–3318. URL: https://aclanthology.org/2022.lrec-1.353/ (cit. on p. 14).

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson (2022). "Shared computational principles for language processing in humans and deep language models". In: *Nature Neuroscience* 25.3, pp. 369–380. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01026-4. URL: https://doi.org/10.1038/s41593-022-01026-4 (cit. on p. 84).

Grundler, Giulia, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni (Oct. 2022). "Detecting Arguments in CJEU Decisions on Fiscal State Aid". In: *Proceedings of the 9th Workshop on Argument Mining*. Ed. by Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics, pp. 143–157. URL: https://aclanthology.org/2022.argmining-1.14/ (cit. on pp. 15, 30, 42).

Hafid, Salim, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze (2022). "SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, pp. 3988–3992. ISBN: 9781450392365. DOI: 10.1145/3511808.3557693. URL: https://doi.org/10.1145/3511808.3557693 (cit. on p. 28).

Hansen, Marcus and Daniel Hershcovich (Dec. 2022). "A Dataset of Sustainable Diet Arguments on Twitter". In: *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*. Ed. by Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 40–58. DOI: 10.18653/v1/2022.nlp4pi-1.5. URL: https://aclanthology.org/2022.nlp4pi-1.5/ (cit. on pp. 27–29, 42).

Hautli-Janisz, Annette, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed (June 2022). "QT30: A Corpus of Argument and Conflict in Broadcast Debate". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 3291–3300. URL: https://aclanthology.org/2022.lrec-1.352/ (cit. on pp. 15, 16).

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre (2022). "An empirical analysis of compute-optimal large language model training". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 30016–30030. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf (cit. on p. 39).

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large language models are zero-shot reasoners". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc. ISBN: 9781713871088 (cit. on p. 39).

Liu, Zhexiong, Meiqi Guo, Yue Dai, and Diane Litman (Oct. 2022). "ImageArg: A Multi-modal Tweet Dataset for Image Persuasiveness Mining". In: *Proceedings of the 9th Workshop on Argument Mining*. Ed. by Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics, pp. 1–18. URL: https://aclanthology.org/2022.argmining-1.1/ (cit. on p. 28).

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi (May 2022). "Cross-Task Generalization via Natural Language Crowdsourcing Instructions". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3470–3487. DOI: 10.18653/v1/2022.acl-long.244. URL: https://aclanthology.org/2022.acl-long.244/ (cit. on p. 114).

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf (cit. on p. 114).

Steimann, Jan, Marc Feger, and Martin Mauve (2022). "Inspiring Heterogeneous Perspectives in News Media Comment Sections". In: *Human Interface and the Management of Information: Visual and Information Design*. Ed. by Sakae Yamamoto and Hirohiko Mori. Cham: Springer International Publishing, pp. 118–131. ISBN: 978-3-031-06424-1. DOI: https://doi.org/10.1007/978-3-031-06424-1_10 (cit. on pp. 8, 10, 16, 109).

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2022a). *Finetuned Language Models Are Zero-Shot Learners*. arXiv: 2109.01652 [cs.CL]. URL: https://arxiv.org/abs/2109.01652 (cit. on pp. 39, 40).

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022b). *Emergent Abilities of Large Language Models*. arXiv: 2206.07682 [cs.CL]. URL: https://arxiv.org/abs/2206.07682 (cit. on p. 39).

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022c). "Chain-of-thought prompting elicits reasoning in large language models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc. ISBN: 9781713871088 (cit. on pp. 39, 115).

Zhou, Yichu and Vivek Srikumar (May 2022). "A Closer Look at How Fine-tuning Changes BERT". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 1046–1061.

DOI: 10.18653/v1/2022.acl-long.75. URL: https://aclanthology.org/2022.acl-long.75/ (cit. on p. 60).

Ajjour, Yamen, Johannes Kiesel, Benno Stein, and Martin Potthast (May 2023). "Topic Ontologies for Arguments". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1411–1427. DOI: 10.18653/v1/2023.findings-eacl.104. URL: https://aclanthology.org/2023.findings-eacl.104/ (cit. on pp. 1, 3, 14, 25).

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314 (cit. on p. 39).

Grison, Thibault, Virginie Julliard, Félix Alié, and Victor Ecrement (2023). "Excessive moderation on Twitter. Case study on the invisibilization of LGBT and TDS content online". en. In: *Réseaux* No 237.1, pp. 119–149. DOI: 10.3917/res.237.0119. URL: https://doi.org/10.3917/res.237.0119 (cit. on p. 8).

Habernal, Ivan, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard (June 2023). "Mining legal arguments in court decisions". In: *Artif. Intell. Law* 32.3, pp. 1–38. ISSN: 0924-8463. DOI: 10.1007/s10506-023-09361-y. URL: https://doi.org/10.1007/s10506-023-09361-y (cit. on pp. 15, 23, 30).

Independent Technology Research, Coalition for (2023). *Imposing Fees to Access the Twitter API Threatens Public Interest Research*. Accessed: 2025-04-28. URL: https://independenttechresearch.org/letter-twitter-api-access-threatens-public-interest-research/ (cit. on pp. 2, 3, 29).

Kung, Po-Nien and Nanyun Peng (July 2023). "Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1317–1328. DOI: 10.18653/v1/2023.acl-short.113. URL: https://aclanthology.org/2023.acl-short.113/ (cit. on p. 115).

Liu, Zhexiong, Mohamed Elaraby, Yang Zhong, and Diane Litman (Dec. 2023). "Overview of ImageArg-2023: The First Shared Task in Multimodal Argument Mining". In: *Proceedings of the 10th Workshop on Argument Mining*. Ed. by Milad Alshomary, Chung-Chi Chen, Smaranda Muresan, Joonsuk Park, and Julia Romberg. Singapore: Association for Computational Linguistics, pp. 120–132. DOI: 10.18653/v1/2023.argmining-1.12. URL: https://aclanthology.org/2023.argmining-1.12/ (cit. on p. 16).

Lopes Cardoso, Henrique, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins (2023). "Argumentation models and their use in corpus annotation: Practice, prospects, and challenges".

In: *Natural Language Engineering* 29.4, pp. 1150–1187. DOI: 10.1017/S1351324923000062 (cit. on pp. 2, 16–18, 20, 21, 33, 42).

Tay, Yi, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler (2023). *UL2: Unifying Language Learning Paradigms*. arXiv: 2205.05131 [cs.CL]. URL: https://arxiv.org/abs/2205.05131 (cit. on p. 40).

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971 (cit. on pp. 39, 40).

Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi (2023). *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*. arXiv: 2205.10625 [cs.AI]. URL: https://arxiv.org/abs/2205.10625 (cit. on p. 115).

Chen, Guizhen, Liying Cheng, Anh Tuan Luu, and Lidong Bing (Aug. 2024). "Exploring the Potential of Large Language Models in Computational Argumentation". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 2309–2330. DOI: 10.18653/v1/2024.acl-long.126. URL: https://aclanthology.org/2024.acl-long.126/ (cit. on pp. 3, 17, 37, 40, 114, 167).

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783 (cit. on p. 17).

Gu, Yuling, Oyvind Tafjord, and Peter Clark (Aug. 2024). "Digital Socrates: Evaluating LLMs through Explanation Critiques". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 5559–5586. DOI: 10.18653/v1/2024.acl-long.302. URL: https://aclanthology.org/2024.acl-long.302/ (cit. on p. 115).

Mezza, Stefano, Wayne Wobcke, and Alan Blair (Aug. 2024). "Exploiting Dialogue Acts and Context to Identify Argumentative Relations in Online Debates". In: *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*. Ed. by Yamen Ajjour, Roy Bar-Haim, Roxanne El Baff, Zhexiong Liu, and Gabriella Skitalinskaya. Bangkok, Thailand: Association for Computational Linguistics, pp. 36–45. DOI: 10.18653/v1/2024.argmining-1.4. URL: https://aclanthology.org/2024.argmining-1.4/ (cit. on p. 15).

Pangakis, Nicholas and Sam Wolken (June 2024). "Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels". In: *Pro-*

ceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024). Ed. by Dallas Card, Anjalie Field, Dirk Hovy, and Katherine Keith. Mexico City, Mexico: Association for Computational Linguistics, pp. 113–131. DOI: 10.18653/v1/2024.nlpcss-1.9. URL: https://aclanthology.org/2024.nlpcss-1.9/ (cit. on p. 115).

Saphra, Naomi, Eve Fleisig, Kyunghyun Cho, and Adam Lopez (June 2024). "First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2310–2326. DOI: 10.18653/v1/2024.naacl-long.128. URL: https://aclanthology.org/2024.naacl-long.128/ (cit. on pp. 3, 17, 36, 39, 86, 115).

Svete, Anej and Ryan Cotterell (June 2024). "Transformers Can Represent n-gram Language Models". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 6845–6881. DOI: 10.18653/v1/2024.naacl-long.381. URL: https://aclanthology.org/2024.naacl-long.381/ (cit. on p. 36).

Torralba, A., P. Isola, and W.T. Freeman (2024). *Foundations of Computer Vision*. Adaptive Computation and Machine Learning series. MIT Press. ISBN: 9780262378666. URL: https://mitpress.mit.edu/9780262048972/foundations-of-computer-vision/ (cit. on pp. 60, 61, 65, 68).

Wang, Xin, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu (2024). "Disentangled Representation Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12, pp. 9677–9696. DOI: 10.1109/TPAMI.2024.3420937 (cit. on p. 68).

Weber, Simon B., Marc Feger, and Michael Pilgermann (2024). "Don't Stop Believin': A Unified Evaluation Approach for LLM Honeypots". In: *IEEE Access* 12, pp. 144579–144587. DOI: 10.1109/ACCESS.2024.3472460 (cit. on pp. 9, 10, 109).

Zelle, Yannick, Thibault Grison, and Marc Feger (2024). "SciTok - A Web Scraping Tool for Social Science Research". In: *HCI International 2023 – Late Breaking Posters*. Ed. by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy. Cham: Springer Nature Switzerland, pp. 103–109. ISBN: 978-3-031-49212-9. DOI: https://doi.org/10.1007/978-3-031-49212-9_14 (cit. on pp. 8, 16, 109).

Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang (2024). *Instruction Tuning for Large Language Models: A Survey*. arXiv: 2308.10792 [cs.CL]. URL: https://arxiv.org/abs/2308.10792 (cit. on pp. 39, 114).

Zhou, Lexin, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo (Oct. 2024a). "Larger and more instructable language models become less reliable". In: *Nature* 634.8032, pp. 61–68. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07930-y. URL: https://doi.org/10.1038/s41586-024-07930-y (cit. on p. 17).

– (Oct. 2024b). "Larger and more instructable language models become less reliable". In: *Nature* 634.8032, pp. 61–68. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07930-y. URL: https://doi.org/10.1038/s41586-024-07930-y (cit. on p. 86).

Braun, Marie and Marc Feger (2025). "TextLabel: A Web Application for Streamlining Text Annotation and Classification". In: *HCI International 2025 Posters*. Ed. by Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy. Cham: Springer Nature Switzerland, pp. 243–250. ISBN: 978-3-031-94171-9. DOI: https://doi.org/10.1007/978-3-031-94171-9_21 (cit. on pp. 10, 109).

Cabessa, Jérémie, Hugo Hernault, and Umer Mushtaq (Jan. 2025). "Argument Mining with Fine-Tuned Large Language Models". In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 6624–6635. URL: https://aclanthology.org/2025.coling-main.442/ (cit. on pp. 3, 17, 22, 37, 40, 114, 167).

DataReportal (Apr. 2025). *Global Social Media Statistics*. https://datareportal.com/social-media-users. Accessed: 20 June 2025 (cit. on p. 13).

Feger, Marc, Katarina Boland, and Stefan Dietze (July 2025). "Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 23900–23915. URL: https://aclanthology.org/2025.acl-long.1164/ (cit. on pp. 1, 7, 9, 10, 14, 17, 25, 36, 42, 71, 83, 86, 105–107, 109, 112, 114, 115, 168).

Iñaki Goñi, Julian (Mar. 2025). "Citizen participation and technology: lessons from the fields of deliberative democracy and science and technology studies". In: *Humanities and Social Sciences Communications* 12.1, p. 287. ISSN: 2662-9992. DOI: 10.1057/s41599-025-04606-4. URL: https://doi.org/10.1057/s41599-025-04606-4 (cit. on p. 1).

Lawsen, A. (2025). *Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. arXiv: 2506.09250 [cs.AI]. URL: https://arxiv.org/abs/2506.09250 (cit. on pp. 17, 115).

Li, Hao, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic (2025). *Large Language Models in Argument Mining: A Survey*. arXiv: 2506.16383 [cs.CL]. URL: https://arxiv.org/abs/2506.16383 (cit. on p. 3).

OpenAI (2025). *ChatGPT*. https://openai.com/index/chatgpt/. Accessed: 21 June 2025 (cit. on p. 17).

Romberg, Julia, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa (Apr. 2025). "Towards a Perspectivist Turn in Argument Quality Assessment". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 7458–7485. DOI: 10.18653/v1/2025.naacl-long.382. URL: https://aclanthology.org/2025.naacl-long.382/ (cit. on pp. 1, 14, 43).

Shojaee, Parshin, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar (2025). *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. URL: https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf (cit. on pp. 17, 115).

Stahl, Maja, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth (July 2025). "ArgInstruct: Specialized Instruction Fine-Tuning for Computational Argumentation". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 11103–11127. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.579. URL: https://aclanthology.org/2025.findings-acl.579/ (cit. on pp. 1, 14, 17, 114).

Steimann, Jan Lukas (June 2025). "What Do Others Think About This Comment? – Recommending Diverse and Relevant User Comments in Comment Sections". Doctoral thesis, Faculty of Mathematics and Natural Sciences, Department of Computer Science. PhD thesis. Düsseldorf, Germany: Heinrich-Heine-Universität Düsseldorf. URL: https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=70313 (visited on 08/05/2025) (cit. on p. 8).

Weber, Simon Benedikt (Apr. 2025). "Critical Care, Critical Defense: Dissecting Hospital Security Challenges to Advance Attack Detection". Doctoral thesis, Faculty of Mathematics and Natural Sciences – Computer Networks (Rechnernetze). PhD thesis. Düsseldorf, Germany: Heinrich-Heine-Universität Düsseldorf. URL: https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=69352 (visited on 08/25/2025) (cit. on p. 9).

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen (2025). *A Survey of Large Language Models*. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223 (cit. on pp. 3, 36, 39, 59, 86).

List of Figures

List of Tables

Eidesstattliche Erklärung
laut §5 der Promotionsordnung vom 15.06.2018

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Ort, Datum

Marc Feger