From the Ivory Tower into the Wild - Analysis of (Mis)information in Online Discourses in the Age of Deep Learning

Inaugural-Dissertation

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

vorgelegt von

Katarina Boland
aus Kassel

Düsseldorf, Juni 2023

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:

27.06.2023

Abstract

The Web has evolved into an ubiquitous platform where citizens have the opportunity to express opinions, interact with each other and share information of any kind. In doing so, they leave digital traces, leading to an abundance of *Online Discourse Data (ODD)*. ODD refers to data published online to share opinions, factual claims and other information with other individuals or organizations. It includes messages posted on social media platforms such as Twitter, blog posts, messages in discussion forums, news articles and editorials, scientific information published online, e.g. in digital publications on pre-print servers or in press releases, and any reactions to those such as comments, likes and shares.

ODD has gained relevance as a valuable source of research data for social scientific studies, offering a longitudinal perspective on dynamically evolving debates. Due to the overwhelming amount of information, processing it typically requires the aid of machines.

At the same time, ODD has been aiding the advancement of computational methods as it can be used to train powerful, data-intensive Deep Learning models. However, the growing popularity of these methods has been accompanied by concerncs regarding transparency, interpretability and reproducibility of the generated findings and the methods themselves. While they have advanced the state-of-the-art performance in many tasks, scholars are still researching why these methods perform well, and recently, if they really do - or whether their impressive scores are caused by artifacts in the data or evaluation protocols. This uncertainty extends to an imprecise use of terminology within and beyond the research community and inflated expectations of what these methods can achieve, especially since the utility for real-world applications beyond research prototypes has been receiving little attention.

This thesis contributes to a clearer understanding of core concepts relating to the analysis of ODD by presenting an extensive multidisciplinary survey and proposing a unifying terminology and model. It further provide insights and tools for the application of state-of-the-art computational methods for Online Discourse Data Analysis in real-world systems to aid fact-checking and combat misinformation. With this, it adds to recent endeavors to assess machine learning methods not merely on the

quality of their predictions, but consider their utility for real-world applications. The presented research shows that efficient unupervised systems can be the better choice for verified claim retrieval applications than approaches dominating the leaderboards of respective shared tasks by fine-tuning large language models. Finally, the thesis contributes to the empirical literature on evidence-based policy-making showing that the analysis of ODD can aid in revealing possible interactions between political decisions and public concerns. It further advances the methodological literature on extracting, filtering and analyzing ODD for social scientific studies.

Acknowledgments

KA Hello ChatGPT!

Hello! How can I assist you today?

KA I'm writing my PhD thesis and I need a text for the acknowledgment section. I have the best and most supportive colleagues and supervisors I could wish for, I loved the time with them, I loved doing this research and I feel like it really made me grow in many ways. And I'm also looking forward a lot to continuing working with them!

Can you help me write a short text to thank them, please?

Certainly! I'd be happy to help you with the acknowledgment section of your PhD thesis. Here's a short text you can use as a starting point:

(...long and cheesy text...)

Feel free to modify and personalize this text to suit your specific circumstances and the individuals you wish to acknowledge. Good luck with your thesis!

KA can you make a short humorous poem out of this?

Certainly! Here's a short and humorous poem based on the content of the original text:

In the realm of research, I embarked on a quest,
With colleagues and supervisors, the very best.
Their support and guidance, beyond compare,
Made my PhD journey fun and rare.

Oh, the countless debates and brainstorming streams,
Fueling my research with ingenious schemes.
We laughed, we pondered, our minds ablaze,
Pushing boundaries in so many ways.

To my supervisors, wise and patient souls,
Who shaped my skills and filled knowledge holes.
Their belief in my talents, I must confess,
Turned me into a confident research tigress.

But let's not forget the department's crew,
A camaraderie that made me feel anew.
With shared experiences and support so fine,
Together we sipped the academic wine.

Now, with gratitude overflowing from my core,
I look ahead to adventures still in store.
Excited to continue this scholarly race,
With this remarkable team, laughter in place.

So here's to the laughter and knowledge we share,
To breakthroughs and discoveries, beyond compare.
With a humorous spirit, we'll conquer it all,
Leaving a mark, as we stand tall.

Feel free to tweak and adjust this poem to suit your personal style and preferences. Enjoy adding a touch of humor to your acknowledgment section!

KA Thank you, I like this! How would you like me to credit you?

You're welcome! I'm glad you liked the poem. As an AI language model, I don't require personal credit, but if you'd like to acknowledge the assistance, you can simply mention "AI assistance provided by ChatGPT" or "Generated with the help of an AI language model." That way, it reflects the AI's contribution to your thesis. Best of luck with your PhD thesis!

KA Definitely a good idea to mention the AI's contribution to my thesis!

Thanks, ChatGPT, you're the best (for this task) :)

Contents

Chapter 1

Introduction

The Web has evolved into an ubiquitous platform where people have the opportunity to express opinions, make claims about facts or beliefs, interact with others and share information of any kind. According to Ortiz-Ospina (2019), at least 3.5 of 7.7 billion people worldwide used the Internet in 2019. Social media platforms were used by one-in-three people and by more than two-thirds of all internet users. With their usage of these and other platforms, users leave digital traces, leading to the creation of large volumes of *Online Discourse Data (ODD)*. ODD refers to data published online to share opinions, factual claims and other kinds of information with other individuals or organizations. It includes messages posted on social media platforms such as Twitter, blog posts, messages in discussion forums, news articles and editorials, scientific information published online, e.g. in digital publications on pre-print servers or in press releases, and any reactions to those such as comments, likes and shares.

This has sparked many new research questions concerning the impact of the digitalization on society and vice versa, such as the interactions between social media usage and the design of particular platforms and algorithms on one side, and political processes, democracy and society as a whole on the other side. Also, ODD has gained relevance as a valuable source of research data for social scientific studies as it gives a glimpse into what's in the public attention at a given point in time, offering a longitudinal perspective on dynamically evolving topics and debates. Possible interactions between events happening in the real world and topics discussed in the public debate can be studied in real-time.

At the same time, technological advances have enabled the use of deep neural architectures for machine learning - powerful, yet data-hungry methods. Together with the abundance of ODD, this has lead to revolutions and paradigm shifts in many scientific disciplines such as Natural Language Processing (NLP) and Computer Vision Chernyavskiy et al. (2021b).

The rapidly increasing interest in Machine Learning and Deep Learning techniques that could since be observed in many research areas (e.g. recommender systems Dacrema et al. (2019) and question answering Crane (2018)), however, did not come without a cost: as Pfeiffer and Hoffmann (2009) detail, increasing popularity of a research field generally bears a high risk of going hand in hand with a decreased reliability of findings published in the scientific literature. A growing body of works has recently shown that Deep Learning and Machine Learning are no exception: results have been shown to be volatile, depending on a large number of factors such as parameter settings, computer hardware and software versions Crane (2018), which increases general concerns regarding lacking reproducibility Dacrema et al. (2019): interpretation of research findings, comparisons of different works and assessment of progress gets more and more difficult with the growing interest in Deep Learning-based methods, especially when they use different, sometimes sub-optimal baselines, and the results are scattered across multiple papers. These problems cannot be tackled by peer review mechanisms alone Rogers and Augenstein (2020) but require changes in research practices Lucic et al. (2022). Moreover, Deep Learning methods have commonly been treated as a black box by users and developers alike Xu et al. (2019), and research is still in the process of attempting to understand which information neural networks such as language models encode Rogers et al. (2020); Niven and Kao (2019) and what their capabilities and limitations are Bang et al. (2023); Qin et al. (2023); Chernyavskiy et al. (2021b); Linzbach et al. (2023). This uncertainty translates to imprecise use of terminology and confusion within and beyond the research community paired with inflated expectations of what these methods can achieve Bender and Koller (2020). The latter problem is reinforced by the use of benchmark data that contains unwanted and hidden statistical and social biases, artificially reducing task difficulty by offering spurious cues to Machine Learning algorithms, making their performance seem more impressive than it is (Kiela et al., 2021; Le Bras et al., 2020). Carried by this hype, more and more methods have been developed that pushed the state-of-the-art further and further regarding prediction quality - yet only recently, other important factors have come into awareness, such as the practical utility of developed solutions, their efficiency and costs regarding energy consumption, their robustness and fairness (Ethayarajh and Jurafsky, 2020; Bender et al., 2021).

Beyond these problems with accidental misinformation, deliberate disinformation also has become a more and more pressing concern with the advancement of digitalization and the rise of social media. Now that everyone can publish information that has the potential to reach a large audience, citizens have the means to actively influence the public discourse Bennett and Pfetsch (2018). With false information spreading faster than the truth (Vosoughi et al., 2018a), the development of efficient fact-checking mechanisms has become an active research area.

Overall, the confusion regarding terminology, methods and data used in ODD analysis and Deep Learning more generally hinders the advancement of methods within and across research communities and their utility for practical applications.

This thesis aims to contribute to a clearer understanding of core concepts relating to the analysis of ODD and provide insights, models and tools for the application of state-of-the-art computational methods for Online Discourse Data Analysis in real-world systems.

1.1 Motivation

The analysis of ODD has been researched in a variety of scientific disciplines. While this bears great potential for collaboration and cross-fertilization of different research endeavors, the lack of shared terminology and models impedes sharing of data and methods. This thesis thus starts with an extensive survey of ODD-related concepts and terminology, research tasks and methods across various related scientific discplines and proposes a model for representing central notions such as claims and their relations (Chapter 2). With this, it contributes to a clearer understanding of the research area and lays the basis to facilitate sharing of data in structured knowledge graphs.

Knowledge graphs of fact-checked claims extracted from online portals can serve a variety of use-cases in Computer Science and beyond. One important use is their ability to serve as a database of verified claims that can be used by automated methods aiding human fact-checkers efficiently verify statements by retrieving similar previously fact-checked claims. While many of such claim retrieval systems have been proposed in the past, their comparison is difficult, as the results are scattered across different publications, evaluations were performed on different datasets and there has been little attention to the utility of the approaches in real-world claim retrieval applications, which do not only require methods to yield satisfactory performance regarding prediction quality, but also impose restrictions on runtime and computational expenses. Chapter 3 of this thesis analyzes the state-of-the-art in claim retrieval regarding practical utility and moreover discusses the need for large language models for this task. Next to the theoretical insights, it contributes tools that can be used for claim retrieval in online claim retrieval applications using an existing knowledge graph, also introduced in this chapter.

The volatility of results depending on factors such as the used software and the consequential importance of provenance information for scientific findings is in the focus of Chapter 4. While the importance of transparency of research has been recognized, explicit links between scientific assets, such as publications, software and datasets are still often lacking. Deep Learning can help extract informal references from publication texts, however, obtaining training data for this task is costly. This chapter proposes a case study to create a weakly-labeled corpus, coined silverstandard, using a combination of weakly supervised classifiers and distant supervision that can be used to train more powerful approaches with transfer learning. This chapter contributes to the practical application of state-of-the-art information extraction approaches to increase the reproducibility of research in realistic settings where large amounts of training data are unavailable.

Finally, Chapter 5 addresses the analysis of ODD as social scientific research data. ODD has gained relevance as research data in social sciences as it gives a glimpse into what's in the public attention at a given point in time, offering a longitudinal perspective on dynamically evolving topics and debates. People discuss what is salient and relevant to them instead of answering pre-defined questions about what researchers deem relevant about a specific issue. Possible interactions between events happening in the real-world and topics discussed in the public debate can be studied in real-time. At the same time, processing ODD is challenging because the amounts of data are huge and typically require the assistance of machines. For those, the unstructured and heterogeneous nature of the data and the often informal language used therein poses challenges. This chapter combines computational methods based on deep learning with manual analyses to analyze the German Twitter discourse during the pandemic to reveal possible interactions between policies and public concerns. With this, it contributes to the empirical literature on evidence-based policy-making and advances the methodological literature on extracting, filtering and analyzing Twitter data to make ODD useable as social scientific research data.

1.2 Publications and Contributions

The publications that form the basis of this thesis and my contributions to them are listed below.

1 (Boland et al., 2022a), basis of Chapter 2
Boland, Katarina, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. "Beyond facts: a survey and conceptualisation of claims in online discourse analysis." Semantic Web 13 (1): 1-35. doi: `https://doi.org/10.3233/SW-212838`. `http://www.semantic-web-journal.net/system/files/swj2838.pdf`. Status: published

Summary:
Analyzing statements of facts and claims in online discourse is subject of a multitude of research areas and central to many tasks such as machine-aided fact-checking. While all these fields are concerned with strongly related notions, such as claims, facts and evidence, terminology and conceptualisations used across and within communities vary heavily, making it hard to assess commonalities and relations of related works, re-use data and methods and work collaboratively in interdisciplinary or multidisciplinary settings. This work surveys the state-of-the-art from a range of fields and across a range of research tasks and applications. On this basis, a conceptual model – *Open Claims* – for claims and related notions is proposed that takes into consideration their

inherent complexity and distinguishes meaning, linguistic representation and context. Its applications for various tasks related to Online Discourse Data analysis and knowledge graphs are showcased.

My contributions:
I conducted an extensive literature survey on different definitions, conceptualizations and models of all notions relevant to the analysis of claims in online discourses and created a multidisciplinary overview comparing them. I reviewed the majority of publications for this purpose and further contributed to surveying and summarizing literature on the related knowledge engineering tasks, selected relevant tasks and mapped them to the proposed model jointly with the co-authors. I had the lead in designing the conceptual model and the methodology of the survey. I wrote the sections (including all subsections) "Methodology", "Facts and Claims - A multidisciplinary Survey of Definitions" and improved them according to the co-authors' feedback. I wrote several of the "Related Knowledge Engineering Tasks" subsections and contributed to all other parts of the paper.

2 (Boland et al., 2019a), basis of Section 2.4.3
Boland, Katarina, Pavlos Fafalios, Andon Tchechmedjiev, Konstantin Todorov, and Stefan Dietze. 2019. "Modeling and Contextualizing Claims." In Proceedings of the Blockchain enabled Semantic Web Workshop (BlockSW) and Contextualized Knowledge Graphs (CKG) Workshop co-located with the 18th International Semantic Web Conference, BlockSW/CKG@ISWC 2019, CEUR Workshop Proceedings 2599. Status: published

Summary:
This paper introduces a conceptual model for representing claims made in Online Discourse Data and their contexts to allow more fine-grained analysis of the data. This model served as the basis for the *Open Claims* model. This work further introduces an implementation of this model that uses established vocabularies such as schema.org, Open Annotation and PROV-O.

My contributions:
My contributions were the literature review and lead for the design of the conceptual model. The paper was written jointly by all co-authors.

3 (Boland et al., 2023a), basis of Chapter 3
Boland, Katarina, Alica Hövelmeyer, Pavlos Fafalios, Konstantin Todorov, Usama Mazhar and Stefan Dietze. "Efficient and Robust Claim Retrieval for Online Fact-Checking Applications" (to appear). Status: submitted to Information Retrieval Journal.

Summary:
Verified claim retrieval is an important task for fact-checking information in ODD using hybrid Artificial Intelligence. It has been approached through numerous, mostly supervised neural approaches, but their utility for actual applications has thus far received little attention. This paper experimentally evaluates state-of-the-art approaches regarding prediction quality, robustness across datasets as well as efficiency with respect to computational effort and runtime for offline computations and during inference. We show that computationally efficient unsupervised approaches can perform on par with costly supervised approaches relying on large fine-tuned language models and can hence be a better choice for practical online claim verification applications.

My contributions:
I had the lead in developing the idea and structure of the paper and designing the experimental setup. I implemented all modifications to the compared approaches and generated all data for those (*RIET Lab, SimBa 2023, UofSheffield, Check_Square variants, NLP&IR@UNED, ...*). Predictions for the *Elasticsearch Baseline* and the unsupervised systems *ClaimLinker* and *TIET*, that required no modifications, were generated by two co-authors. I designed and performed all data analyses and co-supervised the extension of the *ClaimLinker* system and the development of the *SimBa* approach. I wrote all parts of the paper, all co-authors contributed improvements to the texts.

4 (Boland and Krüger, 2019), basis of Chapter 4
Boland, Katarina, and Frank Krüger. 2019. "Distant supervision for silver label generation of software mentions in social scientific publications." In BIRNDL 2019: Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, edited by Muthu Kumar Chandrasekaran, and Philipp Mayr, CEUR workshop proceedings 2414, 15-27. Aachen: RWTH. Status: published

Summary:

While scientific findings lay the foundation of our knowledge, they, too, can be biased and flawed, for example when errors in software packages lead to wrong data analysis outcomes. To increase transparency of research and verifiability of scientific findings, it is thus important to provide as much provenance information as possible. While deep learning-based approaches have been successful in extracting software mentions in scientific publications, training data is sparse and its creation costly. This publication focuses on detecting software mentions in scientific publications in conditions when training data is sparse. We show that by combining even only a small number of weakly supervised approaches with distant supervision, a silver standard corpus can be created that serves as a useful basis for transfer learning.

My contributions:
Both authors jointly designed the experimental setup and performed the data analysis. My contribution was the development and application of the InfoLink tool, the implementation of the different CRF variants and the evaluation of the combined approaches. I mainly wrote the sections "Method", "Evaluation" and "Results", incorporating contributions from the second author, and contributed minor textual improvements to all other sections.

5 (Boland et al., 2023b), basis of Chapter 5
Boland, Katarina, Christopher Starke, Felix Bensmann, Frank Marcinkowski, and Stefan Dietze (to appear). "Public Discourse about COVID-19 Vaccinations: A Computational Analysis of the Relationship between Public Concerns and Policies". Status: submitted to EPJ Data Science.

Summary:
Using traditional survey data alone, dynamic interactions between policy changes and public opinions are hard to investigate. This study draws on ODD, more precisely Twitter data, to analyze which topics are prevalent in the public discourse about COVID-19 vaccinations in German-speaking countries in different points in time and how their salience relates to different phases of the pandemic and to policy events. AI-based methods, partly relying on language models, are used to make ODD accessible as longitudinal data for these investigations. With this, this work adds to the empirical literature on data-driven policy-making and advances the methodological literature on extracting, filtering and analyzing Twitter data for policy research.

My contributions:
The idea for this publication was conceived jointly with the co-authors. I designed the study and performed all data processing steps except the relevance filtering of tweets using keyword search and applying the Sentiment Analysis tool on the tweets, which was executed by one co-author. I also performed the data analysis and interpretation of results with feedback from the co-authors. I contributed to the collection of policy event data and performed the manual data labeling tasks together with the second author. I wrote all parts of the paper and improved them according to the co-authors' feedback except the sections "Introduction" and "Related Work" which were written jointly with the second author.

6 (Schulze and Boland, 2019), basis of Section 2.5.1
Schulze, Heidi, and Katarina Boland. 2019. "Building a categorization schema for automated source typing." 69th Annual International Communication Association (ICA) Conference, 24.05.2019. Status: Conference presentation

Summary:
This work introduces a hierarchical categorization schema for source types in Online Discourse Data for social scientific studies. It differentiates, among others, publishers and authors of claims, scientific and non-scientific sources and different user types such as politicians vs. scientists based on requirements extracted based on a literature review of related communication science and computer science publications.

My contribution:
Both authors reviewed literature from communication sciences and fields related to computer sciences jointly. I had the lead in developing the category scheme in joint discussions with the second author, developed annotation guidelines and supervised the annotations that have been used to refine the model in several iterations.

Further own recent works cited in this thesis (without re-using texts directly):

- (Tchechmedjiev et al., 2019a), cited in Chapter 3
  Tchechmedjiev, Andon, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. "Claim-

sKG: A Knowledge Graph of Fact-Checked Claims." In The Semantic Web – ISWC 2019. ISWC 2019, edited by Chiara Ghidini, Olaf Hartig, and Maria Maleshkova, Lecture Notes in Computer Science 11779, 309-324. Cham: Springer. doi: https://doi.org/10.1007/978-3-030-30796-7_20. Status: published

Summary:
This paper introduces *ClaimsKG* - a structured database which serves as a registry of claims. Basis of the database is a knowledge graph which provides data about claims, metadata (such as their publishing site), automatically annotated involved entities, links to fact-checking articles and normalized truth ratings. *ClaimsKG* is generated through a (semi-)automated pipeline which harvests claims and respective metadata from popular fact-checking sites on a regular basis, lifts data into an RDF/S model, which exploits established schema such as schema.org and NIF, and annotates claims with related entities from DBpedia. With this, *ClaimsKG* represents a resource that can be used for a variety of tasks: as a database of verified claims for claim retrieval systems and benchmark datasets, to train fact-checking models and to facilitate access to fact-checked information for journalists and researchers.

My contribution:
My contribution to this paper was the Use Cases and Exploitation section (researching the respective information and writing the section). I also contributed to this version of *ClaimsKG* by updating extractors for the pipeline.

- (Gangopadhyay et al., 2023), cited in Chapter 3
  Gangopadhyay, Susmita, Katarina Boland, Danilo Dessí, Stefan Dietze, Pavlos Fafalios, Andon Tchechmedjiev, Konstantin Todorov, Hajira Jabeen. 2023. "Truth or Dare: Investigating Claims Truthfulness with ClaimsKG." In Proceedings of the Second International Workshop on Linked Data-driven Resilience Research 2023 co-located with Extended Semantic Web Conference 2023 (ESWC 2023). Hersonissos, Greece, May 28, 2023. Status: published
  (Gangopadhyay et al., 2022), cited in Chapter 3
  Gangopadhyay, Susmita, Katarina Boland, Sascha Schüller, Torodov Konstantin, Andon Tchechmedjiev, Benjamin Zapilko, Pavlos Fafalios, Hajira Jabeen, and Stefan Dietze. 2022. ClaimsKG - A Knowledge Graph of Fact-Checked Claims (August, 2022). GESIS - Leibniz-Institute for the Social Sciences. Data File

Version 1.0.0. doi: https://doi.org/10.7802/2469. Status: published

Summary:
This publication and dataset present the new version of *ClaimsKG* which includes the most recent data contained in the harvested fact-checking portals, adds additional fact-checking portals to the pool of harvested resources, thereby extending the knowledge base with new languages, uses a more recent Named Entity Recognition and Disambiguation tool and improves the pipeline's architecture.

My contribution:
I contributed to the described new version of *ClaimsKG* by performing quality checks for the harvested data and assisting the implementation of the updated architecture and the inclusion of the new Named Entity Recognition tool.

- (Hövelmeyer et al., 2022), cited in Chapter 3
  Hövelmeyer, Alica , Katarina Boland, and Stefan Dietze. 2022. "SimBa at CheckThat! 2022: Lexical and Semantic Similarity-Based Detection of Verified Claims in an Unsupervised and Supervised Way." In CLEF Working Notes 2022, Proceedings of the Working Notes of CLEF 2022- Conference and Labs of the Evaluation Forum, edited by Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, CEUR Workshop Proceedings 3180, 511-531. Aachen: RWTH Aachen. `https://ceur-ws.org/Vol-3180/paper-40.pdf`. Status: published

  Summary:
  This work introduces a state of the art Claim Retrieval approach making use of several partly complementary language models. It features an analysis of the complementarity of lexical and semantic similarity features and embeddings created by different language models.

  My contribution:
  I contributed to the overall idea for the approach and the paper, the design of the experimental setup and the joint supervision of the data analysis. The paper was mainly written by the first author. I contributed to improving the structure and texts.

- (Maliaroudakis et al., 2021a), cited in Chapter 3
  Maliaroudakis, Evangelos, Katarina Boland, Stefan Dietze, Konstantin Todorov, Yannis Tzitzikas, and Pavlos Fafalios. 2021. "ClaimLinker: Linking text to a knowledge graph of fact-checked claims." In WWW '21: Companion Proceedings of the Web Conference 2021, edited by Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, 669–672. New York: ACM. doi: `https://doi.org/10.1145/3442442.3458601`. Status: published

  Summary:
  This paper introduces a web service and API, called *ClaimLinker*, that performs claim retrieval and links arbitrary input texts to verified claims in *ClaimsKG*.

  My contribution:
  I contributed to the design of the *ClaimLinker* method and by adding minor improvements to the paper text.

- (Beretta et al., 2020), cited in Chapter 2
  Beretta, Valentina, Sébastien Harispe, Katarina Boland, Luke Lo Seen, Konstantin Todorov, and Andon Tchechmedjiev. 2020. "Can Knowledge Graph Embeddings Tell Us What Fact-checked Claims Are About?." In Proceedings of the First Workshop on Insights from Negative Results in NLP, co-located with EMNLP 2020, edited by Anna Rogers, João Sedoc, and Anna Rumshisky, 71–75. doi: `https://doi.org/10.18653/v1/2020.insights-1.11`. Status: published

  Summary:
  This paper studies the complementarity of graph and text embeddings to predict the topics of claims in *ClaimsKG*. While both types of embeddings prove to capture complementary information to some extent, the low performance of graph embedding features alone indicate that the model fails to capture topological features pertinent to the topic prediction task relying on a local link prediction objective.

  My contribution:
  I contributed to the design of the experimental setup, the goldstandard creation and the discussion of the data analysis and contributed minor improvements to the text of the paper.

- (Otto et al., 2020), cited in Chapter 4
  Otto, Wolfgang, Andrea Zielinski, Behnam Ghavimi, Dimitar Dimitrov, Narges Tavakolpoursaleh, Karam Abdulahhad, Katarina Boland, and Stefan Dietze. 2020. "Knowledge extraction from scholarly publications: The GESIS contribution to the rich context competition." In Rich search and discovery for research datasets, edited by Julia Lane, Ian Mulvany, and Paco Nathan, 107-127. Los Angeles u.a.: Sage. `https://study.sagepub.com/richcontext`. Status: published

  Summary:
  This publication presents methods to extract informal references to research data in scientific publications and to extract keywords describing the research.

  My contribution:
  I contributed to the discussion of the methodology to detect dataset mentions and contributed minor improvements to the text of the paper.

- (Hienert et al., 2019), cited in Chapter 4
  Hienert, Daniel, Dagmar Kern, Katarina Boland, Benjamin Zapilko, and Peter Mutschke. 2019. "A digital library for research data and related information in the social sciences." In Proceedings of 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 148-157. Piscataway, NJ: IEEE. doi: `https://doi.org/10.1109/JCDL.2019.00030`. Status: published

  Summary:
  This publication introduces an integrated search system for social science information which provides links between research data, publications, survey variables and questions, instruments and tools in order to increase transparency of research and foster re-use of scientific information. A log-based usage study shows that users make active use of the provided link information and search across different information types.

  My contribution:
  I was responsible for the Link Infrastructure consisting of Link Import, Link Enrichment, maintenance and integration of LOD backend and the Link Index. I implemented and conceptualized all components and developed the concept of the architecture together with one co-author. I wrote the text of the respective section and contributed minor improvements to the text of the paper.

- (Gasquet et al., 2019), cited in Chapter 3
  Gasquet, Malo, Darlene Brechtel, Matthäus Zloch, Andon Tchechmedjiev, Katarina Boland, Pavlos Fafalios, Stefan Dietze, and Konstantin Todorov. 2019. "Exploring Fact-checked Claims and their Descriptive Statistics." In ISWC 2019 Satellites: Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019), edited by Mari Carmen Suárez-Figueroa, Gong Cheng, Anna Lisa Gentile, Christophe Guéret, Maria Keet, and Abraham Bernstein, CEUR workshop proceedings, 289-292. Aachen: RWTH. Status: published

  Summary:
  This demo paper introduces web-based tools to make the *ClaimsKG* knowledge graph accessible to non-tech savy users, e.g. journalists and social scientists.

  My contribution:
  I contributed to the use-case/demonstration scenarios included in the paper and contributed minor improvements to the text.

- (Breuer et al., 2022), cited in Chapter 5
  Breuer, Johannes, Felix Bensmann, Katarina Boland, Ran Yu, and Stefan Dietze. 2022. "All public opinions are not equal – Developing and testing a method for assessing the relationship between survey data and Twitter data as measures of public opinion." 12th International Conference on Social Media & Society, 18.07.2022. `https://www.youtube.com/watch?v=D4UOa3AlTCI`. Status: Conference presentation
  (Breuer et al., 2021), cited in Chapter 5
  Breuer, Johannes, Felix Bensmann, Stefan Dietze, Ran Yu, and Katarina Boland. 2021. "Assessing the relationship between survey data and Twitter data as measures of public opinion - A methodological pilot study." ESRA 2021: 9th Conference of the European Survey Research Association, Online, 16.07.2021. `https://www.europeansurveyresearch.org/conf2021/uploads/219/2790/62/Relationship_between_survey_data_and_Twitter_data_as_measures_of_public_opinion_ESRA2021.pdf`. Status: published

  Summary:
  These two works present a pilot study introducing a pipeline for filtering and preprocessing Twitter data to use it as a complement to social scientific survey data.

My contribution:
I contributed to the experimental design and the discussion of the data analysis.

1.3 Structure of this Thesis

The chapters and sections use texts/images from the following publications:

- Chapter 2: publication 1 (Boland et al., 2022a), except
  - Section 2.4.3: publications 2 (Boland et al., 2019a) and 1 (Boland et al., 2022a)
  - Section 2.5.1: publication 1 (Boland et al., 2022a) enriched with information contained in 6 (Schulze and Boland, 2019) regarding source types
  - Section 2.5.3: publication 1 (Boland et al., 2022a) enriched with unpublished information concerning claim relatedness
- Chapter 3: publication 3 (Boland et al., 2023a)
- Chapter 4: publication 4 (Boland and Krüger, 2019)
- Chapter 5: publication 5 (Boland et al., 2023b)

Chapter 2

Facts and Claims - A Multidisciplinary Survey of Definitions

Analyzing statements of facts and claims in online discourse is subject of a multitude of research areas. Methods from natural language processing and computational linguistics help investigate issues such as the spread of biased narratives and falsehoods on the Web. Related tasks include fact-checking, stance detection and argumentation mining. Knowledge-based approaches, in particular works in knowledge base construction and augmentation, are concerned with mining, verifying and representing factual knowledge. While all these fields are concerned with strongly related notions, such as claims, facts and evidence, terminology and conceptualisations used across and within communities vary heavily, making it hard to assess commonalities and relations of related works and how research in one field may contribute to address problems in another. We survey the state-of-the-art from a range of fields in this interdisciplinary area across a range of research tasks. We assess varying definitions and propose a conceptual model—Open Claims—for claims and related notions that takes into consideration their inherent complexity, distinguishing between their meaning, linguistic representation and context. We also introduce an implementation of this model by using established vocabularies and discuss applications across various tasks related to online discourse analysis.

2.1 Introduction

The Web has evolved into an ubiquitous platform where many people have the opportunity to be publishers, to express opinions and to interact with others. It has been widely used as a source to mine and understand online discourse or to extract knowledge. On the one hand, understanding and analyzing societal discourse on the Web are becoming increasingly important issues involving computational methods in natural language processing (NLP) or computational linguistics. Related tasks include fact or claim verification, discourse modeling, stance detection or argumentation mining. In this context, a wide range of interdisciplinary research directions have emerged involving a variety of scientific disciplines including investigations into the spreading patterns of false claims on Twitter (Vosoughi et al., 2018a), pipelines for discovering and finding the stance of claim-relevant Web documents (Wang et al., 2018; Bhatt et al., 2018; Hanselowski et al., 2018a), approaches for classifying sources of news, such as Web pages, pay-level domains, users or posts (Popat et al., 2017), or research into fake news detection (Tschiatschek et al., 2018) and automatic fact-checking (Hassan et al., 2015). In addition, understanding discourse in scholarly and scientific works has been a long-standing research problem throughout the past years (Lauscher et al., 2018b,a; Gonzalez Pinto et al., 2019; Accuosto and Saggion, 2019; Green, 2014; Graves et al., 2014; González Pinto and Balke, 2018; Kirschner et al., 2015; Hyland, 1998; Neves et al., 2019; Green, 2015, 2018; Hernández A. and Gómez, 2014).

On the other hand, knowledge-based approaches, in particular works in knowledge base (KB) construction and augmentation, often are concerned with mining, verifying and representing factual knowledge from the Web. Research in such areas often deploys methods and conceptualisations strongly related to some of the aforementioned computational methods related to claims, e.g. when aiming to verify facts from the Web for augmenting KBs (Yu et al., 2018; Dong et al., 2014). Whereas the focus in knowledge base augmentation is on extracting and formally representing trust-worthy factual statements as an atomic assertion in the first-order-logic sense, research focused on interpreting claims expressed in natural language tends to put stronger emphasis on understanding the context of a claim, e.g. its source, timing, location or its role as an argument as part of (online) discourse. Capturing the meaning of claims requires both to preserve the actual claim utterances as natural language texts as well as structured knowledge about the claims. Utterances often carry a range of assertions and sentiments embedded in complex sentence structures, which are easy to process by humans but are hard to interpret by machines. Preserving structured knowledge about claims, including their contexts and constituents, enables machine-interpretation, discoverability and reuse of claims, for instance, to facilitate research in the aforementioned areas.

Despite these differences, methods in various disparate fields, such as claim/fact verification or fact-checking as well as KB augmentation, tend to be based on similar intuitions and heuristics and are concerned with similar and related notions from different perspectives. Hence, achieving a shared understanding and terminology has become a crucial challenge.

However, both the used terminology and the underlying conceptual models are still strongly diverging, within and across the academic literature and the involved applications (Daxenberger et al., 2017; Torsi and Morante, 2018). For example, "Animals should have lawful rights" is considered a claim in Chen et al. (Chen et al., 2019b) and according to many definitions from the argumentation mining community which define claims as the conclusive parts of an argument. It does not constitute a claim according to the guidelines of the FEVER fact-checking challenge (Thorne et al., 2019) where claims are defined as factoid statements. This claim would also not be eligible for inclusion in a fact-checking portal as it does not contain factual content that can be checked and does not seem check-worthy (although this would depend on the context, such as who uttered the statement and when). The claim might be contained in the ground truth of a topic-independent claim extraction approach, but might only be used to evaluate a topic-dependent approach when it is connected to a given topic (more details in Sect. 2.3).

This heterogeneity poses challenges for the understanding of related works and data by both humans as well as machines and hinders the cross-fertilisation of research across various distinct, yet related fields. Thus, our work aims at facilitating a shared understanding of claims and related terminology across diverse communities as well as the representation of semi-structured knowledge about claims and their context, which is a crucial requirement for advancing, replicating, and validating research in the aforementioned fields.

In order to address the aforementioned problems, this paper makes the following main contributions:

- **An extensive survey** (Sect. 2.3) of related works concerned with defining, understanding and representing online discourse and related notions, most importantly, claims and facts. The survey is the first of its kind, providing a comprehensive overview of definitions, terminology used across various fields and communities.

- **A conceptual model** (Sect. 2.4), which we call *Open Claims Model*, and corresponding terminology of claims and their constituents and context, that is grounded in both scientific literature in related fields such as argumentation mining or discourse analysis as well as the actual practices of representing and sharing claims on the Web, for instance, as part of fact-checking sites. To this end, we also provide an OWL (Web Ontology Language) implementation of the model as well as an RDF/S (Resource Description Framework Schema) implementation that uses state-of-the-art vocabularies, such as *schema.org* and *PROV-O* (Provenance Ontology), in order to facilitate Web-scale sharing, discovery and reuse of claims and their context, for instance through semi-structured Web page markup or as part of dedicated knowledge graphs (KGs) such as *ClaimsKG* (Tchechmedjiev et al., 2019b).

- **An introductory review of related information extraction and knowledge engineering tasks** (Sect. 2.5), involved with the extraction, verification and (inter)linking of claim related data. Our aim is to provide an overview of related state-of-the-art works that may be used for populating a KB of claims and their context according to the proposed conceptual model. This also enables us to discover under-researched areas and challenging directions for future work.

Note that while an earlier version of the conceptual model has been presented in Boland et al. (2019b), the novel contributions of this work include the actual survey of related works in the context of online discourse, a critical review of related tasks, as well as improvements to the model and its implementation facilitated by the substantial survey provided here.

This work is meant to facilitate a shared representation of claims across various communities, as is required for inter-disciplinary research. This includes works aimed at detecting and representing the inherent relations of uttered claims among each other or with represented factual knowledge and other resources, such as web pages or social media posts, e.g., as part of stance detection tasks. Assessing and modeling the similarity of claims, for example, is a challenging task. When two claims are similar to each other, what precisely does this mean? Do they have the same topic but have been uttered to express a different stance? Are they expressing a shared viewpoint but have been uttered by different agents? Do they talk about similar topics but with diverging specificity, i.e. the topic of one claim is a single aspect of the more broad topic of the other one? Or is one claim a part of a more complex claim that includes multiple assertions? Even claims deemed equal with regard to their content may have to be differentiated: they may, for example, be repeated utterances with the same content by the same agent (but at different times), paraphrases (same content but different utterances, also at different times, maybe by different agents) or just duplicates in the respective database. A fine-grained model that allows relating claims and individual claim components allows specifying different dimensions of relatedness and similarity. This also enables more formal and clear definitions for tasks related to detection of claim similarity and relatedness. Use cases involve research into the detection of viewpoints and communities sharing related narratives and viewpoints on the Web (Starbird, 2017), the analysis of quotation patterns involving varied sources and media types or profiling of sources and references used in news media (Niculae et al., 2015), and fact-checking applications, e.g. linking claims to previously fact-checked claims (Shaar et al., 2020a; Maliaroudakis et al., 2021b).

2.2 Methodology of the Survey

In this section, we describe the publication selection and review process employed in this survey. An overview of the workflow is given in Figure 2.1.

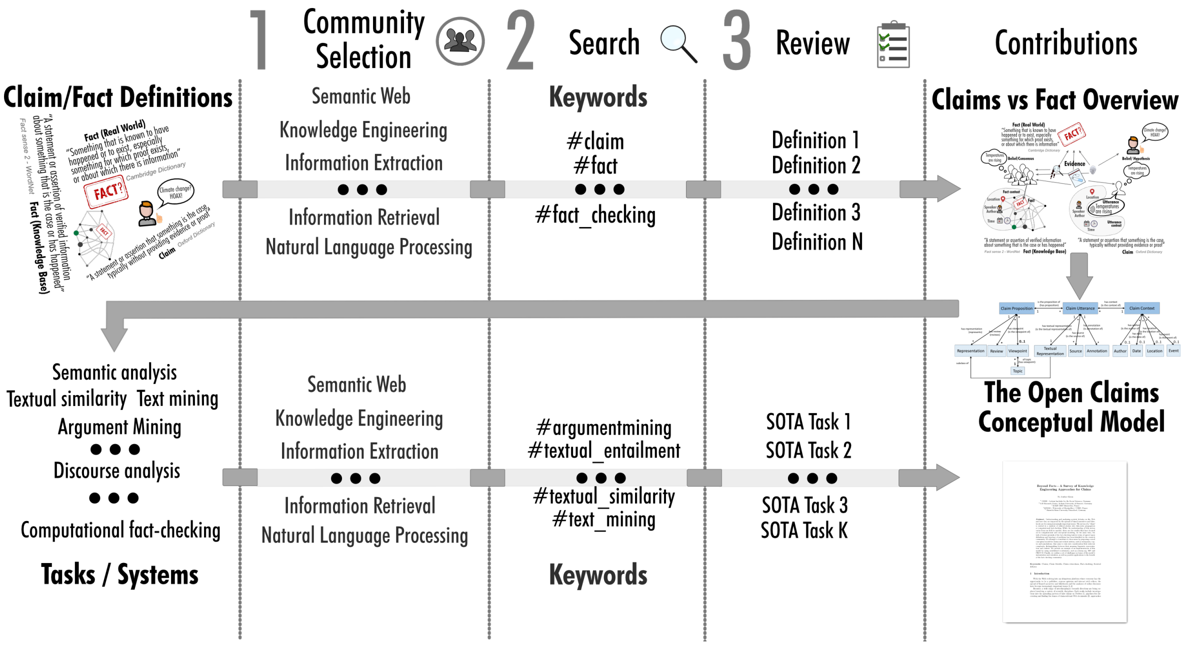


Figure 2.1: Publication selection and review workflow

2.2.1 Selection of Research Fields

First, we identified application areas and research fields involved with claims, facts or relevant concepts.

Application domains include, on the one hand, areas related to natural language claims, which are of concern in fact-checking portals, computational journalism or scientific discourse analysis, for instance, as part of scholarly publications, all involving claims of varying complexity. On the other hand, structured knowledge bases such as Wikidata are used in various applications such as Web search and involve factual statements bound to a predefined grammar relying on triples involving a subject (s), predicate (p) and object (o).

It becomes apparent that a more explicit and clear definition of the concepts of *facts vs. claims* is needed as both are relevant to this survey. Works focusing on claims made in the context of discourse can be found in *argumentation mining*, *argumentation theory*, *discourse modeling*, and *pragmatics*.

Facts are central for *knowledge representation / augmentation* works. With claims not only transporting beliefs or knowledge about factual information, but also conveying subjective information such as opinions, stances or viewpoints, relevant definitions and concepts can also be found in works targeting *stance detection*, *viewpoint extraction* and *opinion mining / sentiment analysis*. Rumours can be considered specific kinds of claims, thus we include definitions from the *rumour detection* field. Finally, retrieval of claims or respectively facts about specific entities is central to *question answering* and *information retrieval* in general, for instance, in the context of *fact retrieval*, *entity summarisation* or *entity retrieval*. Relevant works from these fields are also taken into account.

2.2.2 Search and Review Process

Table 2.1: Core venues analyzed systematically for the survey of fact and claim definitions and related concepts. Related events and workshops that were also considered: Workshop on Argument Mining (ArgMining), Fake News Challenge (FNC), CLEF Lab: CheckThat!, Fact Extraction and VERification (FEVER) Shared Task

Community	**Journals**	**Conferences**
NLP	Computer Speech and Language	ACL, EMNLP, COLING, NAACL-HLT
Web (Mining)	ACM TWEB	WWW, WSDM
IR	Information Retrieval Journal (Springer)	SIGIR, ECIR
AI		AAAI, IJCAI, ECAI
Knowledge-based Systems & Knowledge Graphs	SWJ, TKDE, JWS, Elsevier KBS	ISWC, ESWC, CIKM

Works addressing the aforementioned fields and tasks can be found in a variety of different scientific communities, particularly NLP, Web Mining, Information Retrieval (IR), Knowledge-based Systems and Artificial Intelligence (AI). Based on an initial set of publications from these communities dealing with extraction, verification or linking of claims and facts, found using a keyword-based search, we selected venues from the most relevant papers for systematic screening. Table 2.1 gives an overview of the chosen core journals, conferences, workshops and events. For each of those, we screened the proceedings of the years 2015 - 2019 (incl. 2020 and 2021 to the extent possible at the time of writing and revision preparation) and widened the search beyond these venues using online search engines and databases, also considering pre-prints. Publications cited by relevant publications were also taken into account regardless of their venue. For each publication, we extracted formal and informal definitions and descriptions of the concepts of claims and facts which are the basis for the analysis in Sect. 2.3 and the development of the model introduced in Sect. 2.4. As part of the modeling process, we defined possible relations between the different classes and mapped the generation of information on classes and relations to knowledge engineering tasks (Sect. 2.5). We extended our search in the listed venues and beyond to also cover these tasks. The following set of keywords was used for both steps: *fact-checking, fact checking, fake news, fact verification, argumentation, discourse, pragmatics, logic, knowledge representation, knowledge base augmentation, knowledge base construction, Knowledge-Base Augmentation, stance, viewpoint, claim, opinion mining, sentiment analysis, rumour detection, rumor detection, question answering, information extraction, relation extraction, ontology learning*. This search procedure resulted in a set of 598 publications that we deemed potentially relevant for the topics. Distribution across venues and time periods are displayed in Figures 2.2 and 2.3. Note that not all of these publications contain relevant definitions or ended up being cited in this survey. To maintain readability, both figures only contain venues and years, for which we collected at least 10 publications.

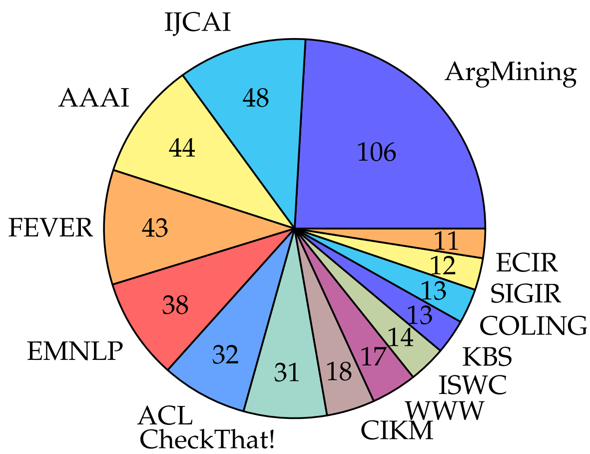


Figure 2.2: Analyzed publications and distribution over venues for all venues with at least 10 publications

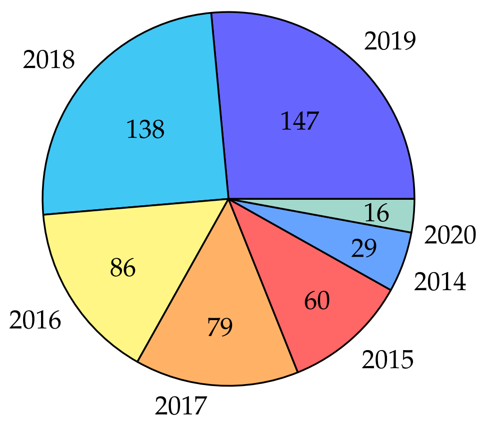


Figure 2.3: Analyzed publications and distribution over years for all years with at least 10 publications

2.2.3 Related Surveys and Conceptualizations

While this is, to the best of our knowledge, the first extensive survey on the conceptualization of facts and claims, several works have looked into different aspects of the problem providing overviews of related work in specific areas related to these aspects.

Konstantinovskiy et al. (2021) present a novel annotation schema and a benchmark for check-worthy claim detection, providing both an overview of claim definitions from other studies and a new definition of a claim that is constructed as a common denominator of existing ones. The novelty is that the definition is cast in the context of a claim being worthy of fact-checking—an important property of an utterance in view of verifying its veracity. The difficulty of identifying and defining fact-check worthiness of a claim is discussed with regard to the different perspectives that can be given to a single claim according to the human annotator's background.

Daxenberger et al. (2017) also take interest in the task of claim identification, but from an argumentation mining perspective, where this task is defined as recognizing argument components in argumentative discourse. The authors propose a qualitative analysis of claim conceptualization in argumentation mining data sets from six different domains ("different domains" here mean different data distributions). They show that the ways in which claims are conceptualized in each of these data sets are largely diverging and discuss and analyze the presumed harmful impact of these divergences on the task of cross-domain claim identification.

Thorne and Vlachos (2018) take a holistic stance on the problem and task of automated fact-checking. They provide an overview of approaches, data sets and methods covering the various steps of the process. This is the first paper of its kind that formulates the ambition to unify the often diverging definitions presented in related works from the fact-checking field by identifying shared concepts, data sets and models. A particularity of the survey is the fact that the authors consider both text-like and structured definitions of claims (e.g. in the form of triples), covering works on knowledge graph building and completion.

Fake news detection is related to fact-checking, but remains a distinct problem. Zhou et al. (2019) provide a definition of fake news and present relevant fundamental theories in various disciplines on human cognition and behaviour that are assumed useful for understanding fake news propagation and detection. Their survey on fake news detection methods is built along four categories of methods: (i) Knowledge-based methods, which verify if the knowledge within the news content matches certified facts; (ii) Style-based methods that look into the form of fake news (e.g., expressing extreme emotions); (iii) Propagation-based methods that are based on online spreading patterns; and (iv) Source-based methods investigating the credibility of sources.

Rumours are often seen as a specific kind of fake news. Zubiaga et al. (2018) provide a survey on rumour identification and resolution, where conflicting and diverging definitions of rumours from related works are discussed, but without making parallels to related notions such as fake news or biased discourse. The main motivations are the assumed impact of social media on rumour generation and spread. The survey focuses on datasets for rumour detection, as well as existing tools for accessing, collecting and annotating social media data for the purposes of automated rumour detection. The authors analyse generic rumour detection systems by breaking them down to their different components and subsequently discussing the related approaches to address the challenges related to each of those components. In that, the paper presents rumour tracking systems, rumour stance classification and veracity classification approaches.

Both the lack of and necessity for shared understanding and conceptualization of

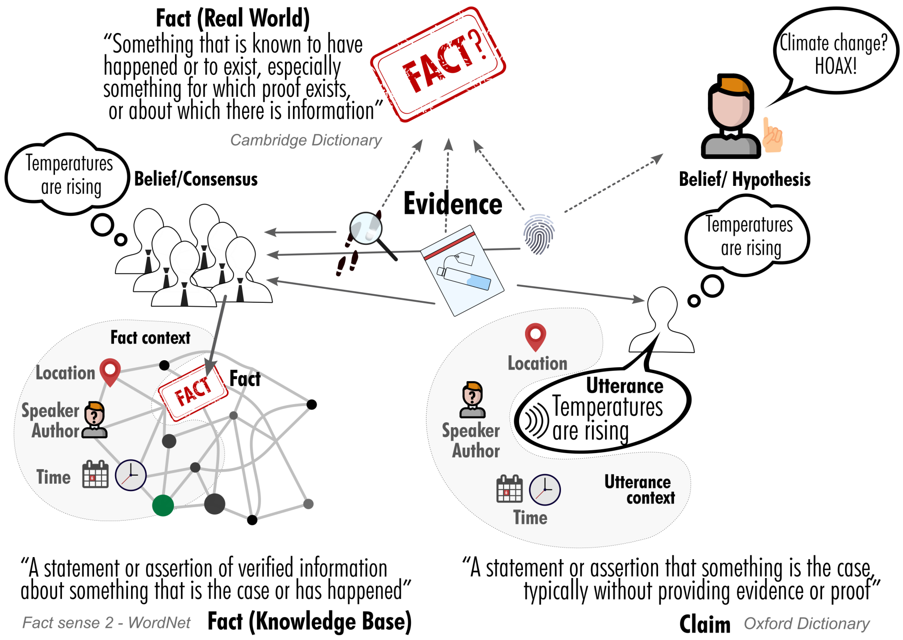


Figure 2.4: An overview of definitions and relations between facts and claims

claims surfaces from all of the above studies, which is underlined as their main motivation. However, the fact that some of these surveys discuss the same notions and refer to the overlapping sets of related work but by using different terminology (like e.g. Zubiaga et al. (2018)) comes to show that these works do not fully contribute to closing the terminological and conceptual gap that exists within and across fields as these studies discuss more narrow concepts of claims/facts used in specific domains rather than aiming at providing a shared view on the overlap and differences between used terminology.

2.3 Definitions and Conceptualizations

While the analysis of facts and claims plays a crucial role for a number of fields, the definitions of these concepts vary and are often left to the intuition of the reader. Existing definitions vary considerably not only across different fields but also within a single community. At the same time, different communities use the same terminology to refer to different concepts. In this section, we expatiate on different concepts for facts and claims, explain commonalities and differences and introduce a selected vocabulary to refer to these and related concepts throughout this paper. An overview is given in Figure 2.4.

2.3.1 Facts

A fact in the everyday use of the term (depicted on the top of Figure 2.4) refers to "A thing that is known or proved to be true"[1], "something that has actual existence"[2], "something that is known to have happened or to exist, especially something for which proof exists, or about which there is information"[3], "something that actually exists; reality; truth"[4], "an event known to have happened or something known to have existed"[5] or "a concept whose truth can be proved"[6]. Note that not everything that is a fact according to this definition can be observed directly; instead, beliefs about them can be formed by observing *evidence*.

Facts in Knowledge Bases

In the semantic web community and the fields of knowledge representation and knowledge base construction / augmentation, facts are seen as the knowledge that is represented in KGs or KBs (Balazevic et al., 2019; Gad-Elrab et al., 2019; Veira et al., 2019; Voskarides et al., 2018; Al-Bakri et al., 2015; Martinez-Rodriguez et al., 2020; Zhu et al., 2017; Padia et al., 2019; Chen et al., 2019a; Zhan, Q et al., 2019; Trisedya et al., 2019; Wang et al., 2019; Rospocher et al., 2016; Augenstein et al., 2016a; Ciampaglia et al., 2015; Shi and Weninger, 2016; Pasquale Minervini et al., 2017; Martinez-Rodriguez et al., 2020; Fionda and Pirrò, 2018; Gerber et al., 2015). More precisely, items in KGs or KBs are coined statements of facts or assertions or triples encoding / representing facts (Ciampaglia et al., 2015; Shi and Weninger, 2016; Pasquale Minervini et al., 2017; Chen et al., 2019a; Veira et al., 2019), with the

[1]Oxford Dictionary; `https://www.lexico.com/en/definition/fact`
[2]Merriam-Webster Dictionary; `https://www.merriam-webster.com/dictionary/fact`
[3]Cambridge Dictionary; `https://dictionary.cambridge.org/dictionary/english/fact`
[4]Dictionary.com; `https://www.dictionary.com/browse/fact`
[5]WordNet; `http://wordnetweb.princeton.edu/perl/webwn?s=FACT`; S3
[6]WordNet; `http://wordnetweb.princeton.edu/perl/webwn?s=FACT`; S4

facts being assumed to be true, can be proven to be true or are likely to hold (Padia et al., 2019; Popat, 2019; Ciampaglia et al., 2015). However, the use of terminology is not consistent: *fact* is often used as synonym for RDF triple (Gerber et al., 2015; Huynh and Papotti, 2018; Zhang et al., 2019a; Padia et al., 2019; Zhu et al., 2017) or for the representation of a fact, respectively assertion, but there is often no distinction made between "fact" and "statement of a fact" (Fionda and Pirrò, 2018; Martinez-Rodriguez et al., 2020; Pasquale Minervini et al., 2017; Rospocher et al., 2016). The interchangeable use of "statement of fact" and "fact" leads to a widespread terminology of "checking whether facts are true" (Syed et al., 2019), implying that facts may not be true. Depending on the precise definition of fact, this might be an oxymoron, i.e. when defining a fact as something that is known to be true. Having the task of fact-prediction as background, some works coin the relations between entities or the paths in a knowledge base as *facts* (Voskarides et al., 2018; Padia et al., 2019). As Gerber et al. (2015) note, facts have a scope, e.g. a temporal one, that determines the context that has to be taken into account in order to judge their validity.

Types of Facts

Several more fine-grained distinctions of different types of facts can be found in the literature. Facts can refer to relations or attributes (Zhang et al., 2019a), or can be attributes of other facts (Voskarides et al., 2018). They can pertain to numerical properties, quotes or other object properties (Huynh and Papotti, 2018). They can be assessed according to their "*check-worthiness*" (Huynh and Papotti, 2018) or *importance* for the containing KB (Voskarides et al., 2018). Another interesting distinction is made by Tsurel et al. (Tsurel et al., 2017) who aim at identifying facts that are suitable to be used as interesting trivia by developing a measure for *trivia-worthiness* that relies on surprise and cohesiveness of the contained information.

Throughout this paper, we will use the term *fact* referring to knowledge that is generally accepted to be true and refer to items in knowledge bases as *statements of facts*.

Facts vs. Evidence

Related to the notion of *fact* is the notion of *evidence*. Evidence is seen as something to support or contradict a claim (Stahlhut, 2019; Aharoni et al., 2014; Rinott et al., 2015). Some works give a more narrow definition relating to their specific use cases, e.g. (Zhan, Q et al., 2019, p. 1) define *evidence* as "text, e.g. web-pages and documents, that can be used to prove if news content is or is not true". As Stahlhut (2019) notes, the task of evidence detection is similar to premise detection in argumentation mining. A premise in argumentation mining is, as (Stab and Gurevych, 2014, p. 1) put it, "a reason given by an author for persuading the readers of the claim". *Evidence* and *premise* directly correspond to each other, as both terms are often used interchangeably (Shnarch et al., 2018; Trautmann et al., 2020; Lippi and Torroni, 2015).

Evidence can be categorized into many different types, such as *expert opinion, anecdote, or study data* (Stahlhut, 2019), or, with slightly different wording, *study, expert or anecdotal* (Aharoni et al., 2014). Walker et al. (2018) distinguish *lay testimony, medical records, performance evaluations, other service records, other expert opinions, other records*. Niculae et al. (2017) include *references* such as URLs or citations as pointers to evidence. Premises can refer to *logos, pathos or ethos* (Hidey et al., 2017). For scientific articles, Mayer et al. (2018b) distinguish the classes *comparative, significance, side-effect, other*.

While some works refer to knowledge found in texts or other resources as *evidence* for a fact (Nie et al., 2019; Rospocher et al., 2016; Fionda and Pirrò, 2018; Augenstein et al., 2016a) and call it *fact* only after the truthfulness has been determined and that knowledge is entered into a knowledge base, other works assume the truthfulness of the mentions and refer to them or the knowledge they represent as *facts* directly (Clancy et al., 2019; Hanselowski et al., 2018c). Very related is the task of *Truth Discovery*. "Truth Discovery aims at identifying facts (true claims) when conflicting claims are made by several sources" (Beretta et al., 2018). In this domain, the terms *data items* and *truths* are used to refer to invalidated mentions of knowledge and the true values respectively (Wang et al., 2016; Xiao et al., 2019).

2.3.2 Claims

A claim is commonly seen as "a statement or assertion that something is the case, typically without providing evidence or proof".[7]

Claims in Argumentation

In line with this definition, works in argumentation mining and argumentation theory focus on claims as the key components of arguments (Daxenberger et al., 2017), as statements that are made to convince others or express someone's views, evaluations or interpretations (Rosenthal and McKeown, 2012; Lippi and Torroni, 2016a; Lugini and Litman, 2018; Hidey et al., 2017).

Claims denominate the conclusion of an argument, the assertion the argument aims to prove or the thesis to be justified (Besnard and Hunter, 2008; Palau and Moens, 2009; Lippi and Torroni, 2015, 2016a; Levy et al., 2018; Lippi et al., 2018; Stab and Gurevych, 2014). Claims correspond to propositions in argumentation models and both terms are often used interchangeably, "The claim is a proposition, an idea which is either true or false, put forward by somebody as true" (Palau and Moens, 2009). As Daxenberger et al. (2017) point out, the exact definition of a claim, even inside the field of argumentation mining, depends on the domain or task and is somewhat arbitrary. Also, as Torsi and Morante (2018) show, related annotation categories are often not well defined.

With the use case of scientific articles in mind, Mayer et al. (2018a) define a claim as a concluding statement made by the author about the outcome of the study. Focusing on debates, (Aharoni et al., 2014, 2), but also (Rinott et al., 2015, 1), define a claim as a "general, concise statement that directly supports or contests the topic". A topic here is defined as "a short, usually controversial statement that defines the subject of interest" or "a short phrase that frames the discussion" respectively. Examples for such topics are "Use of performance enhancing drugs (PEDs) in professional sports" with a claim being "PEDs can be harmful to athletes health" (Rinott et al., 2015, p. 2) or "The sale of violent video games to minors should be banned" with a claim being "Violent

[7] `https://www.lexico.com/en/definition/claim`

video games can increase children's aggression"(Aharoni et al., 2014, p. 3). Note that these definitions diverge from the common definition of a *topic* as the *underlying semantic theme* of a document with a topic being *a probability distribution over terms in a vocabulary* (Blei and McAuliffe, 2007) as used in topic modelling and document classification. There, a *topic* may be represented by terms on a coarse-grained level such as *Health* or *Computers & Internet* (Yang et al., 2016). This concept of a *topic* is also used by Chen et al. (2019b) in their work about discovering perspectives about claims. Also, the second example of a topic can be seen as a claim or stance itself. Durmus et al. (2019) represent topics by tags of pre-defined categories similar to the above described semantic themes plus what they call a *thesis*, corresponding to Aharoni et al. (2014)'s claim-like topics, e.g. "free Press is necessary to democracy.", "All drugs should be legalised.".

In the following, *topic* will be used to refer to the frame of the discussion, as defined by Rinott et al. (2015) while the *underlying semantic theme* will be referred to as the *subject*.

Types of Claims

According to Lippi and Torroni (2016a), there are three different types of claims: 1) *epistemic*, i.e., claims about knowledge or beliefs, 2) *practical*, i.e., claims about actions, alternatives and consequences, and 3) *moral*, i.e., claims about values or preferences. For example, "our survival rate for cancer that used to be some of the worse in Europe now actually is one of the best in Europe, we are changing the NHS and we are improving it" [sic] is an epistemic, "cuts will have to come, but we can do it in a balanced way, we can do it in a fair way" a practical and "I don't want Rebecca, I don't want my own kids, I don't want any of our children to pay the price for this generation's mistake" a moral claim (Lippi and Torroni, 2016a).

Similarly, Schiappa and Nordin (2013); Fierro et al. (2017) differentiate *claims of fact, value* and *p*olicy. Claims of fact state that something is true, i.e. they express a belief about a fact. This corresponds to the epistemic claims according to Lippi and Torroni (2016a)'s taxonomy with claims of value and policy corresponding to moral and practical claims, respectively. Epistemic claims are also referred to as *factoid claims* (Thorne et al., 2018c, 2019) or, more commonly, *factual claims*, e.g. (Levy et al., 2018;

Hassan et al., 2015, 2017a; Mohtaj et al., 2019; Ghanem et al., 2019, 2018; Padia et al., 2018; Konstantinovskiy et al., 2021). However, assessing the *factuality* of a claim may refer to assessing a claim's veracity (Hasanain et al., 2019) rather than assessing whether it is a factual or non-factual claim. Note that all types of claims can be used to express a stance in discourse but not all of them are verifiable.

Some works propose a more fine-grained differentiation of claims according to their use cases, e.g. Lauscher et al. (2018b) distinguish *Own Claims* vs. *Background Claims* vs. *references to Data* for argumentation mining of scientific texts. Hassan et al. (2019) distinguish between the classes *Non-statistical* (e.g. quotes), *Statistical*, *Media* (e.g., photo or video), and *Other*, Zhang et al. (2016) between categorical vs. numerical claims. Park and Cardie (2014) categorize claims according to their verifiablity and distinguish between *unverifiable, verifiable nonexperiential, verifiable experiential* claims with *experiential* referring to whether the claim refers to the writer's personal state or experience or not. Another notion that can be seen as a specific type of claim is a *rumour*. In an attempt to unify various definitions found in works addressing the identification and veracity assessment of rumours, (Zubiaga et al., 2018, p. 1) define rumours as "items of information that are unverified at the time of posting". The authors further distinguish between different types of rumours, with respect to their currentness (emerging vs. longstanding rumours).

Claims vs. Stances vs. Viewpoints

Habernal and Gurevych (2017) explain that the term claim in the context of argumentation theory is a synonym for standpoint or point of view referring to what is being argued about, i.e. the topic. This is in line with Liebeck et al. (2016)'s and Aharoni et al. (2014)'s debate-oriented definition and with (Hidey et al., 2017, p. 4)'s definition of claims as "proposition that expresses the speaker's stance on a certain matter". *Standpoint, point of view and stance* in these definitions do not mean the content of the claim has to be of an unverifiable or of a purely opinionated nature. Stab and Gurevych (2017) see a *stance* as an attribute of a claim.

Stances are usually defined as text fragments representing opinions, perspectives, points of views or attitudes with respect to a target (Zhi et al., 2017; Ghanem et al., 2018; Ghosh et al., 2014; Kotonya and Toni, 2019; Hanselowski et al., 2018a). They can be expressed explicitly or implicitly (Rajendran et al., 2016). Fragments can be messages such as tweets or posts (Joseph et al., 2017; Giasemidis et al., 2020), paragraphs (Potash et al., 2019) or complete articles (Hanselowski et al., 2018a). Joseph et al. (2017) see stances as latent properties of users rather than text fragments. Text fragments can however reveal a user's stance. As Joseph et al. (2017) point out, stance and sentiment are related, but not the same: a negative sentiment of a text can be paired with a positive stance towards a particular target and vice versa. Also the tasks of aspect-based sentiment analysis and stance detection differ, even though both aim at detecting opinions towards a target. For example, a piece of text may express a positive sentiment towards a specific aspect of a person, e.g. their personality, but still argue against this person's claim.

Stance detection has been used to determine opinions on the veracity of claims (Ferreira and Vlachos, 2016; Ma et al., 2018). Stances in these works are similar to what is coined *evidence* in fact-checking works, as described above, although they do not necessarily contain factual information that can be used to verify information. Note that this may be the case for *evidence* as well, depending on the precise definition. The fact that a claim is supported by another entity than the source can be seen as evidence for the claim's truthfulness in itself (cf. *expert*-type evidence).

Stances have been classified into different categories such as *for, against and observing* (Ferreira and Vlachos, 2016), *pro* and *con* (Bar-Haim et al., 2017b) and *none* (Toledo-Ronen et al., 2016), *agree, disagree, discuss*, or *unrelated* (Nadeem et al., 2019). There is also a hierarchical model that classifies the stance of web documents in three levels: first as *related* or *unrelated*, the *related* ones as *taking a stance* or *being neutral*, and those *taking a stance* as *agree* or *disagree* (Roy et al., 2021). Another fine-grained distinction can be found in Hidey et al. (2017), who distinguish *interpretations, rational evaluations, emotional evaluations, agreement* and *disagreement*. As Kotonya and Toni (2019) note, the task of stance classification is closely related to relation-based argumentation mining that determines attack and support relations between argumentative units.

Another related task is that of viewpoint discovery. Thonet et al. (2016) define a viewpoint as "the standpoint of one or several authors on a set of topics". A viewpoint goes beyond a person's stance on a specific subject and represents their global standpoint or side they are taking. As Thonet et al. (2016) explain, a viewpoint in a debate about the building of Israeli communities on disputed lands can for example be summarized as "pro-Palestine" or "pro-Israel". Consequently, Viewpoint Discovery is considered a sub-task of Opinion Mining (Thonet et al., 2016; Quraishi et al., 2018).

Another closely related, but different notion is that of a *perspective* which Chen et al. describe as an argument that constitutes a particular attitude towards a given claim, an opinion in support of a given claim or against it (Chen et al., 2019b). For example, for a claim "Animals should have lawful rights" a perspective would be "Animals are equal to human beings", which would express support for the claim. A *perspective* corresponds to an opinion on a specific *aspect* in a viewpoint. *Perspectives* can be supported by evidence, connected to claims by *supports* or *attacks* relations and can be seen as a specific type of claims that are connected to what Chen et al. coin *argue-worthy* claims.

Claims in Journalism and Fact-checking

Works outside of the area of argumentation focus less on the role of the claim in the context of the discourse and more on the content of the claims.

A very general definition is given by (Zhang et al., 2016, p. 2) for their truth discovery approach: "A claim is defined as a piece of information provided by a source towards an entity".

From a journalistic fact-checking perspective, dedicated platforms focus on statements supported by (a group of) people or organizations that appear news-worthy, check-worthy, significant and verifiable (cf. definitions from, e.g., politifact.com[8], truthorfiction.com[9], or checkyourfact.com[10]). Newsworthiness and significance are not only subjective, both can also vary depending on historical or political context (Graves, 2018).

For other use cases, different definitions or restrictions of what is considered a claim are employed.

[8] https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/
[9] https://www.truthorfiction.com/about/
[10] https://checkyourfact.com/about-us

Automatic fact-checking often constrains the problem by limiting the kinds of claims being checked to focusing on simple declarative statements (*short factoid sentences* (James Thorne and Andreas Vlachos, 2019)) or claims about statistical properties (Vlachos and Riedel, 2015; Thorne and Vlachos, 2017; Graves, 2018). For the Fast & Furious Fact-Check Challenge, four primary types of claims were distinguished and further differentiated into more fine-grained sub-categories:[11] 1) *numerical claims* (involving numerical properties about entities and comparisons among them), 2) *entity and event properties* (such as professional qualifications and event participants), 3) *position statements* (such as whether a political entity supported a certain policy) and 4) *quote verification* (assessing whether the claim states precisely the author of a quote, its content, and the event at which it supposedly occurred). Note that fact-checking portals contain many quoted claims but it is not always clearly marked whether the quote itself is verified (i.e. did the person indeed make the claim?) or the content of the quoted claim (i.e. is the claim allegedly made by the person correct?).

Claims in Information Retrieval and Question Answering

In the area of Information Retrieval and Question Answering, several works focus on retrieving scientific claims and claims in digital libraries. Here, a claim is defined as a statement formulating a problem together with a concrete solution (González Pinto and Balke, 2017) or a sentence in a scientific document that relates two entities given in a query (Gonzalez Pinto et al., 2019; González Pinto and Balke, 2018). More generally, from a database-centric perspective, Wu et al. (2014, 2017) represent a claim as a "parametrized query over a database". This allows to computationally study the impact of modifying a claim (i.e. its parameters) on the result of the query and to thus identify claim properties, such as claim robustness which may serve as evidence to detect potential misleadingness i.e. due to cherry-picking. A related perspective has been proposed by Cohen et al. (2011) in the field of Computational Journalism.

[11] `https://www.herox.com/factcheck/5-practise-claims`

2.3.3 Discussion and Concluding Remarks

In summary, works focusing on the argumentation domain investigate claims in the context of a discourse, i.e. taking their pragmatic role into account. Claims are uttered by the author or speaker to achieve an aim through a speech act (Searle, 1969). In order to recognize the meaning of an utterance and draw conclusions about the intention of the author, the pragmatic context has to be taken into account. A claim often carries a variety of intentional or unintended meanings, where subtle changes in the wording or context can have significant effects on its validity (Graves, 2018). Works in other areas, such as Knowledge Bases and Fact-checking, typically focus on the content of epistemic claims, i.e. rather than trying to analyze intended meanings or messages, they try to find and check evidence for assertions and find facts vs. false claims of fact. Works in the area of information retrieval focus more on the surface of claims, trying to retrieve relevant texts without necessarily analyzing their content or contexts. These differences are reflected in the claim definitions found in the respective works.

Note that due to these different foci, there is a difference in what is referred to as *claim* in argumentation mining vs. in the automatic fact-checking community: what is used as premise or evidence in an argument is often selected as check-worthy *claim* by fact-checking sites, not the evaluative component of the argument that is coined *claim* in argumentation mining. Generally, the distinction of argumentative units such as claims and evidence in argumentation mining is based on the statement's usage or its relations in an argument, while fact-checking classifies statements into claims, stances and other categories considering features inherent to the statement itself (such as their subjectivity), regardless of their connection to the discourse. Thus, what is identified as *claim* in works of one research field or labelled *claim* in a ground truth corpus may or may not be called *claim* in the other, depending on the specific use case and context.

Likewise, some works focus on identifying claims (or other argumentative components) that belong to a pre-defined topic (called *corpus wide topic-dependent* (Shnarch et al., 2018), *context-dependent* (Aharoni et al., 2014), or the *information-seeking* perspective (Trautmann et al., 2020)), while others aim at extracting any units that act as claims for any topic (*closed-domain discourse-level* (Trautmann et al., 2020) or *context/topic-independent*). Using topic-dependent annotations as ground truth for topic-independent extraction approaches leads to impaired precision values (Lippi and Torroni, 2016c).

Lastly, another difference between statements of fact in knowledge bases and claims is that for the former, a certain level of consensus at least in the respective community can be assumed, while claims may only represent the beliefs of one person or be uttered by them to achieve a certain goal such as spreading disinformation. Thus, it makes sense to model truth values for claims while statements in knowledge bases are assumed to be true. There may be errors in knowledge bases, however. Thus, modeling uncertainty or confidence values is applicable for them.

The task of assessing the correctness of a statement of fact is called *fact validation*. The task of assessing the veracity of a claim is called *fact-checking*. Fact-checking has also been modeled as a specific *stance detection* task where the stance of a source or evidence unit towards an epistemic claim is used to assess the claim's veracity. Finding the true values in case of conflicting evidence is the aim of *truth discovery*.

2.3.4 Naming Conventions

To arrive at a more precise usage of terminology, we will, throughout this paper, refer to items in knowledge bases as *statements of facts*, while other mentions or assertions of knowledge, will be referred to as *claims about a fact* that can act as *evidence* about some information being true and its content being a *fact*. An index of all naming conventions followed in this work is given in Table 2.2.

2.4 Conceptual Modeling

In this section, we propose a conceptual model for representing claims and related data as well as an example of an implementation of this model in RDF using established vocabularies.

The conceptual model was informed through the survey described in the previous

Table 2.2: Index of main notions and definitions as discussed in this paper

Term	Definition
Claim	A statement or assertion that something is the case
Fact	A thing that is known or proved to be true
Statement of fact	A statement in a knowledge base
Evidence	Information that can be used to assess the truthfulness of a claim or claimed value or relation
Topic	Phrase describing the frame of the discussion
Subject	Keyword describing the semantic theme
Stance	Support or opposition expressed by a user or text fragment with respect to a given target
Viewpoint	Standpoint of one or several authors on a topic or set of topics

section. In order to derive a conceptual model, we followed the following steps: 1) identification of key concepts to be reflected in the model (e.g. *claim proposition*, *claim utterance*), 2) deriving definitions of these concepts by considering established definitions from the literature, 3) excluding definitions that are inconsistent with each other or not reflecting the required granularity (e.g. we argue that a distinction between *proposition* and *utterance* is important for many NLP and knowledge engineering tasks), and 4) identifying relations between all concepts which are consistent with and/or implied by our definitions. Through this process, we arrive at a conceptual model containing key concepts, relations and definitions which is then implemented in OWL as well as through a dedicated RDF/S data model. We start by giving an overview of the key terminology.

2.4.1 Key Terminology - From Pragmatics to Fact-Checking

For our conceptual model, we follow notions from pragmatics to allow modeling not only a claim in isolation, but also its meaning in a given discourse and its role in communication.

As Green (1996) puts it, "(...) communication is not accomplished by the exchange of symbolic expressions. Communication is, rather, the successful interpretation by an addressee of a speaker's intent in performing a linguistic act."(Georgia M. Green, 1996, p. 1) “Minimally the context required for the interpretation (...) includes the time, place, speaker, and topic of the utterance.” (Georgia M. Green, 1996, p. 2) While this quote refers to the interpretation of indexical expressions (i.e. words like “here” and “now”), the same holds true for the interpretation of the meaning of an utterance in general.

A linguistic act, or speech act following Searle (1969), includes an utterance, a proposition, an illocution and a perlocution. An *utterance* is a grammatically and syntactically meaningful statement. A *proposition* is the semantic content, i.e. meaning. An *illocution* is the intended effect, e.g. persuading the addressee or requesting a service, while a *perlocution* is the achieved effect.

For example, referring to the topic “Brexit”, i) British journalist David Dimbleby said during a topical debate in Dover *“We are going to be paying until 2064, apparently”*[12], and ii) a news article of The Independent on the same topic wrote *(“UK will be paying Brexit “divorce bill” until 2064”*[13]*)*. While the surface forms of these utterances differ, they express the same proposition. At the same time, utterances with equivalent surface forms may be used to express different and even contradicting propositions or viewpoints when embedded in different contexts. Consider the two claims: (i) *“The unemployment rate among Poles in Britain is lower than the unemployment rate among Brits”*, uttered by British public policy analyst and former Labour Party politician David Miliband[14] to argue that immigrants are not a drain on the British welfare system and thus not bad for the British society; (ii) *“EU migrants are MORE likely to have a job in the UK than British citizens”*, written by MailOnline journalist Matt Dathan[15] to make a point that immigrants are taking away British citizens’ jobs and thus are bad for society. The propositions are semantically similar and both utterances aim at persuading the audience (illocutionary act) but the expressed viewpoints are different.

This can only be recognized when taking the context into account which is why

[12] https://fullfact.org/europe/brexit-divorce-bill-2064/

[13] https://www.independent.co.uk/news/uk/politics/uk-brexit-divorce-bill-taxpayers-deadline-treasury-obr-office -budget-responsibility-a8253751.html

[14] https://www.politifact.com/factchecks/2016/jun/26/david-miliband/david-miliband-link-between-jobs-immigration-and-b/

[15] https://www.dailymail.co.uk/news/article-3628840/The-true-cost-open-borders-revealed-EU-migrants-likely-job-UK\-British-citizens.html

we argue that the context should be modeled along with the claim utterance. The importance of contextual information has also been recognized for the task of fact-checking: “Who makes a claim, when they say it, where they say it, and who they say it to, can all affect the conclusion a fact-checker could reach. Whether it’s true to say unemployment is what country or which part of a country a speaker is referring to, and when the speaker makes the claim. An open format for recording public debate should support metadata, including at least the time, the place, the venue or publication, and the speaker.” (Mevan Babakar and Will Moy, 2016).

As outlined in the previous section, we see a *fact* as a conceptual object which represents the current consensual knowledge in a given community about something or someone. While this knowledge is relatively stable, a change of its truth value is possible, for example when flaws in scientific studies are discovered and findings have to be corrected (Rekdal, 2014).

Any verified information about a claim, like who uttered it, when and where, can be considered a *fact*. Facts explicitly uttered by an agent can be modeled as (factual) claims. Facts extracted from a knowledge base can be represented using the same model: provenance information about the knowledge base can be represented as *source*, that is, as part of the *utterance*. The statement of a fact is typically not embedded in a discourse. Thus, certain attributes of the *context*, like the *topic* of the discourse and the *agent*, would remain undefined. Likewise, non-factual claims (e.g. “animals should have lawful rights”) do not have universally accepted truth values, i.e. they are unverifiable, and hence, *verdict* would remain undefined for the respective *proposition*. Therefore, we argue that facts, factual claims and non-factual claims can be represented using the same model.

2.4.2 The Open Claims Conceptual Model

In line with the rationale outlined above, we introduce the *Open Claims* conceptual model, which distinguishes three main components of a claim represented by three central classes: (1) *claim proposition*, (2) *claim utterance*, and (3) *claim context* (Fig. 2.5).

A *claim proposition* is the meaning of a statement or assertion. In the context of fact-checking and argumentation mining, it is usually related to a controversial topic and is supported by one person or a group of people. A *claim proposition* can have been expressed in many different ways and in different contexts, thus it has one or more *claim utterances*. For example, it might have been expressed in different languages, using different words in the same language, or uttered by different persons and/or in different points in time.

In contrast, a specific claim utterance is typically associated with only one claim proposition, i.e., it has a single meaning. However, the claim proposition can be represented in different ways, for example, by selecting a representative utterance with its context, or through a more formal model. Each claim utterance is related to a specific *claim context*, which includes the person who uttered the claim, the time point at which the claim was uttered, the location or the event of the utterance and the topic of the enclosing discourse. The claim context provides information to interpret the claim utterance and thus understand its proposition.

Since explicit information about the perlocution (achieved effect) and illocution (intended effect) of utterances is usually unavailable, we do not consider them in this model. They can, however, easily be added to the model as an extension.

Below, we provide details and the main properties of each of the three main classes (Claim Proposition, Claim Utterance, Claim Context).

An OWL implementation of the Open Claims model is available online.[16] To facilitate data integration with existing relevant datasets, such as ClaimsKG (Tchechmedjiev et al., 2019b), TweetsKB (Fafalios et al., 2018a) and TweetsCOV19 (Dimitrov et al., 2020a), we also provide an RDF/S implementation of the model using existing vocabularies (more below in Sect. **??**).

Claim Proposition

A claim proposition is the meaning of one or more claim utterances in their respective contexts. A claim proposition is associated with i) zero, one or more *representations*,

[16] `https://data.gesis.org/schema/openclaims`

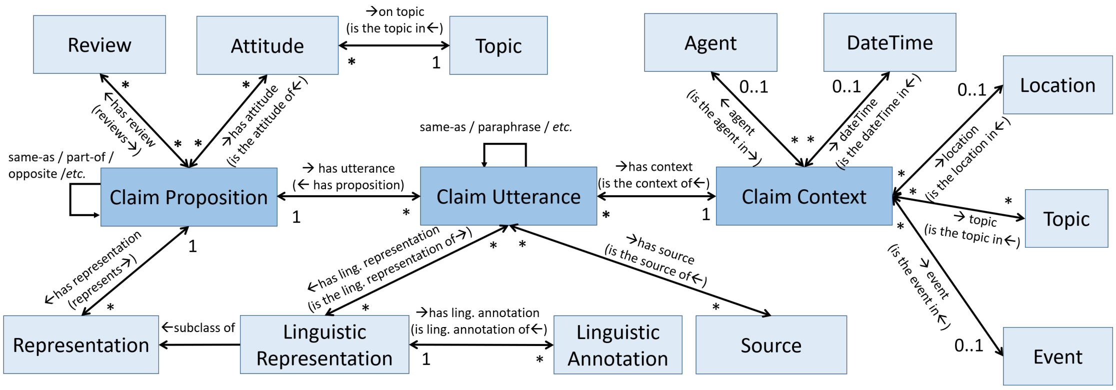


Figure 2.5: The Open Claims Conceptual Model

ii) zero, one or more *reviews*, iii) zero, one or more *attitudes*, and iv) zero, one or more other *claim propositions*.

A *representation* can have the form of free text, e.g. a sentence that describes the proposition as precisely as possible, or be more formal, e.g. a first-order logic model, or the URI of a named graph pointing to a set of RDF statements.

A *review* is a resource (e.g. a document) that analyzes one or more check-worthy claim propositions and provides a verdict about their veracity or trustworthiness. An example of such a review is an article published by a fact-checking organization. Note here that not all factual claims have a clear verdict. For instance, the claim *"the presence of a gun makes a conflict more likely to become violent"* represents hypothesis which can be linked to both supporting and contradicting evidence and is thus difficult to be associated with a single overall correctness score. If a claim is associated to a review which gives a true verdict about its veracity, then the claim can be considered a fact (it represents the current knowledge about something). Non-factual claims are not linked to any reviews and have no verdicts.

An *attitude* is the general opinion (standpoint, support) on a given topic (e.g. a viewpoint), which often underlies a set of specific values, beliefs or principles. For instance, *pro-Brexit* and *anti-Brexit* are two different viewpoints for the Brexit topic. A claim proposition can be associated with several attitudes for different topics. For example, the proposition linked to the claim *"immigrants are taking our jobs"* can support both the *against immigration* (for the Immigration topic) and the *pro-Brexit* attitude (for the Brexit topic).

A claim proposition can also be associated with other claim propositions through some type of relation, e.g. *s*ame-as, opposite, part-of, etc.

Claim Utterance

A claim utterance is the expression of a claim in a specific natural language and form, like text or speech. Among other things, it can be something said by a politician during an interview, a text within a news article written by a journalist, or a tweet posted by a celebrity about a controversial topic. It is associated with i) one or more *linguistic representations* (subclass of *representation*), ii) one or more *sources*, and iii) zero, one or more other *claim utterances* (through relations such as *s*ame-as, paraphrase, etc.).

A *linguistic representation* can be, for example, a text in a specific language that best imprints the claim as it was said/appeared, or a sound excerpt from someone's speech.

A *source* provides evidence of the claim's existence. For instance, it can be the URL of an interview video, a news article, or a tweet, i.e. source here means the medium reporting the utterance, not the originating agent (speaker or author which is part of the context). For this distinction, see also (Newell et al., 2018b). A linguistic representation can have one or more *linguistic annotations* which provide formal linguistic characteristics. For instance, it can be an entity or date mentioned in the text of the claim utterance, the sentiment of the text (e.g. positive, negative, neutral), or the linguistic tone of a speech (like irony). These annotations can enable advanced exploration of the claims (e.g. based on mentioned entities) and can be manually provided by a domain expert or automatically produced using a NLP or speech processing tool (like an *entity linking* (Shen et al., 2015) tool for the case of entity annotations in text).

Links between utterances can be also used to explicate their role in discourse, e.g. by using relations such as *used-as-evidence-for* or *used-as-evidence-against* to model premises, evidence, conclusions and other components and relations in argumentation. Likewise, *supports* and *attacks* relations may hold between utterances to connect stances and their targets. With this, we follow Carstens and Toni (2015) and the discussion in Sect. 2.3 with the notion that whether a statement is of type *evidence* or another type and whether it was uttered to express a stance depends on its usage in the context of a discourse, e.g. its relations, rather than being an inherent property of the statement in isolation.

Claim Context

The *claim context* provides background information about the claim utterance. It is associated with metadata information about the claim utterance and, together with the linguistic representation of the claim utterance, can provide an answer to the *Five W's*: i) *what* was said (linguistic representation of claim utterance), ii) *who* said it (agent; person, group, organisation, etc., making the claim), iii) *when* it was said (date/time the claim was uttered), iv) *where* it was said (location where claim was uttered), and v) *why* it was said (event or activity in the context of which the claim was uttered, and/or the topic of the underlying discourse). The claim context provides the necessary information for interpreting the claim utterance (and thus understanding its proposition).

Instantiation Example

Fig. 2.6 depicts an instantiation example of the proposed conceptual model. The example shows information for two *claim utterances* (in pink background in Fig. 2.6): i) the one by David Dimbleby *("We are going to be paying until 2064, apparently")*, and ii) the one by The Independent *("UK will be paying Brexit "divorce bill" until 2064")*. Both utterances correspond to the same claim proposition (in green background in Fig. 2.6) and each one has its own context information (in yellow background in Fig. 2.6). The linguistic representation of the first claim utterance has been annotated with one *date annotation* (2064) and that of the second claim utterance with one *entity annotation* (United Kingdom).

The claim proposition has two representations, a textual one *("Britain will be paying its Brexit bill for 45 years after leaving the EU")* and a formal one *("cost = {of='Brexit', for='UK' amount=?, until=2064}")*, and supports the *against-Brexit* viewpoint of the

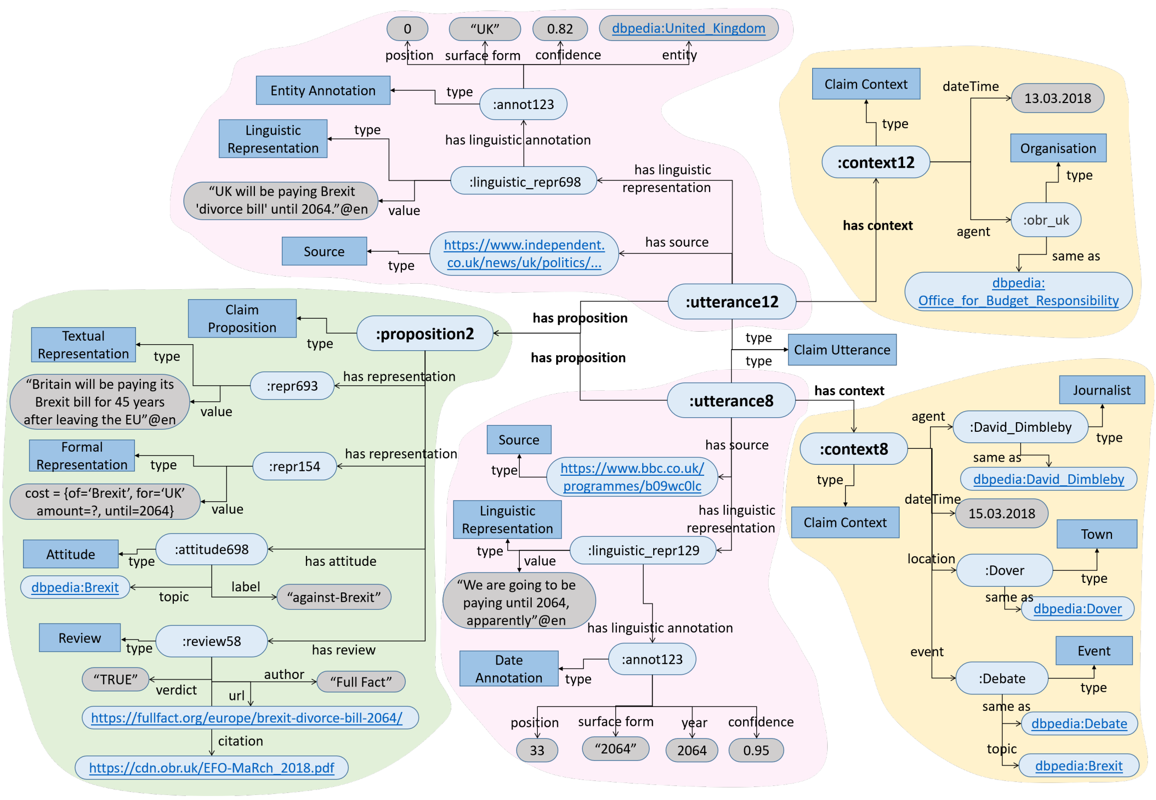


Figure 2.6: Instantiation example of the conceptual model

Brexit topic. In addition, there is a review of this claim proposition with verdict “true”, published by Full Fact (UK’s independent fact-checking organisation). Moreover, we can see the URL of the review article as well as a reference to a document file which provides evidence for its correctness.

The context of each claim utterance provides additional metadata information about the claim. For example, we see that the first utterance was said by *David Dimbleby* on *15.03.2018*, in the context of a *debate* about *Brexit* which took place in *Dover*. For the second claim utterance, the example only represents its agent (*UK Office of Budget Responsibility*) and date (*13.03.2018*).

2.4.3 RDF/S Implementation

We introduce an RDF/S implementation of the proposed conceptual model using established vocabularies, in particular schema.org[17], the Open Annotation (OA) Data Model[18], the Marl Ontology[19], the NLP Interchange Format (NIF)[20], and the PROV Data Model[21]. The selection of these vocabularies was based on the following three main objectives: i) relying on stable term identifiers and persistent hosting, ii) being supported by a community, iii) being extensible.

Fig. 2.7 depicts the proposed schema. For representing the main concepts of our conceptual model, we exploit classes and properties of schema.org, a collaborative, community activity with a mission to maintain and promote a common schema for structured data on the Web and beyond. We make use of the class schema:Claim (currently under integration in schema.org) to describe a *claim utterance*. According to schema.org, this class represents a specific, factually-oriented claim. For the *claim proposition*, we use the class schema:Intangible, a utility class that serves as the umbrella for a number of 'intangible' things. Although this class does not sufficiently reflect the semantics of a claim proposition, it appears to be the most reasonable term for representing a proposition. For the same reason, we use schema:Intangible to describe a *claim context*. An alternative solution is to bypass the *claim context* class and directly link an instance of schema:Claim to instances of the four classes connected to the claim context (*author*, *date*, *location*, *event*). These four classes are described through corresponding schema.org classes: schema:Thing (e.g., a person, an organization, a blog, etc.), schema:Date, schema:Place, schema:Event. For connecting a schema:Claim to a schema:Intangible, we can use the property schema:about or its inverse schema:subjectOf.

For representing a *source*, we use the class schema:CreativeWork (or one of its subclasses). Thereby, we take advantage of its properties and can describe additional information about the source, such as headline, language, keywords, publisher, etc. The *linguistic representation* of a claim utterance, as well as the (preferred) *representation*

[17]https://schema.org/
[18]http://www.openannotation.org/
[19]http://www.gsi.dit.upm.es/ontologies/marl/
[20]https://persistence.uni-leipzig.org/nlp2rdf/
[21]https://www.w3.org/TR/prov-dm/

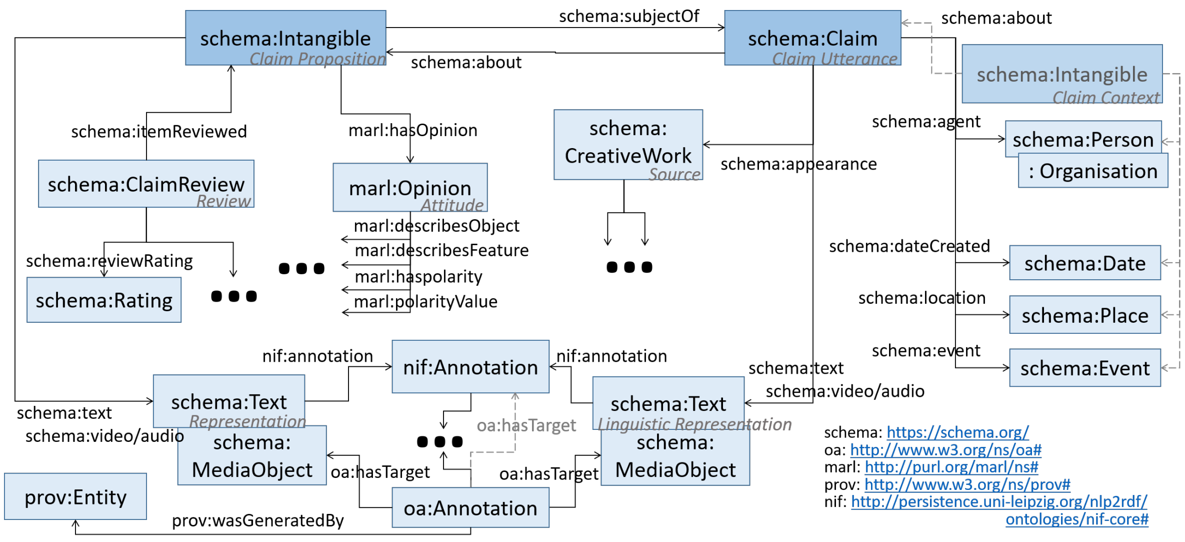


Figure 2.7: RDF/S implementation of the Open Claims Conceptual Model

of a claim proposition, can be described through the class schema:Text (for textual representations) or schema:MediaObject (for image, audio or video representations). For describing *annotations*, we make use of the widely-used OA and NIF data models, while provenance information is represented though the PROV data model. NIF allows us to include detailed information about the outcome of an NLP process on textual representations (like begin/end indexes and confidence scores). The *review* of a claim proposition is described through the class schema:ClaimReview, which in turn is connected to a schema:Rating for assigning a rating score about the veracity of the claim proposition. Finally, we exploit the Marl ontology to represent *attitudes*. Marl is a data schema designed to annotate and describe subjective opinions, and provides the attributes that enable to connect opinions with contextual information.

2.5 Related Knowledge Engineering Tasks

In this section, we review different knowledge engineering and information extraction tasks pertaining to claim related data, like utterances, claim verification scores, claim context information (e.g. who uttered the claim, when and where) and other claim metadata that is described in our Open Claims model. Fig. 2.8 depicts how the below discussed knowledge engineering tasks are mapped to the Open Claims model.

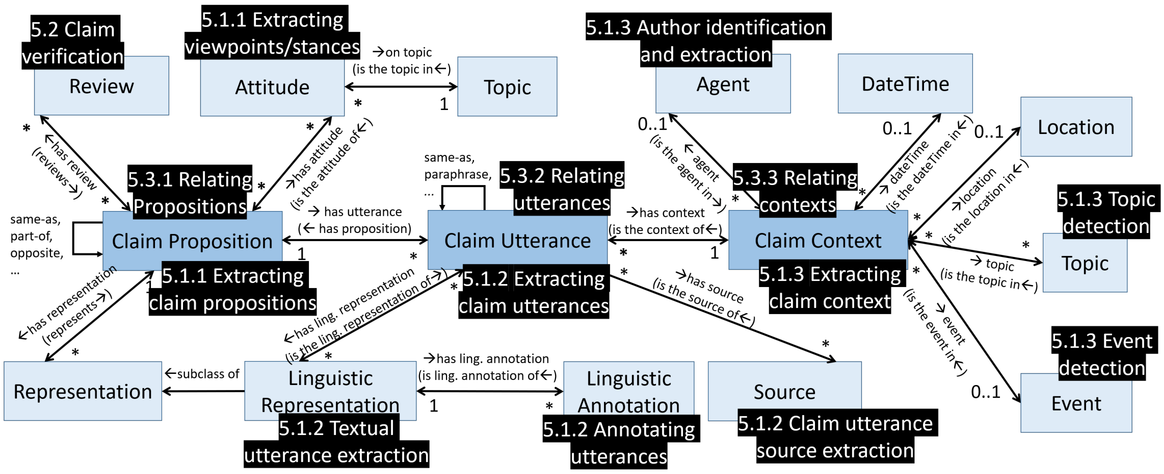


Figure 2.8: The Open Claims model annotated with related knowledge engineering tasks

We identify three main (sometimes overlapping) categories of tasks: *extraction*, *verification*, *interlinking* and position them within the context of our conceptual model. Note that we do not aim to provide an exhaustive overview of those tasks, but rather introduce examples of works of different relevant areas and show how they are positioned with respect to extracting or generating the information and relations suggested by our model. *E*xtraction pertains to detecting statements, utterances and other components and attributes in a corpus of (mainly) textual modality. *V*erification pertains to the assignment of truth ratings or credibility scores to claims or other related components such as information sources. *I*nterlinking, finally, includes a range of tasks that aim at detecting various relations between claims or related components thereof, such as same-as relations, stances or topic-relatedness.

2.5.1 Extraction

Given the complexity and varying definitions of what is or what constitutes a claim, a number of different knowledge extraction approaches can be associated to the tasks in each of the three groups outlined above. We will follow the definition of a claim and its components as given by our model (Sect. 2.4) in order to review the existing techniques for knowledge extraction pertaining to each of these components and attributes. In parallel, we identify challenging problems that are underrepresented in the literature.

Extracting Claim Propositions

The task of extracting a claim proposition can be reformulated as assigning an identifier to a group of statements that are assumed to be semantically equivalent. Our model suggests that the meaning of a claim can be captured both by the means of natural language as well as formal knowledge representation frameworks, e.g. description logics.

Extracting formally represented claim propositions at different levels of formality is of main interest in the field of knowledge extraction, both from unstructured (web pages, social networks) or semi-structured (Wikipedia) sources. Populating and building KBs and thus providing structured knowledge on the Web has been of central interest in the NLP, web, data mining and the semantic web communities over the past decades, focusing on a variety of tasks such as named entity recognition, entity linking, relation extraction or word sense disambiguation. The extensive research in this field has led to a very broad range of works. A comparison of generic information extraction tools and systems is provided by Gangemi (2013), while Martinez-Rodriguez et al. (2020) and Ristoski and Paulheim (2016) focus on semantic web approaches (aiming at the provision of structured knowledge for populating ontologies, linked data and knowledge graphs). The reader may also turn to the book on NLP methods for building the semantic web (Maynard et al., 2016) as well as a recent survey on fact extraction from the web (Weikum et al., 2019).

Relation extraction and ontology learning from text are overviewed by Kumar (2017) and Wong et al. (2012), respectively, while Atefeh and Khreich (2015) dedicate their survey to the task of extracting event-related knowledge. Uren et al. (2006) consider methods that take the inverse approach of annotating documents with entities or statements of facts based on existing knowledge bases. Very closely related to this work is a recent work by Al-Khatib et al. (2020) who extract knowledge encapsulated in arguments to inform a knowledge graph encoding positive and negative effects between concept instances and classifying the consequences as good or bad. For instance, from the claim “Nuclear energy leads to emission decline”, a positive effect of *nuclear energy* on *emission decline* would be extracted and the consequence, *emission* *decline*, rated as *good*. The proposed extraction framework uses a combination of supervised learning and pattern-based approaches.

If we look at textual representations of a claim, the task can be approached by first extracting textual utterances (see below), then grouping them together according to their meaning by the help of textual similarity methods (some of them described in 2.5.3) and then identifying in a cluster of semantically equivalent utterances one that will serve as an identifier for the meaning of the claim. A formal approach to the assignment of textual identifiers to a set of equivalent claims has not been discussed in the literature, to the best of our knowledge, but the task relates closely to the text summarization task, which is surveyed by Lin and Ng (2019).

Extracting viewpoints and stances. Existing computational models (Paul et al., 2010) describe viewpoints via a summarization framework, able to find phrases that best reflect them. In Thonet et al. (2016, 2017), unsupervised topic models are proposed to jointly discover viewpoints, aspects and opinions in text and social media. An unsupervised model for viewpoint detection in online debate forums, proposed in Trabelsi and Zaiane (2018), favors "heterophily" over "homophily" when encoding the nature of the authors' interactions in online debates. With respect to viewpoint detection in social media, the model by Barberá (2015) groups Twitter users along a common ideological dimension based on who they follow. A graph partitioning method that exploits social interactions for the discovery of different user groups (representing distinct viewpoints) discussing about a controversial topic in a social network is proposed in Quraishi et al. (2018), also providing a method to explain the discovered viewpoints by detecting descriptive terms that characterize them.

Our model suggests, in line with the current research, that viewpoints with respect to topics take the form of polarized opinions. Given a controversial topic, for example an issue like *climate change*, viewpoint discovery aims at finding the general viewpoint expressed in a piece of text or supported by a user. This task can indeed be considered a sub-task of opinion mining, which aims to analyze opinionated documents and to infer properties such as subjectivity or polarity. The survey in Pang and Lee (2008) provides a general review of the opinion mining and sentiment analysis tasks. However, for some topics, there may be more than two viewpoints. As of yet, there is limited research that studies these cases.

Viewpoint extraction is closely connected to the *stance detection* problem, a supervised classification problem in NLP where the stance of a piece of text towards a particular *target* is explored. Stance detection has been applied in different contexts, including *social media* (stance of a tweet towards an entity or topic) (Mohammad et al., 2016; Du et al., 2017; Augenstein et al., 2016b; Lai et al., 2017; Sun et al., 2018; Ebrahimi et al., 2016; Xu et al., 2018), *online debates* (stance of a user post or argument/claim towards a controversial topic or statement) (Walker et al., 2012; Sridhar et al., 2015; Bar-Haim et al., 2017a; Guggilla et al., 2016), and *news media* (stance of an article towards a claim) (Pomerleau and Rao, 2017b; Hanselowski et al., 2018a; Bhatt et al., 2018; Wang et al., 2018; Zhan, Q et al., 2019). A recent work by Schiller et al. (2021) details the different and varying task definitions found in previous works, diverging not only with regard to domains, but also classes and number and type of inputs, and introduce a benchmark for stance detection that allows the comparison of models against a variety of heterogeneous datasets. In contrast to the works on viewpoint extraction described previously, works on stance detection focus more on supervised models and textual features (like the sentiment expressed in the text, or the use of polarised words), and less on the structure of the underlying network of users or documents, which can be exploited by unsupervised approaches. For two recent surveys of stance detection works, we refer to Küçük and Can (2020) and Ghosh et al. (2019).

In recent work, Sen et al. (2020) compare untargeted and targeted opinion mining methods (sentiment analysis, aspect-based sentiment analysis, stance detection) to infer approval of political actors in tweets. They show that the compared targeted approaches have low generalizability on unseen and unfamiliar targets and that indirectly expressed stances are hard to detect, and thus identify the need for further research in this area.

Chen et al. (2019b) propose the task of *substantiated perspective discovery* where the goal is to discover a set of perspectives and supporting evidence paragraphs that take a stance to a given input claim, and release a first dataset for this task.

Extracting Claim Utterances

Textual utterance extraction. In this survey, we focus on methods for extracting information from language rather than other modalities such as speech or video. The methods discussed in the literature, with few exceptions, are tailored towards a particular context, topic or type of targeted utterances, usually referred to as *claims* in these works.

Identifying and extracting argumentative components such as claims (also called propositions in these works) or evidence units (also called premises) is a central task in the argumentation mining field (Lawrence and Reed, 2019; Daxenberger et al., 2017). The first survey on the topic by Peldszus and Stede (2013) assumes the availability of an argumentative text and focuses on the problem of analyzing the underlying structure of the presented argument from two perspectives: (1) argument annotation schemes drawing from works in the classical AI field of argumentation and (2) automatic argumentation mining, discussing the first approaches that enhance the historical field with data-centered machine learning approaches. A more recent survey by Lippi and Torroni (2016b) provides a structured view on the existing models, methods, and applications in argumentation mining attempting to draw a single unifying view over a plethora of related sub-tasks and dispersed efforts. The authors define the argumentation mining problem as a pipeline consisting of the detection of argument components in raw text and predicting the structure (or relations) between these components, where the former is of particular interest to the task that we consider in this section. Building on and completing these surveys, Cabrio and Villata (2018) adopt a data-driven perspective of the existing work in argumentation mining with a focus on applications, algorithms, features, and resources for evaluation of state of the art systems. Taking also a data-driven perspective, the difficulty of devising cross-domain claim identification approaches has been discussed and analyzed in (Daxenberger et al., 2017) by using multiple domain-specific data sets. In that, the authors address the analysis of the generalization properties of systems and features across heterogeneous domains and study their robustness across the underlying fields. Shnarch et al. (2018) propose a methodology to combine smaller amounts of high quality labeled data with noisy weakly labeled data to train neural networks for extracting evidence units for given topics.

The extraction of a claim is the first step in a computational fact-checking pipeline, where it is common to see fact verification as a three-step process: (i) detecting/extracting a check-worthy claim, (ii) reviewing the claim with respect to its veracity and (iii) publishing the reviewed claim (Hassan et al., 2017c; Thorne and Vlachos, 2018).[22] In Hassan et al. (2017c), the authors propose a first version of the ClaimBuster tool with a particular focus on the extraction of check-worthy claims. The claim-spotting problem is defined as a two step task, comprising (1) classification of pieces of text as check-worthy or not and (2) their ranking with respect to their check-worthiness. An end-to-end fact-checking platform, including both steps (1) and (2) is presented in a follow up work (Hassan et al., 2017a). To overpass the limitations of using hand-crafted features for claim detection, Hansen et al. (2019) propose a neural check-worthiness ranking model that represents a claim as a set of features, where each word is accounted for by its embedding (capturing its semantics) and its syntactic dependencies (capturing its relation to other terms in the sentence). The extraction of simple claims about statistical properties to be subjected to verification is addressed in Vlachos and Riedel (2015). The authors apply a distantly supervised claim identification approach that relies on approximate matching between numerical values in text and a knowledge base (Freebase). A relevant line of work has been followed in the field of subjectivity analysis of text, proposing approaches which aim at classifying sentences into objective and subjective categories, e.g., (Biyani et al., 2014; Wiebe and Riloff, 2005; Yu and Hatzivassiloglou, 2003). It has been shown in Hassan et al. (2017a) that subjectivity identifiers are limited in discerning factual claims as compared to the method presented by ClaimsBuster.

Annotating Utterances. In our model, we discuss an annotation of utterances based on (1) entities (such as names, dates, locations, etc.) and (2) lower-level linguistic features extracted from the text that can be useful for a number of tasks, such as bias detection or fake-news analysis, as discussed in Rashkin et al. (2017). For (1), one can turn to the literature surrounding (end-to-end) Entity Linking[23], particularly the exhaustive survey in Sevgili et al. (2020). The features in (2) include characteristics of the discourse, such as shades of irony or the overall polarity score of the expression, as well as linguistic or syntactic cues (part-of-speech (POS) tags, syntax, dependencies,

[22] https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

[23] http://nlpprogress.com/english/entity_linking.html

semantic parsing, punctuation or capitalization) that can be indicative of a certain intention. For the identification of such cues, one could turn to NLP annotation pipelines (with standardized annotation type taxonomies). The industrial standard is UIMA (Unstructured Information Management Applications) (Ferrucci and Lally, 2004), a comprehensive meta-framework for inter-operable linguistic annotation. Recent developments in deep approaches to NLP have led to the development of ad-hoc annotation models such as SpaCy.[24]

Claim utterance source extraction. Sources are identified as the media that publishes a claim. Their extraction can be straightforward in many cases (e.g. when the utterance itself is extracted directly from its source). In certain cases identifying the original source may be more challenging and would require tracking down the claim to its original publication by, e.g. following cascades of retweets or identifying and analysing quotations (Starbird et al., 2018; Newell et al., 2018a; Niculae et al., 2015; Vosoughi et al., 2018a).

Extracting Claim Context

This group of approaches deals with annotating a claim with contextual information that helps reply to the questions who uttered the claim when and where. In order to extract a date or a location one can rely on Entity Linking (EL) or Named Entity Recognition (NER) techniques outlined in the previous section. We focus in more detail on the tasks of event detection, topic detection, and author identification and attribution.

Event detection. The event in which a claim was uttered is an important component from the context that defines a claim. An event can be seen as a complex entity defined by a set of attributes, such as a date, persons involved and a location. Following this definition, one can apply the methods described in the previous paragraph in order to extract independently these components to populate an event. However,

[24] `https://spacy.io/`

recent approaches consider an event as an atomic entity that can be detected from web corpora (often social networks) (Hasan et al., 2018; Chen et al., 2019c).

Topic detection. Detecting what claims are about is a challenging issue. If available, context such as the source articles the claim was extracted from, a claim review article, or the discourse the utterance was embedded in, e.g. the given subject in a debating portal, can be considered for claim topic detection. Here standard NLP methods of topic extraction, modeling or detection from text can be employed (Martinez-Rodriguez et al., 2020). However, detecting the topic when only the textual content of a claim utterance can be considered, or when the textual context is sparse, is challenging.

Approaches developed for extracting topics from short text (like tweets and micro-blogs) can be adapted for claim topic modeling (Sriram et al., 2010). However, the complex structure and positioning in a context of elements (such as sources, authors and other entities) has to be taken into consideration when predicting topics of claims. Topics can be seen as groups of equivalent claims (e.g. all claims pertaining to "US immigration policies") situated in a network of contextual entities (e.g. a knowledge graph such as the one given in our model implementation example in Fig. 2.7). Therefore, link prediction methods on knowledge graphs may be used, where a recent work by Beretta et al. (2020) studies the effectiveness of neural graph embedding features for claim topic prediction as well as their complementary with text embeddings. The authors show, however, that state-of-the-art link prediction models fail to capture equivalence structures and transfer poorly to downstream tasks such as claim topic prediction, which may, however, also be connected to the lack of sufficiently large and reliable ground truth data (topic-labeled claims) that would allow to train neural embedding models. This calls for the development of novel methods that surpass the state-of-the-art graph embedding model's reliance on a local link prediction objective, which likely limits the ability of these models to capture more complex relationships (e.g. equivalence cliques between claims, keywords and topic concepts) and the generation of suitable ground truth data.

Author identification and extraction. Identifying the author of an utterance is not trivial (Mevan Babakar and Will Moy, 2016) yet authorship is crucial for interpreting its meaning. Moreover, claims are often quoted by distant sources, e.g. in news articles or other media. The attribution of content to an author[25] is consequently gaining increased attention in the context of the analysis of news articles, e.g. by Newell et al. (2018a); Salway et al. (2017) who build structured databases of claims with extracted quotes and author information. Approaches for quotation extraction and attribution from newspaper articles for both direct and indirect speech usually comprise three different component identification steps: (1) cue phrases signalling the presence of a quotation (e.g. "say" or "criticize") are identified using manually curated word lists (Krestel et al., 2008) or classifiers trained on labelled data (Pareti, 2015; Scheible et al., 2016; Newell et al., 2018a). On this basis, (2) quotation content spans are identified using manually defined syntactic rules (Krestel et al., 2008), conditional random fields (CRFs) (Pareti, 2015; Newell et al., 2018a) or semi-markov models (Scheible et al., 2016). Finally, (3) author entities are identified, typically using sequence models such as CRFs (O'Keefe et al., 2012; Pareti, 2015; Newell et al., 2018a). In that, Newell et al. (2018a) extend O'Keefe et al. (2012)'s sequence-based quote attribution to a two-stage approach using maximum entropy classifiers for connecting cue and content spans and cue and author spans, respectively, allowing multiple content and cue spans to take part in an attribution relation. A different approach is followed by Pavllo et al. (2018) who employ pattern-based bootstrapping to extract quotation-speaker pairs. A recent paper by Jiang et al. (2020) extracts structured information from fact-checking articles, including the "claimant". This corresponds to either the source or the author of the claim, depending on which of those is mentioned in the fact-checks where usually, this distinction is not made.

A fine-grained distinction between sources types and author types (*medium-type sources* and *agent-type sources* in their terminology) is proposed by Schulze and Boland (2019). Their category scheme is based on a literature review of journalism research that analyzes different source types in news media texts and on manual annotations of news articles. The scheme is hierarchical to facilitate automatic classification and ensure its utility for a wide range of use-cases such as the analysis of biases and credibility of news articles. In preliminary experiments, they can show that the automatic classification of source types on the highest level of the hierarchy is feasible with small amounts of annotated data. However, they find the distribution of source

[25] Coined "source" in the respective works; in order to not confuse different terminologies, we are referring to these entities as "authors" in the following text although this diverges from the naming used in the literature in this field.

types to be highly skewed in their sample, which poses challenges for automatic classification of the less frequent types.

2.5.2 Claim Verification

A number of terms, such as fact-checking, truth discovery, claim or fact verification pertain to a large degree to the process of the automatic assignment of a veracity score to a statement uttered by a particular person or a group of people (Thorne and Vlachos, 2018). Note that the analysis of false or mis-information spread, or fake-news detection,[26] defined and surveyed in Sharma et al. (2019), often deal with entire news articles or outlets and are, therefore, broader problems where claim verification can be seen as one of their ingredients.

Claim truthfulness verification is reviewed in Cazalens et al. (2018); Thorne and Vlachos (2018), where (Thorne and Vlachos, 2018) in particular propose to unify diverging definitions of the task and its components from various disciplines, such as NLP, machine learning, knowledge representation, databases, and journalism. Indeed, most of the existing techniques rely on background knowledge sources (e.g. encyclopedic knowledge graphs, such as DBpedia or Freebase) that provide a "truthfulness context" (Hassan et al., 2017c; Thorne and Vlachos, 2017) and a combination of various computational methods in order to infer the veracity scores of a claim either from those background knowledge sources or, more rarely, in a self-dependent manner. In addition, versatile features pertaining to all three main components of our model (meaning, utterance and context) are often considered in a combined manner, making it difficult to break down claim verification approaches along each of these three axes independently.

In certain cases, claims are given a structured form (e.g. triples or database queries), which allows for the verification of entity-centric information by calling on machine learning techniques (Yu et al., 2018). In that, fact verification can be seen as a particular kind of a link prediction or knowledge base augmentation task (Ciampaglia et al., 2015; Shi and Weninger, 2016). In contrast, certain methods apply symbolic inference

[26] "False and often sensational information disseminated under the guise of news reporting", according to Collins English Dictionary.

approaches on KGs in order to infer the truth value of a statement (Beretta et al., 2018), or to identify potential errors (Galárraga et al., 2013). A multitude of features, machine learning models and inference techniques are combined together in the KB construction approach presented in (Dong et al., 2014).

In other cases, statements are taken in their textual form (Popat et al., 2017; Wang, 2017), while again largely machine learning techniques are applied in order to assess their veracity. Training data in the form of examples of true and false claims come either from archives of fact-checked statements (Vlachos and Riedel, 2014; Wang, 2017; Barrón-Cedeño et al., 2018) or from manually labelled (often crowdsourced) collections of claims (Gorrell et al., 2019; Mihaylova et al., 2018; Thorne et al., 2018a). Statistical (topic) models as well as standard NLP filters are used in order to construct a feature space. Note that the majority of approaches based on machine learning rely primarily on highly contextualized features on document/text level, such as words, n-grams, salient entities and topics (Hassan et al., 2015). Additional context- and aspect-related features such as provenance, time and sources are considered in Popat et al. (2017); Vlachos and Riedel (2014). An analysis of news corpora is provided by Rashkin et al. (2017) in an attempt to identify linguistic and stylistic cues that help discriminate facts from false information. In addition, certain approaches, like (Yin et al., 2008), look at how a claim spreads through a crowd or how sources and claims are connected, exploiting social/community-related features.

2.5.3 Interlinking

There exist a variety of types of relations between claims and in particular between their components as introduced in our conceptual model. We consider that the problem of claim relatedness depends on the particular perspective and application context—for example, two claims can be considered contextually similar because they have been uttered at the same event by the same person, but still differ in their meaning and textual expression. Following the main building components of our model, we identify a number of dimensions on which this problem can be studied. One could be interested in relating instances of propositions, utterances or contexts within each of these three groups. These are the kind of relations that will be discussed in this section. Else, one can look into cross-class relations (e.g. establishing the association between an utterance and its author or viewpoint). Such relations result from knowledge extraction processes already discussed in Sect. 2.5.1. Although most of these problems can be considered as challenging with little existing work that approach them directly, we will outline below relevant works.

Relating Propositions

According to our model, the proposition, or meaning, of a claim is materialized via a particular representation (e.g. a natural language or a logical expression) and is further described by its topics to which we associate viewpoints. As discussed in Sect. 2.5.1, different extraction methods can be applied in order to derive those representations. Independently from the particular representation type, we outline three general types of relations that we can establish between proposition instances: *e*quivalence (same-as), similarity and *r*elatedness.

*S*ame-as. The equivalence or identity relation binds together claims that have the exact same meaning. In the case of textual expression of the meaning of a claim, when two propositions are expressed differently although they convey the same message (have the same meaning), we talk of a relation of paraphrase. Paraphrasing detection allows to discover equivalent text fragments that differ (to a given extent), where neural language models are currently largely applied to the task (Liu et al., 2019; Zhang et al., 2019b). In the case of a symbolic or formal expression of a claim (or a fact), we outline works on relation alignment, such as Pereira Nunes et al. (2013).

*S*imilar. Two propositions can be similar to a given degree on a scale between "identical" (represented by the *s*ame-as relation) and "dissimilar". This notion relates to that of semantic similarity discussed, for example, in Gracia and Mena (2008) and tackled in the Semantic Textual Similarity task (Cer et al., 2017; Agirre et al., 2016). A first systematic study on finding similar claims is proposed by Dumani and Schenkel (2019).

*R*elated. Relatedness, as opposed to similarity, covers "any kind of lexical or functional association" (Gracia and Mena, 2008) and is, hence, a more general concept than semantic similarity. Relatedness encloses various relationships, such as meronymy (a relation of composition (part-of) that is such that the meaning of a complex expression relates and can be expressed by the meanings of the parts from which it is constructed), antonymy (opposite meanings, including conflicting / contradicting claims), logical or textual entailment (Dagan et al., 2006), same topic, or any kind of functional relationships or relationships or frequent association. A survey of semantic relatedness methods, evaluation and datasets is given by Zhang et al. (2013); Hadj Taieb et al. (2020). As opposed to logical entailment, textual entailment is understood as a relationship between pairs of text fragments where one entails the other if a human reading the former would be able to infer that the latter holds.

Relating Utterances

Several works address finding equivalent claims in the context of claim verification (Hassan et al., 2017a,c; Maliaroudakis et al., 2021b), where a claim matcher (or linker) is a component of a fact-checking system matching new claims to claims that have already been checked. Shaar et al. (2020a) recently proposed the task of detecting previously fact-checked claims defined as ranking a set of verified claims according to their potential to help verify an input claim. They propose a learning-to-rank approach and release a first dataset for the task. Clustering similar arguments is at the core of the work by Reimers et al. (2019) who use contextualized word embeddings to classify arguments as pro or con and identify arguments that address the same aspect of a topic.

Concerning the matching of text fragments more generally, recent advances in neural NLP and the advent of deep contextualized language models for language understanding, have allowed a renewal state-of-the-art techniques for matching text fragments through the pooling or aggregation of classical (Pagliardini et al., 2018) and contextualized word-embeddings (Liu et al., 2018; Howard and Ruder, 2018) into phrase, sentence or document embeddings (Akbik et al., 2019, 2018) and the computation of distance metrics to find the closest matching utterances.

In the context of the *Open Claims* model, relations between utterances can further be derived from the relations between their constituents. For example, an utterance is a repetition of another utterance when all constituents are equal except for at least one attribute of the context such as the author or the date. An utterance is a paraphrase of another utterance when the propositions are equal but the (linguistic) representations differ.

In general, claim utterances can be equal/similar or dissimilar w.r.t. one or multiple components pertaining to meaning, linguistic realization and context. The precise relation between utterances can be expressed by detailing the combination of similarity judgements along these different dimensions as exemplified by Table **??**. Utterances can be classified as duplicates when all of their components match. Paraphrases share all components but the linguistic realization, which differs and can be dissimilar regarding textual similarity measures. Repetitions are utterances made at different points in time or by different spearkers or at different locations. Relations can be mixed, as illustrated by *repetition_paraphrased*, where two utterances expressing the same meaning can be made in different forms and at different occasions and/or by different speakers. Utterances about the same or similar topics, but expressing otherwise different things, can be detected by comparing the relations between the topic fields.

Other types of relations comprise support/attack relations or pro/con stances. Many works treat this as an extraction and classification task, e.g. classifying an argumentative unit as *evidence* (see Sect. 2.5.1), while others treat this as an argumentative relation extraction task, e.g. relating two units with a *supports* or *attacks* relation (Carstens and Toni, 2015; Opitz and Frank, 2019; Nguyen and Litman, 2016).

Combining all these relations allows to express precise relations between utterances, e.g. two claims may be about the same topic but express opposing stances. Representing more fine-grained relations between claims and their contexts has been shown to be important for a number of use-cases, for instance, to specify the relation between claim and evidence for scientific claim verification when they express different levels of specificity or when there is conflicting evidence for a claim (Wadden et al., 2022) and for retrieving verified claims for utterances that exhibit a part-of relation to one another (Shaar et al., 2022). Using this model allows to go beyond a fuzzy concept of claim similarity or relatedness that conflates many different dimensions.

Relation	same	different / dissimilar	different / similar
duplicate	meaning linguistic realization context.topic context.speaker context.date context.location		
duplicate_ paraphrased	meaning context.topic context.speaker context.date context.location	linguistic realization	
repetition	meaning linguistic realization context.topic	(at least one of:) context.speaker context.date context.location	
repetition_ paraphrased	meaning context.topic	linguistic realization (at least one of:) context.speaker context.date context.location	
same_ topic	context.topic	meaning linguistic realization	
similar_ topic		meaning linguistic realization	context.topic (similarity w.r.t. topic semantics or granularity)
entails_ textual		linguistic realization context.topic	meaning (similarity w.r.t. textual entailment def.)
entails_ logical		linguistic realization context.topic	meaning (similarity w.r.t. logical entailment def.)
homonym	linguistic realization	meaning (at least one of:) context.speaker context.date context.location	
near_ homonym		meaning (at least one of:) context.speaker context.date context.location	linguistic realization

Relating contexts

A context is broken down to its constituents: events, entities, dates, etc. Establishing links among contexts comes down to linking their respective components. For that purpose, one may call upon state-of-the-art approaches to data linking, where, following years of research and practice, a wealth of methodological approaches and tools are currently out there (Nentwig et al., 2016). Among those, property-centric approaches (e.g. (Jentzsch et al., 2010; Ngomo and Auer, 2011)) can be of particular interest in order to establish relations (like identity or overlap) between different contexts, comparing their elements individually by the help of well-suited similarity measures (e.g. measures of similarity between proper names or dates).

2.6 Conclusion

This paper bridges the gap between various disciplines involved with online discourse analysis from a range of perspectives by (a) surveying definitions of claims, facts and related concepts across different research areas and communities, (b) establishing a shared conceptualisation and vocabulary in this context and (c) discussing a range of tasks involved with such notions, for instance, for extracting or interlinking related concepts through NLP techniques. We contribute to a shared understanding of a wide range of disparate yet strongly related research areas, facilitating a deeper understanding of shared methods, approaches and concepts and the potential for reuse and cross-fertilisation across communities. Below, we highlight under-researched areas and potential future directions.

Currently, a framework for claim relatedness and similarity is missing. Several works from different fields appear to deal with the problem from different perspectives, but an approach that takes into consideration the various aspects of a claim, as well as its various representations, as defined in our *O*pen Claims model in order to discover claim relatedness or similarity of different kinds is yet to be proposed. While there are works addressing the extraction of structured information from claims (Jiang et al., 2020), allowing for example the detection of nested claims, current fact-checking methods and sites largely ignore such issues. For instance, a complex claim can have different (or no) truth rating as compared to its constituents. For instance, "Colin Kaepernick says Winston Churchill said, "A lie gets halfway around the world before the truth has a chance to get its pants on.""[27] The claim that Kaepernick uttered this is true, while the claim within the claim is false since Churchill never said that. Using the proposed model, such cases can be modeled precisely and unambiguously.

Given the subtle differences between claims, where meaning often derives from subtext and context, disambiguating claim representations, e.g. when mapping novel claims to knowledge bases of fact-checked claims, appears challenging. Even for humans, deciding on the type of relationship of two claims is a non-trivial task. For example, the claim "Interest on the debt will exceed defense spending by 2022" provides an exact date, while "Interest on debt will exceed security spending" does not provide a date. Can these two claims be considered the same, and, if not, what is their relation? Using the proposed model, such subtle differences can be made explicit. In addition, automated fact-checking has the potential to elevate the problem given its lack of maturity at different steps, where for instance, the classification of half-correct or poorly disambiguated claims as correct may introduce further false claims into the wild.

Similarly, the process of stance detection is challenging as it has been shown to not work well for the minority class, i.e. documents disagreeing with the claim (Hanselowski et al., 2018a; Roy et al., 2021), and for unseen targets (Sen et al., 2020). Little research in viewpoint discovery deals with extracting viewpoints for more than two polarized positions, a topic that could be worthwhile researching for the analysis of debates.

Detecting claim topics and linking those to a specific commonly shared vocabulary or thesaurus of topics (like, e.g., the TheSoZ (Zapilko et al., 2013) or the Unesco[28] thesauri) appears to be a difficult and under-researched topic that promises to enhance claim retrieval, improve search and interoperability across sources, and facilitate access to currently existing or yet to be constructed structured resources of claims (Tchechmedjiev et al., 2019b).

[27] `https://www.politifact.com/factchecks/2017/oct/09/colin-kaepernick/nfls-colin-kaepernick-incorrectly-credits-winston-/`

[28] `http://vocabularies.unesco.org/thesaurus`

Generally speaking, considering the wide variety of methods and datasets involving claims and related notions, adopting a shared and well-defined vocabulary has the potential to significantly increase impact and reuse of research methods and data.

Chapter 3

Verified Claim Retrieval

As detailed in the previous chapter, the publication of structured data about claims allows the uncovering of explicit and implicit relations between claims and related entities (Gasquet et al., 2019). One task that has gained significant relevance is that of relating utterances as a first step in computer-aided fact-checking: with a growing number of fact-checking articles available online, the load of fact-checkers can be minimized by linking check-worthy claims to previously fact-checked ones. By providing access to a database of verified claims from multiple fact-checking portals and offering a unified structure including normalized truth value ratings, ClaimsKG (Tchechmedjiev et al., 2019b; Gangopadhyay et al., 2023) facilitates these efforts. The ClaimLinker web application (Maliaroudakis et al., 2021a) offers a lightweight tool that accepts arbitrary input texts and retrieves matching verified claims in ClaimsKG to support fact-checking for journalists and other users. After its release, the task of detecting previously fact-checked claims, also coined verified claim retrieval, continued to be actively researched, spawning many different methods and research prototypes, yielding new high scores in the respective leaderboards every year. Their utility for online claim retrieval applications, however, which requires the analysis of additional properties such as efficiency and robustness, remains unclear. While large language models have been shown to be successful even in few-shot settings (Brown et al., 2020), their environmental and computational costs are high and may not be worth their performance gain, especially since similar or even better results can often be achieved by smaller, more specialized and less costly models (Bender et al., 2021). In the same vain, despite its impressive performance across a range of different tasks and its general-purpose utility, ChatGPT has been found to be outperformed by comparably simpler models for many tasks (Qin et al., 2023; Bang et al., 2023). In this chapter, we analyze the state of the art in verified claim retrieval with regard to their applicability in real-world claim retrieval applications and investigate the need for large language models and costly fine-tuning operations for this task.

Robust and Efficient Claim Retrieval for Online Fact-Checking Applications

Understanding the veracity of statements is important when consuming information on the Web. Whereas fact-checking sites have provided a large corpus of already verified claims, matching a given utterance to already fact-checked claims remains a challenging task. Verified claim retrieval has been approached through a variety of different methods, among them approaches relying on supervised neural models. Whereas such models tend to perform strongly, they require significant training effort and their robustness towards unseen data distributions may vary heavily. Also, prior works demonstrate the capability of unsupervised models to provide state-of-the-art performance. In this paper, we assess established claim retrieval benchmark datasets and experimentally evaluate and compare different state-of-the-art supervised and unsupervised methods with regard to performance, but also computational effort and run time. We show that unsupervised approaches outperform supervised ones with respect to robustness. While the best state-of-the-art method relies on supervised deep neural networks, its high computational costs make it difficult to use in online fact-checking applications. The best unsupervised method reaches a similar performance and meets efficiency requirements of online application scenarios due to low hardware requirements. Our experiments verify that, due to the nature of the task and data, the choice of pre-trained language models is more important than fine-tuning and that training supervised models on the target data may not be cost-efficient in online claim retrieval applications.

3.1 Introduction

The spread of controversies, biased discourse, and falsehoods on the Web has become an increasingly important issue, from both a societal as well as a research perspective Vosoughi et al. (2018b); Allcott and Gentzkow (2017). False information has been shown to spread rapidly, and even faster than the truth Vosoughi et al. (2018b). Checking the veracity of information before it can spread is thus crucial. However, fact-checking is costly, requiring human expertise and time. As of now, over 100 fact-checking initiatives exist worldwide.[1] As manually fact-checking a single claim may take more than a week Shaar et al. (2022), one of the initial steps for fact-checkers is to find out if a claim has already been verified by another organization.

However, given that a particular claim proposition may occur in the form of diverse utterances Boland et al. (2022b), matching a given statement or utterance to fact-checked claims available from fact-checking portals remains a challenging problem. This problem is known as verified claim retrieval and has been recognised by the *Check That!* initiative Barrón-Cedeño et al. (2020); Shaar et al. (2021); Nakov et al. (2022), that has advanced research in this field by producing benchmarks and baselines and organizing shared tasks.

A variety of approaches has since been proposed. However, those were evaluated on different individual datasets, making it hard to compare their performance, their strengths, and weaknesses, and assess their robustness. Generally, while supervised approaches usually dominate regarding task performance, they tend to overfit the data distributions prevalent in the training corpus at hand, making them less robust on new data and evolved data distributions. Also, they are prone to picking up spurious correlations in the data, causing them to perform well on a given dataset rather than performing well on the task Le Bras et al. (2020). It is thus crucial to understand the characteristics of benchmark datasets, to compare the performance of methods across different datasets, and to understand how well the data reflects the task in the real-world.

Another important aspect that has so far achieved little attention is the question whether proposed approaches are actually useful in practice Ethayarajh and Juraf-

[1] `https://ifcncodeofprinciples.poynter.org/signatories`

sky (2020), in this case, to link unverified statements to a database of fact-checked claims, such as ClaimsKG Tchechmedjiev et al. (2019c); Gangopadhyay et al. (2023), as proposed by Maliaroudakis et al. (2021a) with their ClaimLinker system. This means that beyond the quality of predictions, method performance with respect to run time and computational costs needs to be understood. This involves both efficiency during inference as well as offline computation costs during training, especially since dynamically emerging and evolving vocabulary has proven a significant challenge, requiring frequent updates of models to ensure stable performance over time Amba Hombaiah et al. (2021). This does not only pose challenges for practical reasons, but also leads to high energy consumption, exacerbating environmental issues Strubell et al. (2019), especially when unnecessarily large language models are applied Bender et al. (2021).

With this work, we investigate the state-of-the-art in the retrieval of previously fact-checked claims, assessing used benchmark datasets and the performance of claim retrieval models. In particular, we aim at understanding whether the performance gains on particular datasets justify the required effort and computational and environmental costs for training and fine-tuning large supervised models.

We also compare the efficiency of approaches and their scalability. Given the growing amount of previously fact-checked claims and the importance of fast response times in real-world fact-checking scenarios, fast run time at inference is a crucial criterion.

To investigate these questions, we (a) assess established benchmark datasets for the retrieval of previously fact-checked claims to aid the interpretation of evaluation results, (b) compare 22 (15 supervised and 7 unsupervised) state-of-the-art methods with respect to performance, and (c) replicate and newly generate additional results on more datasets for 8 of these systems (4 supervised and 4 unsupervised) to compare them with respect to overall performance, robustness, scalability and run time.

We find that fine-tuning and supervision can indeed increase performance at the cost of risking overfitting. For instance, the overall best approach, *RIET Lab*, is supervised and relies on large language models. It scores between 0.008 and 0.039 points higher in MAP@5 on all datasets but one than the overall second best approach, unsupervised *SimBa 2023* which uses considerably smaller pre-trained models. However, *SimBa 2023* scores 0.108 points higher on the remaining dataset. The choice of pre-trained models and embeddings turns out to be more crucial than fine-tuning and the high computational costs for supervision may not be worth the performance gain for real applications, given that the performance difference between the best supervised approach and the best unsupervised approach averages to -0,07 points in MAP@5 over all datasets while the unsupervised approach achieves lower retrieval times on a laptop than the supervised approach on a high-performance GPU server, which usually is not available for online fact-checking applications.

The main contributions of our work are:

- an assessment of established benchmark datasets for the retrieval of previously fact-checked claims
- a comparative evaluation of different supervised and unsupervised approaches on established datasets for the claim retrieval task and extensive experimental results w.r.t. robustness of the approaches comparing the performances when training and testing on claims of different discourse types
- the assessment of the applicability of different approaches for real-world claim retrieval applications based on performance, efficiency and computational expenses
- modifications of existing approaches to allow cross-evaluation and application on different datasets. All modifications and all scripts to replicate this study are released to the community.

The paper is structured as follows. We start with the problem definition and introduction of our research questions (Section 3.2), before giving an overview of related work and the state of the art (Section 3.3). Section 3.4 details the compared claim retrieval approaches and our experimental setup. Our data analysis follows in Section 3.5 before we present and discuss the results of our comparative evaluation in Section 4.6. Section 3.7 concludes the paper with a discussion of future work.

The source code and detailed documentation and instructions to reproduce our results are available on GitLab[2]

[2] https://git.gesis.org/bolandka/claimlinking

3.2 Problem Definition and Research Questions

From an application perspective, the task of retrieving previously fact-checked claims represents the initial step of fact-checking, e.g. in journalistic or news contexts, aiming at identifying if a given statement has been already fact-checked by trusted sources and finding fact-checked statements that are relevant to determining the veracity of the input claim.

More formally, given an input query q (e.g. the text of a tweet, or a sentence in a political debate) and a dataset D of previously fact-checked verified claims (with each verified claim denoted as c), the *claim retrieval* task aims at providing a ranked list of top-n claims C from D. Note here that a retrieved claim c consists of not only the text of the utterance but also additional metadata such as the title of the fact-checking article.

For example, consider the sentence from a political debate as input query: *"You know, interest on debt will soon exceed security spending."*

Table 3.1 displays the top-3 list of previously fact-checked claims provided by the ClaimLinker Maliaroudakis et al. (2021a) system.

In this context, the size of the dataset D (number of previously fact-checked claims) is important because it can affect both the retrieval effectiveness and the retrieval efficiency. In particular, a larger dataset might mean more semantically similar claims (which can make retrieval more difficult and affect precision) and higher response times, especially for supervised methods that often require pair-wise comparisons. Operating on a small dataset might lead to more unmatchable input claims, reducing a system's recall.

In this paper, we investigate the following research questions:

1. RQ1: What are the characteristics of the task of retrieving previously fact-checked claims and existing benchmark datasets?

2. RQ2: Which supervised and unsupervised state-of-the-art approaches perform best on these established datasets and why?

c **text**	c **speaker**	c **date**	c **source**	c **verdict**
Within a few years, we will be spending more on interest payments than on national security	Mitch Daniels	11 February 2011	politifact.com	false
In just 17 years, spending for Social Security, federal health care and interest on the debt will exceed all tax revenue	Dave Brat	29 May 2015	politifact.com	mostly true
By 2022, just the interest payment on our debt will be greater than the defense of our country	Joe Manchin	9 May 2018	politifact.com	mostly true

Table 3.1: Top-3 verified claims retrieved by the ClaimLinker system for the input query *"You know, interest on debt will soon exceed security spending."*

3. RQ3: How robust is their performance depending on the training data?
4. RQ4: How efficient are the different approaches and how well do they scale depending on the number of verified claims in the dataset?
5. RQ5: Which approach is the best choice for (online) claim retrieval applications?

3.3 Related Work

In the following, we provide an overview of existing claim retrieval tasks, datasets, approaches and studies giving insights regarding dataset characteristics that are relevant to this work.

3.3.1 Claim retrieval tasks and datasets

A number of shared tasks and datasets have been proposed that evolve around fact-checking and the linking of input claims to information that can aid with fact-checking.

FEVER Thorne et al. (2018d), for Fact Extraction and VERification, is a shared task focusing on checking factual claims. It consists of the retrieval of evidence from Wikipedia and using it to verify the claims. The FEVER dataset Thorne et al. (2018b) consists of 185,445 claims generated by modifying Wikipedia excerpts that are subsequently verified independently from the original part of the text from which they are derived. The contained claims and their corresponding evidence (sentences that support or refute them) are categorized into three classes: Supported, Refuted, or NotEnoughInfo.

Also dealing with the problem of claim verification and the categories Supports, Refutes, NoInfo/NotEnoughInfo, the dataset SciFact Wadden et al. (2020) and its recent extension SciFact Open Wadden et al. (2022) focus on claims and evidence from the scientific domain. Claims are created by human experts based on citing sentences. Corresponding evidence is drawn from the abstracts of the cited publications. SciFact contains 1.4k claims paired with annotated abstracts from a selection of the Semantic Scholar Open Research Corpus S2ORC Lo et al. (2020) containing corresponding evidence. SciFact Open retrieves additional abstracts and evidence annotations for 279 of these claims, arriving at 500k research abstracts, 279 claims, and 460 claim-evidence pairs in the corpus. Automated methods are used to filter potentially relevant abstracts contained in the full S2ORC corpus for manual annotations.

The Fake News Challenge (FNC) Pomerleau and Rao (2017a) is about classifying the stance of news articles to claims (agrees, disagrees, discusses, unrelated). The benchmark data it provides consists of 75,385 labeled headline and article pairs, with the news article headlines serving as claims Hanselowski et al. (2018b).

While these are related tasks, the linked evidence and articles cannot be used to directly infer veracity labels for a claim as is expected for verified claim retrieval. Rather, the contained evidence has to be reviewed and assessed by experts.

The MultiFC Augenstein et al. (2019) and the ClaimsKG Tchechmedjiev et al. (2019c); Gangopadhyay et al. (2023) datasets both provide structured data of and about claims coming from reputed fact-checking portals, where claims are stored together with rich metadata (such as authors, sources, claim reviews and other contextual information, including veracity labels). While both datasets retrieve and store claims from fact-checking portals, they are complementary in a couple of aspects. MultiFC focuses on evidence-based fact-checking in terms of downstream tasks, where via the Google Search API the 10 most highly ranked search results per claim are retrieved and stored. ClaimsKG, on the other hand, provides a rich data model (an RDFS ontology) to represent check-worthy (or fact-checked) claims and related metadata, which is an important effort towards standardization and enables federated access to distributed data, where a specific search engine is provided[3] in addition to a public Sparql endpoint Gasquet et al. (2019). MultiFC contains data in English, ClaimsKG is multilingual, harvesting data from fact-checking portals in English, French, Russian, Spanish, Italian, and Hindi. These datasets can be used to provide a pool of verified claims with additional metadata for fact-checking applications and to extract links to claims that are mentioned in fact-checking articles.

One initiative that directly focuses on the linking of unverified claims to previously fact-checked claims is the CheckThat! Lab Barrón-Cedeño et al. (2020); Shaar et al. (2021); Nakov et al. (2022). From 2020 until 2022, its second task has been part of the CLEF campaign, providing a forum for development and evaluation of systems for the retrieval of verified claims for tweets and, since 2021, statements from political debates. Here, the claim retrieval task is defined as *Given a check-worthy input claim and a set of verified claims, rank the previously verified claims in order of usefulness to fact-check the input claim*Shaar et al. (2021); Nakov et al. (2022) and *Given a check-worthy input claim and a set of verified claims, rank those verified claims, so that the claims that can help verify the input claim, or a sub-claim in it, are ranked above any claim that is not helpful to verify the input claim* Barrón-Cedeño et al. (2020); Shaar et al. (2020b). In 2020, the task has been coined *Verified Claim Retrieval*. In 2021 and 2022, it has been referred to as *Detecting Previously Fact-Checked Claims*. Both titles refer to the same ranking task. The task was offered for multiple languages. The CheckThat! Lab features real claims uttered by humans: Task 2A treats check-worthy tweets as input claims, Task 2B contains claims made in political debates or speeches. In both cases, real fact-checks

[3] `https://data.gesis.org/claimskg-explorer/home`

from fact-checking portals are used to construct the set of verified claims to match against, partly drawn from ClaimsKG.

Since this task definition matches our problem definition, we choose the CheckThat! Lab datasets and the participating systems for our study.

3.3.2 Claim retrieval approaches

Early work has discussed the task of retrieving fact-checked claims as part of an end-to-end fact-checking pipeline, without providing detail or an evaluation Hassan et al. (2017b). Hence, the two ground-laying works in that field are Shaar et al. (2020b) and Shaar et al. (2022). In Shaar et al. (2020b), the authors provide two datasets: one using Politifact articles and one using tweets mentioned in Snopes articles. The latter and the annotation protocol later became the basis for the CLEF CheckThat Lab! claim retrieval datasets. They evaluate a standard information retrieval method (BM25) and provide a definition of the task. In Shaar et al. (2022), the authors extend this method and incorporate global and local contextual information contained within the source texts, i.e. the political debate containing the input claims, and the target texts, i.e. the fact-checking articles containing the verified claims. They find that contextual information, especially on the source-side, helps the claim retrieval performance. A multi-modal approach was proposed by Vo and Lee (2020), whose neural ranking model uses both text and images for the retrieval and recommendation of claims about images, based on mining tweets containing images and links to fact-checking articles. On the same dataset and on data from the Chinese platform Weibo, Sheng et al. (2021) developed and evaluated their memory-enhanced transformer-based approach to select key sentences from fact-checking articles to estimate the relevance of a verified claim for a given input claim.

Systems submitted to the CheckThat! Lab shared claim retrieval tasks range from relying on pre-trained or fine-tuned language models and using supervised methods to efficient index-based systems relying on traditional information retrieval techniques such as BM25 in supervised or unsupervised setups. Some systems performed data augmentation and different data cleaning and pre-processing steps. The most successful approaches relied on transformers, achieving up to 0.956 MAP@5 on the

2022 English tweets dataset. We describe the systems in more detail in Section 3.4.

Several systems used the CheckThat Lab! benchmark without participating in the challenges. Building on the winner of the 2021 competition, Hardalov et al. (2022) proposed a framework using distant supervision and a modified self-adaptive training approach to retrieve fact-checked claims for claims made in tweets. Their silver standard corpus is created by harvesting tweets that contain a link to a fact-check from Snopes, similar to Vo and Lee (2020)'s mining of tweets and fact-checking articles, yet focused on textual claims and implemented an automatic filtering and labeling step. They beat the state-of-the-art performance on the CLEF CheckThat! Lab 2021 English tweet data (0.882) with a MAP@5 score of 0.903. The web-based online claim retrieval tool ClaimLinker Maliaroudakis et al. (2021a) links arbitrary texts to claims in ClaimsKG. It obtains candidate claims through a blocking step based on Elasticsearch's (ES) BM25 scores and re-ranks them using textual similarity measures. Despite its simplicity, it succeeds in retrieving the correct verified claim in the first position (P@1) in 76% of the cases for the CheckThat! Lab tweets dataset of 2020 (83.5% for P@5). MAP scores are not specified by the authors. Mansour et al. (2022) preprocess tweets and enrich them with extracted information from embedded URLs, images, and videos. They experiment with different claim numbers for a BM25-based blocking step and different BERT variants for reranking the results. Their best approaches beat the previous state-of-the-art performance on CheckThat! Lab English tweet data of the years 2020 and 2021 with MAP@5 scores of 0.955 and 0.936, respectively. However, the 2021 system has access to additional metadata. Using the same system as for 2020, the score on 2021 drops to 0.929. This shows that the datasets of the different years have different characteristics or task difficulties, which makes the different systems proposed in the literature hard to compare, even when they use the CheckThat! Lab benchmark and the same evaluation metrics. We will provide more insights both into the differences of the datasets and the performances of the systems in Sections 3.5 and 4.6.

3.3.3 CheckThat! Lab claim retrieval data analyses

Several works investigate characteristics of claim retrieval benchmark datasets that are relevant to assess task difficulty, robustness of methods, and possible biases during evaluation.

Shaar et al. (2020b) include an exploratory analysis of their dataset, giving examples of trivial cases where the verified claim is identical to the input claim, harder cases where they differ regarding lexical choice, and complex input claims with sub-claims, which are verified by two corresponding verified claims. They further distinguish input claim and verified claim pairs into two categories: those that can be matched using simple approximate string matching techniques (Type-1), and all others (Type-2). Analyzing a manually annotated sample of 100 pairs extracted from their politifact debates dataset, they found 48% to be of Type-2. An analysis of word-level TF.IDF-weighted cosine-similarities further revealed their politifact debates data to be harder than their tweet data from Snopes, with 27% of the debates pairs having similarity scores above 0.25 vs. 50% of the tweets data. Shaar et al. (2022) analyzed the debates dataset by Shaar et al. (2020b) in more detail and categorized the pairs of matching claims into 4 categories: (1) clean, (2) clean-hard, (3) part-of and (4) context-dep. (1) and (2) refer to self-contained input claims with a matching verified claim that directly (1) vs. indirectly (2) verifies it, (3) and (4) refer to input claims that are not self-contained and require additional sentences from the discoursive context to fully form an individual claim (3) or need co-reference resolution to contain all relevant information (4). Furthermore, the authors show that different data splits can greatly influence task difficulty because systems can benefit from training and testing on claims drawn from the same debates or from debates about similar topics.

While Shaar et al. (2020b)'s data and annotation protocol later formed the basis of the CheckThat! Lab datasets, the latter contained additional data and different splits, thus the results cannot be transferred to these benchmark datasets. However, they highlight the importance of analyzing the data in more detail to understand and compare model performance.

Mihaylova et al. (2021a) analyzed the number of words and sentences in the different fields for the CheckThat! Lab 2021 data. Hardalov et al. (2022) provide insights on similarities between input tweets and titles and subtitles of matching verified claims, compare Jaccard vs. Cosine similarity scores, and analyze Sentence-BERT-based retrieval performance for different similarity thresholds.

While these analyses add valuable insights for understanding the performance of their models, a comprehensive analysis across datasets and the relation of dataset characteristics and performance of different models is hitherto missing. We aim to close this gap with this study.

3.4 Experimental Setup

In the following, we describe our evaluation protocol including analyzed datasets, evaluated methods, and evaluation metrics.

3.4.1 Datasets

For comparative evaluation, we use the datasets and corresponding training/test splits as provided by CheckThat! Lab Task 2 (*verified claim retrieval*) challenges of the CLEF2020 Barrón-Cedeño et al. (2020), CLEF2021 Shaar et al. (2021) and CLEF2022 Nakov et al. (2022) initiatives in English language. Table 3.2 provides statistics regarding training and test data for each dataset. The dataset from the first edition (in the following referred to as *2020-tweets*) contains claims extracted from tweets as input queries and corresponding verified claims from fact-checking websites as targets. The targets were drawn from the fact-checking platform Snopes, harvested and augmented with data from ClaimsKG Tchechmedjiev et al. (2019c) as described in Barrón-Cedeño et al. (2020) and Shaar et al. (2020b). The data from the 2021 and 2022 challenges each contain one dataset with tweets (Task 2A, in the following referred to as *2021-tweets* / *2022-tweets*), extensions of the previous datasets, and one with a different discourse type: claims from political debates (Task 2B, referred to as *2021-debates* / *2022-debates*).

Claims from political debates were mainly drawn from the fact-checking portal Politifact using the same protocol as described in Shaar et al. (2020b). The ground truth of links between input queries and matching verified claims was constructed using information provided by Politifact overview articles which feature links between claims from political speeches and fact-checks on their portal, and by Snopes, where the review articles often refer to tweets spreading the fact-checked claim.

For the tweet data, there is always exactly one correct verified claim to retrieve per input query. For claims from political debates, the number of correct matches may be greater than one. A verified claim may be the correct target for more than one input query in all datasets. Each dataset contains a set of target verified claims to match all queries against: queries from test, training, and development sets alike. I.e. there always is one set of target verified claims to use both for training and testing. The data for each year consists of the data of the previous year with some additions. The test sets, consisting of a file with input queries, i.e. their IDs and texts, and a file with gold matchings between input query IDs and target verified claim IDs, contain new input queries each year and links to either new targets or links to verified claims that have already be contained in the set of verified claims in previous editions. The training sets also consist of an input query file and a file with matchings between queries and targets. These files did not change after the 2021 edition, i.e. the training sets of *2021-tweets* and *2022-tweets* are the same, and the training sets of *2021-debates* and *2022-debates* are the same. However, the set of target verified claims to match against differs for *2021-tweets* vs. *2022-tweets*, i.e. the latter contains additional candidate claims. Thus, training a model on data from the respective years may yield different results because a model has access to additional negative examples during training. The set of verified claims is the same for *2021-debates* and *2022-debates*. This means that only new input claims were added for the test set, but no new matching verified claims, i.e. the new input queries have links to targets that have already been contained in the set of verified claims in the earlier years. Test sets of a previous year were used as development sets for the upcoming year.

Since there always is one set of target verified claims to use both for training and testing, methods use the set of verified claims that is also used for testing when training and fine-tuning, but paired with different input queries.

The debates data offers fewer positive training examples, more candidates to choose from, and requires finding more than one match for some input claims, which supposedly makes it more difficult than the tweet data.

We investigate the characteristics of the data in more detail in Section 3.5 to address RQ1.

Name	#ITr	#VTr	#ITe	#VTe	#VClaims
2020-tweets	800	800	200	200	10,375
2021-tweets	999	999	202	202	13,825*
2022-tweets	999	999	209	209	14,231
2021-debates	472	562	79*	103	19,250
2022-debates	472	562	65*	83*	19,250*

Table 3.2: Datasets and statistics.*#ITr*: Number of input claims in the training set, *#VTr*: Number of input claim - verified claim pairs in the training set, *#ITe*: Number of input claims in the test set, *#VTe*: Number of input claim - verified claim pairs in the test set, *#VClaims*: Total number of verified claims to match against. *Numbers based on our own counts, diverge from the numbers specified in the task overview paper

3.4.2 Retrieval performance and robustness

For investigating RQ2, we train and test all available methods on all datasets for a comparative evaluation. For RQ3, we perform a cross-dataset evaluation: we apply the trained models to test data of different datasets. This serves 1) to analyze the influence of the training sets without confounding the results with potential differences in test set difficulty and 2) to assess the robustness of the methods and how well-learnt models generalize to data of different distributions and of a different discourse type (tweets vs. claims from political debates). More precisely, we train supervised approaches on each of the *2020-tweets*, *2021-tweets*, *2022-tweets*, and *2021/2-debates* training datasets and apply each of the trained models on all of the test datasets. Since *2021-debates* and *2022-debates* share the same training sets and verified claims, we only train one model per approach on the debates data. We train three models for each neural network-based approach on each dataset to investigate their robustness/stability and report the average scores. With this, we compare performances of the supervised approaches for in-distribution vs. out-distribution training and relate these scores with the performance of unsupervised approaches on each test dataset.

We rely on the performance metrics used in the official CLEF CheckThat! Lab challenges: Mean Average Precision@k (MAP@k) and Precision@k with MAP@5 being the primary measure to rank the performance of systems Nakov et al. (2022). Note that these metrics are applicable both for the single and the multiple target case.

3.4.3 Efficiency

For investigating RQ4, efficiency, and scalability, we analyze runtime complexity and costs for online computations of the different approaches. Online processing encompasses all steps that depend on the input claims and can only be executed at inference time: pre-processing and feature generation for input claims, comparisons with verified claims, and inference. Offline processing refers to pre-processing of verified claims in the database, e.g. computing and storing their embeddings or other features, and includes all training and fine-tuning steps. In order to assess applicability in real applications, we measure the time required for the offline and online processing steps of the best supervised and the best unsupervised approach in our comparison.

We use two different servers: *Server 1* with 4x NVidia GeForce RTX2080Ti (11 GB) GPU, 2 x Intel Xeon 2.1 GHz CPU with 24 available cores and 1.4 TB RAM, and *Server 2* with 1x A6000 GPU (48 GB) and Xeon E5-4667 v3 CPU with 32 available cores and 64 GB RAM. In addition, we examine whether the best unsupervised approach can be run on a laptop. We use a model with an AMD Ryzen 9 6900HS processor with 3.30 GHz Radeon Graphics and 16 GB RAM. We measure the average time needed to link an input claim to a claim in a database depending on the size of the database, i.e. the number of candidate verified claims. We compare the results for database sizes of 1k, 5k vs. 10k claims respectively. The respective samples were drawn randomly from the 2020-tweets dataset with all claims from the smaller sets being included in the larger ones. We do not store and use any intermediate results such as embeddings of already considered input claims and treat all experiments with the different database sizes independently. We measure three runs for each database size and report the averaged run times.

The number of input claims remains constant: we use the 200 input claims from the 2020-tweets test dataset.

3.4.4 Claim retrieval methods and baselines

In the following, we introduce the methods compared in this study: systems and baselines that were part of any of the CLEF CheckThat! Lab claim retrieval challenges (2020, 2021, and 2022) plus the online claim retrieval tool *ClaimLinker*, and an extension of the 2022 participant *SimBa*, *SimBa 2023*. We include here all systems for which publications or working notes are available.

An overview of the systems and their characteristics can be found in Table 3.3. While many systems experiment with the use of different verified claim fields, most use only the texts of the verified claims or the texts plus the titles of the fact-checking articles for prediction. Likewise, some systems experimented with incorporating the context information from the debates transcripts but none used them in their final submissions.

The individual systems are summarized below.

Buster.ai Bouziane et al. (2020) incorporates external data from similar tasks to fine-tune a pre-trained version of RoBERTa. They further clean all tweets and while training, focus on retrieving adversarial negative examples using indexed search to better distinguish negative examples with high similarities.

[9] `https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens`
[10] `https://huggingface.co/sentence-transformers/bert-large-nli-max-tokens`
[11] `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1`
[12] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`
[13] `https://huggingface.co/princeton-nlp/sup-simcse-roberta-large`
[14] `https://github.com/facebookresearch/InferSent`
[15] `https://tfhub.dev/google/universal-sentence-encoder/4`
[16] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`
[17] `https://huggingface.co/princeton-nlp/sup-simcse-roberta-large`
[18] `https://github.com/facebookresearch/InferSent`
[19] `https://tfhub.dev/google/universal-sentence-encoder/4`
[20] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`
[21] `https://huggingface.co/princeton-nlp/unsup-simcse-roberta-base`
[22] `https://huggingface.co/sentence-transformers/sentence-t5-base`
[23] `https://huggingface.co/sentence-transformers/sentence-t5-large`
[24] `https://huggingface.co/EleutherAI/gpt-neo-1.3B`
[25] `https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens`
[26] `https://huggingface.co/bert-base-uncased`

UNIPI-NLE Passaro et al. (2020) bases its predictions on a blocking step using an Elasticsearch (ES) index and word matching techniques and subsequent application of Sentence-BERT models fine-tuned in two steps.

UB_ET Thuma et al. (2020) retrieves the top 1k tweet–claim pairs using parameter-free DPH divergence from the randomness term weighting model in Terrier, computes several features from weighting models (BM25, PL2, and TF-IDF) and then builds a LambdaMart model on top for reranking.

Aschern Chernyavskiy et al. (2021a) uses a fine tuned-sentence BERT model for blocking, then re-ranks the top 20 candidates using a trained LambdaMART reranker with Sentence-BERT and TF.IDF similarity scores.

CrowdChecked Hardalov et al. (2022) adopts Aschern's pipeline and adds modified self-adaptive training and distant supervision using claims and fact-checking URLs mined from social media.

NLytics Pritzkau (2021) used similarities derived from a pre-trained RoBERTa model to rank the candidate target claims. This approach can thus be seen as a baseline for using RoBERTa without further additions.

DIPS Mihaylova et al. (2021b) computes the cosine similarities for each input and verified claim pair using Sentence-BERT embeddings for the fields claim text, review title, review subtitle, and review date, and passes them as a sorted list to a neural network. For the debates data, they use the fields title and text.

BeaSku Skuczyńska et al. (2021) use a triplet loss training method to perform fine-tuning of the Sentence-BERT model and use the scores predicted by that model together with BM25 scores as features to train a rankSVM-based reranker.

bigIR did not submit a working note. However, their approach is described in detail in Mansour et al. (2022), and according to Nakov et al. (2022), the submitted variant retrieves candidates using BM25 on preprocessed and augmented tweets, as described in Section 3.3, and performs reranking using a fine-tuned MPNet model.

Fraunhofer SIT Frick and Vogel (2022) augment the data using back-translation of tweet texts, train classifiers based on cosine-similarities of embeddings derived from fine-tuned sentence transformers for claim texts, review titles and subtitles, respectively, and combine the scores using an SVM as meta classifier. The authors accidentally submitted results generated by averaging the scores instead of using the better-performing meta classifier. We provide the results of their best-performing latter variant (Fraunhofer SIT SVM) in addition to the submitted variant.

For all publicly available submissions, detailed in the following, we additionally provide detailed cross-evaluation experiments and replicate their results. We modified the approaches to allow training and testing on all different datasets. All modified code can be found in our repository. Unless stated otherwise, we used the variants that constituted the teams' primary submissions.

TIET Shukla and Sharma (2020) retrieves the top 1,000 claims for an input tweet based on their BM25 scores and combines those with a similarity score based on the cosine between Sentence-BERT (Base NLI Mean Tokens) embeddings for input and candidate claims.

NLP&IR@UNED Martinez-Rico et al. (2020) uses Universal Sentence Encoder Cer et al. (2018b) embeddings for input tweets and verified claims and train a feed-forward neural network with cosine similarity of embeddings, type-token ratio, average word length, number of verbs/nouns, ratio of content words, and ratio of content tags as features.

TheUofSheffield McDonald et al. (2020) is a point-wise Learning-to-Rank approach that trains a Linear SVM to learn scoring models based on BM25 scores and TF-IDF term weightings for claim texts and titles. Preprocessing consists of removing URLs, emojis, punctuation, and stopwords, converting to lowercase, tokenization, and Porter stemming and lemmatization.

Check_square Cheema et al. (2020) builds a KD-tree based on embeddings for claim texts and titles to retrieve the 1,000 most similar verified claims for an input claim. We report results for two of their variants which performed better than their primary submission: one using Sentence-BERT-Large pre-trained on SNLI with MAX tokens, fine-tuned with triplet loss, and the other one using multilingual DistilBERT-embeddings without fine-tuning (distmult). In both cases, the tweet pre-processing includes removing URLs, emails, phone numbers, and user mentions.

ES Baseline[27] retrieves claims from an ES index purely based on their BM25 scores. To maintain a consistent baseline, we utilized the script of 2021 and built the ES-index for the verified claims of the 2020 edition, which were presented in a different format, as described in the baseline script of 2020[28].

RIET Lab Shliselberg and Dori-Hacohen (2022) uses a fine-tuned Sentence-T5 transformer model for blocking and a fine-tuned generative GPT-Neo model for re-ranking.

SimBa Hövelmeyer et al. (2022) uses an embedding-based blocking step followed by re-ranking based on the combination of multiple embeddings and lexical features. The 2022 challenge submission for the tweets data, **SimBa (tweets)** was unsupervised and combined the cosine similarity scores from the embeddings produced by all-mpnet-base-v2[29]Reimers and Gurevych (2019), SimCSE[30]Gao et al. (2021) and a token overlap count for re-ranking. Blocking was performed by selecting the union of the top 50 verified claims identified by either of the *all-mpnet-base-v2*, SimCSE, InferSent[31]**?**, and Universal Sentence Encoder[32]Cer et al. (2018a) embeddings. For the debates data, a supervised approach was used, **SimBa (debates)**. It used the blocking mechanism to select hard negatives for training a Linear SVC with random undersampling. As features, it relied on SimCSE embeddings, a string distance measure, token overlap, token overlap ratio, and overlaps and overlap ratios of WordNet synsets. We modify this approach, **SimBa 2023**, to use *all-mpnet-base-v2* for blocking and the Braycurtis distance[33] as a similarity measure, since it provided more accurate results than cosine similarity. For re-ranking, it uses *all-mpnet-base-v2*, SimCSE[34], Sentence-T5[35]Ni et al. (2021), Universal Sentence Encoder and the ratio of common words (without preprocessing).

[27] https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/blob/master/task2/baselines/elasticsearch_baseline.py
[28] https://github.com/sshaar/clef2020-factchecking-task2/blob/master/elastic_search_baseline.py
[29] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[30] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large
[31] https://github.com/facebookresearch/InferSent
[32] https://tfhub.dev/google/universal-sentence-encoder/4
[33] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.braycurtis.html
[34] https://huggingface.co/princeton-nlp/unsup-simcse-roberta-base
[35] https://huggingface.co/sentence-transformers/sentence-t5-base

Finally, we include the web-based online claim retrieval tool **ClaimLinker** Maliaroudakis et al. (2021a) in our analyses. ClaimLinker obtains candidate claims through a blocking step based on BM25 scores and re-ranks them using textual similarity measures. For the variant reported here, we added preprocessing to remove special characters, punctuation, and stopwords, split hashtags, and lemmatize all claims using the Stanford CoreNLP Lemmatizer[36]. This step aims to enable the tool to better process tweets. It uses text overlap features (common words, lemmas, entities, words+POS, ngrams, nchargrams) and combines BM25 and textual similarity scores with a weighting of 60/40.

[36] `https://stanfordnlp.github.io/CoreNLP/lemma.html`

Approach	Su.	F.t.	Block	Idx	LMs	Fields
NLP&IR@UNED	✓			✓		vclaim
TIET			1000	✓	bert-base-nli-mean-tokens[4]	vclaim
TheUofSheffield	✓			✓		vclaim, title
ES-Baseline				✓		vclaim, title
Check_square		✓	1000		bert-large-nli-max-tokens[5]	vclaim, title
Check_square distmult			1000		distiluse-base-multilingual-cased[6]	vclaim, title
ClaimLinker			30	✓		vclaim, title
SimBa (tweets)			≥50		all-mpnet-base-v2[7], usup-simcse-roberta-large[8], InferSent fastText[9], UniversalSE[10]	vclaim, title
SimBa (debates)	✓		≥50		all-mpnet-base-v2[11], usup-simcse-roberta-large[12], InferSent fastText[13], UniversalSE[14]	vclaim, title
SimBa 2023			50		all-mpnet-base-v2[15], unsup-simcse-roberta-base[16], sentence-t5-base[17]	vclaim, title
RIET Lab	✓	✓	25		sentence-t5-large[18], gpt-neo-1.3B[19]	vclaim
Buster.ai	✓				RoBERTa*	augmented vclaims
UB_ET	✓		1000	✓		vclaim
UNIPI-NLE	✓	✓	1000	✓	bert-base-nli-mean-tokens[20], bert-base-uncased[21]	vclaim, title
CrowdChecked	✓	✓	100		stsb-bert-base[22]	vclaim, title, subtitle
Aschern	✓	✓	20		stsb-bert-large[23]	vclaim, title, subtitle
NLytics					roberta-base[24]	vclaim
DIPS (tweets)	✓		50		paraphrase-distilroberta-base-v1[25]	vclaim, title, subtitle, date
DIPS (debates)	✓		100		paraphrase-distilroberta-base-v1	title, text, queryID: date
BeaSku	✓	✓			distilbert-base-nli-mean-tokens	vclaim, title, body
BigIR		✓	20	✓	STSb-MPNet-base[26]	vclaim, title
Fraunhofer SIT		✓			all-MiniLM-L6-v2, all-MPNet-Base-v2	vclaim, title, subtitle
Fraunhofer SIT SVM	✓	✓			all-MiniLM-L6-v2, all-MPNet-Base-v2	vclaim, title

Table 3.3: Claim retrieval approaches compared in this study. Su.: approaches using a supervised classifier or ranker. F.t.: approaches fine-tuning language models for the task. Block: Amount of candidates after the blocking step, if any. Idx: Approaches making use of an index such as Elasticsearch. LMs: used language models. *precise variant not specified. Fields: Verified target claim fields considered by the system: vclaim: text of the verified claim, (sub)title/text: (sub)title/text of the verified claim review article. Approaches above the dividing line are considered for additional cross-evaluation experiments.

3.5 Data Analysis

In order to facilitate an informed interpretation of experimental results of state-of-the-art systems and the utilised features, understanding the characteristics of the used datasets is crucial. This is even more important when aiming to understand model robustness and overfitting through cross-dataset evaluation. Thus, in the following, we provide an assessment of the aforementioned widely used benchmark datasets for claim retrieval evaluation.

We draw on characteristics investigated and deemed relevant for the task in previous works (cf. Section 3.4), such as claim and input lengths and token overlaps, and provide extended analyses of these properties for all different datasets, including statistics for inputs and verified claims in test vs. training splits. We add detailed insights into the completeness and correctness of the data and gold data annotations and their match to the task definition and real-world claim retrieval requirements.

3.5.1 Matches per input query and per verified claim

As per task definition, each input query in the tweets datasets has one matching verified claim in *D*. However, we find that in *2020-tweets*, one training query has two matching verified claims. In the *2021/2-tweets* training set, this is the case for two input queries: the query corresponding to the *2020-tweets* query plus a new entry. Moreover, there are duplicates in some of the gold input query - matching target verified claim pair files for the test datasets, as detailed in Table 3.4. Duplicates here mean repeated entries of the same input query - verified claim pair. Thus, the respective input queries implicitly gain a higher weight during evaluation than the ones that are not repeated. There are no duplicates in the input query files. For *2020-tweets*, one test input query (ID 1198) does not have a match to a verified claim in the gold file. Instead, there is a duplicate for another input query (ID 1167). A closer examination of the data reveals that the duplicated line can be corrected by changing the listed query ID with the missing query ID, i.e. both match the same verified claim (see below for statistics regarding multiple matches per verified claim). All other input queries are contained in the gold files. The other tweet datasets do not contain duplicates. However, as Table 3.4 shows, there are many duplicates in the debates test sets. For the debates data, 12.288% of training queries have more than one matching verified claim. For the test queries, this is the case for 16.456% (29.114% without de-duplication) for *2021-debates* and 18.462% (26.154% without de-duplication) for *2022-debates*. The average number of matching verified claims per test query remained stable: 1.165 (1.304 without de-duplication) for *2021-debates* vs. 1.185 (1.277 without de-duplication) for *2022-debates*. For training queries, the ratio is 1.191 (1.191 without de-duplication). All following analyses are performed on the original datasets, without de-duplication, to allow insights into the results reported in the literature which used the data as-is.

For all datasets, one verified claim may be the correct match for multiple input queries. Table 3.4 details the distribution for each dataset and split. For all datasets, the number of positive examples is low. The debates datasets contain fewer positive examples than the tweets datasets. When a verified claim is the correct match for an input query, it in many cases also matches other input queries (∼18-20% for tweets training datasets, ∼12-28% for tweets test, ∼36% for debates training, ∼37-44% for debates test). This is pronounced for the debates data. Therefore, the number of verified claims to choose for matching is low, as detailed in column #c_match, and models may learn to identify relevant verified claims and that each of them has a high probability of matching multiple input queries.

In eight cases, a verified claim is the correct target for input queries in the training set and input queries in the test set, i.e. a model may learn to find matching claims for this specific verified claim in the training phase. This is true for seven verified claims in *2020-tweets* and one claim in *2022-tweets*.

3.5.2 Tweet datasets

For the tweets data, training, and test input queries consist of tweet texts including links to pictures or other URLs, user mentions, hashtags, and all other content allowed in tweets. In addition, they have the user name and mention as well as the date of publication appended to their end, separated from the tweet text by "–". For an example, see Table 3.5. The query with q_id 2 is one of three queries in the *2020-tweets* training set that is missing the appended information. The *2021-tweets* training set

dataset	split	#V	#c	#c_match	#c_multi
2020-tweets	train	801 (801)	10375	657 (6.333%)	115 (17.504%)
2020-tweets	test	200 (199)	10375	150 (1.446%)	42 (28.000%)
2021-tweets	train	999 (999)	13825	782 (5.656%)	159 (20.332%)
2021-tweets	test	202 (202)	13825	143 (1.034%)	31 (21.678%)
2022-tweets	train	999 (999)	14231	782 (5.495%)	159 (20.332%)
2022-tweets	test	209 (209)	14231	180 (1.265%)	21 (11.667%)
2021-debates	train	562 (562)	19250	354 (1.839%)	129 (36.441%)
2021-debates	test	103 (92)	19250	52 (0.270%)	23 (44.231%)
2022-debates	train	562 (562)	19250	354 (1.839%)	129 (36.441%)
2022-debates	test	83 (77)	19250	52 (0.270%)	19 (36.538%)

Table 3.4: Correct matches per verified claim. #V: number of gold verified-input query pairs in total (and without duplicates), #c_match: number of claims in the verified claims file that are the correct match for at least one input query (absolute and relative to #c). #vc_multi: number of claims in the verified claims file that are the correct match for more than one input query (absolute and relative to #c_match)

misses this information in four cases. In all other data, the appended information is present. Therefore, *check-worthy claim* in this task refers to a tweet, which may consist of multiple sentences and a user mention and publication date.

Target verified claims of the *2020-tweets* data consist of the target ID, the text of the verified claim, and the title of the fact-checking article, as illustrated in Table 3.5.

In the 2021 and 2022 editions, the target verified claim data include further information from the fact-checks: review title, subtitle, date, text, author, and only in 2022, url of the review (in JSON format), see Figure 3.1.

As Mihaylova et al. (2021b) observed, the appended date format of the input queries corresponds to the format of the date field in the target verified claims which is supplied for the 2021 and 2022 data, but not for 2020.

Concerning completeness of the verified claim data, the *2021-tweets* and *2022-tweets* data contain some empty or missing fields: subtitles are non-empty for all but 4 claims, authors for all but 26, dates for all but 287. The URLs in *2022-tweets* are only available for 2,266 of the 14,231 target verified claims. All other metadata is available for every target claim. In 9 cases, a *2020-tweets* target claim included a newline which,

```
{
'vclaim_id': 'vclaim-sno-15-homeless-dead-in-chicago',
'vclaim': 'Fifteen homeless people in Chicago
were found dead on the street because of record-low temperatures that hit
the city in late January 2019.',
'title': 'Were 15 Homeless People Found Frozen to Death in Chicago?',
'subtitle': 'A picture of a Canadian man was repurposed online to prod
social media users to think about how "blessed" they are.',
'page_url': 'https://www.snopes.com/fact-check/15-homeless-dead-in-chicago/',
'author': 'Arturo Garcia',
'date': '30 January 2019'
}
```


Figure 3.1: Target verified claim in 2022-tweets


```
8935    The Monsanto Protection Act creates a “precedent-setting
limitation on\njudicial review of genetically-engineered crops.”
Monsanto Protection Act
```

```
{
'title': 'Monsanto Protection Act',
'subtitle': "Does the Monsanto Protection Act create a 'precedent-setting
limitation on judicial review of genetically-engineered crops'?",
'author': 'David Mikkelson',
'date': '13 September 2013',
'vclaim_id': 'vclaim-sno-monsanto-protection-act',
'vclaim': 'The Monsanto Protection Act creates a “precedent-setting
limitation on'
}
```


Figure 3.2: The same target verified claim in the *2020-tweets* (top) vs. *2021/2-tweets* (bottom) datasets

*q*_id	*q*	*c*_id	*c* text	*c* title
1	Trump needs to immediately divest from his businesses and comply with the emoluments clause. Iran could threaten Trump hotels *worldwide* and he could provoke war over the loss of revenue from skittish guests. His business interests should not be driving military decisions. — Ilhan Omar (@IlhanMN) January 6, 2020	394	In January 2020, U.S. Rep. Ilhan Omar advised Iran to attack Trump-branded hotels in the world, thus committing treason.	No, U.S. Rep. Ilhan Omar Didn't Give 'Treasonous' Military Advice to Iran
2	A number of fraudulent text messages informing individuals they have been selected for a military draft have circulated throughout the country this week.	670	The U.S. Army is sending text messages informing people they've been selected for the military draft.	Is US Army Sending Texts About a Military Draft?
3	Fact check: The U.S. Army is NOT contacting anyone regarding the draft. If you are receiving texts, phone calls or direct messages about a military draft, they are not official communications from the U.S. Army pic.twitter.com/3S32De8ekP — U.S. Army CGSC (@USACGSC) January 8, 2020	670	The U.S. Army is sending text messages informing people they've been selected for the military draft.	Is US Army Sending Texts About a Military Draft?

Table 3.5: Matching queries and targets from the 2020-tweets training dataset.

during conversion to JSON from tsv for the *2021-tweets* and *2022-tweets* datasets, caused the claim to be truncated at the newline, see Figure 3.2.

The input claims in test and training sets are ordered and their IDs are constructed depending on this order. Thus, the IDs provide useful information: claims matching the same target are grouped together, e.g. the *2020-tweets* test input queries with the IDs 999 and 1000 both match the verified claim with ID 6094. This is true for all datasets. This fact can potentially be used by supervised models.

```
{
"url": "/factchecks/2020/jan/15/elizabeth-warren/democratic-debates-biggest-
electoral-losers-number/",
"speaker": "Elizabeth Warren",
"vclaim": "\"Look at the men on this stage. Collectively, they have lost 10
elections. The only people on this stage who have won every single election
that they\u2019ve been in are the women, Amy and me.\"",
"truth_label": "true",
"date": "stated on January 14, 2020 in a debate:",
"title": "The Democratic debates\u2019 biggest (electoral) losers, by the
numbers",
"text": <full review text>,
"vclaim_id": "vclaim-pol-17583"
}
```


Figure 3.3: Target verified claim with ID *vclaim-pol-17583* from the *2021-debates* test dataset

3.5.3 Political debates datasets

For the debates data, the target verified claims consist of verified claim text (vclaim) and metadata from fact-checking review articles: URL, speaker, truth label, date, title and review text (truncated here), as shown in Figures 3.3 and 3.4.

Note that the speaker field may contain all kinds of information about a source of a fact-checked claim: it can be the actual speaker of an utterance, as in the example above, or other source information such as "Facebook posts".

*q*_id	*q*	*c*_id
2020_jan_15_441	Look at the men on this stage.	vclaim-pol-17583
2020_jan_15_442	Collectively, they have lost ten elections.	vclaim-pol-17583
20160112_Obama_state_of_the_union_0013	And our businesses have created jobs every single month since it became law.	vclaim-pol-05563

Table 3.6: Input queries and matching target verified claims from the *2021-debates* test (rows 1-2) and training (row 3) sets. The verified claim targets are displayed in full in Figures 3.3 and 3.4

.

The test datasets include transcripts of the speeches, containing not only the text

```
{
"url": "/factchecks/2016/jan/13/barack-obama/business-has-created-jobs-
every-month-obamacare-be/",
"speaker": "Barack Obama",
"vclaim": "\"Our businesses have created jobs every single month since
(Obamacare) became law.\"",
"truth_label": "true",
"date": "stated on January 12, 2016 in his State of the Union speech:",
"title": "Business has created jobs every month since Obamacare became
law, Obama said in State of the Union",
"text": <full review text>,
"vclaim_id": "vclaim-pol-05563"
}
```


Figure 3.4: Target verified claim with ID *vclaim-pol-05563* from the 2021-debates test dataset

of the input claim utterances but also their discoursive contexts. The input claims themselves are individual sentences from the speeches, yet they sometimes refer to entities mentioned in the context rather than the claims themselves, and can depend on one another, as can be seen in Table 3.6. With this, the data can be seen as diverging from the task definition defining *a check-worthy claim* as input, as the individual queries in this dataset are not necessarily check-worthy by themselves and do not constitute individual claims, depending on the precise definition of what a *claim* is (see Boland et al. (2022b) for an overview of diverging definitions), which is left implicit in this task definition.

In both the 2021 and 2022 datasets, the input query IDs contain the date of the debate which can be used as a feature to compare with the verified claim dates (cf. Mihaylova et al. (2021b)). Also, as illustrated in Table 3.6, the IDs sometimes contain the name of the speaker and/or the event during which the claim was uttered, both of which can be learned as a feature to match to verified target claims as well.

3.5.4 Comparison of individual datasets and splits

We next have a closer look at the training and test input and the set of verified claims for all datasets and compare characteristics such as claim length and word overlap. This provides insights into the task difficulty and the suitability of the data to assess the robustness and applicability of models to handle real-world claim retrieval scenarios.

Word overlap

In order to assess the difficulty of linking queries and target claims for the different datasets and the task in general, we measure the ratio of overlapping tokens/lemmas in input queries and targets, for both only the matching targets and all target verified claims including the false matches. For this, we tokenize and lemmatize the input using spacy[37], count the common tokens/lemmas for each pair of input query and target claim and normalize the count by the number of tokens in the smaller of the two sentences. This is because, as detailed in the task definition of 2020 (see Section 3.3), the target may be a sub-claim of the query or vice versa to justify a match and we do not want to give a higher weight to either of the directions in this analysis. We use the original data for the analysis, i.e. we do not perform de-duplication, as the original data is what has been used by the systems we strive to evaluate. As few systems use subtitle information, see Table 3.3, we omit them from this analysis.

As Table 3.7 shows, the input queries and matching verified claims in the debates datasets have a higher token overlap than the ones in the tweets datasets (0.385 vs. 0.342 for train and 0.375 vs. 0.342 for test on average). For the non-matching pairs, overlaps are more frequent in the debates datasets as well (0.156 vs. 0.121 for train and 0.157 vs. 0.113 for test on average). Overlaps for matching queries and targets are generally high both in the training sets (mean: 0.359) and the test sets (mean: 0.355) for all datasets, while the overlaps between queries and all targets are much lower (train mean: 0.135, test mean: 0.130). The *2020-tweets* test set has a considerably lower overlap ratio between input queries and verified claims than all other datasets. While the overlaps are lower for the matching pairs as well, the difference between the overlap ratios of matching and all pairs is greater in this dataset, with the matching pairs having roughly 3.7 times as many overlaps compared to 2.5 (*2021-tweets*), 2.8 (*2022-tweets*), 2.3 (*2021-debates*) and 2.4 (*2022-debates*). A similar pattern can be observed for the titles. In contrast to the claim texts, the overlap ratios are higher in the tweets than in the debates datasets. For both discourse types, overlap ratios are higher for claim texts than for claim review titles.

While the overlap of lemmas naturally is higher than the overlap of tokens, the ratio of overlaps is already high without any pre-processing.

[37] `https://spacy.io`

Dataset	Split	Targets	∩vclaim_T	∩vclaim_L	∩title_T	∩title_L
2020-tweets	train	matches	0.345	0.396	0.252	0.336
2020-tweets	train	all	0.123	0.157	0.058	0.087
2020-tweets	test	matches	0.270	0.338	0.313	0.379
2020-tweets	test	all	0.073	0.100	0.041	0.062
2021-tweets	train	matches	0.341	0.393	0.250	0.334
2021-tweets	train	all	0.121	0.158	0.064	0.098
2021-tweets	test	matches	0.379	0.434	0.237	0.347
2021-tweets	test	all	0.132	0.171	0.067	0.103
2022-tweets	train	matches	0.342	0.393	0.250	0.334
2022-tweets	train	all	0.121	0.158	0.064	0.098
2022-tweets	test	matches	0.379	0.426	0.258	0.339
2022-tweets	test	all	0.134	0.169	0.0651	0.099
2021-debates	train	matches	0.385	0.452	0.197	0.247
2021-debates	train	all	0.156	0.213	0.064	0.098
2021-debates	test	matches	0.371	0.442	0.147	0.189
2021-debates	test	all	0.158	0.220	0.065	0.101
2022-debates	train	matches	0.385	0.452	0.197	0.247
2022-debates	train	all	0.156	0.213	0.064	0.098
2022-debates	test	matches	0.379	0.451	0.159	0.206
2022-debates	test	all	0.157	0.217	0.066	0.104

Table 3.7: Word (token (T) / lemma (L)) overlap between input queries and target verified claims (vclaim) and titles for the different datasets and splits relative to the number of tokens. Split: overlap between targets and queries from training or from test set. Targets: overlaps measured only for matching queries and targets or for all queries and all target verified claim pairs.

The performance of BM25-based baselines adds further insights regarding the similarities of input queries and matching verified claims for the different datasets. As these are included in the comparison of approaches in Section 4.6, we omit these statistics here.

Input query and target verified claim lengths

Figure 3.5 details the number of characters, respective tokens, in verified claims, verified claim review titles, and input claims.

Subfigures (a) and (d) of Figure 3.5 and Table 3.8 show that the distributions of verified claim lengths are similar in the different tweet datasets, with the mean of tokens ranging between 19.0 and 19.6, while debate claims are longer on average (22.7), but vary more strongly with a standard deviation of 9.6 vs. 7.4 - 7.5 for tweets. Verified claims matching input training and test queries are longer than the average of all verified claims for *2021-tweets* and *2022-tweets*, while the test input query matches are shorter for *2020-tweets* and the debates datasets. Even though *2021-debates* and *2022-debates* have different test queries, they are linked to the same targets (cf. Section 3.4) and thus share the same numbers.

The length of verified target claim titles (cf. Subfigures (c) and (f) of Figure 3.5 and Table 3.5) fluctuates more strongly across datasets (mean #tokens: 9.0 for *2020-tweets*, 10.4 for *2021-tweets* and *2022-tweets*, standard deviation 4.4 - 4.8 for *2020-tweets*) and they are longer in the debates data (mean #tokens: 13.6, standard deviation 4.0). Verified claim titles matching training queries are longer than all others for the tweets datasets and for *2021-tweets* and *2022-tweets*, the titles of claims matching test queries are longer than the average, too, but not for *2020-tweets*. Titles contain overall fewer tokens than claim texts.

dataset	**targets**	**vclaim_mean**	**vclaim_std**	**title_mean**	**title_std**
2020-tweets	all	19.012	7.385	9.030	4.799
2020-tweets	train	21.046	6.958	13.192	3.698
2020-tweets	test	19.347	5.987	9.440	3.825
2021-tweets	all	19.501	7.353	10.361	4.370
2021-tweets	train	21.166	7.068	13.249	3.671
2021-tweets	test	24.357	9.585	12.469	2.918
2022-tweets	all	19.617	7.483	10.411	4.353
2022-tweets	train	21.160	7.081	13.249	3.674
2022-tweets	test	23.244	10.623	11.817	2.990
2021/2-debates	all	22.650	9.554	13.562	4.004
2021/2-debates	train	22.000	8.944	13.672	3.308
2021/2-debates	test	21.654	9.476	12.865	3.000

Table 3.8: Length of verified claims and their titles. Targets train/test: Claims c that match at least one input query in the training/test set

Similarly, the length of training input claims (cf. Subfigures (b) and (e) of Figure 3.5 and Table 3.9) is similar for the tweets datasets (mean #tokens: 42.3, standard deviation 15.8 - 15.9) with them being complete tweet texts with appended information. Debates training input claims are shorter (mean #tokens: 24.1, standard deviation

dataset	split	#tokens_mean	#tokens_std
2020-tweets	train	42.244	15.931
2020-tweets	test	29.755	5.815
2021-tweets	train	42.301	15.752
2021-tweets	test	44.540	16.765
2022-tweets	train	42.301	15.752
2022-tweets	test	45.459	16.645
2021-debates	train	24.107	16.342
2021-debates	test	22.684	15.008
2022-debates	train	24.107	16.342
2022-debates	test	24.292	15.407

Table 3.9: Input query lengths in training and test splits.

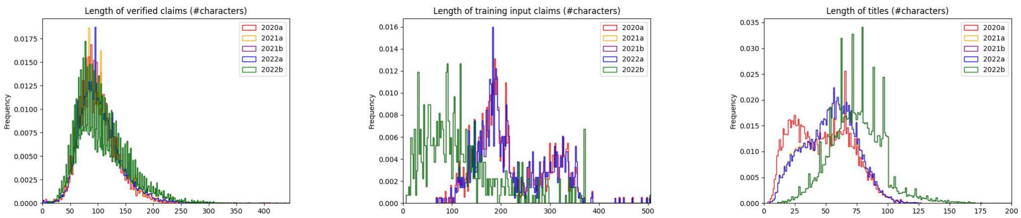
16.3), as they are individual statements uttered in a speech or debate.

A comparison of the last two columns of the figure and the corresponding Table 3.9 reveals that the distributions regarding input query length differs across training and test splits for all datasets except *2022-debates*, most strongly for the *2020-tweets* data. Striking is the difference between test and training input queries in *2020-tweets*, where the test queries are much shorter than the training queries (29.8 vs. 42.2) and have less variance regarding their length (5.8 vs. 15.9).

Many training input queries and target claims contain a very low or high number of tokens (cf. Subfigures (d) and (e) of Figure 3.5). Sentences with less than three tokens (excluding punctuation) can hardly constitute self-contained check-worthy or verifiable claims, as they should typically at least contain a subject, predicate, and object.

However, we count a small number of verified claims (29 for *2020-tweets*, 25 for *2021/2-tweets*), 9 for *2021/2-debates*) and 2 training input queries for the debates data with less than three non-punctuation tokens. None of these verified claims are correct matches for any training or test queries.

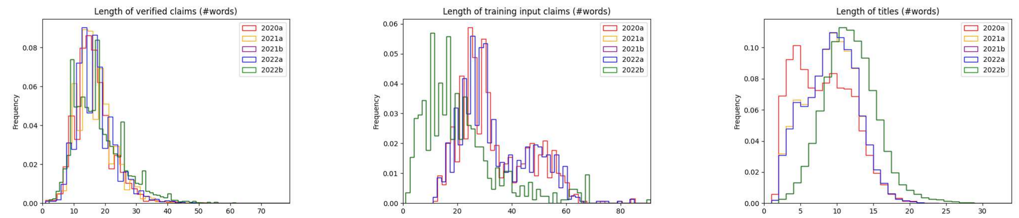
We examined the short queries and targets in more detail and found the latter to be due to extraction errors, e.g. as for the target claim from the 2021-tweets corpus in Figure 3.6 which consists merely of the word "The" or the target claim from *2022-tweets* in Figure 3.7 which has the verdict in the vclaim field. Such errors are hard



(a) Length (#characters) of verified claims

(b) Length (#characters) of input claims

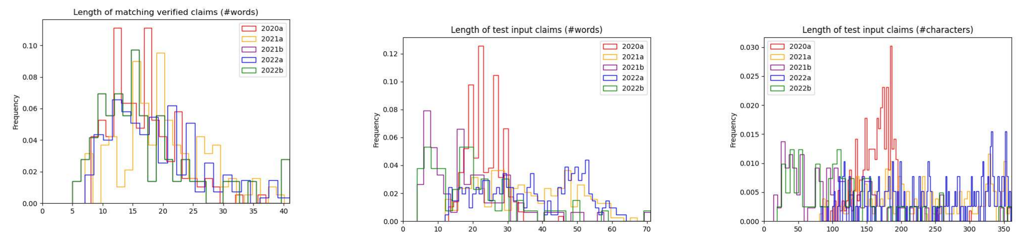
(c) Length (#characters) of verified claim titles



(d) Length (#words) of verified claims

(e) Length (#words) of input claims

(f) Length (#words) of verified claim titles



(g) Length (#words) of matching verified test claims

(h) Length (#words) of test input claims

(i) Length (#characters) of test input claims

Figure 3.5: Length of target verified claims, training and test input claims, and target verified claim titles in the different datasets.

```
{
  "title": "3 Musketeers",
  "subtitle": "Rumor: The names of the Milky Way and 3
  Musketeers candy bars were inadvertently switched.",
  "author": "David Mikkelson",
  "date": "13 June 2015",
  "vclaim_id": "vclaim-sno-3-musketeers",
  "vclaim": "The"
}
```


Figure 3.6: Short vclaim in the set of 2021-tweets target verified claims due to an extraction error.


```
{
  "title": "Maryland School Bans Marine Veteran Over Daughter\u2019s
  Homework",
  "subtitle": "Rumor claims a Marine was banned from the premises of
  his daughter's school for objecting to a homework assignment about Islam.",
  "author": "David Mikkelson",
  "date": "2 November 2014",
  "vclaim_id": "vclaim-sno-marine-banned",
  "vclaim": "mixture"
}
```


Figure 3.7: Short vclaim in the set of 2022-tweets target verified claims due to an extraction error.

to avoid when automatically extracting information from articles of fact-checking portals, as their structure is not fully consistent and the reviewed claims are not always explicitly listed, see, for instance, the fact-check at `https://www.snopes.com/fact-check/citibank/` which is part of *2020-tweets* where it is represented as

```
{"target_id": "4187", "target": "Citibank customers.",
"title": "Citibank Phish"}.
```

The short input queries in the debates dataset are due to the fact that they may refer to the context, as outlined above and in Table 3.6, e.g. ID 20161019_3pres_0423: "Wrong", and ID 20160303_GOP_michigan_0575: "Ten times". This is in line with Shaar et al. (2022)'s analysis of Shaar et al. (2020b)'s base debates dataset, where they found many claims that are not self-contained.

With this, they do not represent input that is expected in a real claim retrieval system, as users can typically be expected to enter a claim with its context. Note that, however, the context is accessible in the transcript files and systems participating in the Check That! Lab challenge may use them to enrich the input queries. When processed without additional context, debates input queries contain much less information (tokens and characters) than the tweets datasets.

As Figure 3.5 reveals, there is a long tail of very long claims and queries. We define outliers as those queries and claims whose token counts exceed the mean $+2*$ standard deviation. The results of the analysis are listed in Table 3.10.

Dataset	**Targets**	**Threshold**	**Count**
2020-tweets	all	33.8	328 (3.161%)
2020-tweets	train	34.9	25 (3.805%)
2020-tweets	test	31.3	8 (5.333%)
2021-tweets	all	34.2	486 (3.515%)
2021-tweets	train	35.3	25 (3.197%)
2021-tweets	test	43.5	5 (3.497%)
2022-tweets	all	34.6	543 (3.816%)
2022-tweets	train	35.3	25 (3.197%)
2022-tweets	test	44.5	10 (5.556%)
2021/2-debates	all	41.8	877 (4.556%)
2021/2-debates	train	39.9	19 (5.367%)
2021/2-debates	test	40.6	3 (5.769%)

Table 3.10: Long target verified claims classified as outliers regarding their numbers of tokens.

As Table 3.10 details, 3-4% of all target verified claims and those matching input training queries in the tweets datasets exceed the maximum expected length. The debates claims are not only longer on average, they also contain more long outliers. For titles, we find between 2 and 3% to be outliers, and a proportionally slightly higher ratio of 4% for all test datasets except *2020-tweets*.

Outliers with fewer tokens than expected account for less than 1% in all datasets. We thus omit these numbers.

Dataset	Targets	Threshold	Count
2020-tweets	all	18.6	266 (2.564%)
2020-tweets	train	20.6	20 (3.044%)
2020-tweets	test	17.1	4 (2.667%)
2021-tweets	all	19.1	227 (1.642%)
2021-tweets	train	20.6	25 (3.197%)
2021-tweets	test	18.3	6 (4.196%)
2022-tweets	all	19.1	235 (1.651%)
2022-tweets	train	20.6	25 (3.197%)
2022-tweets	test	17.8	7 (3.889%)
2021/2-debates	all	21.6	626 (3.252%)
2021/2-debates	train	20.3	8 (2.260%)
2021/2-debates	test	18.9	2 (3.846%)

Table 3.11: Long target verified claim titles classified as outliers regarding their number of tokens.

3.5.5 Summary

The CheckThat! Lab claim retrieval datasets feature input queries and targets with many direct lexical overlaps. We can thus expect that claim retrieval methods relying on simple word overlap metrics can achieve high scores, even without lemmatization or similar pre-processing. Likewise, embedding-based systems can be expected to perform strongly when the vocabulary matches the dataset contents. Overlaps for queries with verified claim texts are higher than with titles and titles on average contain fewer tokens than the verified claims. Thus, titles might be less helpful when used in systems relying on lexical similarities to a large degree, but they could add more focused information with semantically similar concepts, as they usually are abstractive summaries of the claims.

Check-worthiness detection and claim retrieval are separate tasks. According to the task definition, input queries for the claim retrieval task are check-worthy claims or tweets. However, indeed input queries are not always individual, self-contained claims that constitute check-worthy claims in themselves as also visible in Shaar et al. (2020b)'s analysis on their dataset that formed the basis of the Check That! Lab data. Therefore, models training on this data may not be able to distinguish essential parts of queries and targets, i.e. those that contain the check-worthy information, but only learn to find any matching information. Especially in light of the importance of word overlaps, the length of queries and verified claims can play a large role. Very short, long, or noisy claims can lead to overfitting. We also find erroneous claims in the datasets due to the difficulties of automatic harvesting from semi-structured fact-checking portals that can potentially increase the task difficulty. However, this mirrors data for real claim retrieval applications. Thus, investigating the robustness of methods regarding noise can yield valuable insights into the applicability of systems for applications.

The tweets datasets are similar to one another, and so are the debates datasets. *2020-tweets* stands out regarding word overlaps: here, word overlaps can be expected to be a better feature to distinguish matches from non-matches than for other datasets. This dataset also has shorter test queries with low variance regarding their length but high variance regarding the length of the training queries. This dataset contains less information for the verified claims, e.g. no date field. As dates are appended to the input tweets, models extracting this data for matching with the targets are expected to yield better results for 2021 and 2022 tweets data than for 2020.

In all datasets, the IDs give away some information: 1) their ordering (succeeding input queries may link to the same target) and 2) info about dates, events, speakers that can be extracted from them and used for linking (debates). When comparing the performance of different systems, it should be checked whether they make use of such information to gain a more realistic assessment of their performance in real-world systems.

In real-world applications, there may be no correct match for an input query while in the given data, there always is one match per input query for the tweets datasets and at least one match per input query for the debates data. Nevertheless, the data can be used to develop methods that distinguish relevant from non-relevant claims. While the task is formulated as a ranking task, the gold data distinguishes relevant verified claims from irrelevant ones without ranking multiple valid matches.

3.6 Evaluation of Claim Retrieval Approaches

In this section, we present the evaluation of the claim retrieval systems following the experimental setup and research questions outlined above. In order to ensure comparability to the values specified in the literature, we use the original gold files including duplicates. However, we also measure the impact of the duplicates separately in order to reveal possible biases. While replicating the results for the systems, we found that the official results were generated using a fixed version of the *2020-tweets* gold file (cf. Section 3.5.1). We therefore use this fixed version for our comparative evaluation.

3.6.1 Performance on different datasets

Tables 3.19 and 3.20 detail the results of all approaches for training and testing on data of the same year and discourse type. Tables 3.14, 3.15, 3.16, 3.17 and 3.18 detail the performance by dataset. The MAP@k scores for *UofSheffield* remain constant for all values of k because this system returns only the top-1 matching verified claim per input query instead of a ranked list.

When only considering approaches that have participated in the challenge of a respective year (marked with $\star$), we can see that the top score was always achieved by a supervised method. Adding all approaches under investigation, we find that the unsupervised *SimBa 2023* outperforms the best supervised approach for *2020-tweets*, *BigIR* (cf. Table 3.14). The winner of the *2020-tweets* competition was *Buster.AI* with a score of 0.929. This approach ranks third in the overall comparison for this dataset. Surprising is the relatively weak performance of *RIET Lab*, which is the best-performing approach on all other datasets. This is due to one out of three runs where the model failed to learn maximally meaningful features: while the best run scored 0.957, the worst run achieved 0.801. For *2020-tweets*, but not the other datasets, two index-based approaches yield the top scores regarding precision (*BigIR* and *ES Baseline*), but not MAP, which includes recall. Also, the precision scores are generally higher than for the other datasets. The data analysis (Section 3.5) showed that *2020-tweets* features a relatively low token overlap of non-matching input queries and verified claims and a much higher overlap ratio in matching pairs. Thus, simple token matching techniques serve well to find matches with high precision. At the same time, absolute overlap ratios are lower than in the other tweet datasets, which hints at the presence of many pairs for which semantic similarity measures are required. In line with this, we find that models focusing strongly on embedding-based similarities, such as *SimBa 2023* and *Check_square: distmult*, yield both higher precision and MAP scores on *2020-tweets* than on the other datasets even though the *ES Baseline* has its highest performance on this dataset, closely followed by *2022-tweets*.

For *2021-tweets*, the winning approach is *RIET Lab* (0.956), followed by *SimBa 2023* (0.945), *BirIR* (0.936), and the extension of the winner of the 2021 competition, *Aschern* (0.883): CrowdChecked (0.903). For the latest 2022 edition, the best performance on the tweets data is again achieved by *RIET Lab* with a score of 0.956 (cf. Table 3.16). *SimBa 2023* ranks second with 0.926, achieving comparable results to *AI Rational* (0.922) and *BigIR* (0.921). To the best of our knowledge, none of the approaches compared here uses information from the IDs as features, except *DIPS*, which extracts the dates from the input claims Mihaylova et al. (2021b), and no approach exploits the order of input queries. Except for *SimBa 2023*, all approaches in the top 5 for the tweets data are supervised.

This is different for the debates data, where most unsupervised approaches outperform the supervised ones (cf. Tables 3.17 and 3.18), with the exception of *RIET Lab*, which is the best-performing approach for both 2021 and 2022. The unsupervised *SimBa 2023* outperforms the supervised *SimBa debates* system of 2022. In line with the observations from the data analysis (Section 3.5), claim retrieval is harder for the debates data than for the tweets data with the best-performing method achieving scores of $\sim$0.5 while for the English tweets, the top scores reach $\sim$0.96 MAP@5. The difficulty of the *2021-debates* and *2022-debates* test sets is similar, with the latter being easier with MAP@5 scores ranging from 0.238 - 0.501 vs. 0.196 - 0.480 for *2021-debates*.

The *ES Baseline* performs much lower than on the tweets data, yet the index-based approaches *ClaimLinker*, *TIET*, and *ES Baseline* work relatively well for the debates data compared to other systems. Their scores are generally similar to one another, except for *2020-tweets*. As detailed in Section 3.5, the difference between token/lemma overlap ratios in the negative vs. positive examples of the test set is the highest for this dataset. Thus, text overlap features seem to be particularly beneficial, which may explain *ClaimLinker*'s performance bonus compared to the *ES Baseline*.

Across datasets, the performance of the supervised approaches fluctuates considerably, even within the tweet datasets. Striking is the performance drop for *UofSheffield* for *2021-tweets*, and, to a lesser degree, *2022-tweets* (cf. Table 3.20), *NLP&IR@UNED*'s drop for *2021-tweets* as well as *RIET Lab*'s lower performance and high fluctuation on *2020-tweets*. The highest standard deviation for the supervised approaches is obtained by *NLP&IR@UNED* for *2022-tweets* with a maximum MAP@5 score of 0.506, a minimum of 0.235 and a standard deviation of 0.112 for all three runs.

Our data analysis in Section 3.5 revealed many outliers regarding verified claim length. Investigating the predictions in more detail, we find that *NLP&IR@UNED* has a strong bias of classifying very long claims (i.e. claims with more than 34 and 42 tokens, which are the average outlier ranges for the tweets and debates datasets, respectively, cf. Section 3.5.4) as correct matches for input queries (see Table 3.12). To generate these numbers, we counted for each verified claim with excessive length how frequently their corresponding IDs occur at the top position and within the top-5 matches of the predictions files.

Dataset	2020-tweets		2021-tweets		2022-tweets		2021-deb		2022-deb	
Top-k	**1**	**5**	**1**	**5**	**1**	**5**	**1**	**5**	**1**	**5**
System										
Gold	**7**	**7**	**25**	**25**	**24**	**24**	**5**	**5**	**2**	**2**
ClaimLinker	5	10	22	33	20	32	6	17	4	11
TIET	7	27	26	107	30	109	5	37	3	28
SimBa 2023	7	21	24	77	26	98	6	28	3	23
UofSheffield	6	6	4	4	20	20	3	3	3	3
Check_square: distmult	3	29	25	111	27	127	8	26	5	21
Check_square run 1	6	32	15	63	19	77	5	27	6	17
Check_square run 2	3	32	21	70	18	71	6	24	7	21
Check_square run 3	3	29	18	77	17	74	6	21	6	19
NLP&IR@UNED run 1	15	176	79	382	150	719	5	17	14	87
NLP&IR@UNED run 2	8	69	90	539	174	746	8	73	6	39
NLP&IR@UNED run 3	31	341	89	381	192	804	4	40	8	56
RIET Lab run 1	6	30	25	85	25	92	10	27	6	21
RIET Lab run 2	8	30	26	106	26	96	9	26	5	19
RIET Lab run 3	3	26	25	87	25	94	12	29	6	20

Table 3.12: Number of verified claims with more than 34 (tweets) / 42 (debates) tokens in the systems' top-5 predictions in C for the respective test sets.

Dataset	**2020-tweets**		**2021-tweets**		**2022-tweets**	
Top-k	**1**	**5**	**1**	**5**	**1**	**5**
System						
ClaimLinker	**1**	**1**	0	0	0	0
TIET	0	**1**	0	0	0	0
SimBa 2023	0	0	0	**1**	0	0
UofSheffield	0	0	**7**	**7**	**2**	**2**
Check_square: distmult	**1**	**26**	0	0	0	**2**
Check_square run 1	0	0	0	0	0	**1**
Check_square run 2	0	0	**1**	**1**	0	**2**
Check_square run 3	0	0	0	0	0	**4**
RIET Lab run 1	0	0	0	0	0	**4**
RIET Lab run 2	0	0	0	0	0	**2**
RIET Lab run 3	0	0	0	0	0	**3**

Table 3.13: Number of verified claims with less than 3 tokens (excluding punctuation) in the systems' top-5 predictions in C for the respective test sets. Systems not including any of such claims are omitted in this table.

As revealed by Table 3.12, the counts vary heavily in between runs for *NLP&IR@UNED* and the counts are generally high. The other systems do not exhibit a similar bias, regardless of whether they use only vclaims or vclaims plus titles. All systems but *UofSheffield* rightfully include a smaller number of long claims in their predictions for *2020-tweets* than for the other tweet datasets. As outlined in Section 3.5, *2020-tweets* has considerably shorter test input queries with shorter matching verified claims, but the length of the matching training verified claims does not differ from the other tweets datasets. Note that *UofSheffield* only returns one claim per input query, and the gold files only list the top positions, thus their top-5 counts equal their top-1 counts. The erroneous and context-dependent short claims (cf. Section 3.5.4) also influence the results for some systems. As Table 3.13 shows, all supervised systems except *NLP&IR@UNED*, which is biased towards long claims, include some of the short claims in their predictions for *2020-tweets*. *UofSheffield* and *Check_square* additionally include them for *2021-tweets*. *UofSheffield*'s bias is particularly strong, as this system returns those claims at the top position in 9 cases. The unsupervised systems *Claim-Linker*, *TIET*, and *SimBa 2023* include only one short claim in total each, for either *2020-tweets* or *2021-tweets*. The unsupervised *Check_square: distmult* system returns a high number of short claims in its top-5 predictions for *2020-tweets* and two of them for *2022-tweets*. The difference to the supervised variant is the choice of embeddings: fine-tuned Sentence-BERT embeddings in *Check_square*, pre-trained multilingual DistilBERT embeddings in *Check_square: distmult*. While it overall achieves higher performance than the variant using fine-tuned BERT embeddings, the fine-tuning using triplets that include hard negatives with high cosine similarity seems to be effective in increasing robustness to this kind of noise.

This analysis supports the interpretation that the supervised approaches may learn spurious correlations in this data and may thus be sensitive to noise and less stable regarding their performance when applied to real data in claim retrieval systems than unsupervised approaches. At the same time, lower performance on the given data does not necessarily mean lower task performance. Instead, a system may be worse at picking up spurious correlations, as explained by Le Bras et al. (2020). We find hints for this in our comparison of *Check_square* using fine-tuned BERT vs. pre-trained DistilBERT embeddings and the greater preference of the latter for choosing long claims as matches for *2020-tweets*. A related observation was made by Skuczyńska et al. (2021) who found that fine-tuning using titles did not improve their system, but improved its robustness. In general, using the more abstractive titles can intuitively be expected to help a system deal with paraphrased input claims, and many systems use them in addition to the vclaim texts, cf. Table 3.3. However, the methodology of dataset collection might have introduced biases, e.g. fact-checkers listing tweets or statements they checked in their articles might be likely to phrase their summarizing claims in a similar way. In this case, the importance of abstractive titles as an additional information source might be pronounced in real-world data. However, the correlations present in the data may also reflect properties of real-world data, i.e. input queries and matching verified claims naturally share many tokens, such as entities mentioned in their most common form.

As the different behavior of the systems is not always visible using all of the datasets, this analysis highlights that testing on the different dataset is important to assess model performance and biases for this task.

Our results generally verify the observations made in the literature that the choice of language models for this task greatly impacts the results: For *Check_square*, the authors found on the *2021-tweets* data that multilingual DistilBERT embeddings without fine-tuning outperformed fine-tuned monolingual Sentence-BERT models. For the former, fine-tuning using their triplet loss methodology hurts performance while it helps in the case of Sentence-BERT Cheema et al. (2020). Similarly, Skuczyńska et al. (2021) found on the same dataset that their RoBERTa model fine-tuned on triplets performed worse than monolingual DistilBERT without fine-tuning. We could replicate that *Check_square*'s variant using multilingual DistilBERT embeddings without fine-tuning outperforms the fine-tuned BERT embeddings on all datasets but *2022-tweets*. For this dataset, both variants perform on par. Similarly, according to Shaar et al. (2020b), Sentence-BERT and Universal Sentence Encoder performed much better in their system and on their data than BERT, and Sentence-BERT better than Universal Sentence Encoder. In the same vain, Passaro et al. (2020) found on *2020-tweets* that using a standard BERT model to represent sentences resulted in low performance while pre-trained Sentence-BERT performed a lot higher and further benefited from fine-tuning. Mansour et al. (2022) found that BERT performs poorly in comparison with other models such as RoBERTa, MiniLM, or MPNet in their *BigIR* system, which they tested for *2020-tweets* and *2021-tweets*. They find MPNet to yield the best performance. MPNet also yielded the best performance in Hövelmeyer et al. (2022)'s analysis on the *2022-tweets* data, followed by SimCSE-

RoBERTA-large and Universal Sentence Encoder. InferSent's performance turned out to be considerably worse, both using fastText and GloVe. The two dominating systems, *RIET Lab* and *SimBa 2023*, both use sentence transformers for blocking: fine-tuned Sentence-T5-Large for *RIET Lab*, pre-trained MPNet-Base for *SimBa 2023*. Sentence-T5 is also used for re-ranking by *SimBa 2023* in its smaller pre-trained base variant. *Check_square* and *TIET* rely on standard Sentence-BERT embeddings. Their relatively weak performance is in line with these findings and can thus most likely be attributed to the language model choice. Consistent with this, their performance on *2021-tweets* is lower than the baseline RoBERTa performance shown by *NLytics*. Also, while *CrowdChecked* benefits from the additional training data (cf. Hardalov et al. (2022)), it relies on Sentence-BERT and is outperformed by approaches using better-performing language models. This supports the finding that the language model is the most important factor in solving this task, more important than acquiring additional training data and developing sophisticated training/fine-tuning strategies. The latter finding is also supported by the comparison of *SimBa 2023* and *RIET Lab*, whose performance is close, even though *RIET Lab* applies considerable training and fine-tuning effort on very large language models while *SimBa 2023* employs high-performing language models in their pre-trained base variants without any supervision. For *SimBa 2023* and its choice of language models combined with lexical features, we find no disadvantage regarding model robustness.

Finally, we repeated the analysis using the de-duplicated gold files. While the scores deviate for datasets with duplicates, *2021-debates* and *2022-debates*, cf. Section 3.5, e.g. on *2022-debates* 0.478 on the de-duplicated gold data vs. 0.501 for *RIET Lab* and 0.461 on the de-duplicated gold data vs. 0.481 for *SimBa 2023*, the order of the models regarding their performance and the differences between the datasets remain unchanged. We therefore omit details of these scores.

Rank	Sup.?	Method	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
1		SimBa 2023	0.945	0.328	0.197	**0.945**	**0.964**	**0.964**
2	✓	BigIR*	**0.950**	*	*	*	*	0.955
3	✓	Buster.AI*⋆	0.895	0.320	0.195	0.897	0.926	0.929
4	✓	UNIPI-NLE*⋆	0.875	0.315	0.193	0.877	0.907	0.912
5	✓	UB_ET*⋆	0.815	0.307	0.186	0.818	0.862	0.864
6	✓	RIET-Lab	0.827	0.297	0.179	0.827	0.856	0.857
7	✓	NLP&IR@UNED*⋆	0.805	0.300	0.185	0.807	0.851	0.856
8		NLP&IR@UNED	0.798	0.294	0.179	0.798	0.837	0.840
9		Check_square: distmult	0.795	0.292	0.181	0.795	0.829	0.836
10		ClaimLinker	0.810	0.283	0.170	0.810	0.828	0.828
11	✓	UofSheffield	0.815	0.272	0.163	0.815	0.815	0.815
12	✓	UofSheffield*⋆	0.805	0.270	0.162	0.807	0.807	0.807
13	✓	Check_square	0.758	0.280	0.173	0.758	0.797	0.802
14		TIET	0.745	0.278	0.180	0.745	0.785	0.800
15		ES Baseline	0.793	**0.344**	**0.217**	0.682	0.782	0.794
16		TIET*⋆	0.740	0.267	0.164	0.743	0.768	0.773
17	✓	Check_square*⋆	0.650	0.247	0.152	0.652	0.690	0.695
18		ES Baseline 2020*⋆	0.472	0.249	0.156	0.470	0.601	0.609

Table 3.14: Results for *2020-tweets*. Sup.? = Supervised/fine-tuned system? Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Systems marked with ⋆ participated in the challenge of that year

Rank	Sup.?	Method	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
1	✓	RIET-Lab	**0.926**	**0.329**	**0.199**	**0.926**	**0.955**	**0.956**
2		SimBa 2023	0.921	0.322	0.195	0.921	0.943	0.945
3	✓	BigIR*	0.911	*	*	*	*	0.936
4	✓	CrowdChecked*	*	*	*	*	*	0.903
5	✓	Aschern*⋆	0.861	0.300	0.182	0.861	0.880	0.883
6		NLytics*⋆	0.738	0.289	0.179	0.738	0.792	0.799
7	✓	DIPS*⋆	0.728	0.282	0.177	0.728	0.778	0.787
8		TIET	0.683	0.271	0.173	0.683	0.741	0.753
9		ES Baseline	0.703	0.262	0.164	0.703	0.741	0.749
9		ES Baseline 2021*⋆	0.703	0.262	0.164	0.703	0.741	0.749
10		ClaimLinker	0.708	0.256	0.159	0.708	0.735	0.742
11		Check_square: distmult	0.673	0.257	0.167	0.673	0.717	0.732
12	✓	Check_square	0.630	0.251	0.160	0.630	0.684	0.695
13	✓	NLP&IR@UNED	0.470	0.234	0.156	0.470	0.575	0.592
14	✓	UofSheffield	0.292	0.097	0.058	0.292	0.292	0.292

Table 3.15: Results for *2021-tweets*. Sup.? = Supervised/fine-tuned system? Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Systems marked with ⋆ participated in the challenge of that year

Rank	Sup.?	Method	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
1	✓	RIET-Lab*⋆	*	0.322	**0.194**	**0.943**	**0.955**	**0.956**
2	✓	RIET-Lab	**0.928**	**0.324**	**0.194**	0.928	0.948	0.948
3		SimBa 2023	0.904	0.319	0.191	0.904	0.926	0.926
4	✓	AI Rational*	*	0.313	0.190	0.904	0.919	0.922
5	✓	BigIR*⋆	*	0.316	0.189	0.900	0.921	0.921
6		SimBa (tweets)*⋆	*	0.314	0.190	0.876	0.905	0.907
7	✓	Fraunhofer SIT SVM*⋆	0.752	0.275	0.170	0.752	0.787	0.794
8		ES Baseline	0.756	0.276	0.168	0.756	0.789	0.791
9		ClaimLinker	0.751	0.265	0.161	0.751	0.771	0.773
10		TIET	0.708	0.274	0.174	0.708	0.760	0.771
11	✓	Check_square	0.705	0.266	0.167	0.705	0.745	0.754
12		Check_square: distmult	0.703	0.266	0.166	0.703	0.746	0.752
13	✓	UofSheffield	0.679	0.226	0.136	0.679	0.679	0.679
14	✓	Fraunhofer SIT*⋆	0.557	0.221	0.141	0.557	0.601	0.610
15	✓	NLP&IR@UNED	0.215	0.184	0.134	0.215	0.354	0.381

Table 3.16: Results for *2022-tweets*. Sup.? = Supervised/fine-tuned system? Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Systems marked with ⋆ participated in the challenge of that year

Rank	Sup.?	Method	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
1	✓	RIET-Lab	**0.414**	**0.205**	**0.127**	**0.414**	**0.475**	**0.480**
2		SimBa 2023	0.405	0.177	0.109	0.411	0.439	0.441
3		ClaimLinker	0.346	0.158	0.095	0.353	0.385	0.385
4		TIET	0.316	0.148	0.091	0.316	0.351	0.353
5		ES Baseline	0.304	0.144	0.091	0.304	0.339	0.346
5		ES Baseline 2021*	0.304	0.143	0.091	0.304	0.339	0.346
6	✓	DIPS*⋆	0.266	0.143	0.099	0.278	0.313	0.328
7	✓	Beasku*⋆	0.253	0.139	0.101	0.266	0.308	0.327
8		Check_square: distmult	0.266	0.131	0.089	0.259	0.295	0.304
9	✓	Check_square	0.236	0.105	0.074	0.224	0.247	0.259
10	✓	NLP&IR@UNED	0.181	0.101	0.066	0.171	0.216	0.221
11		NLytics*⋆	0.165	0.101	0.068	0.171	0.210	0.215
12	✓	UofSheffield	0.203	0.072	0.043	0.196	0.196	0.196

Table 3.17: Results for *2021-debates*. Sup.? = Supervised/fine-tuned system? Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Systems marked with ⋆ participated in the challenge of that year

Rank	Sup.?	Method	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
1	✓	RIET-Lab	0.441	**0.222**	**0.137**	0.415	**0.497**	**0.501**
2		SimBa 2023	**0.462**	0.190	0.120	**0.446**	0.474	0.481
3	✓	SimBa (debates)*⋆	*	0.190	0.126	0.408	0.446	0.459
4		ClaimLinker	0.385	0.169	0.105	0.369	0.408	0.411
5		TIET	0.338	0.154	0.095	0.323	0.365	0.367
6		ES Baseline	0.323	0.154	0.099	0.308	0.353	0.361
7		Check_square: distmult	0.323	0.149	0.098	0.300	0.336	0.347
8	✓	Check_square	0.277	0.135	0.084	0.249	0.298	0.300
9	✓	NLP&IR@UNED	0.226	0.113	0.082	0.215	0.249	0.261
10	✓	UofSheffield	0.246	0.087	0.052	0.238	0.238	0.238

Table 3.18: Results for *2022-debates*. Sup.? = Supervised/fine-tuned system? Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Systems marked with ⋆ participated in the challenge of that year

Method	Dataset	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
Check_square: distmult	2020-tweets	0.790	0.292	0.181	0.792	0.827	0.834
	2021-tweets	0.673	0.257	0.167	0.673	0.717	0.732
	2022-tweets	0.703	0.266	0.166	0.703	0.746	0.752
	2021-debates	0.266	0.131	0.089	0.259	0.295	0.304
	2022-debates	0.323	0.149	0.098	0.300	0.336	0.347
SimBa 2023	**2020-tweets**	**0.940**	0.328	0.197	**0.943**	**0.962**	**0.962**
	2021-tweets	**0.921**	**0.322**	**0.195**	**0.921**	**0.943**	**0.945**
	2022-tweets	**0.904**	**0.319**	**0.191**	**0.904**	**0.926**	**0.926**
	2021-debates	**0.405**	**0.177**	**0.109**	**0.411**	**0.439**	**0.441**
	2022-debates	**0.462**	**0.190**	**0.120**	**0.446**	**0.474**	**0.481**
SimBa (tweets)*	2022-tweets*	*	0.314	0.190	0.876	0.905	0.907
ClaimLinker	2020-tweets	0.800	0.282	0.170	0.800	0.820	0.820
	2021-tweets	0.693	0.256	0.160	0.693	0.727	0.735
	2022-tweets	0.742	0.265	0.161	0.742	0.766	0.768
	2021-debates	0.333	0.124	0.074	0.333	0.346	0.346
	2022-debates	0.323	0.133	0.083	0.323	0.362	0.365
TIET	2020-tweets	0.745	0.278	0.180	0.745	0.785	0.800
	2021-tweets	0.683	0.271	0.173	0.683	0.741	0.753
	2022-tweets	0.708	0.274	0.174	0.708	0.760	0.771
	2021-debates	0.316	0.148	0.091	0.316	0.351	0.353
	2022-debates	0.338	0.154	0.095	0.323	0.365	0.367
NLytics*	2021-tweets*	0.738	0.289	0.179	0.738	0.792	0.799
	2021-debates	0.165	0.101	0.068	0.171	0.210	0.215
ES Baseline	2020-tweets	0.793	0.344	0.217	0.682	0.782	0.794
	2021-tweets*	0.703	0.262	0.164	0.703	0.741	0.749
	2021-tweets	0.703	0.262	0.164	0.703	0.741	0.749
	2022-tweets	0.756	0.276	0.168	0.756	0.789	0.791
	2021-debates*	0.304	0.143	0.091	0.304	0.339	0.346
	2021-debates	0.304	0.144	0.091	0.304	0.339	0.346
	2022-debates	0.323	0.154	0.099	0.308	0.353	0.361
ES Baseline 2020*	2020-tweets	0.472	0.249	0.156	0.470	0.601	0.609

Table 3.19: Performance of the unsupervised approaches on the different datasets. Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Boldface marks the best scores on a dataset for the supervised approaches, underlined scores are the best scores for all approaches, including the supervised ones (cf. Table 3.20). Blue color marks the best scores of an approach across datasets.

Method	Dataset	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
Check_square	2020-tweets*	0.650	0.247	0.152	0.652	0.690	0.695
	2020-tweets	0.753	0.280	0.173	0.756	0.794	0.799
	2021-tweets	0.630	0.251	0.160	0.630	0.684	0.695
	2022-tweets	0.705	0.266	0.167	0.705	0.745	0.754
	2021-debates	0.236	0.105	0.074	0.224	0.247	0.259
	2022-debates	0.277	0.135	0.084	0.249	0.298	0.300
RIET-Lab	2020-tweets	0.822	0.297	0.179	0.824	0.853	0.854
	2021-tweets	**0.926**	**0.329**	**0.199**	**0.926**	**0.955**	**0.956**
	2022-tweets*	*	0.322	**0.194**	**0.943**	**0.955**	**0.956**
	2022-tweets	**0.928**	**0.324**	**0.194**	0.928	0.948	0.948
	2021-debates	**0.414**	**0.205**	**0.127**	**0.414**	**0.475**	**0.480**
	2022-debates	**0.441**	**0.222**	**0.137**	**0.415**	**0.497**	**0.501**
UofSheffield	2020-tweets*	0.805	0.270	0.162	0.807	0.807	0.807
	2020-tweets	0.810	0.272	0.163	0.812	0.812	0.812
	2021-tweets	0.292	0.097	0.058	0.292	0.292	0.292
	2022-tweets	0.679	0.226	0.136	0.679	0.679	0.679
	2021-debates	0.203	0.072	0.043	0.196	0.196	0.196
	2022-debates	0.246	0.087	0.052	0.238	0.238	0.238
NLP&IR@UNED	2020-tweets*	0.805	0.300	0.185	0.807	0.851	0.856
	2020-tweets	0.795	0.294	0.179	0.797	0.835	0.838
	2021-tweets	0.470	0.234	0.156	0.470	0.575	0.592
	2022-tweets	0.215	0.184	0.134	0.215	0.354	0.381
	2021-debates	0.181	0.101	0.066	0.171	0.216	0.221
	2022-debates	0.226	0.113	0.082	0.215	0.249	0.261
BigIR*	**2020-tweets***	**0.950**	*	*	*	*	**0.955**
	2021-tweets*	0.911	*	*	*	*	0.936
	2022-tweets*	*	0.316	0.189	0.900	0.921	0.921
SimBa (debates)*	2022-debates*	*	0.190	0.126	0.408	0.446	0.459
AI Rational*	2022-tweets*	*	0.313	0.190	0.904	0.919	0.922
Buster.AI*	2020-tweets*	0.895	0.320	0.195	0.897	0.926	0.929
UNIPI-NLE*	2020-tweets*	0.875	0.315	0.193	0.877	0.907	0.912
UB_ET*	2020-tweets*	0.815	0.307	0.186	0.818	0.862	0.864
CrowdChecked*	2021-tweets	*	*	*	*	*	0.903
Aschern*	2021-tweets*	0.861	0.300	0.182	0.861	0.880	0.883
DIPS*	2021-tweets*	0.728	0.282	0.177	0.728	0.778	0.787
	2021-debates*	0.266	0.143	0.099	0.278	0.313	0.328
Beasku*	2021-debates*	0.253	0.139	0.101	0.266	0.308	0.327
Fraunhofer SIT*	2022-tweets*	0.557	0.221	0.141	0.557	0.601	0.610
Fraunhofer SIT SVM*	2022-tweets*	0.752	0.275	0.170	0.752	0.787	0.794

Table 3.20: Performance of the supervised approaches on the different datasets. Numbers for approaches marked with * have been taken from the respective publications. Scores not detailed therein are marked with * in the respective columns. All other numbers we generated/replicated for this study. Boldface marks the best scores on a dataset for the supervised approaches, underlined scores are the best scores for all approaches, including the unsupervised ones (cf. Table 3.19). Blue color marks the best scores of an approach across datasets.

3.6.2 Cross-dataset evaluation

The cross-evaluation results are displayed in Tables 3.21 and 3.22. Since we reported averaged results in the previous subsection, the performance of a trained model on the test dataset of the same type and year does not correspond to the values in Table 3.20. We chose a random run (e.g. always the model trained in the last run) rather than selecting a specific model. As the results show, training an approach on a dataset of a different year or type does not necessarily hurt performance. Instead, we find that models that perform well on one dataset perform well on others, too. E.g. NLP&IR@UNED and UofSheffield learnt their most useful model on *2020-tweets*. This yields the best results on all datasets. The same can be observed for RIET Lab with its models for *2021-tweets* and *2022-tweets*. Check_square learnt its least useful model on *2021-debates* which scores worse than all models learnt on any other dataset for all test data, including *2022-debates* which contains data of the same discourse type. Together with the insights from the previous subsection which showed that individual training runs led to great variation in performance, this finding highlights that the supervised approaches are vulnerable to overfitting, making their performance hard to predict. However, when a model is learnt successfully for this task, it generalizes well across data of different discourse types. Training on *2020-tweets* generally yields the best results, with the exception of RIET Lab and Check_square, for which the model trained on 2022-tweets achieves better scores on the *2020-tweets* test set and the models learnt on *2021-tweets* and on *2020-tweets* perform on par on the *2021-debates* test set. As outlined in Section 3.5, tweet datasets have a higher number of positive training examples than the debates datasets and a smaller number of verified claims to match against. At the same time, retrieval of matching pairs depends largely on token overlaps for both the tweets and the debates data. These characteristics explain why models learnt on the tweets data perform better than models learnt on the debates datasets, even when applied to data of the other discourse type. *2020-tweets* has a smaller training set than the other tweet datasets, but also a smaller number of non-matching verified claims, i.e. negative training examples, and the ratio of positive vs. negative examples has a lesser bias towards negative examples. Unless a model overfits the data, training on *2020-tweets* thus yields the best models.

To control for biases introduced by the duplicates in the gold files, we repeated the evaluation with the de-duplicated versions. As in the dataset-wise comparison in the previous subsection, we found no biases and omit the scores.

3.6.3 Efficiency

The systems we compare exhibit the following runtime complexities for the online computing steps, with D denoting the size of the verified claim database, i.e. the number of verified candidate claims to match against, and $C\{n\}$ unknown constants, see Table 3.23.

Note that using the Sentence Transformer, embeddings are projected to Euclidean space which enables search with $O(i+D)$ inferences with i being the number of input claims, instead of $O(i*D)$ pair-wise comparisons (cf. Shliselberg and Dori-Hacohen (2022)).

For real applications, especially web-based ones, the magnitudes of the constants matter.

As detailed in Subsections 3.6.1 and 3.6.2, RIET Lab is the approach with the highest performance on most datasets but has very high online and offline computational costs for training and fine-tuning and requires the availability of a high-performance GPU server. SimBa 2023 achieves competitive scores, and its offline computations are relatively cheap, consisting only of computing embeddings using pre-trained language models and other textual features for the verified claims.

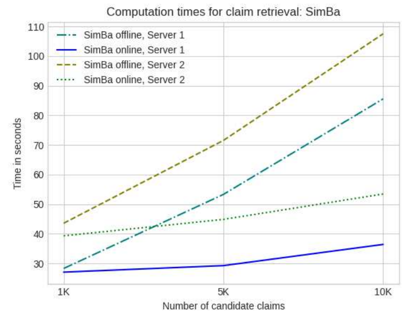
In order to answer RQ5, we now investigate the question whether the two best performing approaches, supervised RIET Lab and unsupervised SimBa 2023, are suitable for online fact-checking applications where response times matter. We experimentally measure their offline computation and online retrieval times.

Figure 3.8 shows the time required to link the set of 200 claims. Table 3.24 details the average time required to retrieve verified claims for one input claim, depending on dataset size, illustrating both total runtimes and scalability.

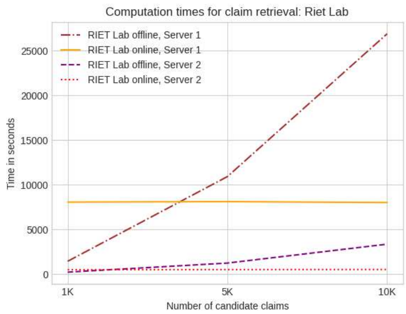
The offline computations for *RIET Lab* consist of training of the transformer-based candidate selection blocking step, and fine-tuning the GPT-Neo 1.3 Billion parameter

Method	Dataset (training → test)	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
Check_square	2020-tweets → 2020-tweets	0.755	0.277	0.170	0.757	0.789	0.794
	2021-tweets → 2020-tweets	0.735	0.272	0.170	0.738	0.772	0.780
	2022-tweets → 2020-tweets	0.795	0.290	0.177	0.797	0.829	0.832
	2021-debates → 2020-tweets	0.640	0.245	0.155	0.642	0.683	0.692
	2020-tweets → 2021-tweets	0.733	0.274	0.172	0.733	0.771	0.781
	2021-tweets → 2021-tweets	0.619	0.231	0.151	0.619	0.653	0.667
	2022-tweets → 2021-tweets	0.663	0.248	0.157	0.663	0.701	0.712
	2021-debates → 2021-tweets	0.475	0.203	0.131	0.475	0.538	0.548
	2020-tweets → 2022-tweets	0.727	0.274	0.168	0.727	0.770	0.775
	2021-tweets → 2022-tweets	0.679	0.255	0.157	0.679	0.718	0.722
	2022-tweets → 2022-tweets	0.713	0.276	0.171	0.713	0.760	0.766
	2021-debates → 2022-tweets	0.560	0.220	0.138	0.560	0.605	0.612
	2020-tweets → 2021-debates	0.253	0.143	0.091	0.247	0.304	0.309
	2021-tweets → 2021-debates	0.253	0.143	0.091	0.247	0.304	0.310
	2022-tweets → 2021-debates	0.241	0.135	0.089	0.247	0.293	0.302
	2021-debates → 2021-debates	0.228	0.110	0.073	0.215	0.249	0.258
	2020-tweets → 2022-debates	0.292	0.159	0.102	0.277	0.337	0.343
	2021-tweets → 2022-debates	0.292	0.154	0.102	0.277	0.331	0.342
	2022-tweets → 2022-debates	0.277	0.149	0.105	0.277	0.322	0.340
	2021-debates → 2022-debates	0.262	0.128	0.086	0.231	0.279	0.291
RIET-Lab	2020-tweets → 2020-tweets	0.765	0.285	0.171	0.767	0.805	0.805
	2021-tweets → 2020-tweets	0.940	**0.332**	**0.199**	0.943	**0.967**	0.967
	2022-tweets → 2020-tweets	**0.945**	0.330	**0.199**	**0.948**	**0.967**	**0.968**
	2021-debates → 2020-tweets	0.910	**0.332**	**0.199**	0.912	0.950	0.950
	2020-tweets → 2021-tweets	0.881	0.317	0.195	0.881	0.913	0.919
	2021-tweets → 2021-tweets	0.926	0.330	**0.199**	0.926	0.956	0.958
	2022-tweets → 2021-tweets	0.926	**0.332**	**0.199**	0.926	0.958	0.958
	2021-debates → 2021-tweets	0.871	0.315	0.190	0.871	0.908	0.909
	2020-tweets → 2022-tweets	0.919	0.322	0.195	0.919	0.942	0.944
	2021-tweets → 2022-tweets	0.933	0.324	0.194	0.933	0.951	0.951
	2022-tweets → 2022-tweets	0.928	0.324	0.194	0.928	0.946	0.946
	2021-debates → 2022-tweets	0.866	0.308	0.186	0.866	0.892	0.893
	2020-tweets → 2021-debates	0.418	0.181	0.119	0.411	0.444	0.454
	2021-tweets → 2021-debates	0.481	0.207	0.124	0.481	0.514	0.514
	2022-tweets → 2021-debates	0.468	0.207	0.127	0.468	0.507	0.510
	2021-debates → 2021-debates	0.418	0.207	0.129	0.418	0.482	0.488
	2020-tweets → 2022-debates	0.492	0.205	0.135	0.462	0.499	0.511
	2021-tweets → 2022-debates	0.554	0.231	0.138	0.531	0.573	0.573
	2022-tweets → 2022-debates	0.538	0.231	0.142	0.515	0.563	0.566
	2021-debates → 2022-debates	0.477	0.231	0.142	0.454	0.529	0.533

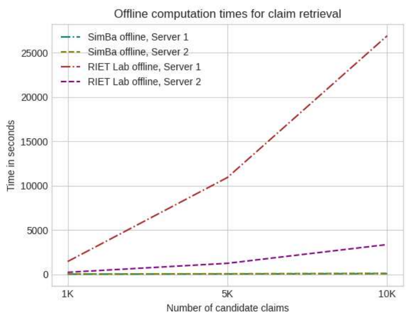
Table 3.21: Cross-evaluation results for supervision and fine-tuning. Boldface marks the best score on a test dataset across all approaches (including those listed in Table 3.22), underline the best score across all datasets and approaches, blue color the best scores for an approach for each test dataset.



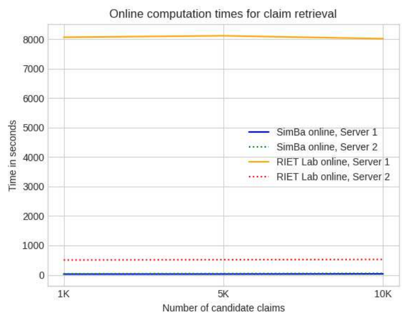
(a) Average retrieval runtime (in seconds) per input claim for different dataset sizes: SimBa 2023



(b) Average retrieval runtime (in seconds) per input claim for different dataset sizes: RIET Lab



(c) Average offline retrieval runtime (in seconds) per input claim for different dataset sizes



(d) Average online retrieval runtime (in seconds) per input claim for different dataset sizes

Figure 3.8: Offline and online runtimes to retrieve targets for 200 input queries for *SimBa 2023* and *RIET Lab*.

Method	Dataset (training → test)	P@1	P@3	P@5	MAP@1	MAP@3	MAP@5
UofSheffield	2020-tweets → 2020-tweets	0.810	0.272	0.163	0.812	0.812	0.812
	2021-tweets → 2020-tweets	0.640	0.215	0.129	0.642	0.642	0.642
	2022-tweets → 2020-tweets	0.775	0.260	0.156	0.777	0.777	0.777
	2021-debates → 2020-tweets	0.795	0.267	0.160	0.797	0.797	0.797
	2020-tweets → 2021-tweets	0.752	0.251	0.150	0.752	0.752	0.752
	2021-tweets → 2021-tweets	0.292	0.097	0.058	0.292	0.292	0.292
	2022-tweets → 2021-tweets	0.698	0.233	0.140	0.698	0.698	0.698
	2021-debates → 2021-tweets	0.723	0.241	0.145	0.723	0.723	0.723
	2020-tweets → 2022-tweets	0.742	0.247	0.148	0.742	0.742	0.742
	2021-tweets → 2022-tweets	0.301	0.100	0.060	0.301	0.301	0.301
	2022-tweets → 2022-tweets	0.679	0.226	0.136	0.679	0.679	0.679
	2021-debates → 2022-tweets	0.689	0.230	0.138	0.689	0.689	0.689
	2020-tweets → 2021-debates	0.304	0.118	0.071	0.310	0.310	0.310
	2021-tweets → 2021-debates	0.228	0.084	0.051	0.228	0.228	0.228
	2022-tweets → 2021-debates	0.278	0.105	0.063	0.278	0.278	0.278
	2021-debates → 2021-debates	0.203	0.072	0.043	0.196	0.196	0.196
	2020-tweets → 2022-debates	0.338	0.123	0.074	0.331	0.331	0.331
	2021-tweets →2022-debates	0.262	0.092	0.055	0.254	0.254	0.254
	2022-tweets → 2022-debates	0.338	0.123	0.074	0.323	0.323	0.323
	2021-debates → 2022-debates	0.246	0.087	0.052	0.238	0.238	0.238
NLP&IR@UNED	2020-tweets → 2020-tweets	0.805	0.297	0.180	0.807	0.843	0.846
	2021-tweets → 2020-tweets	0.445	0.187	0.152	0.445	0.497	0.543
	2022-tweets → 2020-tweets	0.070	0.063	0.107	0.070	0.118	0.199
	2021-debates → 2020-tweets	0.690	0.277	0.171	0.690	0.752	0.758
	2020-tweets → 2021-tweets	0.713	0.285	0.180	0.713	0.776	0.787
	2021-tweets → 2021-tweets	0.475	0.234	0.160	0.475	0.579	0.602
	2022-tweets → 2021-tweets	0.307	0.205	0.145	0.307	0.446	0.471
	2021-debates → 2021-tweets	0.569	0.219	0.137	0.569	0.611	0.617
	2020-tweets →2022-tweets	0.842	0.292	0.179	0.842	0.858	0.862
	2021-tweets → 2022-tweets	0.584	0.273	0.171	0.584	0.691	0.700
	2022-tweets → 2022-tweets	0.344	0.223	0.152	0.344	0.485	0.506
	2021-debates → 2022-tweets	0.612	0.220	0.134	0.612	0.635	0.637
	2020-tweets → 2021-debates	0.253	0.122	0.086	0.234	0.272	0.283
	2021-tweets → 2021-debates	0.215	0.089	0.061	0.196	0.213	0.219
	2022-tweets → 2021-debates	0.152	0.068	0.043	0.146	0.156	0.159
	2021-debates → 2021-debates	0.215	0.105	0.071	0.203	0.241	0.248
	2020-tweets → 2022-debates	0.308	0.149	0.105	0.269	0.323	0.337
	2021-tweets → 2022-debates	0.277	0.118	0.080	0.246	0.274	0.282
	2022-tweets → 2022-debates	0.185	0.082	0.055	0.177	0.190	0.195
	2021-debates → 2022-debates	0.262	0.123	0.083	0.231	0.274	0.284

Table 3.22: Cross-evaluation results for supervision and fine-tuning. Boldface marks the best score on a test dataset across all approaches (including those listed in Table 3.21), underline the best score across all datasets and approaches, blue color the best scores for an approach for each test dataset.

model as reranker. For *SimBa 2023*, offline computations refer to the generation of three different embeddings plus tokenization for all verified claims. As shown in Figure 3.8a and Table 3.24, the transformer-based blocking step in *RIET Lab* is more

Level	System	Runtime	Constant costs
$O(1)$	ES Baseline	$C1 * f(1)$	$C1$ = ES look-up
$O(k)$	ClaimLinker	$C2 * f(1) + C3 * f(k)$ with $k = max(30, D)$ and $C2 < C3$	$C2$ = ES look-up, $C3$ = pre-processing and textual similarity features
	TIET	$C4 * f(1) + C5 * f(k)$ with $k = max(10000, D)$ and $C4 < C5$	$C4$ = ES look-up, $C5$ = embedding-based similarities
$O(D)$*	RIET Lab	$C6 * f(D) + C7 * f(k)$ with $k = max(25, D)$ and $C6 < C7$	$C6$ = inference using sentence transformer, $C7$ = GPTNeo-based reranking
	SimBa 2023	$C8 * f(D) + C9 * f(k)$ with $k = max(50, D)$ and $C8 < C9$	$C8$ = embedding-based similarities, $C9$ = 3 embedding-based similarities plus lexical text overlap
$O(D)$	UofSheffield	$C10 * f(D)$	$C10$ = combination of BM25 and TF-IDF scores, pre-processing, cosine similarities
	Check_square	$C11 * f(D)$	$C11$ = embedding-based similarities
	NLP&IR@UNED	$C12 * f(D)$	$C12$ = embedding-based similarities, pre-processing, textual similarity features, inference using neural network

Table 3.23: Runtime complexities. Complexity levels: $O(1)$ - Constant runtime, $O(k)$ - Linear runtime w.r.t. k (constant iff $k < D$), $O(D)$* - Linear runtime with reduced C = reduced slope, $O(D)$ - Linear runtime.

efficient than *SimBa 2023*'s pair-wise comparisons regarding complexity, but *RIET Lab's* constant costs are more than 200 times higher on Server1 and ∼10 times higher on Server2 for a database size of 10k. The differences regarding the runtimes for Server1 vs. Server2 are due to the fact that *RIET Lab* benefits greatly from GPUs, but the GPUs on Server1 do not meet the requirements regarding available memory and thus the computations were performed on CPUs only. *SimBA 2023*, on the other hand, is able to use all GPUs and CPUs on both servers for its computations. Offline computations (i.e. updates of the model) are very costly for *RIET Lab*, while for *SimBa 2023*, they can even be computed on a laptop. As Table 3.24 reveals, *SimBa 2023* is faster on a laptop than *RIET Lab* is on a high-performance GPU server. Online retrieval times for *SimBa 2023* are fast enough for online applications on all hardware setups. For *RIET Lab*, a server with a high-capacity GPU is required.

Method	Hardware	Time (1K claims)	Time (5K claims)	Time (10K claims)
RIET Lab (offline)	Server 1	7.243	54.594	134.444
RIET Lab (online)	Server 1	38.659	39,651	41,270
SimBa 2023 (offline)	Server 1	0.141	0.266	0.428
SimBa 2023 (online)	Server 1	0.135	0.146	0.182
RIET Lab (offline)	Server 2	1.174	6.238	16.814
RIET Lab (online)	Server 2	2.533	2.579	2.631
SimBa 2023 (offline)	Server 2	0.218	0.358	0.538
SimBa 2023 (online)	Server 2	0.197	0.224	0.267
SimBa 2023 (offline)	Laptop	1.131	4.845	9.385
SimBa 2023 (online)	Laptop	0.324	0.335	0.380

Table 3.24: Average retrieval runtime per claim (in seconds) for different dataset (D) sizes.

3.6.4 Discussion

To answer RQ1, our data analysis in Section 3.5 reveals that the benchmark datasets for the claim retrieval task are characterized by a high rate of token overlaps, which explains the relatively high performance of baseline methods relying on BM25 and other lexical similarity scores, as well as the dominance of embedding-based methods (cf. Section 4.6). As the benchmarks are built using real data, they contain noise, such as very long and very short claims, partly due to extraction errors. As this reveals vulnerabilities of methods regarding overfitting and allows to better measure their robustness, we argue that this actually makes the data more useful to assess the suitability of methods in real-world systems, where they will be suspected to this kind of noise as well. Data from different discourse types, tweets, and political debates, are similar regarding characteristics such as ratio of token overlaps. The increased task difficulty for the debates data can to a large part be attributed to the lacking context information in its input queries, a challenge that would typically not arise for our use-case, a real-world claim retrieval system with queries entered by users. Therefore, these different datasets are not very well suited to investigate the general generalizability of models across different data distributions. The tweets datasets are limited in the sense that there is always one matching verified claim for each input query, while in real systems drawing on multiple fact-checking portals, we can expect anything from zero to multiple potential matching verified claims, since different fact-checking portals may check the same check-worthy claims. The debates data loosens these restrictions, allowing multiple matching targets. However, this data contains input queries that are part of check-worthy claims, but do not represent check-worthy claims themselves and are not the type of input that is excepted in claim retrieval-applications. Therefore, training and testing models on this data may not give insights into their ability to detect the essential, check-worthy part of verified claims, but instead limits the task to detecting overlapping parts in input queries and verified claims. With this, the data does not allow to distinguish methods that may differ regarding this aspect. However, as users are not expected to enter any non-check-worthy claims, this may not impede the performance of models trained on this data. Finally, whether the high ratio of overlaps and the small impact of using the more abstractive title information can be attributed to data selection biases or is inherent to the claim retrieval task, remains unclear.

Investigating RQ2, we find that the highest performance is achieved by supervised methods on all datasets but *2020-tweets* (cf. Tables 3.19 and 3.20). The best-performing approach overall is transformer- and GPTNeo-based *RIET Lab*, which achieves up to 0.956 MAP@5 on *2022-tweets*, followed by the unsupervised *SimBa 2023*, which achieves the top score on *2020-tweets* with 0.962. While supervised/fine-tuned systems also generally dominated the leaderboards of the CLEF CheckThat! Lab challenges, their performance turned out to fluctuate considerably between and within different datasets and is hard to predict, e.g. with *RIET Lab* performing ∼0.1 points worse on *2020-tweets* than on the other tweet datasets, with individual runs ranging from 0.801 to 0.957 in MAP@5 and *UofSheffield*'s performance diverging by ∼0.5 points between their *2020-tweets* and *2021-tweets* models (cf. Tables 3.19), despite the similarity of the datasets (cf. Section 3.4). As our cross-evaluation results in Tables 3.21 and 3.22 show, this difference is caused by what the model learned rather than by differences in the test sets. This and our analyses showing that noisy data may greatly bias the systems' performance (cf. Tables 3.12 and 3.13) highlight the importance of the availability of large amounts of training data to avoid overfitting, especially in the presence of noisy data. Acquiring such data, however, is costly, and so are training and fine-tuning using large language models. Unsupervised methods reach a similar or even better performance, with *SimBa 2023* achieving the top rank on the *2020-tweets* data with a MAP@5 score of 0.964 and most unsupervised methods outperforming most supervised ones, except *RIET Lab*, on the debates datasets (cf. Tables 3.17 and 3.18). *SimBa 2023* is the best unsupervised method, ranking second behind *RIET Lab* on all datasets except *2020-tweets*. On this dataset, it achieves the top score (cf. Tables 3.14, 3.15, 3.16, 3.17, 3.18).

Regarding RQ3, our results show that supervised methods for this task generalize well across the used datasets of different discourse types, i.e. models learned on tweets data work well for retrieving claims from debates and vice versa. However, as the models are prone to overfitting, their performance depends heavily on the training data and fluctuates strongly within the same discourse type (cf. Tables 3.21 and Tables 3.22): E.g. *NLP&IR@UNED* and *UofSheffield* learned their most useful model on *2020-tweets* which then yields the best results on all datasets, including those featuring claims from political debates. In general, training models on the *2020-tweets* dataset yields the best-performing models for most systems. This is the dataset where token overlaps are the most indicative to distinguish matching and non-matching pairs, the variance in input query length is highest in the training set and lowest in the test set. As the datasets are relatively small, however, supervised methods may overfit their models: e.g. *RIET Lab*'s worst and best model trained on *2020-tweets* yields a MAP@5 score of 0.801 vs. 0.957, respectively. The debates data features fewer positive training examples than the tweets data, shorter inputs (i.e. less information), and the imbalance of negative vs. positive examples is pronounced. Training on this data yields the lowest performance. Finally, the cross-evaluation confirms that not only the training data for the debates data is more difficult, but also the test sets are harder than those of the tweets datasets: models trained on the tweets data that perform well across datasets still yield lower scores on the debates test datasets. These have a higher number of non-matching verified claims to choose from, multiple correct matches, the input queries contain less words and characters than the tweet input queries and are less self-contained (cf. Section 3.5).

Section 3.6.3, Figure 3.8 and Table 3.24 answer RQ4: methods relying on an index structure naturally exhibit constant low retrieval time. The use of blocking techniques can render approaches using costly computations competitive in regards to efficiency without yielding bad performance (cf. Table 3.19).

Efficiency can be greatly improved by performing costly computations for all claims in the dataset offline, storing them for comparisons with new input claims, and computing the costly features only for the latter. In combination with blocking, this can cap and reduce the retrieval time of approaches making heavy use of costly features. We show that even though the best approach *RIET Lab*'s online retrieval has higher efficiency than the second best approach *SimBa 2023*'s in terms of run time complexity, the absolute costs of the former are too high to allow its application for online claim retrieval. Also, its offline computational costs are high and it requires the availability of a high-end server both for training and inference. Thus, to answer RQ5, our analysis suggests that the unsupervised *SimBa 2023* is well-suited for online fact-checking applications, with very low requirements regarding hardware as it can even be run on a laptop, while its performance is close to *RIET Lab*'s (+0.108, -0.011, -0.008, -0.039, -0.019 on *2020-tweets*, *2021-tweets*, *2022-tweets*, *2021-debates* and *2022-debates*, respectively), and, as outlined above, generally more stable. From this, we conclude that costly fine-tuning and supervision of neural models is not necessary for the task of retrieving fact-checked claims and that an unsupervised method such as *SimBa 2023* that exploits pre-trained language models such as Sentence-T5 and MPNet is overall the best choice for (online) claim retrieval systems.

3.7 Conclusions and Future Work

We presented an analysis of benchmark datasets for the claim retrieval task and an evaluation of the performance of different supervised and unsupervised approaches on different data to compare their robustness and scalability and extract insights on which methods work best and why.

We showed that an unsupervised system that exploits the complementarity of different embeddings achieves state-of-the-art performance on par with the best supervised method and outperforms all others while being highly efficient, scalable and robust.

Our study has several limitations: we focused on the retrieval performance for the English language. In future work, it would be interesting to compare the performance when applied to different languages and in multilingual settings.

Moreover, the usefulness of different language models was shown to vary considerably. It remains unclear which properties exactly render a model useful for this task and how biases introduced by language models (cf. Bender et al. (2021)) affect claim retrieval, e.g. yielding higher performance for certain topics or claims expressing stances uttered with preference by advocates of particular political leanings.

Thus, as part of our future work, we plan to investigate in more detail the different semantics that are encoded in different language models and their informative value and biases for this task.

Despite the fact that the datasets belonging to the same discourse type, tweets vs. political debates, are continuous extensions of their previous versions, we found relatively large fluctuations in systems' performance scores when using either of them. At the same time, the debate data differs from the tweets data in more aspects than the genre and features input queries that do not match the task definition entirely. Regarding characteristics such as the importance of word overlaps, the tweets and debates datasets are similar. Thus, our comparative analysis provides insights on the performance of the different methods on different datasets and their robustness, but cannot provide insights on the generalizability of methods regarding application on different discourse types or out-of-distribution data more generally. Relating to this, our data analysis investigated several characteristics that have been recognized as important factors in previous works and in a top-down fashion drawing on domain insights. Yet, the data may contain additional biases introduced by the methodology of dataset creation, which may not be captured by our analysis Le Bras et al. (2020) and which may artificially reduce task difficulty Kiela et al. (2021). Finally, the data diverges from data that can be expected in real-world claim retrieval systems in the sense that it does not feature ranked lists with multiple similar claims verifying an input query, as can be expected when drawing on data from multiple fact-checking portals. Thus, our study lays the groundwork to understand the state-of-the-art in claim retrieval and put existing methods into perspective, but further research is needed to fully understand task performance in real-world systems.

To tackle these issues, we aim to build on this study and collect real user-generated queries and relevance assessments by integrating different claim retrieval methods into the *ClaimLinker* application to link real input queries to claims from multiple fact-checking portals, enhance the existing benchmark data and more directly evaluate the usefulness of generated predictions.

Chapter 4

Claims in Scientific Discourse

While automated methods can assist in reducing the overwhelming information that has to be reviewed by humans, fact-checking still requires human fact-checkers to do research and rely on knowledge provided by trustworthy sources, e.g. scientific publications. However, these, too, may be flawed. Provenance information is crucial in order to reproduce and check scientific findings. For instance, many scientific investigations in empirical sciences rely on software for a range of different tasks including statistical data analyses, data pre-processing and data presentation. The choice of software may have a great influence not only on the research process but also on the derived findings, with even different versions of the same software potentially impacting the generated results Crane (2018). In order to increase transparency of research and verifiability of findings, knowledge of the used software thus is crucial. Likewise, any research data that was used needs to be accessible for the same reasons. Information systems offering access to linked information about publications, datasets and other important entities can help increase the transparency of research (Hienert et al., 2019) and can also be used to efficiently fact-check information in scientific publications and beyond. However, explicit links between publications, used software and research data are often not available. In addition, software and datasets are, unlike literature, often not cited in a standardized way (Mathiak and Boland, 2015) which makes the automatic generation of links difficult. While recent Named Entity Recognition (NER) approaches based on deep learning yield excellent results for a wide range of related use-cases and tasks, such as for detecting informal dataset mentions Otto et al. (2020), they typically require large sets of annotated data which may be hard to acquire. Since large sets of annotated data are not always available, we investigat in the following the use of weakly supervised approaches with distant supervision to create silver labels to train supervised software mention extraction methods using transfer learning.

Distant supervision for silver label generation of software mentions in social scientific publications

4.1 Introduction

Today, software is used for a variety of tasks in all steps of the research process, e.g. from data collection and data analysis to presentation and dissemination of findings. Therefore, it can shape both the process and the outcomes of scientific investigations in significant ways. Eklund et al., for instance, discovered inflated false-positive rates during analysis of FMRI data when using standard FMRI analysis software packages Eklund et al. (2016). Therefore, research findings relying on analyses using these packages may be systematically flawed. Another problem that was recently identified concerns the automatic formatting of dates in Excel which is shown to mistakenly convert gene names Zeeberg et al. (2004) which may introduce errors into datasets. Provenance information including knowledge about the software that is involved in scientific investigations thus is crucial to create understandable, traceable, and reproducible research that meets the requirements of open science and enables the implementation of recently proposed mechanisms for quality control and reproducibility Howison and Bullard (2016). Links between software, created datasets, and research findings would enable explicit modelling of provenance information and tracing of biases and errors throughout all stages of the research process. Also, assessing the usage of software in scientific publications could serve as a basis for rewarding software as research output Pan et al. (2015) further advancing open science. However, such links are not easily identifiable. While software citation standards exist (e.g., by FORCE11 Smith et al. (2016)), none of them has yet become universally established in scientific publications. Some researchers include only the name of the software, others use the name including information about the manufacturer and the version. This complicates automated extraction of such statements and thereby the automatic detection of links. Manual analyses of software mentions were done previously Howison and Bullard (2016); Nangia and Katz (2017), but were limited to reduced sets of publications (90 and 40, respectively) due to the high costs of manual annotations.

Recently, deep neural networks have gained increasing interest in the domain of NER Lample et al. (2016); Beltagy et al. (2019), which software mention identification can be seen as. The application of neural models for NER provides outstanding recognition results but requires a large training corpus with labelled entities. The provision of such labelled data is often the bottleneck when it comes to neural NER, as it is typically done in a manual process by different annotators. Different approaches have been proposed to overcome this issue, as for instance semi-supervised learning Zhou (2017) and distant supervision Choi et al. (2018). Another approach is the usage of a so called silver standard corpus (SSC) Rebholz-Schuhmann et al. (2010), which in contrast to a gold standard corpus (GSC) is created by automatic labelling by a combination of different classifiers. The quality of SSCs is much lower than the quality of GSCs. However, recent work showed that neural NER can be improved by transfer learning, where the network is first trained on the SSC and later on a GSC Giorgi and Bader (2018). This reduces the necessary size of the GSC, while at the same time increasing the recognition performance of the classifier.

The objective of this case-study is to investigate whether and how weakly supervised classifiers can be employed to create a SSC for the extraction of software mentions from scientific publications which can later be used for transfer learning. We apply three weakly supervised classifiers on a small manually created GSC in order to create *silver labels* which are then used for training a supervised classifier in order to predict the *gold labels* of the GSC.

The remainder of this paper is structured as follows. We first provide an overview of current approaches to NER in general and software mention identification in particular in Section 4.2. The applied weakly supervised classifiers for named entity extraction are described in Section 4.3, our method for combining them in Section 4.4. The GSC is introduced in Section 4.5. Section 4.6 presents the evaluation and discussion of results before we conclude with Section 4.7.

4.2 Related work

Recent approaches to extracting software mentions from scientific publications can be divided into three groups: manual extraction, rule-based, and supervised machine learning-based approaches. Manual approaches, sometimes called content analysis, typically work on small corpora with less than 100 articles or focus on particular software. Li et al. analysed the usage of the statistical software R Li et al. (2017) and LAMMPS Li et al. (2016) in 400 articles, while Nangia and Katz Nangia and Katz (2017), and Howison Howison and Bullard (2016) concentrated on software in general in 40 and 90 articles, respectively. An automatic approach to software mention identification is implemented in the BioNerDS system by Duck et al. Duck et al. (2015), who used a rule-based system based on syntactic token features and a dictionary of known software names. In later work, they employed a post-processing based on supervised machine learning which resulted in recognition rates of .67 F1. Another rule-based system to automatically identify mentions of R was implemented by Li and Yan Li and Yan (2018). Due to the particular focus on R and a dictionary of R packages, they were able to reach recognition rates of .94 F1. Pan et al. Pan et al. (2015) introduced an iterative bootstrapping approach to software mention identification which achieves .58 F1. Additionally, approaches exist that analyse references to software and code based on the URL to repositories Allen et al. (2018); Russell et al. (2018) However, there are currently no other supervised approaches for the identification of usage statements for software in scientific publications. One reason might be the lack of a dataset of sufficient size and quality.

For the related and similar task of extracting dataset references from scientific literature, we again find both semi-supervised and rule-based systems as well as supervised approaches. Boland et al. Boland et al. (2012) employed a pattern-based iterative bootstrapping algorithm, named InfoLink, and were able to reach a precision of up to 1 with a very low recall of .3 on the downside. Another semi-supervised approach is introduced by Ghavimi et al. Ghavimi et al. (2016), which used a dictionary of dataset names and employed similarity scores for identification with a recognition rate of .85 F1. Lu et al. used supervised learning to identify datasets by use of a training set of 1,000 sections that were obtained by active learning and achieved a precision of .82 and a recall of .59 Lu et al. (2012).

NER on scientific texts has been used with other targets in the literature with a particular interest in biomedical publications, as for instance for the identification of drugs, genes, proteins, and diseases Campos et al. (2012). This was also fostered by the BioCreative challenges that addressed gene names or chemicals and drug names. In this vein, Luo at al. employed neural models in order to identify chemical names from scientific texts with .91 F1 on a labelled corpus of 10,000 (training set: 3,500) abstracts with 84,355 labelled entities Luo et al. (2017). In detail, they used a BiLSTM-CRF including an additional attention layer with character-, word-, and dictionary embeddings and other linguistic features. Chemical names are different to software names when it comes to lexical structure as they typically exhibit combinations of characters, numbers, and special characters while software names are often composed from words from the lexicon. Beside the domain of scientific publications, neural NER methods have reached superior Lample et al. (2016); Beltagy et al. (2019) recognition rates but often require large training sets with several thousands of labelled entities. In addition, often target entities with high occurrences are chosen, which make even small training sets more effective. As shown in Section 4.5, software mention statements, particularly in the social sciences, are very rare, which requires even larger training sets. One way to overcome this problem is the use of distant supervision to create large annotated corpora which enable the training of sophisticated methods for even very fine-grained entity typing tasks Choi et al. (2018). A different approach to overcome the lack of large training datasets is the application of transfer learning, which allows to transfer trained concepts, e.g. between different application domains or languages. Giorgi and Bader recently illustrated the benefit of transfer learning in biomedical NER Giorgi and Bader (2018) on datasets with a small number of labels, increasing the recognition rate substantially. They transferred a neural model for NER from a noisy SSC to a GSC, which lead to significant increases in the recognition rates.

To summarize, semi-supervised approaches can achieve high precision but suffer from low recall. Supervised approaches produce more reliable results but require large sets of labelled training data. The application of distant supervision and transfer learning allows the automatic creation of labelled datasets and exploiting them for pre-training of more high-performance supervised methods.

4.3 Weakly supervised Named Entity Extraction

To overcome the data acquisition bottleneck for labelled corpora, we choose a small selection of openly available named entity extraction tools for the creation of silver labels. As described in the related work section, weakly supervised tools naturally suffer from relatively low recall. However, since they implement different algorithms and use different features, we expect the different tools to produce diverging annotations, potentially complementing each other when combined.

4.3.1 BioNerds

Bioinformatics Named Entity Recogniser for Databases and Software (BioNerds) Duck et al. (2015) is a rule-based system for recognition of software and databases from scientific publications in the domain of bioinformatics. Beside hard coded rules, it employs a dictionary of software and database names collected from Wikipedia, Bioconductor, and other sources. BioNerds implements a scoring system where the sum of the scores of the different features is used to decide upon the type of the entity, when a particular threshold is exceeded. The highest scores are provided by the dictionary matches, but also matches of Hearst patterns or positive head nouns achieve positive scores. Furthermore, the occurrence of a URL, a reference or a version number is considered as positive hint. Negative scores are provided, for instance, for matches with the English dictionary, negative head nouns or partial word matches. The threshold to be exceeded in order to be classified positively was selected to be slightly below the score of a match with the dictionary of known entities. As a result, known entities are, given a positive context, almost certainly recognised.

4.3.2 InfoLink

InfoLink Boland et al. (2012) is a weakly supervised iterative pattern-based bootstrapping approach developed for extracting dataset references from (social) scientific publications. Initially, seed words are searched in the corpus to identify patterns from their surrounding contexts. By alternating application of pattern identification and entity extraction, the dictionary of entities is increased iteratively. InfoLink relies on the surface form, i.e. the surrounding words of seed mentions, with some heuristics to normalize years and numbers and a frequency-based pattern scoring mechanism. Patterns consist of regular expressions and Lucene queries for increased efficiency.

4.3.3 Spied

The Stanford Pattern-based Information Extraction and Diagnostics Gupta and Manning (2014) (SPIED) system also implements a semi-supervised approach to named entity recognition. In the main, it operates similarly to InfoLink but includes different and more complex scoring mechanisms and features such as edit distance-based features, distributional similarity, and TF-IDF weighting, the patterns include POS rather than relying solely on surface strings.

4.4 Method

We first apply each weakly supervised tool separately on the corpus to retrieve a list of patterns and terms classified as software mentions. We create one BIO[1] file for each tool and corpus. For this purpose, we search all retrieved terms in the input texts and treat each occurrence as a software mention. For InfoLink, we receive, in addition to the list of terms, as output a list of regular expression patterns that can easily be applied on the input texts without requiring additional pre-processing. We create a second BIO file for InfoLink searching the patterns in the input texts. Since this has the potential to disambiguate software mentions from homonymous other entities, we use these predictions in our combined classifier but keep the term search variant for comparison. Weakly supervised approaches depend to a large part on the usefulness of their given seeds. Since our aim is to generate a silver standard for conditions where no or little training data is available, we do not use knowledge on the distribution of software mentions in the training data to construct

[1]The BIO format is a common format for annotated texts in named entity recognition. For each token, either a **B**egin, **I**n, or **O**utside tag is provided signalling whether the token belongs to an entity of interest (as its first token (**B**) or a subsequent one (**I**) or whether it is not part of any entity to annotate (**O**)).

Table 4.1: Features used for the CRFs.

dependency tag, fine-grained POS tag, coarse POS tag, surface form, lemma, is_alpha, is_stop, shape, sentence length, sentence number, word number

a seed set. Instead, we use Wikidata for distant supervision. Since we are mainly interested in finding software that is used for processing and analysing data for social scientific publications to gain provenance information on generated data and findings, we query Wikidata for all instances belonging to the classes "statistical package" or "mathematical software". Note that while it is also possible to use an extensive list of all known software names, this would introduce more noise due to the fact that software names often consist of common nouns (see Section 4.2) while at the same time providing little extra information relevant to our use-case. We instead rely on the weakly supervised approaches for expanding the list of software names. We incorporate all language variants and alternative names listed in Wikidata. This results in a list of 47 software names of which 10 and 8 are mentioned in the training and test set at least once, respectively. In the second step, we combine the predictions of all tools and use their majority votes as silver labels.

As supervised approach, we model the extraction of software mentions as a sequence labelling task using Conditional Random Fields (CRFs). The CRF is trained on the silver labels and may use the tools' individual predictions and additional output as features. Additional output are confidence values for InfoLink and BioNerds as well as information on the employed rules for BioNerds. Adding to that, we permit the CRF to use a small number of simple features as additional cues. These are listed in Table 4.1.

The threshold for accepting or rejecting patterns has to be set manually for InfoLink. Since we do not want to rely on annotated data to do parameter tuning, we use the configuration which was optimal for the extraction of dataset references Boland et al. (2012).

Table 4.2: Number of articles with the given numbers of software mentions.

# software	0	1	2	3	4	5	6	>6	sum
# articles	45	46	43	21	12	7	6	12	192
% articles	23.4	24.0	22.4	10.9	6.3	3.6	3.1	6.3	100

4.5 Dataset and Preprocessing

In order to measure the quality of our approach, we created a GSC of articles from the social sciences from PLoS[2]. Out of all articles having the keyword "*Social sciences*", we randomly selected 200. Following Duck et al. (2015), we automatically extracted all "Methods and Materials" sections as software mentions are expected to primarily occur here. 8 articles were removed from the set as they did not contain a "Methods and Materials" section. The resulting texts were annotated with the brat annotation software Stenetorp et al. (2012) by six annotators that were instructed to annotate software names without mentions of additional information such as producers or versions. For about 10% of the sentences which were randomly selected from the sentences of all annotators, a second annotation was obtained in order to assess the quality of the annotation. The inter-rater agreement was computed using Cohen's κ and reached *almost perfect* agreement of κ=.82. Overall, 462 (263 unique) software mention statements were found across all articles during annotation of the articles. Their distribution is detailed in Table 4.2. The number of articles that contained no software mentions at all was 45 (23 %). Table 4.3 lists the 10 most common software names including their frequencies in the training and test set. Note that software may be listed multiple times but with different spellings, e.g. for "Matlab" and "SPSS". Since our aim at this point is the identification of software mention statements rather than their disambiguation and linking, we do not align these different variants. Table 4.4 lists the number of unique software mentions that occurred at least *n* times in the corpus.

The annotated corpus was split into sentences using the Stanford NLTK Sentence Splitter Bird et al. (2009), resulting in 12,480 sentences. Afterwards, a white space based tokenisation was done, resulting in 347,544 tokens. The annotated token sequence was finally represented as BIO sequence. 462 of these tokens were annotated

[2] `https://www.plos.org/`

Table 4.3: The 10 most common software mentions and their numbers of occurrences overall and in the training and test set, respectively. The ✓ signals whether the wikidata seeds contain the software name.

software	*SPSS*	*MATLAB*	*SAS*	*Stata*	*R*	*Prism*	*SPM8*	*Matlab*	*PLINK*	*MEGA*
# overall	17	16	15	13	12	12	11	10	9	9
# train	12	13	9	9	10	11	5	5	9	9
# test	5	3	6	4	2	1	6	5	-	-
wikidata	✓	✓	✓	✓	✓	-	-	✓	-	-

Table 4.4: Number of different software mentions occurring with the respective frequencies.

# occurrence	≥ 13	≥ 12	≥ 11	≥ 10	≥ 9	≥ 6	≥ 5	≥ 4	≥ 3	≥ 2	≥ 1
# software	1	2	3	4	8	9	11	14	21	52	215

with the *begin* and 120 with the *in* tag, the remaining with the *outside* tag. From this corpus, we created a training and test set with 75 and 25 percent of articles, respectively.

4.6 Evaluation

4.6.1 Metrics

To measure the performance of the software mention detection task, we distinguish between exact and partial matches and compute precision, recall and F-measure considering each of these. Here, exact match means that the entire name of the software was recognised with the correct range, while partial matches signal that a certain overlap between the label and the prediction exists. We used the SemEval

2013 evaluation script[3] by David Batista.

4.6.2 Experimental setup

We measure the performance of the weakly supervised approaches, individually and in combination, as well as the direct distant supervision using labels from Wikidata and assess their applicability for transfer learning by using the silver labels as ground truth for training a CRF (Silver CRF). For combining the predictions of the weakly supervised approaches and creating silver labels, we test three different methods:

1. *majority*: majority vote of predicted labels
2. *conservative*: tokens are only labelled as belonging to a software mention if all classifiers agree on it belonging to this category
3. *greedy*: tokens are labelled as belonging to a software mention when at least one classifier labels it as such

The conservative and greedy conditions are expected to max out precision and recall, respectively. As an upper bound, we train a CRF on the gold labels of our GSC (Gold CRF). To evaluate the robustness of the approach with respect to seed selection and give insights on the usefulness of the Wikidata seeds, we illustrate the effects of choosing different seed sets for the weakly supervised approaches. For this, we create bins for software mentions depending on their number of occurrences in the training set. The intuition behind that is that the most frequently mentioned software names will also be the most well-known which can be identified without requiring the consultation of external knowledge sources. The less frequent a mention is, the less likely it will be incorporated into a seed set when the occurrence of software mentions in the corpus is not known in advance which is typically the case. Finally, we test the effects of using silver labels and outputs of the weakly supervised approaches as additional features for the gold CRF. For the weakly supervised approaches and the

[3]The original script can be obtained from `https://github.com/davidsbatista/NER-Evaluation/blob/7de8a231d5fd94ced0ef10c42971a30cd3b744b3/ner_evaluation/ner_eval.py`. (We adjusted the calculation of the overlapping range by an offset of 1 and added calculation of F1 scores.)

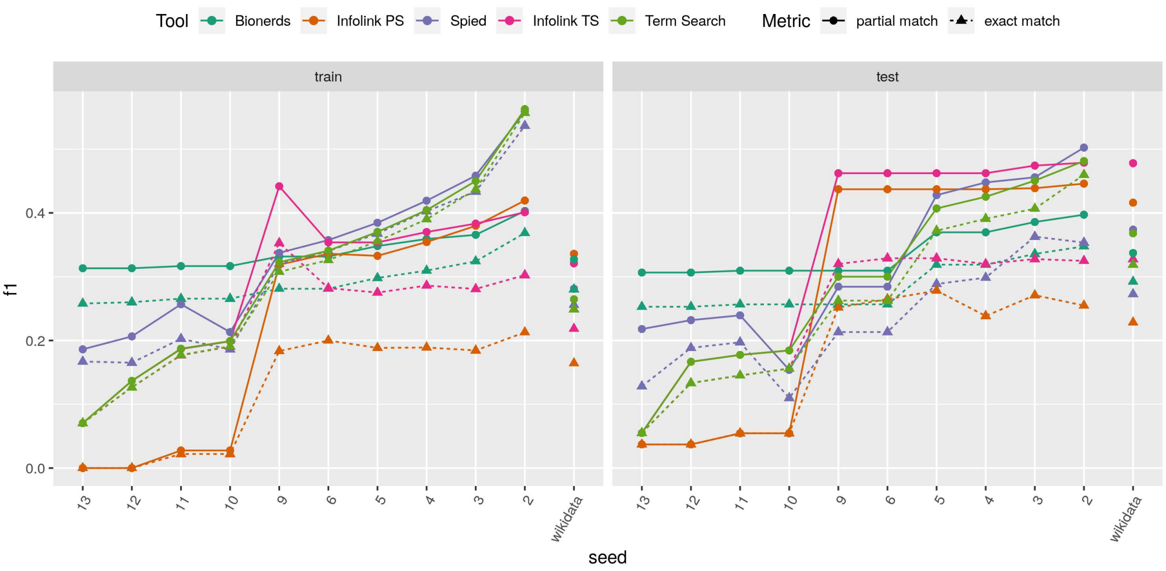


Figure 4.1: F-scores of the weakly supervised tools with distant supervision using different seed sets.

direct labelling of software mentions using Wikidata supervision, we evaluate both the performance on the training and the test set. The CRFs are trained on the training and evaluated on the test set.

4.6.3 Results

The performance of the weakly supervised tools with distant supervision and the influence of the choice of seeds is illustrated in Figure 4.1. The X axis represents the different seed sets used; 13 describes the set containing all software mentions occurring at least 13 times in the training set (the maximum number), 12 all mentions with at least 12 mentions and so forth (see Table 4.4). Wikidata represents the seed set obtained by querying Wikidata. As expected, performance generally increases when seeds are added. Especially when the number of seeds exceeds a certain threshold (4 and 9 in this case when only seeds occurring at least 10 or 6 times are used, respectively), there is a significant increase in performance. At the same time, adding seeds can harm performance for the pattern induction approaches as ambiguous and rare mentions may increase the likelihood of generating deficient extraction patterns. The seed set obtained from Wikidata leads to performance which is close to the optimal seed set which shows that distant supervision using lists of well-known software for seeding the algorithms is a feasible approach. The numbers for term search show the impact of the used seeds for comparison. When near-complete information on mentioned software is available, there is no or little gain from applying weakly supervised approaches in addition to searching the known names directly. However, even then precision may suffer from ambiguous names that may refer to software or other entities, such as with the software package “R” which has a high influence when used in a set with only 3 other less ambiguous seeds (set of seeds >=10 mentions). The performance of the two pattern generating approaches (SPIED and InfoLink) on the training set is considerably worse than on the test set. An analysis of the induced patterns reveals that this is due to the higher number of ambiguous software names in the former, more precisely, the high number of occurrences of the software “R” which causes the generation of deficient patterns. For InfoLink, the pattern search variant succeeds in disambiguating software mentions from homonyms not referring to software as reflected by its higher precision compared to the term search variant. However, many software mentions are missed reducing recall considerably. InfoLink yields the best results for partial matches on both the training and test set. Yet, it also has the highest divergence in scores for exact vs. partial matches reflecting its strength in detecting mentions but its weakness in determining the exact boundaries of the matches. This is caused by its relying on surface features rather than incorporating knowledge gained from linguistic features such as POS tags.

The results for the combination of the different tools and their usage for silver standard generation are illustrated in Figure 4.2. The upper bound for the classifier (Gold CRF) reaches .54 F1 on the test set. The majority vote silver labels obtain .41 F1 with the recall being closer to the upper bound than the precision. The greedy variant achieves the same F-score but is biased towards maximizing recall at the cost of precision yielding higher recall values than the Gold CRF. The conservative variant suffers from low recall causing its F-score to be low (.2) while achieving a higher precision than the Gold CRF. The combination of the weakly supervised approaches with distant supervision outperforms the direct creation of silver labels from the Wikidata software names and the application of the approaches individually. The Silver CRFs achieve lower scores than the direct application of the weakly supervised approaches on the test set. We attribute this to the higher difficulty of the training set

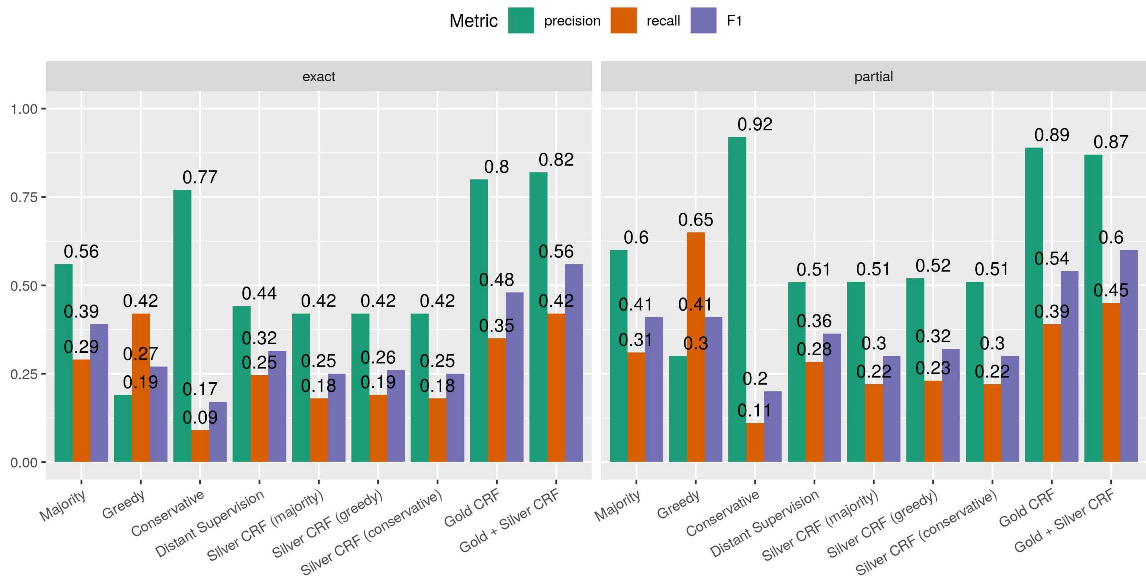


Figure 4.2: Comparison of the performances of the different classifiers on the test set using the Wikidata software names as seeds / for distant supervision.

which results in decreased performance for the pattern induction approaches. These noisy labels are used for training the classifier which is then applied on the test set while the weakly supervised approaches are applied on the easier test set directly. Finally, the best result is achieved by feeding the silver labels as additional features to the Gold CRF. While this has a slightly negative impact on precision, it increases recall by a higher magnitude resulting in .6 F1 with a still very high precision of 0.87.

4.7 Conclusion and Outlook

We investigated the use of weakly supervised classifiers and Wikidata for distant supervision for the extraction of software mentions from social scientific publications without requiring manual annotations. We compared the generation of silver labels by directly labelling mentions according to the Wikidata information to using them as seeds for different information extraction tools. We can show that in doing so, a silver standard with relatively high-precision annotations can be created that may serve to pre-train more powerful algorithms using transfer learning. With each classifier using different features and scoring mechanisms, their combination yields the best results showing that they partly complement each other. Furthermore, we show that predictions of weakly supervised classifiers may provide useful features for supervised methods which leads to good results even when using on a small training set. In this case-study, we employed a small set of basic features for the supervised approaches to demonstrate the feasibility of the approach. In future work, we will use more sophisticated features and supervised classifiers with transfer learning to exploit the generated SSC and extract software mentions from larger collections.

Chapter 5

Online discourse data for analyzing public attention

Irrespective of the truthfulness and correctness of statements made online, they can be a valuable source of information for social scientific studies. Many recent works investigate the usage of ODD, e.g. messages posted on Twitter, to measure public opinion (Breuer et al., 2021, 2022, for instance). The sheer volume of information requires the assistance of computational methods to filter and pre-process the data. At the same time, social scientific expertise is needed to guide the investigations. In this chapter, we use a pipeline of automated methods including language model-based relevance filtering and topic detection of tweets, sentiment analysis and time series analysis to use ODD to investigate salient topics and sentiments in the public debate about vaccinations during the time of the pandemic in German-speaking countries to get insights into public concerns and possible interactions with political decisions.

Public Discourse about COVID-19 Vaccinations: A Computational Analysis of the Relationship between Public Concerns and Policies

Societies worldwide have witnessed growing rifts separating advocates and opponents of vaccinations and other COVID-19 countermeasures. With the rollout of vaccination campaigns, German-speaking regions exhibited much lower vaccination uptake than other European regions. While Austria, Germany, and Switzerland caught up over time, it remains unclear which factors contributed to these changes. Scrutinizing public discourses can help to shed light on the intricacies of vaccine hesitancy among the German-speaking population. These insights are valuable for policy-makers tasked with making far-reaching decisions. On the hand, policies need to effectively curb the spread of the virus. On the other hand, they need to respect fundamental civic liberties and minimize undesired consequences. This balancing act requires empirical insights into citizens' concerns and behaviors. Dynamic interactions between policy changes and public opinions are hard to investigate using traditional survey data alone. This study, thus, draws on Twitter data to analyze the topics prevalence in the public discourse. It further maps the topics to different phases of the pandemic and policy changes to identify potential drivers of change in public attention. We use a hybrid pipeline to detect and analyze 199,207 vaccination-related tweets using topic modeling, sentiment analysis, and a minimum of social scientific domain knowledge to analyze the discourse about vaccinations in the light of the COVID-19 pandemic in the so-called DACH region (Germany, Austria, Switzerland).

We show that skepticism regarding the severity of the COVID-19 virus and towards efficacy and safety of vaccines were among the prevalent topics in the discourse on Twitter but that the most attention was given to debating the theme of freedom and civic liberties. Especially during later phases of the pandemic, when implemented policies restricted the freedom of unvaccinated citizens, increased vaccination uptake could be observed. At the same time, increasingly negative and polarized sentiments emerge in the discourse. This suggests that these policies might have effectively attenuated vaccination hesitancy but were not successfully dispersing citizens' doubts and concerns.

5.1 Introduction

The outbreak of the COVID-19 pandemic has fundamentally disrupted societies around the world. To protect the general public and particularly vulnerable groups, governments introduced policies to limit the spread of this infectious disease. These policies included mask mandates, closing schools and the retail sector, curfews, strict lockdowns, and contact restrictions. While many of these measures are assumed to have successfully slowed the spread of the pandemic and contributed to saving lives, they also had detrimental side effects. For instance, a stalled economy led to widespread unemployment (Blustein and Guarino, 2020), lockdown policies strained people's mental health, and were accompanied by an increase in domestic violence (Piquero et al., 2021). This balancing act forced governments to make difficult trade-off decisions in designing policies with maximum effectiveness and minimal invasiveness. A prime example of this tightrope walk is the COVID-19 vaccination strategies. On the one hand, empirical evidence suggested that widespread vaccination uptake ranked among the most effective means to protect the population from getting infected or hospitalized (Andrews et al., 2022). Thus, encouraging vaccination uptake was primarily seen as a promising strategy to speed up the reopening of society. Yet, on the other hand, invasive policies to promote or even coerce widespread vaccine uptake, such as vaccine mandates or limiting rights for unvaccinated citizens, mark severe encroachment of civil liberties and spurred a significant backlash among the citizenry (Bardosh et al., 2022). For instance, invasive policies increased polarization among citizens, undermining social peace and democracy (Jiang et al., 2021). Therefore, considering citizens'concerns is crucial when designing adequate vaccination policies and minimizing negative side-effects on society. In particular, the German-speaking so-called DACH region (Germany, Austria, Switzerland) exhibited higher vaccine hesitancy than other European countries at the start of the pandemic (Z et al., 2022).

While polling is currently the dominant strategy for governments to retrieve citizens' attitudes, opinions, and behaviors (Rothmayr and Hardmeier, 2002), online public discourses are gaining ground as an additional source of information (Ceron and Negri, 2015, 2016; Rubinstein et al., 2016). Appropriate data are often retrieved from social networking sites (SNS) such as Twitter, Facebook, or YouTube and then analyzed using Natural Language Processing (NLP) methods. While this type of data has been criticized for its lack of representativeness and oversampling of younger, privileged, and tech-savvy people (Hargittai, 2020), they also offer several distinct characteristics that distinguish them from traditional survey data. First, discourse data are highly dynamic, cheap, and contain information about billions of users (Ceron et al., 2014). Second, they can be analyzed ex-post, thereby shedding light on longitudinal fluctuations of public opinion about real-life events (Conrad et al., 2021). Third, they allow for a more nuanced understanding of the ambivalence in public opinion (Foad et al., 2021) by providing insights into the discursive construction of contentious issues (Burnap et al., 2015).

Hence, many scholars have argued that survey and discourse data can complement each other (Buntain et al., 2016; Diaz et al., 2016; Hil, 2020; Stier et al., 2020), particularly in highly dynamic policy contexts such as the COVID-19 pandemic. Indeed, attitudes drawn from surveys and online discourses show similar long-term trends, with discourse data sentiments being more prone to short-term fluctuations (Pasek et al., 2020) and polarized opinions about COVID-19 measures being more pronounced in survey data (Reiter-Haas et al., 2022).

Especially in times of crisis, continuously monitoring how public opinion changes is crucial for policy-makers to make informed decisions. For instance, discourse data can help policy-makers assess salient topics and concerns that may lead to vaccine hesitancy, evaluate the level of polarization of suggested policies, or retrieve information about side effects in real-time.

Our study investigates the potential of online discourse data to inform policy-making about the vaccination strategy in the DACH region. We combined automated methods with manual analyses to harvest suitable datasets from Twitter, filtered relevant tweets discussing vaccinations during the time frame from 01.01.2020 to 31.01.2022, and extracted relevant content. Specifically, we propose a semi-automatic analysis pipeline consisting of tweet filtering, sentiment analysis, and topic modeling to trace the rapid change of topics and public sentiment in the vaccination discourse. We further shed light on how the evolution of topics and sentiments relate to important policy events outlined in the national vaccination strategies of the three countries under investigation. Thus, we formulate the following research questions:

- RQ1: How does the vaccination discourse in DACH countries on Twitter evolve (in terms of tweet frequencies and sentiments)?

- RQ2: Which topics and themes were prevalent in the discourse? To what sentiments were they connected?

- RQ3: How did the topics, themes, and associated sentiments evolve?

- RQ4: How was this related to different phases of the pandemic and policy events?

Gaining insights into these research questions adds to the empirical literature on data-driven policy-making. Moreover, this study further advances the methodological literature on extracting, filtering, and analyzing Twitter data for policy research. Our enriched dataset is available[1] to help advance further research.

We find that when the COVID-19 vaccines were first authorized, the debate on Twitter focused on a range of topics, including *side-effects of individual vaccines* and *vaccinations in general* but also *freedom and civic liberties*. During later phases of the pandemic, when different policies restricting the freedom of unvaccinated citizens were publicly discussed and later implemented, the attention increasingly shifted away from medical and other concerns towards questions of *freedom and civic liberties*. At the same time, vaccination uptake increased. This finding may indicate that these policies might have been an essential factor in attenuating vaccination hesitancy - either due to the imposed restrictions for the unvaccinated or/and the decreased attention to medical concerns. However, while vaccination hesitancy decreased, the discourse, which was connected to more negative than positive sentiments from the start, did not become more positive but, in fact, more polarized. This hints at the concerns being ignored by the citizens rather than having been resolved.

[1] `https://git.gesis.org/bolandka/vaccinationdiscourse`

5.2 Related Work

5.2.1 Leveraging Discourse Data for Policy Making

Discourse data can inform all stages of the policy cycles, from identifying the most pressing issues to evaluating policies after implementation. They have been successfully used to provide insights into citizens' preferences for different policy alternatives (Ceron and Negri, 2015, 2016). This information is vital to the formulation phase of public policy and contributes to more responsive policy-making. Regarding specific policy areas, citizens' tweets were used in financial policy to predict consumers' inflation expectations (Angelico et al., 2022) and stock market indicators such as the Dow Jones, the NASDAQ, and the S&P 500 (Bollen et al., 2011; Zhang et al., 2011). Another use case for policy-relevant insights gained from SNS data pertains to emergency response in natural disasters. For instance, studies have shown that crowdsourcing people's reactions on Twitter is a reliable, cheap, and scalable approach to detect extreme real-world events such as earthquakes (Van Quan et al., 2017; Poblete et al., 2018; Sakaki et al., 2010).

Insights gained from discourse data have also proven to be particularly fruitful with regard to health policies, for instance, to geographically locate illnesses such as allergies, obesity, and insomnia (Paul and Dredze, 2021). Furthermore, Twitter data can track and even forecast the spread of infectious diseases such as Influenza epidemics (Aramaki et al., 2011) ; Yang et al., 2021), and, more recently, COVID-19 (Klein et al., 2021). In addition, the Tweets analyzed in these studies could also be used to gauge public interest or concern about health-related events (Signorini et al., 2011), key information for policy-makers to introduce effective countermeasures.

5.2.2 Twitter for Analyzing Public Opinion during the COVID-19 pandemic

Several studies harvested data from Twitter to gain insights into public opinion about the pandemic. Jing and Ahn (2021) show that political actors in the US strategically used Twitter to establish partisan narratives about the pandemic. Moreover, verified

Twitter users in Italy used their prominence to spread misinformation about the pandemic on the social media platform, especially those users associated with the right and center-right wing political community (Caldarelli et al., 2021). Another study from the Chinese context suggests that public attention to the pandemic was largely influenced by specific events and measures introduced by the government (Cui and Kertész, 2021). Wang et al. (2021) classified tweets as conforming with one of the four health belief models constructs using machine learning. They found that scientific events (e.g., publications) and nonscientific events (e.g., political speeches) seemed to have a comparable influence on health belief trends on Twitter. Several studies investigate prevalent topics and themes in the Twitter discourse about the COVID-19 pandemic (Xue et al., 2020; Al-Ramahi et al., 2021; Boon-Itt and Skunkan, 2020). While Xue et al. (2020) identified 13 topics in tweets through topic modeling and categorized them manually into five overarching themes (e.g., public health measures to slow the spread of the pandemic; social stigma associated with COVID-19), Boon-Itt and Skunkan (2020) (the COVID-19 pandemic emergency, how to control COVID-19, reports on COVID-19) and Al-Ramahi et al. (2021); Boon-Itt and Skunkan (2020) (constitutional rights and freedom of choice; conspiracy theory, population control, and big pharma; fake news, fake numbers, and fake pandemic) each find three overarching themes. All three studies developed lists of predefined hashtags to identify relevant tweets. While this procedure has its merits, it also risks missing other relevant tweets of the debate or inserting a selection bias. Doogan et al. (2020) assigned 131 automatically generated topics to 22 non-pharmaceutical interventions grouped into seven categories: Personal Protection, Social Distancing, Testing and Tracing, Gathering Restrictions, Lockdown, Travel Restrictions, and Workplace Closures. They found that the proportion of intervention-related topics varied between the six investigated countries. The relationship between tweet frequencies and case numbers was statistically significant only for two countries. While less restrictive interventions gained widespread support, more restrictive ones were perceived differently across countries. Mohamed Ridhwan and Hargreaves (2021) used LDA to estimate a good number of topics to generate using Gibbs Sampling Dirichlet Multinomial Mixture, arriving at 35 topics. These were then manually labeled and assigned to 12 themes, such as lockdowns, social distancing, and travel and border restrictions.

Many recent works are concerned with sentiment analysis or emotion detection in tweets about the pandemic. Naseem et al. (2021) benchmark different sentiment analysis methods on their COVIDSENTI sentiment dataset, which consists of 90,000 COVID-19-related tweets covering February and March 2020, and find the transformer-based BERT to outperform all other methods. Their analysis shows that negative sentiments were prevalent and that intensity and frequency of negative sentiments were high before mid of March 2020 but decreased after. Müller et al. (2020) release COVID-Twitter-BERT, pre-trained on a large corpus of English COVID-19 tweets, which achieves considerable improvement over the BERT-LARGE base model on the target domain. Shofiya and Abidi (2021) use a hybrid approach based on SentiStrength and an SVM classifier for sentiment analysis and find that most users in Canada expressed neutral sentiment towards social distancing measures. Xue et al. (2020) use scores based on a word-emotion association lexicon, and find as dominant emotions for their selected topics anticipation that measures can be taken, followed by mixed feelings of trust, anger, and fear related to different topics, especially when discussing new COVID-19 cases and deaths. Also using a lexicon-based approach, Mathur et al. (2020) find that a high number of tweets is related to trust, which they interpret as people's confidence in the ability to fight COVID-19 and policies taken by authorities. At the same time, fear and sadness are also prevalent. They find the number of tweets with positive vs. negative emotions to be almost equal. Mohamed Ridhwan and Hargreaves (2021) use a Recurrent Neural Network for emotion classification and the lexicon-based tool VADER for sentiment analysis of COVID-19-related tweets relating to Singapore based on user information and geo-tags. Topics relating to measures such as social distancing and the encouragement to stay at home and to wear masks were coupled with positive sentiments. In contrast, negative sentiments dominated the discourse about travel and border restrictions. Overall, policies by the Singapore government were coupled with positive sentiments. The authors conclude that the citizenry supported anti-COVID-19 measures.

5.2.3 Twitter for Analyzing Public Opinion Towards COVID-19 vaccines

Several studies zoom in more and specifically investigate public opinion about COVID-19 vaccines using Twitter data. Hu et al. (2021) structure the pandemic into phases according to pre-selected key events relating to the vaccination roll-out in the United States (US). The authors draw on geo-tagged tweets and find strong changes in public sentiment and emotion for the different phases. They conclude that social events and public announcements by influential entities may impact public opinion on COVID-19 vaccines considerably. Fazel et al. (2021) investigate how positive vs. negative sentiments develop in relation to major news announcements about vaccines in the United Kingdom. They find that each announcement was associated with a short-term decrease in negative sentiment and that tweets with negative sentiment toward vaccines were posted by a smaller number of individuals. The high engagement created by negative tweets decreased gradually in the course of the vaccination campaign. Both studies start with the assumption that specific events drive change in public opinion and, thus, conducted a targeted search for events, such as public announcements and news coverage. The work by Muric et al. (2021) analyzes reasons for vaccine hesitancy. Using a manually created list of keywords, the authors create a dataset of Twitter posts and accounts expressing a strong anti-vaccine stance. The findings indicate that vaccine hesitancy is fueled by misinformation originating from websites with questionable credibility. Furthermore, Bonnevie et al. (2021) quantifies the rise of vaccine opposition on Twitter four months before and four months after the spread of the virus in the US. With a manually created keyword list, they collect tweets discussing vaccinations. The authors manually derive conversation themes of vaccination opposition from them, arriving at eleven themes, including adverse health impacts, policies and politics, and disease prevalence. They show that the frequency of these themes changed over time. Sattar and Arifuzzaman (2021) perform sentiment analysis on tweets discussing vaccinations and propose a model to forecast vaccination uptake. The predicted numbers approximated the actual numbers for the US (Ritchie et al., 2020). Herrera-Peco et al. (2021) analyze a COVID-19 antivaccination campaign in Spanish tweets, one week before and one week after the European Medicines Agency announced the authorization of the Pfizer BioNTech vaccine. They find that attacks against vaccine safety were the most frequent antivaccine message. Moreover, the authors also find conspiracy theories, such as presenting the vaccine as a means of manipulating the human genetic code.

Our studies draw on the insights these studies put forth and apply them to the German-speaking context. Moreover, we deviate from these studies by introducing data-driven approaches to identify relevant tweets, peaks, and change points, inserting as little prior knowledge and assumptions as possible into the analyses.

5.3 Methods

5.3.1 Dataset and Preprocessing

We examine tweets within the time span of 01.01.2020 - 31.01.2022 using the *TweetsKB* (Fafalios et al., 2018b) pipeline. TweetsKB is a large-scale knowledge base of annotated tweets harvested using the Twitter streaming API. Since 2013, a random 1% sample of the Twitter stream has been harvested. Their metadata and information automatically extracted from the tweets, such as entities, sentiments, hashtags, and user mentions, is accessible in RDF format. The data from 2013 until December 2020 is available at `https://data.gesis.org/tweetskb/`. We use the same pipeline to harvest tweets for the above-mentioned period, resulting in 12,297,163 tweets. For our analyses, we analyze the textual content of tweets, ignoring pictures and links to videos or other content.

Relevance Filtering

We extract relevant tweets by filtering for time ("01.01.2020 - 31.01.2022"), language ("German"), and topic ("vaccinations"). We use the timestamps provided by Twitter to filter tweets created during our desired time frame. Furthermore, we draw on Twitter's language tag to filter German tweets. Ensuring that the tweets address the relevant topic of vaccinations is not trivial. It requires the complex procedure of creating a seed list with relevant search strings. Commonly, researchers rely on manually created seed lists for hashtags or search strings to identify relevant tweets, e.g., Mohamed Ridhwan and Hargreaves (2021); Buntain et al. (2018); Muric et al. (2021); Bonnevie et al. (2021); Xue et al. (2020); Al-Ramahi et al. (2021); Herrera-Peco et al. (2021). However, manual seed list creation is costly and lowers the reproducibility of results as seed lists for similar topics exhibit a high variance with unknown effects on generated results. For example, the keyword list used to filter COVID-19 - related tweets in Chen et al. (2020) comprises 80[2] keywords. Dimitrov et al. (2020b) expanded this list to include 268 keywords for their TweetsCov-19

[2] the list is updated by the authors: `https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt`, last update to date was on 11/28/2021

dataset, while the multilingual keyword list used by Imran et al. (2022) for the TBCOV corpus includes more than 800 terms. Moreover, manually curated lists may inadequately capture vocabulary mismatch problems, emerging new terms and are prone to biases - e.g., focusing on certain topics or frames while neglecting others.

Since we are interested in exploring different topics assuming as little prior knowledge as possible, we followed an automatic query term expansion approach to generate a list of search terms (*seed list*). Starting with an initial query keyword (“Impfung”, English: “vaccination”), we extracted all tweets that contain this keyword as a single token word while not applying case sensitivity. We created a set of candidate terms from this set of tweets by collecting and lemmatizing all verbs, adjectives, nouns, and proper nouns using the Spacy POS tagger (Explosion, 2022).

Next, we reduced the set of candidates by removing all terms that fall under a given limit of semantic similarity compared to the query keyword. To determine said similarity, we used pre-trained word embeddings from Fasttext.cc trained on Wikipedia and Common Crawl (Bojanowski et al., 2017); precisely, we used the German dataset with 300 dimensions (fasttext.cc, 2023). On this embedding, we computed the cosine-similarities between the query keyword ("Impfung") and the candidate keywords. The similarities range between -1 and 1. Visual inspection of the candidate terms indicates that a minimum cosine-similarity of 0.6 is required to retrieve meaningful results.

The remaining candidate terms were then sorted by the number of their co-occurrence with the query keyword, and we selected the top 30 terms for the seed list.

Table 5.1 displays the resulting sets of seed terms.

This automated procedure suggested including terms referring to other viruses than Corona, e.g., the swine flu. As we assumed such discourses to relate to discourse about COVID-19 in the selected time frame, we do not exclude these keywords from our list.

To construct our final set of tweets discussing vaccinations, we searched for all keywords in the set in all tweets harvested by our TweetsKB pipeline in the specified time frame. We then extracted all tweets mentioning at least one of the keywords

Seed Term	**Translation**
impfung	vaccination
impfen	to vaccinate
impfstoff	vaccine
geimpfte	(the) vaccinated
impfungen	vaccinations
infektion	infection
impfpflicht	compulsory vaccination
geimpft	vaccinated
impfschutz	immunization protection
impftermin	vaccination appointment
impfschäden	vaccination damages
impfschaden	vaccination damage
immunisierung	immunization
impfkampagne	vaccination campaign
masern	measles
impfnebenwirkungen	vaccination side-effects
impfling	freshly vaccinated person / seed chrystal
erstimpfung	primary vaccination
grippeimpfung	influenza vaccination
impftermine	vaccination appointments
impfreaktionen	reactogenicities
impfreaktion	reactogenicity
impf	vaccination-
auffrischungsimpfung	booster injection
zwangsimpfung	compulsory vaccination
impfaktion	vaccination event
schweinegrippe	swine flu
impfstoffs	vaccine (direct object)
grundimmunisierung	fundamental immunization
impfbereitschaft	vaccination willingness

Table 5.1: Automatically generated seed list for filtering tweets discussing vaccinations and added English translations

in their tweet texts or hashtags. This resulted in a set of 201,705 tweets. Removing all tweets written in a language other than German according to the language tag provided by Twitter, resulted in our final set of 199,207 tweets. Revisiting the list of keywords, we expect the seed term "Infektion" (engl. infection) to add noise to our data because such tweets do not necessarily address vaccinations. We excluded 7457 tweets (3.89%) due to the occurrence of this term (alone) from our analysis.

Sentiment Analysis

We use the automatic tool SentiStrength to identify tweet sentiments. It is tailored for the analysis of short social media texts (Thelwall et al., 2012) and measures the strength of both positive and negative sentiments in a tweet on a scale from 1 to 5. Thus, every tweet has one score specifying the intensity of the negative and one score specifying the intensity of the positive sentiment.

Based on the automatically assigned sentiment scores and the tweets' timestamps, we generate time series data, accumulating all sentiments for one day using four different approaches: 1) summing up all sentiment scores (positive and negative intensity scores) per day (SUM), 2) normalizing the summed up score by the number of tweets (REL) and 3) counting the number of positive (POS) and 4) negative (NEG) tweets for each day. A tweet is considered positive when the intensity of its positive sentiment is higher than the intensity of its negative sentiment and vice versa. Note that for generating the plots, we translate the sentiment scores to intervals of 0 to 4 and -4 to 0, respectively. Using the SUM and REL metrics, intensities of sentiments are being regarded, while for POS and NEG, sentiment intensities are translated to positive, negative, and neutral/mixed labels without any information on intensity. All metrics except REL represent the frequency of tweets in addition to the sentiments. Note that by summing up intensity scores, we do not differentiate between tweets that have a neutral sentiment, i.e. no negative and no positive sentiment, and tweets that have a mixed sentiment with both negative and positive sentiments being equally strong.

5.3.2 Topic Modeling

We use BERTopic (Grootendorst, 2022), a recent transformer-based topic modeling technique, to derive topics from the tweet texts in an unsupervised manner, i.e., all topics are derived from the data without relying on any prior knowledge. We decided to use embedding-based topic modeling instead of "traditional" topic modeling techniques such as LDA (Blei et al., 2003). This approach enables us to exploit information about semantic relationships among words, represented in embeddings generated on large data volumes instead of relying on the distribution of words in our tweets alone. Traditional techniques cluster co-occurring words to find topics and then proceed to identify these topics in a set of documents. Input is typically pre-processed (e.g., using lemmatization), and information about sentence structure is disregarded in favor of treating documents as bags of words. Embedding-based methods such as BERTopic, in contrast, typically do not preprocess or otherwise alter the input. They consider the semantic similarity of documents using embeddings to cluster similar documents into topics and try to find typical terms that characterize them in a separate step. BERTopic allows using custom embeddings. We use embeddings-paraphrase-multilingual-MiniLM-L12-v2 sentence transformers model [3], a multilingual sentence-transformers model which maps sentences and paragraphs to a 768-dimensional dense vector space. It was trained on parallel data for 50+ languages and proved useful for semantic search and clustering. Using these embeddings allows us to find similarities in sentences within one language or across languages, a valuable property for German tweets that may use English terms or quote English content. We use BERTopic's default algorithms UMAP (McInnes et al., 2018) to reduce the dimensionality of the document embeddings, and HDBSCAN (McInnes et al., 2017) for document clustering.

We compute topics for the complete set of tweets and then classify them into negative and positive tweets to gain insights into their occurrence in different contexts. Each tweet is assigned to precisely one topic, with one noisy residual category for all tweets that do not fit into any of the topic clusters with high probability. Note that, in principle, one tweet may be assigned to more than one topic based on the calculated probabilities of a tweet belonging to any cluster. Due to the limited length of tweets,

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

we keep the standard procedure of assigning only the most prevalent topic.

Since we use the 1% Twitter API, one tweet in our corpus mentioning a topic represents a much larger discourse. We thus decided to keep the standard value of ten documents for the minimum topic size and set the number of topics to 150 to enable a fine-grained analysis while maintaining a number of topics that is feasible to review manually. By grouping topics into themes, we perform a manual merging step later on. Therefore we prefer a high number of topics at this step to prevent information loss.

For optimizing topic representation after extraction, we set the ngram range to 1,2 and the diversity to 1.0. However, we did not rely on the extracted topic representations alone when interpreting the clusters. Instead, we manually assigned labels to each topic by examining the tweets in the respective clusters. For this, the first two authors of this paper (one Computer Scientist and one Political Scientist) labeled all clusters independently and discussed their results. For some of the topics, the labels diverged regarding their precise wording, but not regarding the perceived content. Final labels were assigned by both authors jointly. For 11 topics we failed to find suitable labels as the tweets seemed too heterogeneous. We excluded these clusters from our analysis.

5.3.3 Phases of the pandemic and policy events

To relate the evolution of the discourse to different phases of the pandemic, we refer to the classification provided by the Robert Koch Institute (RKI) [4], the German government's central scientific biomedicine institution. The phases from the beginning of the pandemic until the end of the time under investigation in this study are classified as listed in Table 5.2. To compare the vaccination uptake in Germany, Austria, and Switzerland for each of the different phases, we add vaccination ratios provided by Our World in Data (Ritchie et al., 2020). Even though the RKI classification refers to the spread of the virus in Germany, Desson et al. (2020) show that the German-speaking countries faced similar epidemiological situations during the pandemic. We

[4] https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2022/Ausgaben/10_22.pdf?__blob=publicationFile

Phase	**Begin**	**End**	**% D**	**% A**	**% CH**	**% UK**
Sporadic cases	27.01.2020	24.02.2020				
Wave 1	02.03.2020	18.05.2020				
Summer plateau 2020	18.05.2020	28.09.2020				
Wave 2	28.09.2020	01.03.2021	5.2	5.2	6.4	30
Wave 3	01.03.2021	14.06.2021	49	48	44	62
Summer plateau 2021	14.06.2021	02.08.2021	63	60	55	70
Wave 4	02.08.2021	27.12.2021	75	74	69	77
Wave 5	27.12.2021	31.01.2022*	77	76	70	78

Table 5.2: Phases of the pandemic as classified by the RKI (own translation of the phase labels; calendar weeks mapped to dates) with added vaccination ratio statistics provided by Our World in Data. %D/A/CH/UK: percentage of vaccinated people in Germany/Austria/Switzerland/UK, respectively, at the end of the phase *end of the investigated time frame

draw on official websites (e.g., Bundestag.de, zusammengegencorona.de, sozialministerium.at) and Wikipedia to identify events relating to vaccination policies in the DACH countries, such as the licensing of new vaccines. We arrive at a list of 57 events: 22 for Germany (Table 5.3), 16 for Switzerland (Table 5.5), and 19 for Austria (Table 5.4). The tables show that the three countries partly issued similar policies at similar times. These do not fully coincide with the pandemic phases. To analyze the discourse about policy events, we derive *policy phases* by grouping similar policy events (see Table 5.6).

Date	Event	Source	RKI Phase	Policy Phase
09/11/2020	Leopoldina publishes position paper on vaccine distribution	`https://www.leopoldina.org/`	Wave 2	I
17/12/2020	First publication of the vaccination recommendation (STIKO)	`https://edoc.rki.de/handle/176904/7579`	Wave 2	II
18/12/2020	Vaccination sequence published in the Federal Gazette	`bundesanzeiger.de`	Wave 2	II
21/12/2020	Authorization of BioNTech vaccine	`https://impfdashboard.de/`	Wave 2	II
27/12/2020	Vaccination start	`https://impfdashboard.de/`	Wave 2	II
06/01/2021	Authorization of Moderna vaccine	`https://impfdashboard.de/`	Wave 2	II
29/01/2021	Authorization of AstraZeneca vaccine	`https://impfdashboard.de/`	Wave 2	II
11/03/2021	Authorization of Johnson & Johnson vaccine	`https://impfdashboard.de/`	Wave 3	II
15/03/2021	Halt of AstraZeneca vaccinations due to safety concerns	`https://www.bundesregierung.de`	Wave 3	II
25/03/2021	Resumption of AstraZeneca vaccinations	`https://www.bundesregierung.de`	Wave 3	II
07/04/2021	Nationwide vaccination in doctors' offices	`https://impfdashboard.de/`	Wave 3	II
06/05/2021	Nationwide suspension of priority groups for AstraZeneca	`https://www.ndr.de/`	Wave 3	III
28/05/2021	Authorization of BioNTech vaccine for youth	`https://investors.biontech.de/`	Wave 3	IV
07/06/2021	Official end of vaccination priority groups	`https://www.tagesschau.de`	Wave 3	IV
23/08/2021	Implementation of 3G Rule restricting access to facilities for unvaccinated and untested individuals	`https://www.bundesregierung.de`	Wave 4	IV
05/11/2021	Health ministry decides on offering booster shots	`https://www.bundesregierung.de/`	Wave 4	V
26/11/2021	Authorization of BioNTech vaccine for children	`https://www.pei.de/`	Wave 4	V
01/12/2021	Vaccination stop for AstraZeneca vaccine	`https://www.zusammengegencorona.de/`	Wave 4	V
10/12/2021	Adoption of facility-based mandatory vaccination	`https://www.bgbl.de/`	Wave 4	V
21/12/2021	STIKO recommends shorter time span before booster shot	`https://www.rki.de/`	Wave 4	V
21/12/2021	EU Commission decides on limited validity of vaccination certificates	`http://data.europa.eu/`	Wave 4	V
22/02/2022	Authorization of Novavax vaccine	`https://www.pei.de/`	Wave 5	V

Table 5.3: List of policy events: Germany

Date	Event	Source	RKI Phase	Policy Phase
25/11/2020	Chancellor's office statement on ethical issues of vaccination published	https://www.bundeskanzleramt.gv.at/	Wave 2	I
21/12/2020	Authorization of BioNTech vaccine	https://www.basg.gv.at/	Wave 2	II
26/12/2020	Publication of priority groups	https://web.archive.org/	Wave 2	II
27/12/2020	Start of vaccinations	https://www.wienerzeitung.at/	Wave 2	II
06/01/2021	Authorization of Moderna vaccine	https://www.basg.gv.at/	Wave 2	II
28/01/2021	Regulation on the procedure of Corona vaccination enters into effect	https://www.ots.at/	Wave 2	II
29/01/2021	Authorization of AstraZeneca vaccine	https://www.basg.gv.at/	Wave 2	II
01/02/2021	Vaccination plan published	https://www.sozialministerium.at/	Wave 2	II
11/03/2021	Authorization of Johnson &d Johnson vaccine	https://www.basg.gv.at/	Wave 3	II
25/05/2021	BioNTech recommended for persons older than 11	https://www.basg.gv.at/	Wave 3	IV
23/07/2021	Moderna recommended for persons older than 11	https://www.basg.gv.at/	Summer plateau 2021	IV
08/10/2021	BioNTech booster recommended for from 6 months after second jab	https://www.basg.gv.at/	Wave 4	IV
23/10/2021	Introduction of lockdown for the unvaccinated upon reaching 600 patients in intensive care	https://www.wienerzeitung.at/	Wave 4	IV
29/10/2021	Moderna booster recommended for from 6 months after second jab	https://www.basg.gv.at/	Wave 4	IV
15/11/2021	Nationwide lockdown for the unvaccinated	https://www.wienerzeitung.at/	Wave 4	V
25/11/2021	Authorization of BioNTech vaccine for children aged 5 - 11	https://www.basg.gv.at/	Wave 4	V
17/12/2021	Johnson & Johnson booster recommended for from 6 months after second jab	https://www.basg.gv.at/	Wave 4	V
20/12/2021	Authorization of Novavax vaccine	https://www.basg.gv.at/	Wave 4	V
20/1/2022	Introduction of compulsory vaccinations for all citizens older than 17	https://www.wienerzeitung.at/	Wave 5	V

Table 5.4: List of policy events: Austria

Date	Event	Source	RKI Phase	Policy Phase
17/12/2020	Vaccination sequence published by the BAG	https://www.bag.admin.ch/	Wave 2	II
19/12/2020	Authorization of BioNTech vaccine	https://www.admin.ch/	Wave 2	II
23/12/2020	Start of vaccinations	https://www.srf.ch/	Wave 2	II
12/01/2021	Authorization of Moderna vaccine	https://www.swissmedic.ch/	Wave 2	II
13/01/2021	Vaccination costs covered by statutory health insurance approved	https://www.bag.admin.ch/	Wave 2	II
21/01/2021	Zurich is one of the first regions to start involving primary care doctors	https://telebasel.ch/	Wave 2	II
03/02/2021	Swissmedic requests further data for approval for AstraZeneca	https://www.swissmedic.ch/	Wave 2	II
22/03/2021	Authorization of Johnson & Johnson vaccine	https://www.swissmedic.ch/	Wave 3	II
22/04/2021	Work on international vaccination certificate begins	https://www.bag.admin.ch/	Wave 3	III
01/06/2021	Those who have recovered should also be vaccinated	srf.ch/	Wave 3	IV
04/06/2021	Legal basis for issuance of vaccination certificates created	https://www.bag.admin.ch/	Wave 3	IV
26/10/2021	Booster vaccination recommended for persons older than 65	https://www.bag.admin.ch/	Wave 4	IV
04/11/2021	Authorization of AstraZeneca not further pursued	https://www.swissmedic.ch/	Wave 4	V
26/11/2021	Booster vaccination recommended for the general population	https://www.bag.admin.ch/	Wave 4	V
10/12/2021	Authorization of BioNTech vaccine for children	https://www.bag.admin.ch/	Wave 4	V
21/12/2021	Recommendation to shorten the time until the booster vaccination	https://www.bag.admin.ch/	Wave 4	V

Table 5.5: List of policy events: Switzerland

Policy Phase	Begin	End	Description
I	01/11/2020	10/12/2020	Beginning of the official COVID-19 vaccination policies
II	10/12/2020	15/04/2021	Publishing of vaccination strategies, authorization of the vaccines and vaccination start, halt, and resumption of AstraZeneca vaccinations in Germany
III	15/04/2021	15/05/2021	Suspension of priority groups for AstraZeneca vaccines in D; international vaccination certificate preparations in CH
IV	15/05/2021	01/11/2021	Vaccine recommendations for specific age groups, access restrictions for unvaccinated persons in Germany
V	01/11/2021	30/01/2021	Booster shot recommendations and authorizations of vaccines for children; AstraZeneca vaccination stop in Germany; lockdowns for unvaccinated under certain conditions in Austria

Table 5.6: Policy phases for the DACH countries

5.3.4 Detection of Peaks in Tweet Frequencies

We define a peak as a point in time where the respective value deviates from the expected interval (mean +/- standard deviation (std)) by more than 1.5 times the expected maximum or minimum value.

$$((mean+std)+(|(mean+std)|*1.5)>peak<((mean-std)-(|(mean-std)|*1.5|)$$

5.3.5 Detection of Change Points

For detecting points in which the tweet frequency changes in the time series data, arguably due to shifts in public attention or opinion, we use the Python library ruptures (Truong et al., 2020). It includes several offline change detection methods for non-stationary signals. We opt for the Pelt (Penalized change point detection) search algorithm, which does not require setting a fixed number of change points in advance. This implementation computes the segmentation, which minimizes the constrained sum of approximation errors for a given model and penalty level (Killick et al., 2012). We use Pelt with the ruptures standard parameters.

5.3.6 Detection of Trends

We employ the Mann-Kendall test (Mann, 1945; Kendall, 1975) to determine whether significant trends in the data regarding sentiments and tweet frequencies exist. We use the Original Mann-Kendall test supplied by Hussain and Mahmud (2019)[5]. This non-parametric test does not consider serial correlation or seasonal effects. The standard alpha significance level is set at 0.05.

5.4 Data Analysis and Results

5.4.1 Evolution of Vaccination Discourse in DACH Countries

To investigate RQ1, we analyze the development of tweet frequencies and sentiments in the vaccination discourse over time and relate them to the general German Twitter discourse.

As Figure 5.1 illustrates, before December 2020, only very few tweets mention any vaccination-related terms. This affirms that the vaccination discourse captured by our automatically generated seed list is indeed driven by COVID-19 vaccinations. Figure 5.2 plots the REL sentiment scores, i.e., they are normalized with regard to the number of tweets. Both figures reveal that the overall vaccination discourse shows stronger negative than positive sentiments. Moreover, the discourse becomes slightly more negative over time. Thus, the negative sentiments were more negative than the positive sentiments were positive. Both the plotted sentiment and tweet frequencies hint at strong fluctuations over time.

Trend analysis using the Mann-Kendall test reveals a significant decreasing trend both for the relative overall sentiment and the relative negative sentiment intensities.

[5] `https://pypi.org/project/pymannkendall/`

We find no trend for the relative positive sentiment intensities. This means that the negative sentiments became more negative over time while the intensity of positive sentiments remained constant.

However, while the summed-up sentiment intensities (SUM) are more negative than positive, the number of predominantly positive tweets is slightly higher than the number of predominantly negative tweets: 53,176 positive tweets (26.69%) and 49,342 (24.77%) negative tweets of 199,207 tweets in total (including tweets with neutral/mixed sentiment). Thus, negative sentiments seem to be expressed with higher intensity than positive tweets. Yet, with a mean of -0.09, the average relative sentiment is close to neutral/mixed. The number of positive and negative tweets as well as the overall tweet frequency increase significantly with time.

We further investigate whether the negative sentiments are inherent to the vaccination discourse or due to an overall negative German Twitter discourse. For that, we analyze the sentiments for all German tweets harvested with our pipeline during the investigated time frame. Figure 5.3 relates the sentiments in the German vaccination tweets to the sentiments in German tweets of all topics in the same time frame. Depicted sentiments are the REL scores. The strong fluctuations in sentiment at the beginning of the year 2020 for the vaccination-related tweets can be attributed to the relatively low number of tweets in that time frame (see Figure 5.1). Similarly, the vaccination sentiments seem to exhibit higher fluctuations due to the smaller number of tweets compared to the general Twitter discourse. The results show that the discourse about vaccinations is more negative than the general discourse in German tweets. In the latter, the sentiment was overall more positive than negative, both in terms of summed-up and relative sentiment intensities, with a mean of 0.05 for the REL score (as compared to -0.09 for the vaccination tweets) and in terms of numbers of tweets with 1,758,776 (14.30%) negative and 3,015,915 (24.53%) positive of a total of 12,297,163 tweets. There is also a significant negative trend for the general German Twitter discourse, which is caused by both the negative sentiments becoming more negative and the positive sentiments becoming less positive, according to the Mann-Kendall trend analysis. Also for the general discourse, numbers for negative, positive, and all tweets show a significant increasing trend.

While significant trends can be observed over time, both the tweet frequencies and sentiments also fluctuate heavily at different points in time. Also, while the summed-

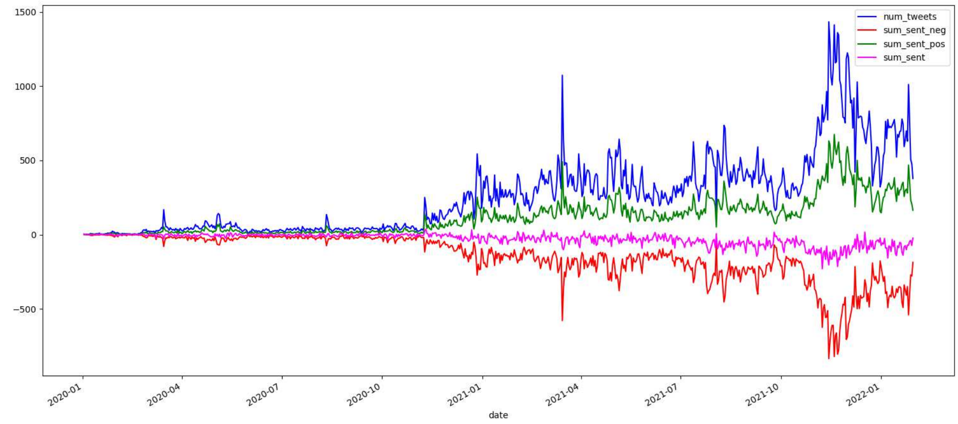


Figure 5.1: Frequencies and summed up sentiments of vaccination tweets over time. The blue line indicates the number of tweets on a given day, the green line the summed-up positive, the red line the summed-up negative, and the magenta line the overall summed-up sentiment intensities, respectively.

up positive and negative sentiments are relatively close to balancing each other out, the increasing trend for positive sentiments and the decreasing trend for negative sentiment scores suggest that the discourse is indeed rather emotional and increasingly so.

To get insights into what happened at different points in time, the following sections investigate in more detail which topics were discussed, in general, and when tweet frequencies increase, i.e. which topics are in the focus of attention, and which topics were responsible for positive and negative sentiments and sentiment trends.

5.4.2 Topics, Sentiments, and Themes

To answer RQ2, we first investigate the topics generated for all tweets as described in the *Topic Modeling* Section and rank them by their frequencies.

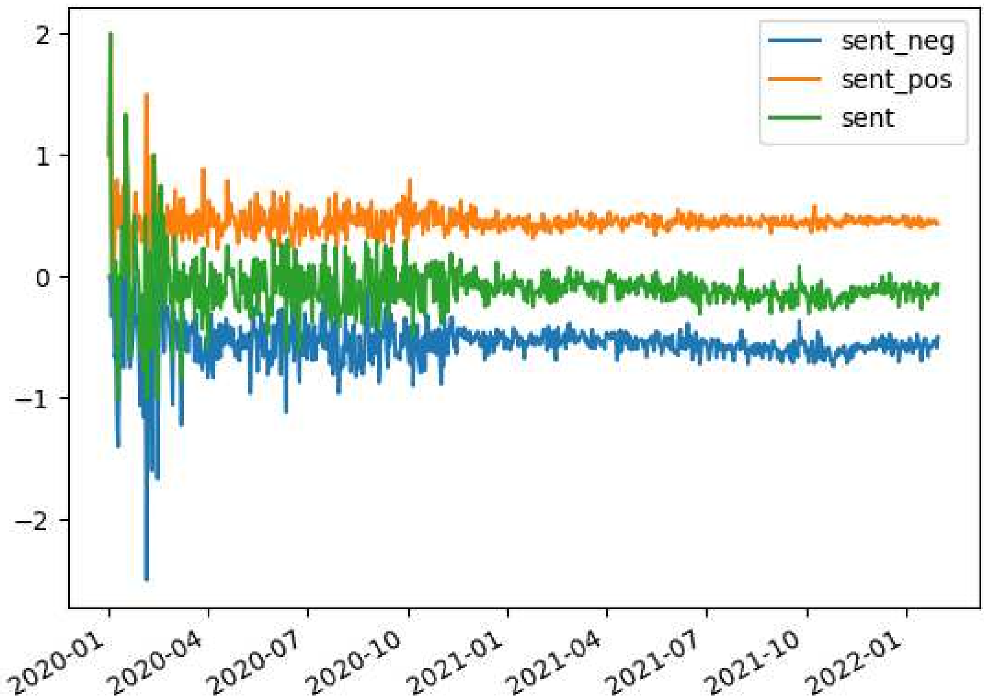


Figure 5.2: Relative sentiment of vaccination tweets over time

Topics and Associated Sentiments

The top 30 topics are depicted in Table 5.7.

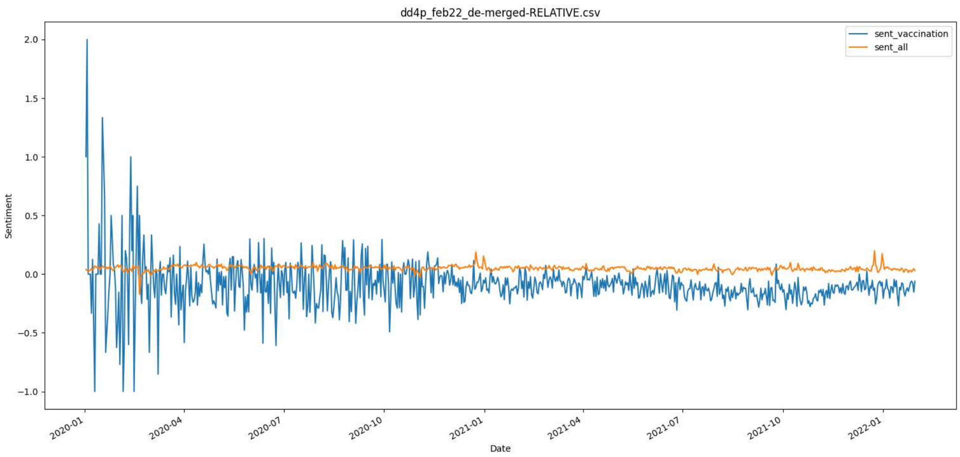


Figure 5.3: Sentiment in German tweets vs. sentiment in German vaccination tweets

Rank	**Topic Label**	#	♡
1	Children	8892	-0.12
2	Anecdotes: Experience with Corona vaccination	2210	-0.17
3	(Prominent) vaccinated and unvaccinated men	2058	-0.07
4	Situation in Germany and comparisons	1796	-0.07
5	'Do (not) get vaccinated'	1738	0.16
6	Corona and other flu viruses	1600	-0.25
7	Corona in Israel	1533	-0.20
8	Regulations	1476	-0.04
9	AstraZeneca vaccine	1425	-0.01
10	Duration of vaccination protection	1382	0.04
11	Lockdowns	1132	-0.10
12	Mutations and virus spread when vaccinated	1118	-0.11
13	Basic rights	1093	0.14
14	Practical implementation	1055	0.39
15	Propaganda and fake news	1021	-0.34
16	Vaccinations for elderly people	988	-0.13
17	Compulsory vaccination	938	-0.15
18	Metadiscussion about Twitter vaccination discourse	761	-0.15
19	Masks and mask mandate	729	-0.15
20	Vaccine efficacy for Omicron	676	-0.17
21	Compulsory vaccination at work	651	-0.06
22	SARSCoV2	635	-0.12
23	Demonstrations and protests	633	-0.04
24	Merkel	625	-0.07
25	Vaccinations for children: medical views	611	-0.08
26	mRNA vaccines	567	-0.02
27	Immune system	550	-0.19
28	Statistics about vaccination uptake and policies	535	-0.08
29	#AllesInDenArm	476	-0.01
30	Austria	472	0.00

Table 5.7: Top 30 topics (complete time interval. # Number of tweets ♡Average sentiment per tweet (REL)

The vaccination discourse covers a wide range of topics (Table 5.7). The most frequently discussed topic addresses the question of whether *children* should be vaccinated, how they can be protected, their role in transmitting the virus, and their role in the pandemic more generally. A high number of tweets share personal experiences with Corona vaccinations (*Anecdotes: Experience with Corona vaccination*), be it the authors' own or those of their peers, and discussing (prominent) (un)vaccinated persons, mostly men (*(prominent) vaccinated and unvaccinated men*). Analyzing the topic clusters, we find that the topic modeling algorithm encoded the gender of the individuals. The first topic featuring anecdotal information focuses on females, while this one mainly contains tweets about males, many prominent individuals among them. Specific influential individuals are discussed in other topics. The highest rankings are Germany's chancellor at the time, Angela *Merkel*, and Germany's current health minister Karl *Lauterbach*, ranking 24th and 31st, respectively. Since the beginning of the pandemic, Lauterbach frequently appeared in German TV shows and interviews, voicing his opinion about measures, policies, and possible developments. A country-level view of the pandemic ranks fourth (*Situation in Germany and comparisons*), focusing on the situation in Germany and comparing it with other countries. Other topics focusing on the country level are those debating Corona and the measures taken in Israel (*Corona in Israel*, rank 7) and *Austria* (rank 30). Twitter has also been used frequently for mobilizing others: a topic cluster containing calls to get vaccinated or not (*'Do (not) get vaccinated'*) ranks fifth, the *#AllesInDenArm* ("everythingIntoTheArm") campaign 29th. Prominent individuals and ordinary Twitter users used this hashtag to communicate their vaccination status and motivate others to get vaccinated. Comparing COVID-19 infections with the flu and discussing experiences with the swine flu and flu vaccinations (*Corona and other flu viruses*) also received significant attention, ranking this topic sixth. This topic also includes tweets discussing the severity of COVID-19 infections and whether the classification as a pandemic is justified.

Rules and *Regulations* are also widely discussed. This topic contains tweets about the so-called 2G and 3G rules which restricted access to certain facilities to vaccinated, recovered, or negatively tested individuals (rank 8). Other regulations that were frequently debated were *Lockdowns* (rank 11), *Masks and mask mandates* (19). Discourses about *Compulsory vaccinations* reached rank 17, and about *Compulsory vaccinations at work* specifically rank 21. Relatedly but assigned to a separate topic, we find a more general debate about freedom, personal responsibilities, restrictions, and *Basic rights* (13). Closely related is the topic of *Demonstrations* and protests (23). The vaccine that seemed to spark the most discussions is the *AstraZeneca vaccine*, ranking 9th among all topics. *mRNA vaccines*, i.e., gene-based vaccines, belong to the topic cluster ranking 26th, which evolves primarily around their mode of operation. Other vaccines are clustered into separate topics, too, but do not rank among the 30 most frequent topics (Sputnik: 42, BioNTech: 56, Novavax: 81, Johnson & Johnson: 100 and, when referred to as 'J&J' 148, Moderna: 101). Tweets dealing with the efficacy of vaccinations can be found in many topics in the top 30. The *Duration of vaccination protection* ranks 10th. *Mutations and virus spread when vaccinated* at rank 12 refers to the extent to which the virus can be spread by the vaccinated and whether vaccinations pander to the evolution of mutations. *Vaccine efficacy for Omicron* is discussed at rank 20. A more general debate about the influence of vaccinations and the Coronavirus on the *Immune system* ranks 27th. This topic also includes debates about immunization through a COVID-19 infection compared to immunization through vaccination. *Vaccinations for elderly people* are discussed in the topic ranking 16th, *Vaccinations for children: medical views* at rank 25. Note that the latter topic focuses on medical considerations while other aspects can be found in the more general *Children* topic ranking first. The *Practical implementation* of the vaccination rollout also receives much attention on Twitter (rank 14). This topic includes tweets regarding opportunities for getting vaccinated.

Meta discussions about the news coverage, *Propaganda and fake news* and about the debate on Twitter (*Metadiscussion about Twitter vaccination discourse*) both make it into the top 20 topics (ranks 15 and 18, respectively). Topics 22 and 28 include heterogeneous collections of tweets: 22 using the *SARSCoV2* term and hashtag (discussing a range of topics), 28 focusing on *Statistics about vaccination uptake and policies*.

The Mann-Kendall test reveals significant positive trends regarding the number of tweets for all of the top 30 topics, i.e., all of them gain increasing attention over time.

Themes

As outlined above, many topics refer to similar issues with varying levels of granularity, e.g., the duty to vaccinate in general (17) vs. the duty to vaccinate at work (21). This restricts the informative value of the frequency rankings.

Thus, we manually identify more general themes to investigate their salience over time and their relations to vaccination policy events. For this, we adopt the same workflow as for generating topic labels: each of the first two authors examines all topic labels and maps them to themes. On this basis, we arrive at the following final set of themes that include at least three topics: 1) freedom and civic liberties, 2) safety and side effects of vaccinations, 3) effectiveness of vaccinations, 4) mobilization, 5) details about the vaccination campaign, 6) conspiracy theories, 7) country comparisons, 8) influential individuals and their stances or behaviors, 9) specific vaccines and 10) data about the pandemic.

Tables 5.8 and 5.9 list the respective themes, the included topics, their summed-up frequencies, and average relative sentiment scores.

Theme	#	♡	Included topics
freedom and civic liberties	9975	-0.06	8 "Regulations", 11 "Lockdowns", 13 "Basic rights", 17 "Compulsory vaccination", 19 "Masks and mask mandate", 23 "Demonstrations and protests", 21 "Compulsory vaccinations at work", 34 "Compulsory vaccination II", 46 "Political parties on compulsory vaccination", 50 "Vaccination as solidarity", 60 "Vaccination as a personal free choice", 79 "Police and police officers", 82 "Travel", 84 "Nazis", 86 "Fascism, vaccination as suppression", 94 "Privileges for the vaccinated and enforcement of vaccinations", 107 "Compulsory vaccination compared to road safety", 112 "Concerts and musicians", 117 "Compulsory vaccination debate", 118 "Spahn's compulsory vaccination statement", 119 "Restaurant visits", 121 "Footballers and professional athletes", 125 "Discrimination of unvaccinated persons", 129 "Corona policies", 137 "Democracy, dictatorship, society", 144 "Compulsory vaccinations for all"
safety and side effects	5437	-0.17	16 "Vaccinations for elderly people", 25 "Vaccinations for children: medical views", 35 "Deaths due to or with Corona vaccination", 36 "Risks for pregnant women and infertility", 40 "Immediate vaccination side-effects", 48 "EMA", 49 "Side effects and risks with and without vaccination", 63 "Deaths after vaccination", 65 "Myocarditis risk after vaccination", 69 "AstraZeneca for specific age groups", 76 "Vaccination side effects", 83 "AstraZeneca vaccination stop", 91 "Safety of vaccinations", 104 "Effects of vaccinations on the menstrual cycle", 109 "Myocarditis risks", 110 "Pregnancy and risks", 127 "Side effects of vaccinations II", 134 "EMA recommendations and authorizations", 142 "Development of vaccines and their tests", 147 "Allergies and allergic reactions"
effectiveness	5040	-0.09	10 "Duration of vaccine protection", 12 "Mutations due to vaccinations, spread of the virus when vaccinated", 20 "Vaccine efficacy for Omicron", 27 "Immune system", 32 "vaccination protection", 98 "Vaccination protection and efficacy", 115 "Number of vaccinated people in hospitals", 116 "Infections after being vaccinated", 136 "Anecdotes of vaccinations and infections", 140 "Virus variants and mutations"

Table 5.8: Mapping of topics to themes. # Number of tweets; ♡average REL sentiment; *Included topics* ranks of the topics regarding their frequencies in all tweets and their manually assigned labels

Theme	#	♡	Included topics
mobilization	4982	0.12	5 ""Do (not) get vaccinated"", 29 "#AllesInDenArm", 41 "Appeals to get vaccinated", 47 ""I am vaccinated"", 54 ""I will not be vaccinated"", 68 "Opinion about own vaccination", 70 "Congratulations to being vaccinated", 71 "Communication of free vaccination appointments", 80 ""Got the second vaccination"", 85 "Booking of vaccination appointments", 89 "Disputations between vaccinated and unvaccinated persons", 97 "Vaccination appointments for children", 108 "Vaccination status updates", 124 "Personal reasons for or against getting vaccinated", 131 "Disputes", 141 "Booster shots", 143 "Calls to sign petitions"
vaccination campaign	2730	0.20	14 "Practical implementation", 33 "Vaccine purchase EU", 58 "Costs, monetary incentives and penalties", 59 "Priority groups", 64 "Bratwurst incentives", 66 "Who pays", 90 "Apps and digital vaccination certificates", 120 "Booked and free vaccination appointments"
conspiracy theories	1601	-0.18	15 "Propaganda and fake news", 37 "Bill Gates and vaccinations", 74 "Chips and implants"
country comparisons	6229	-0.13	4 "Situation in Germany (also in comparison with other countries)", 7 "Corona in Israel", 30 "Austria", 43 "Russia", 61 "Africa", 62 "Italy", 67 "France", 73 "Portugal", 75 "Vaccination of children in specific regions", 77 "China", 88 "Patent clearance", 92 "Great Britain", 105 "Israel", 106 "Global distribution of vaccines", 126 "Gibraltar", 138 "Switzerland"
influential individuals	4471	-0.08	3 "(prominent) vaccinated and unvaccinated men", 24 "Merkel", 31 "Lauterbach", 39 "Kimmich", 55 "Trump and Biden", 72 "Soeder", 96 "Sucharit Bhakdi", 111 "Kubicki", 128 "Politicians", 149 "BioNTech's founder"
specific vaccines	3276	0.01	9 "AstraZeneca vaccine", 26 "mRNA / gene-based vaccines", 42 "Sputnik vaccine", 56 "BioNTech", 81 "Novavax vaccine", 100 "Johnson & Johnson vaccine", 101 "Moderna vaccine", 130 "AstraZeneca vaccine II", 145 "J&J vaccine"
data about the pandemic	754	0.08	44 "Statistics and headlines", 78 "Statistics about vaccination statuses", 113 "Statistics about the number of vaccinations", 133 "Statistics about vaccination rates in Germany"

Table 5.9: Mapping of topics to themes. # Number of tweets; ♡average REL sentiment; *Included topics* ranks of the topics regarding their frequencies in all tweets and their manually assigned labels

Freedom and civic liberties contains tweets that discuss mandates and restrictions and debate force vs. free choice when it comes to Corona regulations. This theme has the highest number of tweets followed by *country comparisons*, which bundles tweets comparing the situation in different countries. The third largest theme is *safety and side effects*. This theme comprises tweets centering around the topic of short- and long-term side effects of vaccinations and risks connected to being vaccinated vs. unvaccinated. *Effectiveness* designates tweets discussing the efficacy and usefulness of vaccinations, such as the duration of the offered protection or the virulence of vaccinated individuals. This theme also received high attention. In the *mobilization* theme, we merge all topics containing messages motivating others to get vaccinated or to not get vaccinated. The stances, roles, or behaviors of authorities, leaders, or other influential individuals are discussed broadly as reflected by the sixth largest theme, *influential individuals*. There is a high interest in discussing *specific vaccines*. Details about the *vaccination campaign* refer to different aspects of the vaccination strategy. This theme measures how policy events are reflected in the Twitter discourse. The theme *conspiracy theories* contains tweets discussing different theories, e.g., concerning Bill Gates' motives regarding vaccinations, but also sarcastic tweets and tweets discussing news coverage and perceived propaganda on a meta-level. Therefore, many tweets belonging to this theme cannot be interpreted as a high level of belief in conspiracies or media distrust. Instead, it signals high attention to these topics. Twitter is also used to share *data about the pandemic*, e.g., ratios of vaccinated persons or any other statistics.

This analysis reveals that while many tweets concern health-related issues (safety and side effects, effectiveness, and specific vaccines), a very high number of tweets focus on the country level and the effects of policies on society. Overall, Twitter users seem to be similarly concerned about their freedom and civic liberties and health-related concerns.

The Mann-Kendall test reveals that all themes exhibit a significant increasing trend regarding their tweet frequencies over the whole time span under investigation reflecting increased attention to vaccination-related discourse as a whole.

5.4.3 Topic and Theme Sentiments Over Time

To address RQ3, we next analyze the frequencies and sentiments of topics and themes over time.

Topic sentiments: complete time interval

The average sentiment scores support the analysis in Section *Evolution of Vaccination Discourse in DACH Countries*. The sentiments for many topics are neither very negative nor positive when averaged over the whole time span, with a few exceptions. The topics with the most positive sentiments of the top 30 were *Practical implementation* (0.39), *Do (not) get vaccinated* (0.17), *Basic rights* (0.14) and *Duration of vaccination protection* (0.04). These are also the only topics with an overall positive sentiment. The *Practical implementation* topic contains many tweets with people celebrating their vaccination appointments, *Do (not) get vaccinated* many calls to action, and *Basic rights* many references to positive concepts like freedom, privileges, and rights. The latter, however, contains many critical voices regarding vaccinations and compulsory vaccinations. The most negative of the top 30 topics were *Propaganda and fake news* (-0.34), followed by *Corona and other flu viruses* (-0.25), *Corona in Israel* (-0.2), *Immune system* (-0.19), *Vaccine efficacy for Omicron* (-0.17) and *Anecdotes: Experience with Corona vaccination* (-0.17). Also, all regulations and potential regulations (*Regulations*, *Lockdowns*, *Compulsory vaccination*, *Compulsory vaccination at work*, *Masks*) are connected to negative sentiments. While sentiments are not to be interpreted as stances, i.e. negative sentiments do not necessarily signal disapproval, these negative scores suggest that these topics were coupled with a focus on negative aspects in the discussion.

The Mann-Kendall test reveals significant negative trends regarding the relative sentiment for the following seven of the top 30 topics: *(Prominent) vaccinated and unvaccinated men*, *Vaccinations for elderly people*, *Corona in Isreal*, *Anecdotes: Experience with Corona vaccination*, *Immune system*, *Vaccine efficacy for Omicron*, *Propaganda and fake news*.

The only topics with significant positive trends are *"Do (not) get vaccinated"* and *Practical implementation*. No significant trends are found for the remaining topics.

When analyzing only the relative positive respectively negative sentiment scores of all tweets on these topics, we observe a significant positive trend for positive sentiments and a significant negative trend for relative negative sentiments for all topics except *Corona and other flu viruses*, for which there is no significant trend for the relative negative sentiments. This finding suggests that the discourse became more emotional and polarised over time. We will check next whether this also holds true for the vaccination discourse beyond the most prominent individual topics.

Theme sentiments: complete time interval

When analyzing only the relative positive respectively negative sentiment scores of all tweets, we observe a significant positive trend for positive sentiments and a significant negative trend for relative negative sentiments for all themes. While the impact on the consolidated sentiment varies across themes, this indicates that the vaccination discourse as a whole gets more emotional and polarised over time. The highest mean relative sentiment (i.e. the most positive sentiment) is connected to *details of the vaccination campaign* (0.20) followed by *mobilization* (0.12), which include the most positive individual topics, as outlined in the previous subsection. Together with *data about the pandemic* (0.08) and *specific vaccines* (0.01), these are the only themes with non-negative average sentiment scores. These are also the only themes with an overall positive trend. For all other themes, the negative trend is stronger than the positive one and they become significantly more negative. The lowest (i.e. most negative sentiment) is connected to *conspiracy theories* (-0.18), *safety and side effects* (-0.17), and *country comparisons* (-0.13).

We analyze the evolution of tweet frequencies for themes in more detail as part of the following section.

5.4.4 Relation to Phases and Policy Events

R	Topic	#	%	♡
	Sporadic cases			
1	country comparisons	16	34.04%	0.19
2	effectiveness	6	12.77%	0.33
3	conspiracy theories	5	10.64%	-0.20
3	mobilization	5	10.64%	0.20
4	influential individuals	4	8.51%	0.75
5	details of the vaccination campaign	3	6.38%	-1.33
5	data about the pandemic	3	6.38%	0.00
5	freedom and civic liberties	3	6.38%	0.00
6	safety and side effects	1	2.13%	0.00
6	specific vaccines	1	2.13%	0.00
	Wave 1			
1	conspiracy theories	176	24.14%	0.06
2	influential individuals	154	21.12%	-0.14
3	effectiveness	104	14.27%	-0.05
4	freedom and civic liberties	95	13.03%	-0.01
5	country comparisons	70	9.60%	0.17
6	mobilization	43	5.90%	0.19
7	data about the pandemic	39	5.35%	-0.05
8	safety and side effects	29	3.98%	0.03
9	details of the vaccination campaign	13	1.78%	0.31
10	specific vaccines	6	0.82%	0.33
	Summer Plateau 2020			
1	country comparisons	202	24.16%	-0.04
2	freedom and civic liberties	137	16.39%	0.05
3	effectiveness	104	12.44%	0.11
4	conspiracy theories	92	11.00%	-0.02
5	influential individuals	91	10.89%	0.09
6	mobilization	62	7.42%	0.35
7	specific vaccines	50	5.98%	-0.08
8	safety and side effects	42	5.02%	-0.17
9	details of the vaccination campaign	30	3.59%	0.13
10	data about the pandemic	26	3.11%	0.08
	Wave 2			
1	freedom and civic liberties	1139	17.27%	-0.08
2	country comparisons	1101	16.69%	-0.06
3	influential individuals	865	13.12%	0.06
4	safety and side effects	812	12.31%	-0.18
5	specific vaccines	738	11.19%	-0.05
6	effectiveness	610	9.25%	-0.10
7	mobilization	532	8.07%	0.16
8	details of the vaccination campaign	403	6.11%	0.06
9	conspiracy theories	266	4.03%	-0.10
10	data about the pandemic	129	1.96%	0.06

Table 5.10: Theme frequency, all tweets (different phases of the pandemic). *R* Rank; # Number of tweets; % Relative number of tweets; ♡Average sentiment per tweet (REL)

R	Topic	#	%	♡
	Wave 3			
1	safety and side effects	1611	17.55%	-0.14
2	freedom and civic liberties	1390	15.14%	-0.04
3	specific vaccines	1384	15.08%	-0.01
4	country comparisons	1254	13.66%	-0.06
5	mobilization	995	10.84%	0.19
6	influential individuals	789	8.60%	-0.02
7	effectiveness	711	7.75%	-0.09
8	details of the vaccination campaign	687	7.48%	0.12
9	conspiracy theories	220	2.40%	-0.10
10	data about the pandemic	138	1.50%	0.03
	Summer Plateau 2021			
1	freedom and civic liberties	742	18.10%	-0.09
2	country comparisons	675	16.46%	-0.26
3	mobilization	569	13.88%	0.14
4	details of the vaccination campaign	504	12.29%	0.32
5	safety and side effects	487	11.88%	-0.24
6	effectiveness	359	8.76%	-0.17
7	influential individuals	324	7.90%	-0.10
8	specific vaccines	234	5.71%	-0.03
9	conspiracy theories	130	3.17%	-0.23
10	data about the pandemic	76	1.85%	0.17
	Wave 4			
1	freedom and civic liberties	5140	27.77%	-0.07
2	effectiveness	2362	12.76%	-0.09
3	country comparisons	2341	12.65%	-0.20
4	mobilization	2335	12.61%	0.07
5	safety and side effects	2073	11.20%	-0.14
6	influential individuals	1756	9.49%	-0.16
7	details of the vaccination campaign	958	5.18%	0.20
8	specific vaccines	702	3.79%	0.13
9	conspiracy theories	543	2.93%	-0.27
10	data about the pandemic	301	1.63%	0.13
	Wave 5			
1	freedom and civic liberties	1403	28.84%	-0.06
2	effectiveness	831	17.08%	-0.09
3	country comparisons	622	12.79%	-0.04
4	influential individuals	513	10.54%	-0.11
5	mobilization	483	9.93%	0.10
6	safety and side effects	428	8.80%	-0.30
7	specific vaccines	189	3.88%	0.04
8	conspiracy theories	180	3.70%	-0.39
9	details of the vaccination campaign	167	3.43%	0.46
10	data about the pandemic	49	1.01%	-0.02

Table 5.11: Theme frequency, all tweets (different phases of the pandemic). *R* Rank; # Number of tweets; % Relative number of tweets; ♡Average sentiment per tweet (REL)

Next, we analyze the attention to themes during different pandemic phases and in relation to policy events to answer RQ4.

Relation to pandemic phases

Tables 5.10 and 5.11 list the frequencies and connected sentiments of all themes during different phases of the pandemic. Since the phases differ in their durations and the number of tweets has been increasing overall over time, we also list the relative number of tweets as a percentage of tweets belonging to any of the themes.

We exclude the phase *Sporadic cases* from the following analysis as it did not contain enough tweets to derive meaningful rankings.

The themes *freedom and civic liberties* and *country comparisons* were prominent throughout all phases of the pandemic: they have the highest (i.e. the top) mean ranks across all time intervals (1.71 and 2.86, respectively) and a high rank stability with standard deviations of 1.11 and 1.35, respectively. For *freedom and civic liberties*, we observe an increased frequency, both absolute and relative, in the last two phases, i.e. it received more attention during the later phases than during the early ones. Both themes have their lowest rank during the first wave and their second lowest during Wave 3. The top 3 themes *effectiveness* (mean rank 4.14) also has its lowest rank in Wave 3, i.e. there it received its least attention. This phase was dominated by the *safety and side effects* theme (mean rank 5.29). Ranking third in Wave 3, behind *freedom and civic liberties*, we find *specific vaccines*. In other phases, *specific vaccines* received attention, too, but to a lesser degree (mean rank 6.86). *Safety and side effects*, *specific vaccines* and *effectiveness* show the greatest fluctuations in ranks with 2.43, 2.27, and 2.12 standard deviation, respectively. Only *conspiracy theories* fluctuates more (3.21).

The development of the highly fluctuating themes *Safety and side effects*, *Effectiveness*, *Specific vaccines* and the dominating theme *Freedom and civic liberties* is illustrated in Figure 5.4. This analysis shows that attention to topics in Wave 3 differs from the other phases and that directly vaccine-related themes were the most unstable. We will investigate possible reasons in more detail in section 5.4.4.

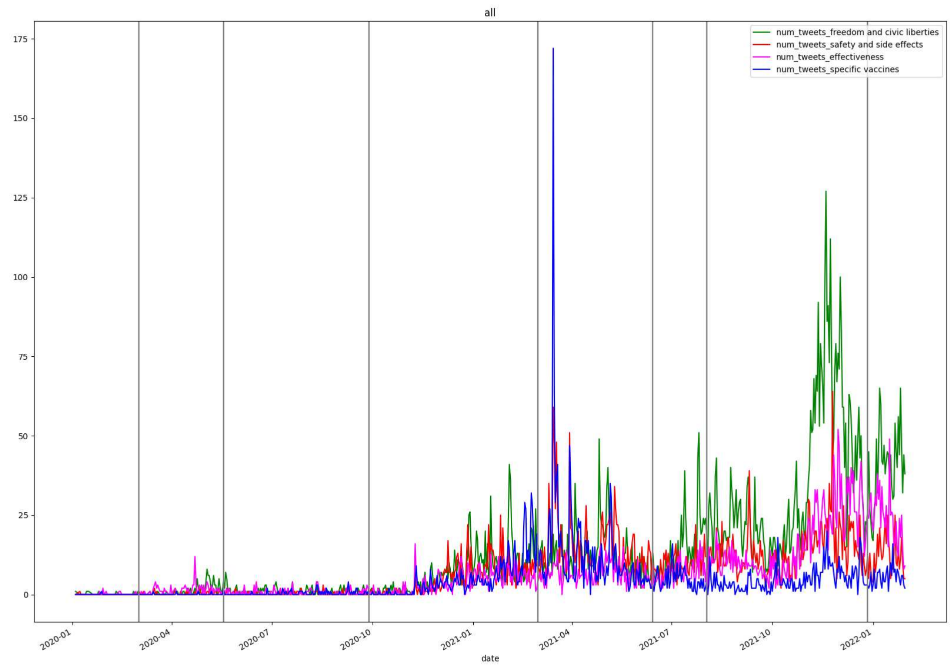


Figure 5.4: Tweet frequencies of the themes *Freedom and civic liberties*, *Safety and side effects*, *Effectiveness* and *Specific vaccines*. Grey lines mark the start of a pandemic phase

Relation to policy events

In order to reveal possible connections to policy actions, we investigate the relation of change points and peaks for the most fluctuating and dominant themes in 5.4. First, we relate them to the policy phases introduced in Table 5.6 with Figure 5.5.

Compared to Figure 5.4, the development of the themes seems to align better with the policy phases than with the pandemic phases as classified by the RKI.

This is further supported by the analysis of change points and peaks, as visualized in Figures 5.6, 5.7, 5.8, 5.9.

Considering the policy phases outlined in Table 5.6, phase I from November until

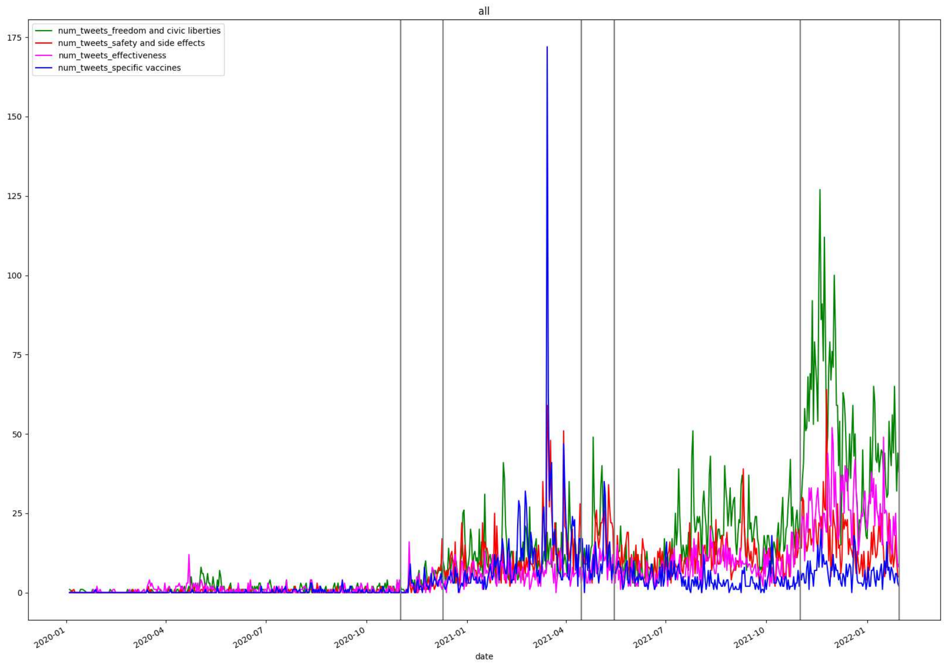


Figure 5.5: The selected themes over different policy phases. Grey lines mark the beginning of a policy phase

mid-December 2020 marks the beginning of official COVID-19 vaccination policies in the DACH region. This period went along with change points in tweet numbers for *Specific vaccines*, which started to be discussed frequently after this point.

In phase II starting mid-December 2020, the first vaccination recommendations and strategies for COVID vaccinations were published and the first vaccines were authorized before vaccinations finally started on December 23 (Switzerland) and 27 (Austria and Germany). This went along with changes in tweet frequencies according to the change point analysis for *Effectiveness*, *Safety and side effects*, and *Freedom and civic liberties* which received increasing attention from then on. In the same phase on March 15, 2021, AstraZeneca vaccinations were halted in Germany due to safety concerns (cf. Table 5.3) before they were resumed on March 25. We observe peaks for the themes *Safety and side effects* and *Specific vaccines* at the exact same time.

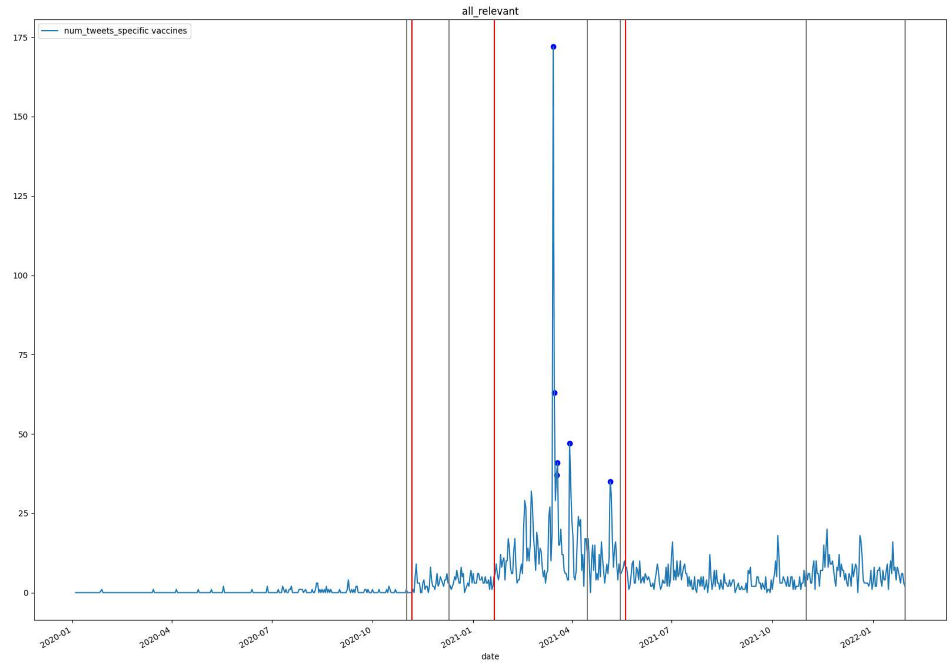


Figure 5.6: Change points and peaks in tweet frequency for the *Specific vaccines* theme. Grey lines mark the policy phases, red lines the change points, blue dots peaks. Change points occur at 06/11/2020, 20/01/2021, 20/05/2021. Peaks are detected at 15/03/2021, 16/03/2021, 18/03/2021, 19/03/2021, 30/03/2021, 06/05/2021.

On May 6, 2021, during the third phase, vaccinations with the unpopular AstraZeneca vaccine were possible for all individuals, regardless of priority group membership. A peak for *Specific vaccines* can be found on the same day.

Policy phase IV contained few policy events and few topic rank fluctuations, change points, and peaks. *Freedom and civic liberties* was the dominant theme. Restrictions for unvaccinated persons were discussed and finally implemented in Germany in August 2021.

November 2021 marked the month of booster recommendations and authorizations of vaccines for children. Also, on November 15, 2021, there was a nationwide lockdown

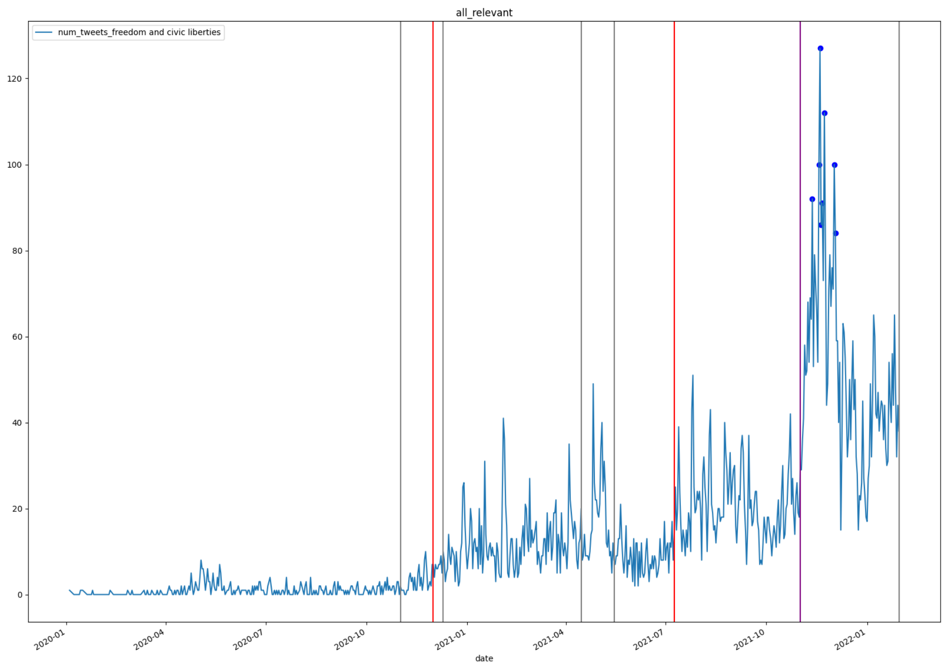


Figure 5.7: Change points and peaks in tweet frequency for the *Freedom and civic liberties* theme. Grey lines mark the policy phases, red lines the change points, purple lines policy phases and change points at the exact same day, blue dots peaks. Change points occur at 01/12/2020, 09/07/2021, 01/11/2021. Peaks are detected at 12/11/2021, 18/11/2021, 19/11/2021, 20/11/2021, 21/11/2021, 23/11/2021, 02/12/2021, 03/12/2021.

for unvaccinated persons in Austria. This and the following month contained the all-time peaks of the *Freedom and civic liberties* and the *Effectiveness* themes.

This analysis suggests that the high fluctuation of attention to the themes *Safety and side effects*, *Effectiveness*, and *Specific vaccines* were related to specific policy actions.

While vaccine safety and individual vaccines had been in public attention from the start, the decision to halt AstraZeneca vaccinations in Germany seemed to have had a big impact on the focus of public attention, fueling the controversy. However, soon other topics were again discussed more broadly. The theme of freedom and civic liberties received more attention early on and only lost its dominating position

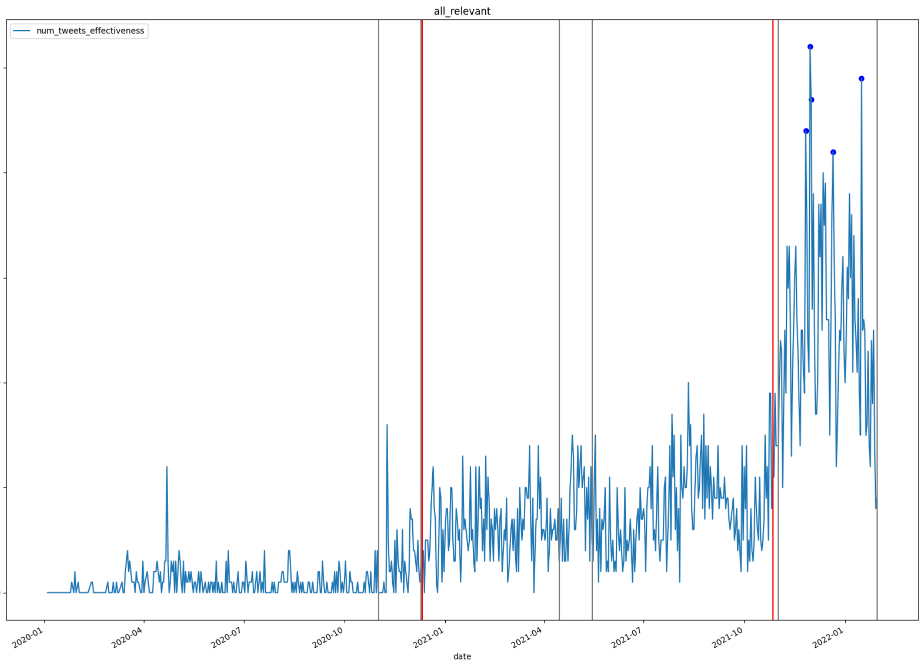


Figure 5.8: Change points and peaks in tweet frequency for the *Effectiveness* theme. Grey lines mark the policy phases, red lines the change points, blue dots peaks. Change points occur at 11/12/2020, 27/10/2021. Peaks are detected at 26/11/2021, 30/11/2021, 01/12/2021, 21/12/2021, 16/01/2022.

shortly during the peak of the AstraZeneca debate. Shortly after, it resumed being the focus of attention. Especially during the later phases of the pandemic, also restrictions for unvaccinated persons were debated, this theme was debated strongly. The recommendation to receive booster shots and the debates about the recommended time between the second and third vaccination seemed to have sparked a contested debate about the overall vaccine effectiveness. Concerns about safety and side effects also resurfaced at that time.

Overall, the debate on Twitter evolved strongly around the question of civic liberties. The public opinion about vaccinations in the DACH region did not seem to have become more positive or the concerns fewer, yet the willingness to get vaccinated increased. This analysis suggests that this can most likely be attributed to the incen-

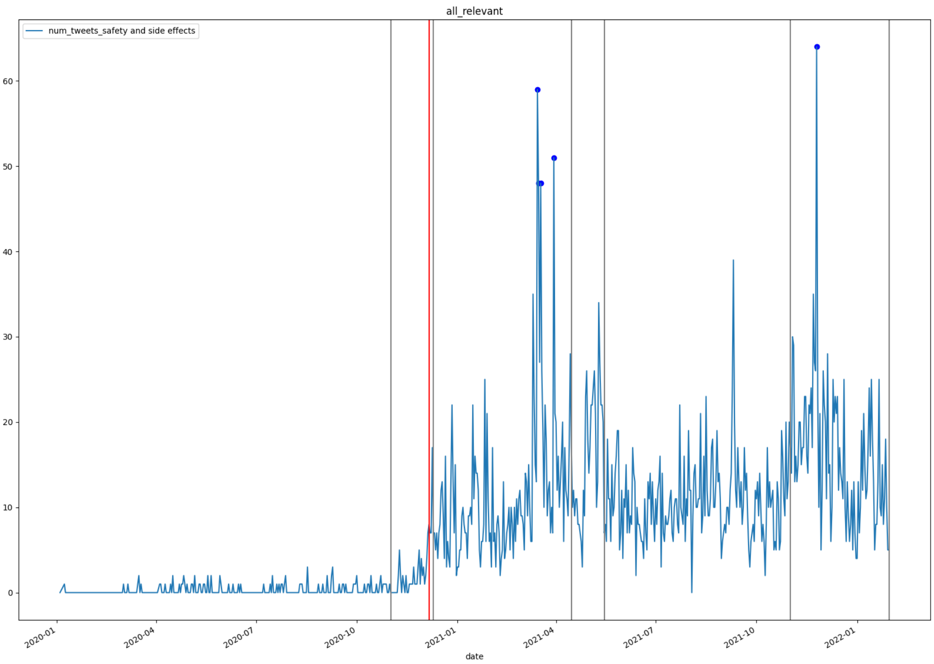


Figure 5.9: Change points and peaks in tweet frequency for the *Safety and side effects* theme. Grey lines mark the policy phases, red lines the change points, blue dots peaks. Change points occur at 06/12/2020. Peaks are detected at 15/03/202, 16/03/2021, 18/03/2021, 30/03/2021

tives to get vaccinated or possible restrictions when not getting vaccinated. Another possible interpretation is that the discussion of these topics might have contributed to removing the medical concerns from the focus of attention, thereby removing an important driver of vaccine hesitancy: as Solís Arce et al. (2021) find in their survey-based study for low- and middle-income countries: "Vaccine acceptance is explained mainly by an interest in personal protection against COVID-19, whereas concerns about side effects are the most common reasons for hesitancy." However, the increasing polarization of the debate suggests that this came at the price of deepening the rifts between proponents and opponents of vaccinations, further politicizing the issue.

5.5 Summary and Discussion

Our empirical study used NLP methods to detect and analyze 199,207 tweets about COVID-19 vaccinations in the DACH region (Austria, Germany, Switzerland). The results reflect that the topic was controversially discussed: we find that the total number of tweets about this important societal issue increased over time, and the sentiments in the discourse became both more polarized and more negative (RQ1), cf. Figure 5.1. Generally, discourse about COVID-19 vaccinations has been significantly more negative than the average discourse on Twitter during the same time period (Figure 5.3).

Investigating RQ2 and RQ3, we find that the Twitter discourse data reveal fluctuations in the topics and themes (cf. 5.8 and 5.9) that are at the center of public attention: while medical concerns such as the safety and side-effects of vaccinations were prominently discussed early in the debate and concerning specific events (cf. Tables 5.10 and 5.11, Figures 5.5, 5.6, 5.8, 5.9), the focus increasingly shifted to a discussion of broader societal concerns: especially those regarding freedom and civic liberties (cf. Tables 5.10 and 5.11, Figures 5.5 and 5.7). At the same time, vaccination acceptance and uptake were low early in the pandemic and increased over time (Desson et al. (2020), Table 5.2). Our investigations into RQ4 give insights into possible drivers of these changes: Figures 5.4 and 5.5 show that shifts in the discourse align better with policy phases than pandemic phases. Figures 5.6, 5.7, 5.8 and 5.9 illustrate that attention peaks to themes were related to policy events such as halting vaccinations with AstraZeneca or incentivizing vaccinated persons. Thus, policies implementing incentives to get vaccinated or restrictions for unvaccinated people seem to have been successful in increasing vaccination uptake, either by the incentives/restrictions themselves or by removing the focus of public attention from medical concerns. However, based on our findings in RQ3, these policies did not increase citizens' positive sentiments about vaccinations in general (cf. Figure 5.2, Tables 5.10 and 5.11) but rather increased polarization (cf. page 177).

Moreover, these findings suggest that information campaigns about medical concerns might have been helpful in addressing citizens' concerns during the early stages of the pandemic. These concerns might have never been cleared for many COVID-19 vaccination critics. Instead, debates about compulsory vaccinations and benefits for vaccinated people sparked a discussion about freedom and civic liberties, which received more attention than medical concerns. This demonstrates that insights into the different reasons for vaccine hesitancy among the citizenry are crucial to design and implement adequate policy responses. While information campaigns about side-effects may be effective in addressing medical concerns, such campaigns would hardly convince people who oppose vaccinations because they feel that their individual freedom is being violated.

It is in line with this interpretation that vaccine hesitancy in the DACH region decreased compared to other European countries, but sentiments did not become more positive (RQ1). The situation in other countries received considerable attention during all phases of the pandemic (RQ3, cf. Tables 5.10 and 5.11). This suggests that policies and debates in other countries may strongly influence citizens' opinions and behaviors. Citizens seek orientation beyond the borders of their country. This finding indicates the need for international solutions and cooperation.

Lastly, our findings suggest that analyzing online discourse data can yield valuable insights for policy-makers regarding topics of interest and attention to public concerns in highly dynamic contexts such as the COVID-19 pandemic. Online discourses can be a fruitful data source in addition to traditional survey data. The findings presented in this study contain relevant information about the possible relationship between policy events and public opinion that could inform political decision-makers. Our analyses suggest that the changes in public attention align better with different policy phases than with phases reflecting the infection rates alone. Yet, our analysis has several limitations. First, Twitter users are not representative of the whole population. Therefore, analyzing tweets can serve to analyze fluctuations and tendencies but should not be interpreted as a representation of general public opinion. Second, while we tried to ingest as little prior knowledge into our analyses as possible, opting for a primarily data-driven approach, our analysis is influenced by the choice of policy events and the segmentation into pandemic and policy phases. We did not investigate other events beyond infection rates and policies that may influence or relate to the discourse, such as news or social media discussions. Third, the assignment of themes was based on automatically generated topics but is still subjective. Different abstraction levels would have also been valid. The same applies to the generation of topics as such. The generated topics are not entirely selective; e.g., topics in the cluster "individual vaccines", e.g., the AstraZeneca topic, contain tweets that also discuss side-effects and vice versa. The same is true for the topic clusters discussing the effectiveness of vaccinations and the Omicron variant. To not produce too much noise, we decided to assign each tweet to the most probable cluster and not assign any cluster for low-confidence assignments. For future work, we will investigate the effects of assigning tweets to multiple clusters, controlling for the noise generated by different thresholds and parameters, and assessing topic cluster stability in different settings. Last, our generated data can be analyzed further to draw more detailed insights on additional topics relating to the formation and change of public opinion related to COVID-19 vaccinations. For example, while the attitudes and behaviors of influential individuals appeared to play an essential role in the public discourse on Twitter, it would be interesting to differentiate between different types of individuals, such as politicians or celebrities, advocates and opponents of vaccinations, and genuine vs. false information in their statements to gain more insights on the role of issues of trust and misinformation.

Chapter 6

Conclusion

This thesis presented a multidisciplinary view on key concepts and tasks relating to the analysis of ODD. The model that is introduced on the basis of an extensive literature survey allows a fine-grained and precise representation of properties and relations of claims and related notions, contributing to a shared understanding across related research disciplines and facilitating data sharing and collaboration.

The analysis of the utility of state-of-the-art machine learning methods for the task of verified claim retrieval offers practical recommendations for the choice of algorithms and tools for fact-checking applications. This work moreover adds to a deeper understanding of the state of the art in this field, including further insights on the performance of different language models.

The thesis further shows that weakly supervised systems can aid in making powerful methods utilizable in lowe-resource settings, such as mining scientific publications to increase the transparency and reproducibility of science.

Finally, by combining computational methods and manual analyses, ODD can be used as research data to investigate important societal issues and potentially aid policy-makers in designing adequate responses in times of crisis.

It is still a long way to understand the functionality of recent deep learning methods and the biases encoded by their models and in ODD. Yet with a growing number of works and initiatives fostering transparency and raising awareness to practical concerns, deep learning research has started its way out of the ivory tower to learn to deal with ODD in the wild.

Bibliography

(2020). *Big data meets survey science: A collection of innovative methods.* Wiley Series in Survey Methodology. John Wiley & Sons, Ltd., 1 edition.

Accuosto, P. and Saggion, H. (2019). Transferring Knowledge from Discourse to Arguments: A Case Study with Scientific Abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W19-4505`.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/S16-1081`.

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (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. `https://www.doi.org/10.3115/v1/W14-2109`.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled Contextualized Embeddings for Named Entity Recognition. 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 724–728, Minneapolis, Minnesota. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/N19-1078.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Al-Bakri, M., Atencia, M., Lalande, S., and Rousset, M.-C. (2015). Inferring Same-as Facts from Linked Data: An Iterative Import-by-Query Approach. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 9–15.

Al-Khatib, K., Hou, Y., Wachsmuth, H., Jochim, C., Bonin, F., and Stein, B. (2020). End-to-end argumentation knowledge graph construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):7367–7374. https://www.doi.org/10.1609/aaai.v34i05.6231.

Al-Ramahi, M., Elnoshokaty, A., El-Gayar, O., Nasralah, T., and Wahbeh, A. (2021). Public discourse against masks in the covid-19 era: Infodemiology study of twitter data. *JMIR Public Health Surveill*, 7(4):e26780.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Allen, A., Teuben, P. J., and Ryan, P. W. (2018). Schroedinger's code: A preliminary study on research source code availability and link persistence in astrophysics. *The Astrophysical Journal Supplement Series*, 236(1):10.

Amba Hombaiah, S., Chen, T., Zhang, M., Bendersky, M., and Najork, M. (2021). Dynamic language models for continuously evolving content. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2514–2524, New York, NY, USA. Association for Computing Machinery.

Andrews, N., Stowe, J., Kirsebom, F., Toffa, S., Rickeard, T., Gallagher, E., Gower, C., Kall, M., Groves, N., O'Connell, A.-M., et al. (2022). Covid-19 vaccine effectiveness against the omicron (b. 1.1. 529) variant. *New England Journal of Medicine*, 386(16):1532–1546.

Angelico, C., Marcucci, J., Miccoli, M., and Quarta, F. (2022). Can we measure inflation expectations using twitter? *Journal of Econometrics*, 228(2):259–277.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Atefeh, F. and Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1):132–164.

Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Augenstein, I., Maynard, D., and Ciravegna, F. (2016a). Distantly supervised Web relation extraction for knowledge base population. *Semantic Web*, 7(4):335–349. `https://www.doi.org/10.3233/SW-150180`.

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016b). Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D16-1084`.

Balazevic, I., Allen, C., and Hospedales, T. (2019). TuckER: Tensor Factorization for Knowledge Graph Completion. 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 5184–5193, Hong Kong, China. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D19-1522`.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017a). 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.

Bar-Haim, R., Edelstein, L., Jochim, C., and Slonim, N. (2017b). Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/W17-5104.

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91. https://www.doi.org/10.1093/pan/mpu011.

Bardosh, K., de Figueiredo, A., Gur-Arie, R., Jamrozik, E., Doidge, J., Lemmens, T., Keshavjee, S., Graham, J. E., and Baral, S. (2022). The unintended consequences of covid-19 vaccine policy: why mandates, passports and restrictions may cause more harm than good. *BMJ Global Health*, 7(5):e008684.

Barrón-Cedeño, A., Elsayed, T., et al. (2020). Overview of CheckThat! 2020 — automatic identification and verification of claims in social media. In *Proceedings of the 11th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 215–236, Cham. Springer International Publishing.

Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., and Haouari, F. (2020). Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Atanasova, P., Zaghouani, W., Kyuchukov, S., Da San Martino, G., and Nakov, P. (2018). Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 2: Factuality. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CLEF 2018 Working Notes, page 13.

Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. and Koller, A. (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*, pages 5185–5198, Online. Association for Computational Linguistics.

Bennett, W. L. and Pfetsch, B. (2018). Rethinking Political Communication in a Time of Disrupted Public Spheres. *J. of Communication*, 68(2):243–253.

Beretta, V., Harispe, S., Boland, K., Lo Seen, L., Todorov, K., and Tchechmedjiev, A. (2020). Can Knowledge Graph Embeddings Tell Us What Fact-checked Claims Are About? In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 71–75, Online. Association for Computational Linguistics.

Beretta, V., Harispe, S., Ranwez, S., and Mougenot, I. (2018). Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage. In Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E., editors, *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science, pages 652–668, Cham. Springer International Publishing. `https://www.doi.org/10.1007/978-3-030-00671-6_38`.

Besnard, P. and Hunter, A. (2008). *Elements of Argumentation*. The MIT Press. `https://doi.org/10.7551/mitpress/9780262026437.001.0001`.

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., and Mittal, A. (2018). Combining Neural, Statistical and External Features for Fake News Stance Identification. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1353–1357, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. `https://www.doi.org/10.1145/3184558.3191577`.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*.

O'Reilly Media Inc.

Biyani, P., Bhatia, S., Caragea, C., and Mitra, P. (2014). Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*, 69:170–178. `https://www.doi.org/10.1016/j.knosys.2014.04.048`.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 121–128, Red Hook, NY, USA. Curran Associates Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blustein, D. L. and Guarino, P. A. (2020). Work and unemployment in the time of covid-19: The existential experience of loss and fear. *Journal of Humanistic Psychology*, 60(5):702–709.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boland, K., Fafalios, P., Tchechmedjiev, A., Dietze, S., and Todorov, K. (2022a). Beyond facts. *Semantic Web*, 1(13):1–35.

Boland, K., Fafalios, P., Tchechmedjiev, A., Dietze, S., and Todorov, K. (2022b). Beyond facts - a survey and conceptualisation of claims in online discourse analysis. *Semantic Web*, 13(5):793–827.

Boland, K., Fafalios, P., Tchechmedjiev, A., Todorov, K., and Dietze, S. (2019a). Modeling and contextualizing claims. In *Proceedings of the Blockchain enabled Semantic Web Workshop (BlockSW) and Contextualized Knowledge Graphs (CKG) Workshop co-located with the 18th International Semantic Web Conference, BlockSW/CKG@ISWC 2019, Auckland, New Zealand, October 27, 2019*. CEUR-WS.org.

Boland, K., Fafalios, P., Tchechmedjiev, A., Todorov, K., and Dietze, S. (2019b). Modeling and Contextualizing Claims. In *Second International Workshop on Contextualized Knowledge Graphs (CKG2019) @ ISWC*.

Boland, K., Hövelmeyer, A., Fafalios, P., Todorov, K., Mazhar, U., and Dietze, S. (2023a). Robust and efficient claim retrieval for online fact-checking applications. *Information Retrieval Journal*. submitted (ID: 963b1532-a693-4b2d-8921-1728cc7af37e).

Boland, K. and Krüger, F. (2019). Distant supervision for silver label generation of software mentions in social scientific publications. In Chandrasekaran, M. K. and Mayr, P., editors, *BIRNDL 2019: Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*, pages 15–27. RWTH, Aachen.

Boland, K., Ritze, D., Eckert, K., and Mathiak, B. (2012). Identifying references to datasets in publications. In *International Conference on Theory and Practice of Digital Libraries*, pages 150–161. Springer.

Boland, K., Starke, C., Bensmann, F., Marcinkowski, F., and Dietze, S. (2023b). Public discourse about covid-19 vaccinations: A computational analysis of the relationship between public concerns and policies. *EPJ Data Science*. submitted (ID: EPDS-D-23-00098).

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Bonnevie, E., Gallegos-Jeffrey, A., Goldbarg, J., Byrd, B., and Smyser, J. (2021). Quantifying the rise of vaccine opposition on twitter during the covid-19 pandemic. *Journal of Communication in Healthcare*, 14(1):12–19.

Boon-Itt, S. and Skunkan, Y. (2020). Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill*, 6(4):e21978.

Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., and Sadeq, A. (2020). Team buster.ai at checkthat! 2020 insights and recommendations to improve fact-checking. In Cappellato, L., Eickhoff, C., Ferro, N., and Névéol, A., editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Breuer, J., Bensmann, F., Boland, K., Yu, R., and Dietze, S. (2022). All public opinions are not equal – developing and testing a method for assessing the relationship between survey data and twitter data as measures of public opinion.

Breuer, J., Bensmann, F., Dietze, S., Yu, R., and Boland, K. (2021). Assessing the relationship between survey data and twitter data as measures of public opinion - a methodological pilot study.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Buntain, C., McGrath, E., and Behlendorf, B. (2018). Sampling social media: Supporting information retrieval from microblog data resellers with text, network, and spatial analysis. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Buntain, C. L., McGrath, E., Golbeck, J., and LaFree, G. (2016). Comparing social media and traditional surveys around the boston marathon bombing. In *Microposts 2016 Workshop Proceedings, CEUR Vol-1691*.

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., and Sloan, L. (2015). Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108.

Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 5427–5433, Stockholm, Sweden. AAAI Press. `https://doi.org/10.24963/ijcai.2018/766`.

Caldarelli, G., De Nicola, R., Petrocchi, M., Pratelli, M., and Saracco, F. (2021). Flow of online misinformation during the peak of the covid-19 pandemic in italy. *EPJ data science*, 10(1):34.

Campos, D., Matos, S., and Oliveira, J. L. (2012). Biomedical named entity recognition: a survey of machine-learning tools. In *Theory and Applications for Advanced Text Mining*. IntechOpen.

Carstens, L. and Toni, F. (2015). Towards relation based Argumentation Mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/W15-0504`.

Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., and Tannier, X. (2018). A Content Management Perspective on Fact-Checking. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 565–574, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/S17-2001`.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018a). Universal sentence encoder. *CoRR*, abs/1803.11175.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al. (2018b). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ceron, A., Curini, L., Iacus, S. M., and Porro, G. (2014). Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New media & society*, 16(2):340–358.

Ceron, A. and Negri, F. (2015). Public policy and social media : how sentiment analysis can support policy-makers across the policy cycle. volume 10, page 309–338.

Ceron, A. and Negri, F. (2016). The "social side" of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, 8(2):131–147.

Cheema, G. S., Hakimov, S., and Ewerth, R. (2020). Check_square at checkthat! 2020 claim detection in social media via fusion of transformer and syntactic features. In Cappellato, L., Eickhoff, C., Ferro, N., and Névéol, A., editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chen, E., Lerman, K., and Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273.

Chen, M., Zhang, W., Zhang, W., Chen, Q., and Chen, H. (2019a). Meta Relational Learning for Few-Shot Link Prediction in Knowledge Graphs. 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 4216–4225, Hong Kong, China. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D19-1431`.

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., and Roth, D. (2019b). Seeing things from a different angle:discovering diverse perspectives about claims. 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 542–557. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/N19-1053`.

Chen, X., Wang, S., Tang, Y., and Hao, T. (2019c). A bibliometric analysis of event detection in social media. *Online Information Review*, 43(1):29–52. Publisher: Emerald Publishing Limited.

Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2021a). Aschern at clef checkthat! 2021: lambda-calculus of fact-checked claims. In *CLEF (Working Notes)*.

Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2021b). Transformers: "the end of history" for natural language processing? In Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., and Lozano, J. A., editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 677–693, Cham. Springer International Publishing.

Choi, E., Levy, O., Choi, Y., and Zettlemoyer, L. (2018). Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96. Association for Computational Linguistics.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PLOS ONE*, 10(6). `https://www.doi.org/10.1371/journal.pone.0128193`.

Clancy, R., Ilyas, I. F., and Lin, J. (2019). Scalable Knowledge Graph Construction from Text Collections. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 39–46, Hong Kong, China. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D19-6607`.

Cohen, S., Li, C., Yang, J., and Yu, C. (2011). Computational Journalism: A Call to Arms to Database Researchers. In *5th Biennial Conference on Innovative Data Systems Research (CIDR '11)*, pages 148–151, Asilomar, California, USA.

Conrad, F. G., Gagnon-Bartsch, J. A., Ferg, R. A., Schober, M. F., Pasek, J., and Hou, E. (2021). Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review*, 39(4):489–508.

Crane, M. (2018). Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Cui, H. and Kertész, J. (2021). Attention dynamics on the chinese social media sina weibo during the covid-19 pandemic. *EPJ data science*, 10(1):8.

Dacrema, M. F., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA. Association for Computing Machinery.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Lecture Notes in Computer Science, pages 177–190, Berlin, Heidelberg. Springer. https://www.doi.org/10.1007/11736790_9.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (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. https://www.doi.org/10.18653/v1/D17-1218.

Desson, Z., Lambertz, L., Peters, J. W., Falkenbach, M., and Kauer, L. (2020). Europe's covid-19 outliers: German, austrian and swiss policy responses during the early stages of the 2020 pandemic. *Health Policy and Technology*, 9(4):405–418. The COVID-19 pandemic: Global health policy and technology responses in the making.

Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., and Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLOS ONE*, 11(1):1–21.

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., and Dietze, S. (2020a). TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2991–2998, New York, NY, USA. Association for Computing Machinery.

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., and Dietze, S. (2020b). Tweetscov19 - a knowledge base of semantically annotated tweets about the covid-19 pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2991–2998, New York, NY, USA. Association for Computing Machinery.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 601–610, New York, New York, USA. ACM Press. https://www.doi.org/10.1145/2623330.2623623.

Doogan, C., Buntine, W., Linger, H., and Brunt, S. (2020). Public perceptions and attitudes toward covid-19 nonpharmaceutical interventions across six countries: A topic modeling analysis of twitter data. *J Med Internet Res*, 22(9):e21419.

Du, J., Xu, R., He, Y., and Gui, L. (2017). Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 3988–3994, Melbourne, Australia. AAAI Press. https://doi.org/10.24963/ijcai.2017/557.

Duck, G., Kovacevic, A., Robertson, D. L., Stevens, R., and Nenadic, G. (2015). Ambiguity and variability of database and software names in bioinformatics. *Journal of biomedical semantics*, 6(1):29.

Dumani, L. and Schenkel, R. (2019). A Systematic Comparison of Methods for Finding Good Premises for Claims. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 957–960, New York, NY, USA. Association for Computing Machinery.

Durmus, E., Ladhak, F., and Cardie, C. (2019). Determining Relative Argument Specificity and Stance for Complex Argumentative Structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/P19-1456.

Ebrahimi, J., Dou, D., and Lowd, D. (2016). Weakly Supervised Tweet Stance Classification by Relational Bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1017, Austin, Texas. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/D16-1105.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, page 201602413.

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Explosion (2022). German - spacy models documentation. `https://spacy.io/models/de`. Version 2.3.2, Accessed: 2022-07-08.

Fafalios, P., Iosifidis, V., Ntoutsi, E., and Dietze, S. (2018a). TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets. In *European Semantic Web Conference*, pages 177–190. Springer. `https://doi.org/10.1007/978-3-319-93417-4_12`.

Fafalios, P., Iosifidis, V., Ntoutsi, E., and Dietze, S. (2018b). Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*, pages 177–190. Springer.

fasttext.cc (2023). Word vectors for 157 languages. `https://fasttext.cc/docs/en/crawl-vectors.html`.

Fazel, S., Zhang, L., Javid, B., Brikell, I., and Chang, Z. (2021). Harnessing twitter data to survey public attention and attitudes towards covid-19 vaccines in the uk. *Scientific Reports*, 11(23402).

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/N16-1138`.

Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Fierro, C., Fuentes, C., Pérez, J., and Quezada, M. (2017). 200K+ Crowdsourced Political Arguments for a New Chilean Constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/W17-5101.

Fionda, V. and Pirrò, G. (2018). Fact Checking via Evidence Patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3755–3761, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization. https://www.doi.org/10.24963/ijcai.2018/522.

Foad, C. M., Whitmarsh, L., Hanel, P. H., and Haddock, G. (2021). The limitations of polling data in understanding public support for covid-19 lockdown policies. *Royal Society Open Science*, 8(7):210678.

Frick, R. A. and Vogel, I. (2022). Fraunhofer sit at checkthat! 2022: Ensemble similarity estimation for finding previously fact-checked claims. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 05-08, 2022*, CEUR Workshop Proceedings.

Gad-Elrab, M. H., Stepanova, D., Urbani, J., and Weikum, G. (2019). ExFaKT: A Framework for Explaining Facts over Knowledge Graphs and Text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*, pages 87–95, Melbourne VIC, Australia. ACM Press. https://www.doi.org/10.1145/3289600.3290996.

Galárraga, L. A., Teflioudi, C., Hose, K., and Suchanek, F. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 413–422, New York, NY, USA. Association for Computing Machinery.

Gangemi, A. (2013). A Comparison of Knowledge Extraction Tools for the Semantic Web. In Cimiano, P., Corcho, O., Presutti, V., Hollink, L., and Rudolph, S., editors, *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, pages 351–366, Berlin, Heidelberg. Springer. https://www.doi.org/10.1007/978-3-642-38288-8_24.

Gangopadhyay, S., Boland, K., Dessi, D., Dietze, S., Fafalios, P., Tchechmedjiev, A., Todorov, K., and Jabeen, H. (2023). Truth or dare: Investigating claims truthfulness with claimskg. In *Proceedings of the Second International Workshop on Linked Data-*

driven Resilience Research 2023 co-located with Extended Semantic Web Conference 2023 (ESWC 2023).

Gangopadhyay, S., Boland, K., Schüller, S., Konstantin, T., Tchechmedjiev, A., Zapilko, B., Fafalios, P., Jabeen, H., and Dietze, S. (2022). Claimskg - a knowledge graph of fact-checked claims (august, 2022). gesis - leibniz-institute for the social sciences. data file version 1.0.0.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Gasquet, M., Brechtel, D., Zloch, M., Tchechmedjiev, A., Boland, K., Fafalios, P., Dietze, S., and Todorov, K. (2019). Exploring fact-checked claims and their descriptive statistics. In *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26-30, 2019*. CEUR-WS.org.

Georgia M. Green (1996). *Pragmatics and Natural Language Understanding*. Lawrence Erlbaum Associates, Hilldale, NJ, 2 edition.

Gerber, D., Esteves, D., Lehmann, J., Bühmann, L., Usbeck, R., Ngomo, A.-C. N., and Speck, R. (2015). DeFacto - Temporal and Multilingual Deep Fact Validation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35(P2). `https://www.doi.org/10.1016/j.websem.2015.08.001`.

Ghanem, B., Glavas, G., Giachanou, A., Paolo, S., Rosso, P., and Rangel, F. (2019). UPV-UMA at CheckThat! Lab: Verifying Arabic Claims using a Cross Lingual Approach. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, CLEF 2019 Working Notes, page 10, Lugano, Switzerland.

Ghanem, B., Rosso, P., and Rangel, F. (2018). Stance Detection in Fake News A Combined Feature Representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5510`.

Ghavimi, B., Mayr, P., Lange, C., Vahdati, S., and Auer, S. (2016). A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use*, 36(3-4):171–187.

Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing Argumentative Discourse Units in Online Interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/W14-2106`.

Ghosh, S., Singhania, P., Singh, S., Rudra, K., and Ghosh, S. (2019). Stance Detection in Web and Social Media: A Comparative Study. *arXiv:2007.05976 [cs]*, 11696:75–87. `https://www.doi.org/10.1007/978-3-030-28577-7_4`.

Giasemidis, G., Kaplis, N., Agrafiotis, I., and Nurse, J. R. C. (2020). A semi-supervised approach to message stance classification. *IEEE Transactions on Knowledge and Data Engineering*, 32(1):1–11. `https://www.doi.org/10.1109/TKDE.2018.2880192`.

Giorgi, J. M. and Bader, G. (2018). Transfer learning for biomedical named entity recognition with neural networks.

Gonzalez Pinto, J. M., Wawrzinek, J., and Balke, W.-T. (2019). What Drives Research Efforts? Find Scientific Claims that Count! In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 217–226, Champaign, IL, USA. IEEE. `https://www.doi.org/10.1109/JCDL.2019.00038`.

González Pinto, J. M. and Balke, W.-T. (2017). Offering Answers for Claim-Based Queries: A New Challenge for Digital Libraries. In Choemprayong, S., Crestani, F., and Cunningham, S. J., editors, *Digital Libraries: Data, Information, and Knowledge for Digital Lives*, Lecture Notes in Computer Science, pages 3–13, Cham. Springer International Publishing. `https://www.doi.org/10.1007/978-3-319-70232-2_1`.

González Pinto, J. M. and Balke, W.-T. (2018). Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries. In Méndez, E., Crestani, F., Ribeiro, C., David, G., and Lopes, J. C., editors, *Digital Libraries for Open Knowledge*, volume 11057 of *Lecture Notes in Computer Science*, pages 257–269. Springer International Publishing, Cham. https://www.doi.org/10.1007/978-3-030-00066-0_22.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., and Derczynski, L. (2019). SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 845–854, Minneapolis, Minnesota, USA,. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/S19-2147.

Gracia, J. and Mena, E. (2008). Web-Based Measure of Semantic Relatedness. In *Proceedings of the 9th international conference on Web Information Systems Engineering*, WISE '08, pages 136–150, Berlin, Heidelberg. Springer-Verlag.

Graves, H., Graves, R., Mercer, R., and Akter, M. (2014). Titles That Announce Argumentative Claims in Biomedical Research Articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99, Baltimore, Maryland. Association for Computational Linguistics. https://www.doi.org/10.3115/v1/W14-2113.

Graves, L. (2018). Understanding the Promise and Limits of Automated Fact-Checking. Factsheet, Reuters Institute, University of Oxford.

Green, N. (2014). Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland. Association for Computational Linguistics. https://www.doi.org/10.3115/v1/W14-2102.

Green, N. (2015). Identifying Argumentation Schemes in Genetics Research Articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, Denver, CO. Association for Computational Linguistics. https://www.doi.org/10.3115/v1/W15-0502.

Green, N. (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. https://www.doi.org/10.18653/v1/W18-5213.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guggilla, C., Miller, T., and Gurevych, I. (2016). CNN- and LSTM-based Claim Classification in Online User Comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2740–2751, Osaka, Japan. The COLING 2016 Organizing Committee.

Gupta, S. and Manning, C. D. (2014). Improved pattern learning for bootstrapped entity extraction. In *Computational Natural Language Learning (CoNLL)*.

Habernal, I. and Gurevych, I. (2017). Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179. `https://www.doi.org/10.1162/COLI_a_00276`.

Hadj Taieb, M. A., Zesch, T., and Ben Aouicha, M. (2020). A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, 53(6):4407–4448.

Hanselowski, A., Pvs, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., and Gurevych, I. (2018a). A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 16, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., and Gurevych, I. (2018b). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., and Gurevych, I. (2018c). UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5516`.

Hansen, C., Hansen, C., Alstrup, S., Grue Simonsen, J., and Lioma, C. (2019). Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 994–1000, New York, NY, USA. Association for Computing Machinery.

Hardalov, M., Chernyavskiy, A., Koychev, I., Ilvovsky, D., and Nakov, P. (2022). CrowdChecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, AACL-IJCNLP '22, pages 266–285, Online only. Association for Computational Linguistics.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38(1):10–24.

Hasan, M., Orgun, M. A., and Schwitter, R. (2018). A survey on real-time event detection from the Twitter data stream. *Journal of Information Science*, 44(4):443–463.

Hasanain, M., Suwaileh, R., Elsayed, T., Barron-Cedeno, A., and Nakov, P. (2019). Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, CLEF 2019 Working Notes, page 15.

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., and Yu, C. (2015). The Quest to Automate Fact-Checking. *Proceedings of the 2015 Computation+ Journalism Symposium*.

Hassan, N., Arslan, F., Li, C., and Tremayne, M. (2017a). Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1803–1812, New York, NY, USA. Association for Computing Machinery.

Hassan, N., Yousuf, M., Mahfuzul Haque, M., A. Suarez Rivas, J., and Khadimul Islam, M. (2019). Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-checking. In *Companion Proceedings of The 2019 World*

Wide Web Conference, WWW ’19, pages 1001–1006, San Francisco, USA. Association for Computing Machinery.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., et al. (2017b). Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., and Tremayne, M. (2017c). ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Hernández A., M. and Gómez, J. M. (2014). Survey in sentiment, polarity and function analysis of citation. In *Proceedings of the First Workshop on Argumentation Mining*, pages 102–103, Baltimore, Maryland. Association for Computational Linguistics. https://www.doi.org/10.3115/v1/W14-2115.

Herrera-Peco, I., Jiménez-Gómez, B., Romero Magdalena, C. S., Deudero, J. J., García-Puente, M., Benítez De Gracia, E., and Ruiz Núñez, C. (2021). Antivaccine movement and covid-19 negationism: A content analysis of spanish-written messages on twitter. *Vaccines*, 9(6).

Hidey, C., Musi, E., Hwang, A., Muresan, S., and McKeown, K. (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. https://www.doi.org/10.18653/v1/W17-5102.

Hienert, D., Kern, D., Boland, K., Zapilko, B., and Mutschke, P. (2019). A digital library for research data and related information in the social sciences. In *Proceedings of 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 148–157. IEEE, Piscataway, NJ.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/P18-1031.

Howison, J. and Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9):2137–2155.

Hu, T., Wang, S., Luo, W., Zhang, M., Huang, X., Yan, Y., Liu, R., Ly, K., Kacker, V., She, B., and Li, Z. (2021). Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective. *J Med Internet Res*, 23(9):e30854.

Hussain, M. and Mahmud, I. (2019). pymannkendall: a python package for non parametric mann kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556.

Huynh, V.-P. and Papotti, P. (2018). Towards a Benchmark for Fact Checking with Knowledge Bases. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 1595–1598, Lyon, France. ACM Press. https://www.doi.org/10.1145/3184558.3191616.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. John Benjamins Publishing. https://doi.org/10.1075/pbns.54.

Hövelmeyer, A., Boland, K., and Dietze, S. (2022). Simba at checkthat! 2022: Lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 05-08, 2022*, CEUR Workshop Proceedings, pages 511–531.

Imran, M., Qazi, U., and Ofli, F. (2022). Tbcov: Two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *Data*, 7(1).

James Thorne and Andreas Vlachos (2019). Adversarial attacks against Fact Extraction and VERification. *CoRR*, abs/1903.05543.

Jentzsch, A., Isele, R., and Bizer, C. (2010). Silk – Generating RDF Links while publishing or consuming Linked Data. In *9th International Semantic Web Conference (ISWC'10)*.

Jiang, S., Baumgartner, S., Ittycheriah, A., and Yu, C. (2020). Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. In *Proceedings of The Web Conference 2020*, pages 1592–1603, Taipei Taiwan. ACM. `https://www.doi.org/10.1145/3366423.3380231`.

Jiang, X., Su, M.-H., Hwang, J., Lian, R., Brauer, M., Kim, S., and Shah, D. (2021). Polarization over vaccination: Ideological differences in twitter expression about covid-19 vaccine favorability and specific hesitancy concerns. *Social Media+ Society*, 7(3):20563051211048413.

Jing, E. and Ahn, Y.-Y. (2021). Characterizing partisan political narrative frameworks about covid-19 on twitter. *EPJ data science*, 10(1):53.

Joseph, K., Friedland, L., Hobbs, W., Lazer, D., and Tsur, O. (2017). ConStance: Modeling Annotation Contexts to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124, Copenhagen, Denmark. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D17-1116`.

Kendall, M. (1975). *Rank Correlation Methods*. Charles Griffin, London, 4 edition.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Killick, R., Fearnhead, P., and Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. In *Proceedings of the*

2nd Workshop on Argumentation Mining, pages 1–11, Denver, CO. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/W15-0501`.

Klein, A. Z., Magge, A., O'Connor, K., Flores Amaro, J. I., Weissenbacher, D., and Gonzalez Hernandez, G. (2021). Toward using twitter for tracking covid-19: A natural language processing pipeline and exploratory data set. *J Med Internet Res*, 23(1):e25314.

Konstantinovskiy, L., Price, O., Babakar, M., and Zubiaga, A. (2021). Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice*, 2(2). `https://doi.org/10.1145/3412869`.

Kotonya, N. and Toni, F. (2019). Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W19-4518`.

Krestel, R., Bergler, S., and Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Kumar, S. (2017). A Survey of Deep Learning Methods for Relation Extraction. *arXiv:1705.03645 [cs]*. arXiv: 1705.03645.

Küçük, D. and Can, F. (2020). Stance Detection: A Survey. *ACM Computing Surveys*, 53(1):12:1–12:37.

Lai, M., Hernández Farías, D. I., Patti, V., and Rosso, P. (2017). Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. In Sidorov, G. and Herrera-Alcántara, O., editors, *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 155–168, Cham. Springer International Publishing.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Lauscher, A., Glavaš, G., and Eckert, K. (2018a). ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5203`.

Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018b). 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. `https://www.doi.org/10.18653/v1/W18-5206`.

Lawrence, J. and Reed, C. (2019). Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818. `https://www.doi.org/10.1162/coli_a_00364`.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Levy, R., Bogin, B., Gretz, S., Aharonov, R., and Slonim, N. (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.

Li, K., Lin, X., and Greenberg, J. (2016). Software citation, reuse and metadata considerations: An exploratory study examining lammps. *Proceedings of the Association for Information Science and Technology*, 53(1):1–10.

Li, K. and Yan, E. (2018). Co-mention network of r packages: Scientific impact and clustering structure. *Journal of Informetrics*, 12(1):87–100.

Li, K., Yan, E., and Feng, Y. (2017). How is r cited in research outputs? structure, impacts, and citation standard. *Journal of Informetrics*, 11(4):989–1002.

Liebeck, M., Esau, K., and Conrad, S. (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)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W16-2817`.

Lin, H. and Ng, V. (2019). Abstractive Summarization: A Survey of the State of the Art. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9815–9822. `https://www.doi.org/10.1609/aaai.v33i01.3301981`.

Linzbach, S., Tressel, T., Kallmeyer, L., Dietze, S., and Jabeen, H. (2023). Decoding prompt syntax: Analysing its impact on knowledge retrieval in large language models. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1145–1149, New York, NY, USA. Association for Computing Machinery.

Lippi, M., Mamei, M., Mariani, S., and Zambonelli, F. (2018). An Argumentation-Based Perspective Over the Social IoT. *IEEE Internet of Things Journal*, 5(4):2537–2547. Conference Name: IEEE Internet of Things Journal.

Lippi, M. and Torroni, P. (2015). Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 185–191, Buenos Aires, Argentina. AAAI Press.

Lippi, M. and Torroni, P. (2016a). Argument mining from speech: detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2979–2985, Phoenix, Arizona. AAAI Press.

Lippi, M. and Torroni, P. (2016b). Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Lippi, M. and Torroni, P. (2016c). MARGOT: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303. `https://www.doi.org/10.1016/j.eswa.2016.08.050`.

Liu, L., Ren, X., Shang, J., Gu, X., Peng, J., and Han, J. (2018). Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

1215–1225, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D18-1153`.

Liu, X., He, P., Chen, W., and Gao, J. (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. `https://www.doi.org/10.18653/v1/P19-1441`.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M., and Srivastava, D. (2012). A dataset search engine for the research document corpus. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1237–1240. IEEE.

Lucic, A., Bleeker, M., Bhargav, S., Forde, J., Sinha, K., Dodge, J., Luccioni, S., and Stojnic, R. (2022). Towards reproducible machine learning research in natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–11, Dublin, Ireland. Association for Computational Linguistics.

Lugini, L. and Litman, D. (2018). Argument Component Classification for Classroom Discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5208`.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., and Wang, J. (2017). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Ma, J., Gao, W., and Wong, K.-F. (2018). Detect Rumor and Stance Jointly by Neural Multi-task Learning. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 585–593, Lyon, France. ACM Press. `https://www.doi.org/10.1145/3184558.3188729`.

Maliaroudakis, E., Boland, K., Dietze, S., Todorov, K., Tzitzikas, Y., and Fafalios, P. (2021a). Claimlinker: Linking text to a knowledge graph of fact-checked claims. In *Companion Proceedings of the Web Conference 2021*, pages 669–672.

Maliaroudakis, E., Boland, K., Dietze, S., Todorov, K., Tzitzikas, Y., and Fafalios, P. (2021b). ClaimLinker: Linking Text to a Knowledge Graph of Fact-checked Claims. In *Companion Proceedings of the Web Conference 2021 (WWW'21 Companion)*. ACM. `https://doi.org/10.1145/3442442.3458601`.

Mann, H. (1945). Non-parametric tests against trend. *Econometrica*, 13:163–171.

Mansour, W., Elsayed, T., and Al-Ali, A. (2022). Did i see it before? detecting previously-checked claims over twitter. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 367–381, Berlin, Heidelberg. Springer-Verlag.

Martinez-Rico, J. R., Araujo, L., and Martinez-Romo, J. (2020). Nlp&ir@ uned at checkthat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs. In *CLEF (Working Notes)*.

Martinez-Rodriguez, J. L., Hogan, A., and Lopez-Arevalo, I. (2020). Information extraction meets the Semantic Web: A survey. *Semantic Web*, 11(2):255–335. `https://www.doi.org/10.3233/SW-180333`.

Mathiak, B. and Boland, K. (2015). Challenges in matching dataset citation strings to datasets in social science. *D-Lib Magazine*, 1/2(21).

Mathur, A., Kubde, P., and Vaidya, S. (2020). Emotional analysis using twitter data during pandemic situation: Covid-19. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 845–848.

Mayer, T., Cabrio, E., Lippi, M., Torroni, P., and Villata, S. (2018a). Argument Mining on Clinical Trials. In *COMMA 2018 - 7th International Conference on Computational Models of Argument Proceedings*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, pages 137 – 148, Warsaw, Poland.

Mayer, T., Cabrio, E., and Villata, S. (2018b). Evidence Type Classification in Randomized Controlled Trials. In *Proceedings of the 5th Workshop on Argument Mining*, pages 29–34, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5204`.

Maynard, D., Bontcheva, K., and Augenstein, I. (2016). *Natural Language Processing for the Semantic Web*. Morgan & Claypool Publishers. `https://doi.org/10.2200/S00741ED1V01Y201611WBE015`.

McDonald, T., Dong, Z., Zhang, Y., Hampson, R., Young, J., Cao, Q., Leidner, J. L., and Stevenson, M. (2020). The university of sheffield at checkthat! 2020: Claim identification and verification on twitter. In *CLEF (Working Notes)*.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861.

Mevan Babakar and Will Moy (2016). The State of Automated Factchecking - How to make factchecking dramatically more effective with technology we have now. Technical report, Full Fact.

Mihaylova, S., Borisova, I., Chemishanov, D., Hadzhitsanev, P., Hardalov, M., and Nakov, P. (2021a). DIPS at checkthat! 2021: Verified claim retrieval. In Faggioli, G., Ferro, N., Joly, A., Maistro, M., and Piroi, F., editors, *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 558–571. CEUR-WS.org.

Mihaylova, S., Borisova, I., Chemishanov, D., Hadzhitsanev, P., Hardalov, M., and Nakov, P. (2021b). Dips at checkthat! 2021: verified claim retrieval. In *CLEF 2021 Working Notes*, pages 558–571.

Mihaylova, T., Nakov, P., Marquez, L., Barron-Cedeno, A., Mohtarami, M., Karadzhov, G., and Glass, J. (2018). Fact Checking in Community Forums. In *arXiv:1803.03178 [cs]*, AAAI 2018, pages 5309–5316. arXiv: 1803.03178.

Mohamed Ridhwan, K. and Hargreaves, C. A. (2021). Leveraging twitter data to understand public sentiment for the covid-19 outbreak in singapore. *International Journal of Information Management Data Insights*, 1(2):100021.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/S16-1003`.

Mohtaj, S., Himmelsbach, T., Woloszyn, V., and Moller, S. (2019). Using External Knowledge Bases and Coreference Resolution for Detecting Check-Worthy Statements. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, CLEF 2019 Working Notes, page 8, Lugano, Switzerland.

Muric, G., Wu, Y., and Ferrara, E. (2021). Covid-19 vaccine hesitancy on social media: Building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR Public Health Surveill*, 7(11):e30642.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Nadeem, M., Fang, W., Xu, B., Mohtarami, M., and Glass, J. (2019). FAKTA: An Automatic End-to-End Fact Checking System. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83, Minneapolis, Minnesota. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N19-4014`.

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J. M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G. K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y. S., and Beltrán, J. (2022). The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 416–428, Berlin, Heidelberg. Springer-Verlag.

Nangia, U. and Katz, D. S. (2017). Understanding software in research: Initial results from examining nature and a call for collaboration. In *2017 IEEE 13th International Conference on e-Science (e-Science)*, pages 486–487.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., and Kim, J. (2021). Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015.

Nentwig, M., Hartung, M., Ngonga Ngomo, A.-C., and Rahm, E. (2016). A survey of current Link Discovery frameworks. *Semantic Web*, 8(3):419–436. `https://www.doi.org/10.3233/SW-150210`.

Neves, M., Butzke, D., and Grune, B. (2019). Evaluation of Scientific Elements for Text Similarity in Biomedical Publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135, Florence, Italy. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W19-4515`.

Newell, C., Cowlishaw, T., and Man, D. (2018a). Quote Extraction and Analysis for News. In *Data Science, Journalism & Media @KDD 2018 International Conference on Knowledge Discover and Data Mining*.

Newell, E., Margolin, D., and Ruths, D. (2018b). An Attribution Relations Corpus for Political News. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 8, Miyazaki, Japan. European Language Resources Association (ELRA).

Ngomo, A.-C. N. and Auer, S. (2011). LIMES: A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2312–2317, Barcelona, Catalonia, Spain. AAAI Press.

Nguyen, H. and Litman, D. (2016). Context-aware Argumentative Relation Mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/P16-1107`.

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., and Yang, Y. (2021).

Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *CoRR*, abs/2108.08877.

Niculae, V., Park, J., and Cardie, C. (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. `https://www.doi.org/10.18653/v1/P17-1091`.

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 798–808, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nie, Y., Chen, H., and Bansal, M. (2019). Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866. `https://www.doi.org/10.1609/aaai.v33i01.33016859`.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, Jeju Island, Korea. Association for Computational Linguistics.

Opitz, J. and Frank, A. (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. `https://www.doi.org/10.18653/v1/W19-4503`.

Ortiz-Ospina, E. (2019). The rise of social media. *Our World in Data*.

https://ourworldindata.org/rise-of-social-media.

Otto, W., Zielinski, A., Ghavimi, B., Dimitrov, D., Tavakolpoursaleh, N., Abdulahhad, K., Boland, K., and Dietze, S. (2020). Knowledge extraction from scholarly publications. In Lane, J., Mulvany, I., and Nathan, P., editors, *Rich search and discovery for research datasets*, pages 107–127. Sage, Los Angeles u.a.

Padia, A., Ferraro, F., and Finin, T. (2018). Team UMBC-FEVER : Claim verification using Semantic Lexical Resources. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 161–165, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5527`.

Padia, A., Kalpakis, K., Ferraro, F., and Finin, T. (2019). Knowledge Graph Fact Prediction via Knowledge-Enriched Tensor Factorization. *Journal of Web Semantics*. `https://www.doi.org/10.2139/ssrn.3331039`.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/N18-1049`.

Palau, R. M. and Moens, M.-F. (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, pages 98–107, New York, NY, USA. Association for Computing Machinery.

Pan, X., Yan, E., Wang, Q., and Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, 9(4):860–871.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pareti, S. (2015). *Attribution: A Computational Approach*. PhD Thesis, Edinburgh.

Park, J. and Cardie, C. (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. `https://www.doi.org/10.3115/v1/W14-2105`.

Pasek, J., McClain, C. A., Newport, F., and Marken, S. (2020). Who's tweeting about the president? what big survey data can tell us about digital traces? *Social Science Computer Review*, 38(5):633–650.

Pasquale Minervini, Volker Tresp, Claudia d'Amato, and Nicola Fanizzi (2017). Adaptive Knowledge Propagation in Web Ontologies. *2*, 12(1):28. `https://www.doi.org/10.1145/3105961`.

Passaro, L. C., Bondielli, A., Lenci, A., and Marcelloni, F. (2020). Unipi-nle at check-that! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers. In *CLEF (Working Notes)*.

Paul, M. and Dredze, M. (2021). You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):265–272.

Paul, M. J., Zhai, C., and Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76, USA. Association for Computational Linguistics.

Pavllo, D., Piccardi, T., and West, R. (2018). Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, ICWSM 2018, pages 231–240. arXiv: 1804.02525.

Peldszus, A. and Stede, M. (2013). From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Pereira Nunes, B., Mera, A., Casanova, M. A., Fetahu, B., P. Paes Leme, L. A., and Dietze, S. (2013). Complex Matching of RDF Datatype Properties. In Decker, H., Lhotská, L., Link, S., Basl, J., and Tjoa, A. M., editors, *Database and Expert Systems Applications*, Lecture Notes in Computer Science, pages 195–208, Berlin, Heidelberg. Springer. `https://www.doi.org/10.1007/978-3-642-40285-2_18`.

Pfeiffer, T. and Hoffmann, R. (2009). Large-scale assessment of the effect of popularity on the reliability of research. *PLOS ONE*, 4(6):1–4.

Piquero, A. R., Jennings, W. G., Jemison, E., Kaukinen, C., and Knaul, F. M. (2021). Domestic violence during the covid-19 pandemic-evidence from a systematic review and meta-analysis. *Journal of criminal justice*, 74:101806.

Poblete, B., Guzmán, J., Maldonado, J., and Tobar, F. (2018). Robust detection of extreme events using twitter: Worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, 20(10):2551–2561.

Pomerleau, D. and Rao, D. (2017a). The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news.

Pomerleau, D. and Rao, D. (2017b). Fake News Challenge Stage 1 (FNC-I): Stance Detection.

Popat, K. (2019). *Credibility Analysis of Textual Claims with Explainable Evidence*. Doctoral Thesis, Saarbrücken. `https://dx.doi.org/10.22028/D291-30005`.

Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. (2017). Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 1003–1012, Perth, Australia. ACM Press. `https://www.doi.org/10.1145/3041021.3055133`.

Potash, P., Ferguson, A., and Hazen, T. J. (2019). Ranking Passages for Argument Convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W19-4517`.

Pritzkau, A. (2021). Nlytics at checkthat! 2021: Detecting previously fact-checked claims by measuring semantic similarity. In *CLEF 2021 Working Notes*, pages 582–591.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver?

Quraishi, M., Fafalios, P., and Herder, E. (2018). Viewpoint Discovery and Understanding in Social Networks. In *Proceedings of the 10th ACM Conference on Web Science*, pages 47–56. `https://www.doi.org/10.1145/3201064.3201076`.

Rajendran, P., Bollegala, D., and Parsons, S. (2016). Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39, Berlin, Germany. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W16-2804`.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D17-1317`.

Rebholz-Schuhmann, D., Yepes, A. J. J., Mulligen, E. M. V., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., and Hahn, U. (2010). CALBC Silver Standard Corpus. *Journal of Bioinformatics and Computational Biology*, 08(01):163–179.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *CoRR*, abs/1908.10084.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (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. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/P19-1054`.

Reiter-Haas, M., Klösch, B., Hadler, M., and Lex, E. (2022). Polarization of opinions on covid-19 measures: Integrating twitter and survey data. *Social Science Computer Review*, 0(0):08944393221087662.

Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science*, 44(4):638–654. https://www.doi.org/10.1177/0306312714535679.

Rinott, R., Dankin, L., Alzate Perez, C., Khapra, M. M., Aharoni, E., and Slonim, N. (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. https://www.doi.org/10.18653/v1/D15-1050.

Ristoski, P. and Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36:1–22. https://www.doi.org/10.1016/j.websem.2016.01.001.

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., and Roser, M. (2020). Coronavirus pandemic (covid-19). *Our World in Data*. https://ourworldindata.org/coronavirus.

Rogers, A. and Augenstein, I. (2020). What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rosenthal, S. and McKeown, K. (2012). Detecting Opinionated Claims in Online Discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, Palermo, Italy. IEEE. https://www.doi.org/10.1109/ICSC.2012.59.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building Event-Centric Knowledge Graphs from News. *Journal of Web Semantics*, 37-38:132–151. https://www.doi.org/10.1016/j.websem.2015.12.004.

Rothmayr, C. and Hardmeier, S. (2002). Government and polling: Use and impact of polls in the policy-making process in switzerland. *International Journal of Public Opinion Research*, 14(2):123–140.

Roy, A., Fafalios, P., Ekbal, A., Zhu, X., and Dietze, S. (2021). Exploiting stance hierarchies for cost-sensitive stance detection of Web documents. *Journal of Intelligent Information Systems*. `https://doi.org/10.1007/s10844-021-00642-z`.

Rubinstein, M., Meyer, E., Schroeder, R., Poel, M., Treperman, J., van Barneveld, J., Biesma-Pickles, A., Mahieu, B., Potau, X., and Svetachova, M. (2016). Ten use cases of innovative data-driven approaches for policymaking at eu level. *Report prepared in the context of: Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking*.

Russell, P. H., Johnson, R. L., Ananthan, S., Harnke, B., and Carlson, N. E. (2018). A large-scale analysis of bioinformatics code on GitHub. *PLOS ONE*, 13(10):e0205898.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 851–860, New York, NY, USA. Association for Computing Machinery.

Salway, A., Meurer, P., Hofland, K., and Reigem, Ø. (2017). Quote Extraction and Attribution from Norwegian Newspapers. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 293–297, Gothenburg, Sweden. Association for Computational Linguistics.

Sattar, N. S. and Arifuzzaman, S. (2021). Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa. *Applied Sciences*, 11(13).

Scheible, C., Klinger, R., and Padó, S. (2016). Model Architectures for Quotation Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/P16-1164`.

Schiappa, E. and Nordin, J. P. (2013). *Argumentation: Keeping Faith with Reason*. Pearson Education. Google-Books-ID: ZOn8nQEACAAJ.

Schiller, B., Daxenberger, J., and Gurevych, I. (2021). Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*.

Schulze, H. and Boland, K. (2019). Building a categorization schema for automated source typing.

Searle, J. R. (1969). *Speech acts. An essay in the philosophy of language*. Cambridge University Press, Cambridge.

Sen, I., Flöck, F., and Wagner, C. (2020). On the Reliability and Validity of Detecting Approval of Political Actors in Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1413–1426, Online. Association for Computational Linguistics.

Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2020). Neural Entity Linking: A Survey of Models based on Deep Learning. *arXiv:2006.00575 [cs]*. arXiv: 2006.00575.

Shaar, S., Alam, F., Da San Martino, G., and Nakov, P. (2022). The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020a). That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/2020.acl-main.332`.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020b). That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaar, S., Haouari, F., Mansour, W., Hasanain, M., Babulkov, N., Alam, F., Martino, G. D. S., Elsayed, T., and Nakov, P. (2021). Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In *CLEF 2021 Working Notes*.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):21:1–21:42.

Shen, W., Wang, J., and Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Sheng, Q., Cao, J., Zhang, X., Li, X., and Zhong, L. (2021). Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. 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 5468–5481, Online. Association for Computational Linguistics.

Shi, B. and Weninger, T. (2016). Discriminative Predicate Path Mining for Fact Checking in Knowledge Graphs. *arXiv:1510.05911 [cs]*, 104:123–133. `https://www.doi.org/10.1016/j.knosys.2016.04.015`.

Shliselberg, M. and Dori-Hacohen, S. (2022). Riet lab at checkthat! 2022: Improving decoder based re-ranking for claim matching. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 05-08, 2022*, CEUR Workshop Proceedings.

Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., Aharonov, R., and Slonim, N. (2018). Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/P18-2095`.

Shofiya, C. and Abidi, S. (2021). Sentiment analysis on covid-19-related social distancing in canada using twitter data. *International Journal of Environmental Research and Public Health*, 18(11).

Shukla, U. and Sharma, A. (2020). Tiet at clef checkthat! 2020: Verified claim retrieval. In *CLEF (Working Notes)*.

Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLOS ONE*, 6(5):1–10.

Skuczyńska, B., Shaar, S., Spenader, J., and Nakov, P. (2021). Beasku at checkthat! 2021: fine-tuning sentence bert with triplet loss and limited data. In *CLEF 2021 Working Notes*, pages 639–647.

Smith, A. M., Katz, D. S., and and, K. E. N. (2016). Software citation principles. *PeerJ Computer Science*, 2:e86.

Solís Arce, J., Warren, S., Meriggi, N., Scacco, A., McMurry, N., Voors, M., Syunyaev, G., Malik, A., Aboutajdine, S., Adeojo, O., Anigo, D., Armand, A., Asad, S., Atyera, M., Augsburg, B., Awasthi, M., Ayesiga, G., Bancalari, A., Björkman Nyqvist, M., Borisova, E., Bosancianu, C., Cabra García, M., Cheema, A., Collins, E., Cuccaro, F., Farooqi, A., Fatima, T., Fracchia, M., Galindo Soria, M., Guariso, A., Hasanain, A., Jaramillo, S., Kallon, S., Kamwesigye, A., Kharel, A., Kreps, S., Levine, M., Littman, R., Malik, M., Manirabaruta, G., Mfura, J., Momoh, F., Mucauque, A., Mussa, I., Nsabimana, J., Obara, I., Otálora, M., Ouédraogo, B., Pare, T., Platas, M., Polanco, L., Qureshi, J., Raheem, M., Ramakrishna, V., Rendrá, I., Shah, T., Shaked, S., Shapiro, J., Svensson, J., Tariq, A., Tchibozo, A., Tiwana, H., Trivedi, B., Vernot, C., Vicente, P., Weissinger, L., Zafar, B., Zhang, B., Karlan, D., Callen, M., Teachout, M., Humphreys, M., Mobarak, A., and Omer, S. (2021). Covid-19 vaccine acceptance and hesitancy in low- and middle-income countries. *Nat Med.*, 27(8):1385–1394.

Sridhar, D., Foulds, J., Huang, B., Getoor, L., and Walker, M. (2015). Joint Models of Disagreement and Stance in Online Debate. 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 116–125, Beijing, China. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/P15-1012`.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842, New York, NY, USA. Association for Computing Machinery.

Stab, C. and Gurevych, I. (2014). Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/D14-1006`.

Stab, C. and Gurevych, I. (2017). Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659. `https://www.doi.org/10.1162/COLI_a_00295`.

Stahlhut, C. (2019). Interactive Evidence Detection: train state-of-the-art model out-of-domain or simple model interactively? In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 79–89, Hong Kong, China. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D19-6613`.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 11, pages 230–239. Number: 1.

Starbird, K., Arif, A., Wilson, T., Koevering, K. V., Yefimova, K., and Scarnecchia, D. (2018). Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains. In *Proceedings of the International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *EACL 2012, 13th Conference of* *the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 102–107.

Stier, S., Siegers, P., Breuer, J., Thorson, K., Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Soc. Sci. Comput. Rev.*, 38(5):503–516.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sun, Q., Wang, Z., Zhu, Q., and Zhou, G. (2018). Stance Detection with Hierarchical Attention Network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Syed, Z. H., Röder, M., and Ngomo, A.-C. N. (2019). Unsupervised Discovery of Corroborative Paths for Fact Validation. In Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., and Gandon, F., editors, *The Semantic Web – ISWC 2019*, Lecture Notes in Computer Science, pages 630–646, Cham. Springer International Publishing. `https://www.doi.org/10.1007/978-3-030-30793-6_36`.

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., and Todorov, K. (2019a). Claimskg. In Ghidini, C., Hartig, O., and Maleshkova, M., editors, *The Semantic Web – ISWC 2019. ISWC 2019*, pages 309–324. Springer, Cham.

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., and Todorov, K. (2019b). ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., and Gandon, F., editors, *The Semantic Web – ISWC 2019*, Lecture Notes in Computer Science, pages 309–324, Cham. Springer International Publishing. `https://www.doi.org/10.1007/978-3-030-30796-7_20`.

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., and Todorov, K. (2019c). Claimskg: A knowledge graph of fact-checked claims. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings*, pages 309–324, Berlin, Heidelberg. Springer.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Thonet, T., Cabanac, G., Boughanem, M., and Pinel-Sauvagnat, K. (2016). VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery. In Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., and Silvello, G., editors, *Advances in Information Retrieval*, volume 9626 of *Lecture Notes in Computer Science*, pages 533–545. Springer International Publishing, Cham. `https://www.doi.org/10.1007/978-3-319-30671-1_39`.

Thonet, T., Cabanac, G., Boughanem, M., and Pinel-Sauvagnat, K. (2017). Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 87–96, New York, NY, USA. Association for Computing Machinery.

Thorne, J. and Vlachos, A. (2017). An Extensible Framework for Verification of Numerical Claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain. Association for Computational Linguistics.

Thorne, J. and Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/N18-1074.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018b). 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)*, pages 809–819.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018c). The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/W18-5501.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018d). The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/D19-6601.

Thuma, E., Motlogelwa, N. P., Leburu-Dingalo, T., and Mudongo, M. (2020). Ub_et at checkthat! 2020: Exploring ad hoc retrieval approaches in verified claims retrieval. In Cappellato, L., Eickhoff, C., Ferro, N., and Névéol, A., editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Toledo-Ronen, O., Bar-Haim, R., and Slonim, N. (2016). Expert Stance Graphs for Computational Argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123, Berlin, Germany. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/W16-2814.

Torsi, B. and Morante, R. (2018). Annotating Claims in the Vaccination Debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56, Brussels, Belgium.

Association for Computational Linguistics. https://www.doi.org/10.18653/v1/W18-5207.

Trabelsi, A. and Zaiane, O. R. (2018). Unsupervised Model for Topic Viewpoint Discovery in Online Debates Leveraging Author Interactions. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, pages 425–433. Association for the Advancement of ArtificialIntelligence.

Trautmann, D., Daxenberger, J., Stab, C., Schütze, H., and Gurevych, I. (2020). Fine-Grained Argument Unit Recognition and Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI' 20. https://doi.org/10.1609/aaai.v34i05.6438.

Trisedya, B. D., Weikum, G., Qi, J., and Zhang, R. (2019). Neural Relation Extraction for Knowledge Base Enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/P19-1023.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake News Detection in Social Networks via Crowd Signals. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 517–524, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tsurel, D., Pelleg, D., Guy, I., and Shahaf, D. (2017). Fun Facts: Automatic Trivia Fact Extraction from Wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 345–354. https://www.doi.org/10.1145/3018661.3018709.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4(1):14–28. https://www.doi.org/10.1016/j.websem.2005.10.002.

Van Quan, N., Yang, H.-J., Kim, K., and Oh, A.-R. (2017). Real-time earthquake detection using convolutional neural network and social data. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 154–157.

Veira, N., Keng, B., Padmanabhan, K., and Veneris, A. (2019). Unsupervised Embedding Enhancements of Knowledge Graphs using Textual Associations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5218–5225, Macao, China. International Joint Conferences on Artificial Intelligence Organization. `https://www.doi.org/10.24963/ijcai.2019/725`.

Vlachos, A. and Riedel, S. (2014). Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics. `https://www.doi.org/10.3115/v1/W14-2508`.

Vlachos, A. and Riedel, S. (2015). Identification and Verification of Simple Claims about Statistical Properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/D15-1312`.

Vo, N. and Lee, K. (2020). Where are the facts? searching for fact-checked information to alleviate the spread of fake news. *arXiv preprint arXiv:2010.03159*.

Voskarides, N., Meij, E., Reinanda, R., Khaitan, A., Osborne, M., Stefanoni, G., Kambadur, P., and de Rijke, M. (2018). Weakly-supervised Contextualization of Knowledge Graph Facts. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*, pages 765–774. `https://www.doi.org/10.1145/3209978.3210031`.

Vosoughi, S., Roy, D., and Aral, S. (2018a). The spread of true and false news online. *Science*, 359(6380):1146–1151. `https://www.doi.org/10.1126/science.aap9559`.

Vosoughi, S., Roy, D., and Aral, S. (2018b). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L. L., and Hajishirzi, H. (2022). SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Walker, M. A., Anand, P., Abbott, R., and Grant, R. (2012). Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 592–596, USA. Association for Computational Linguistics.

Walker, V. R., Foerster, D., Ponce, J. M., and Rosen, M. (2018). Evidence Types, Credibility Factors, and Patterns or Soft Rules for Weighing Conflicting Evidence: Argument Mining in the Context of Legal Rules Governing Evidence Assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels, Belgium. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/W18-5209`.

Wang, C., Yan, M., Yi, C., and Sha, Y. (2019). Capturing Semantic and Syntactic Information for Link Prediction in Knowledge Graphs. In Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., and Gandon, F., editors, *The Semantic Web – ISWC 2019*, Lecture Notes in Computer Science, pages 664–679, Cham. Springer International Publishing. `https://www.doi.org/10.1007/978-3-030-30793-6_38`.

Wang, H., Li, Y., Hutch, M., Naidech, A., and Luo, Y. (2021). Using tweets to understand how covid-19–related health beliefs are affected in the age of social media: Twitter data analysis study. *J Med Internet Res*, 23(2):e26302.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Short Papers), pages 422–426, Vancouver, Canada. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/P17-2067.

Wang, X., Sheng, Q. Z., Yao, L., Li, X., Fang, X. S., Xu, X., and Benatallah, B. (2016). Empowering Truth Discovery with Multi-Truth Prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, pages 881–890, Indianapolis, Indiana, USA. ACM Press. https://www.doi.org/10.1145/2983323.2983767.

Wang, X., Yu, C., Baumgartner, S., and Korn, F. (2018). Relevant Document Discovery for Fact-Checking Articles. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 525–533, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Weikum, G., Hoffart, J., and Suchanek, F. (2019). Knowledge Harvesting: Achievements and Challenges. In Steffen, B. and Woeginger, G., editors, *Computing and Software Science: State of the Art and Perspectives*, Lecture Notes in Computer Science, pages 217–235. Springer International Publishing, Cham.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pages 486–497, Berlin, Heidelberg. Springer-Verlag.

Wong, W., Liu, W., and Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4):20:1–20:36.

Wu, Y., Agarwal, P. K., Li, C., Yang, J., and Yu, C. (2014). Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600. https://www.doi.org/10.14778/2732286.2732295.

Wu, Y., Agarwal, P. K., Li, C., Yang, J., and Yu, C. (2017). Computational fact checking through query perturbations. *ACM Transactions on Database Systems*, 42(1):1–41.

Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., and Zhang, A. (2019). Towards Confidence Interval Estimation in Truth Discovery. *IEEE Transactions on Knowledge* *and Data Engineering*, 31(3):575–588. https://www.doi.org/10.1109/TKDE.2018.2837026.

Xu, C., Paris, C., Nepal, S., and Sparks, R. (2018). Cross-Target Stance Classification with Self-Attention Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/P18-2123.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable ai: A brief survey on history, research areas, approaches and challenges. In Tang, J., Kan, M.-Y., Zhao, D., Li, S., and Zan, H., editors, *Natural Language Processing and Chinese Computing*, pages 563–574, Cham. Springer International Publishing.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., and Zhu, T. (2020). Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *J Med Internet Res*, 22(11):e20550.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/N16-1174.

Yin, X., Han, J., and Yu, P. S. (2008). Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.

Yu, H. and Hatzivassiloglou, V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. https://www.doi.org/10.3115/1119355.1119372.

Yu, R., Gadiraju, U., Fetahu, B., Lehmberg, O., Ritze, D., and Dietze, S. (2018). KnowMore – knowledge base augmentation with structured web markup. *Semantic Web*, 10(1):159–180. https://www.doi.org/10.3233/SW-180304.

Z, D., ad Otten T, K. L., JW, P., and F., P. (2022). Finding the way forward: Covid-19 vaccination progress in germany, austria and switzerland. *Health Policy Technol.*, 11(2).

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2013). TheSoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web*, 4(3):257–263. `https://www.doi.org/10.3233/SW-2012-0081`.

Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C., and Weinstein, J. N. (2004). Mistaken identifiers: Gene name errors can be introduced inadvertently when using excel in bioinformatics. *BMC Bioinformatics*, 5(1):80.

Zhan, Q, Liang, S, Lipani, A, Ren, Z, and Yilmaz, E (2019). From Stances' Imbalance to Their Hierarchical Representation and Detection. In *Proceedings of WWW '19: The World Wide Web Conference*, WWW '19, pages 2323–2332, San Francisco, CA, USA. `https://doi.org/10.1145/3308558.3313724`.

Zhang, H., Li, Q., Ma, F., Xiao, H., Li, Y., Gao, J., and Su, L. (2016). Influence-Aware Truth Discovery. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, pages 851–860, Indianapolis, Indiana, USA. ACM Press. `https://www.doi.org/10.1145/2983323.2983785`.

Zhang, Q., Sun, Z., Hu, W., Chen, M., Guo, L., and Qu, Y. (2019a). Multi-view Knowledge Graph Embedding for Entity Alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5429–5435, Macao, China. International Joint Conferences on Artificial Intelligence Organization. `https://www.doi.org/10.24963/ijcai.2019/754`.

Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26:55–62. The 2nd Collaborative Innovation Networks Conference - COINs2010.

Zhang, Z., Gentile, A. L., and Ciravegna, F. (2013). Recent advances in methods of lexical semantic relatedness – a survey. *Natural Language Engineering*, 19(4):411–479. `https://www.doi.org/10.1017/S1351324912000125`.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019b). ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics. `https://www.doi.org/10.18653/v1/P19-1139`.

Zhi, S., Sun, Y., Liu, J., Zhang, C., and Han, J. (2017). ClaimVerif: A Real-time Claim Verification System Using the Web and Fact Databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, pages 2555–2558, Singapore, Singapore. ACM Press. `https://www.doi.org/10.1145/3132847.3133182`.

Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake News: Fundamental Theories, Detection Strategies and Challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 836–837, New York, NY, USA. Association for Computing Machinery.

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

Zhu, H., Xie, R., Liu, Z., and Sun, M. (2017). Iterative Entity Alignment via Joint Knowledge Embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4258–4264, Melbourne, Australia. International Joint Conferences on Artificial Intelligence Organization. `https://www.doi.org/10.24963/ijcai.2017/595`.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys*, 51(2):32:1–32:36.

Versicherung an Eides Statt

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.

Düsseldorf, 01. Juni 2023 Katarina Boland

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich habe zum Zeitpunkt der Abgabe dieser Dissertation keinerlei vorherige erfolglose oder erfolgreiche Promotionsversuche unternommen.

Düsseldorf, 01. Juni 2023 Katarina Boland

Hier die Hülle

mit der CD/DVD einkleben

This CD contains:

- a *pdf*-version of this PhD thesis
- the LaTeX- and image source data of this PhD thesis
- the source code that was developed for this PhD thesis
- the data that was generated, used and analyzed