Money talks –
Three essays on the influence and relevance of Investor Sentiment from Social Media on capital markets

Inaugural-Dissertation

to obtain the degree of Doctor of Business Administration
(doctor rerum politicaum – Dr. rer. pol.)

submitted to the

Faculty of Business Administration and Economics
Heinrich-Heine University Düsseldorf

presented by

Philipp Stangor, M.Sc.

Research Associate at the Chair of Business Administration,
esp. Financial Services
Heinrich Heine University Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf, Germany

Supervisor: Prof. Dr. Christoph J. Börner

Düsseldorf, 3rd of June 2024

Dedicated to my wife Saskia &
all what's about to come…

Acknowledgements

First and foremost, I would like to thank my doctoral advisor, Prof. Dr. Christoph J. Börner, for his professional and personal support at all times during the creation of this work. I have considered it a privilege to be able to freely choose the topics for each project that sparked my research interest. In mutual trust, the work was never marked by any kind of pressure, which always allowed me to freely achieve my own goals.

I would also like to thank Prof. Dr. Barbara E. Weißenberger for taking on the co-supervision. I wrote my first major academic paper, my bachelor's thesis, under her supervision at her chair, which brings a beautiful sense of closure for me.

During my time at the chair, I had the opportunity to get to know a number of colleagues who made this time a remarkable one that I greatly enjoyed and during which I was able to grow personally. In particular, I would like to thank our last team, consisting of Anne-Marie Ossig, Julia Schedrina, Jonas Krettek, John H. Stiebel, and Lars M. Kürzinger, for helping bringing this to the finish line with many professional and personal conversations. I owe my greatest thanks to the latter, who, as a co-author and constant companion throughout my years at the university, has been an unwavering support.

Last but definitely not least, I thank my family – my parents, my sister, and especially my wife, Saskia – who have always believed in me and supported me with love unconditionally in all respects.

Düsseldorf in June 2024

Table of content

List of tables

List of figures

List of abbreviations

AAII	American Association of Individual Investors
Adj.	Adjusted
AG	Aktiengesellschaft
AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
Bal. Acc.	Balanced Accuracy
BERT	Bidirectional Encoder Representations from Transformers
CBOW	Continuous-Bag-of-Words
CEFD	Closed-End Fund Discount
CEO	Chief Executing Officer
CV	Count Vectors
D.C.	District of Columbia
Diff	Difference
EFA	Eastern Finance Association
e.g.	exempli gratia
EL	EmoLex
ELMo	Embeddings from Language Model
EMH	Efficient Market Hypothesis
esp.	especially
et al.	et alia
GI	(Harvard) General Inquirer
GloVe	Global Vectors (for Word Representation)

H	Hypothesis
HAC	Heteroskedasticity- and autocorrelation-consistent (standard errors)
HE	Henry (dictionary)
HeiCAD	Heine Center for Artificial Intelligence and Data Science
HIT	Human Intelligence Task
HVB	HypoVereinsbank
i.a.	inter alia
i.e.	id est
IDF	Inverse Document Frequency
II	Investors Intelligence
Inc.	Incorporated
IPO	Initial Public Offering
IS	Investor Sentiment
LASER	Language-Agnostic Sentence Representations
LM	Loughran/McDonald (dictionary)
LSTM	Long Short-Term Memory
Max	Maximum
Min	Minimum
MLM	Masked Language Model
MTurk	Mechanical Turk
NASDAQ	National Association of Securities Dealers Automated Quotations
NAV	Net Asset Value
NLP	Natural Language Processing
NMFD	Net Mutual Fund Redemptions
No.	Number

NRC	Nation Research Council (Canada)
NSP	Next Sentence Prediction
Obs.	Observations
PCA	Principal Component Analysis
p.m.	post meridiem
RNN	Recurrent Neural Network
Sd	Standard deviation
SEC	(United States) Securities and Exchange Commission
SG	Skip-Gram
S&P	Standard & Poor's
TBED	Text-Based Emotion Detection
TF	Term Frequency
US	United States
USA	United States of America
USE	Universal Sentence Encoder
VADER	Valence Aware Dictionary and sEntiment Reasoner
XLM	Cross-lingual Language Models

List of symbols

Latin symbols

a	Indirect effect of mediation analysis
$Adj.R^2$	Adjusted determination coefficient
b	Indirect effect of mediation analysis
Be	Bearish (tweets)
Bu	Bullish (tweets)
c	Direct effect of mediation analysis
C	Cost
$Closing$	Closing price
CS	Cosine Similarity
$E(x)$	Expected value of x
EM	Emotional (dictionary)
F_Sen	Financial Sentiment
i	Counter
	Intercept *(in the context of mediation analysis)*
$Intraday$	Intraday return (also R in Tables)
lim	Limes
m	Counter
Med	Median
$\min(x)$	Minimum of x
N	Number
	Negative financials, no tweets *(in the context of vignettes in Section 2)*
$N(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ

NN	Negative financials, negative tweets
NP	Negative financials, positive tweets
$Opening$	Opening price
p	p-Value
P	Positive financials, no tweets *(in the context of vignettes in Section 2)*
P_t	Price at time t
PN	Positive financials, negative tweets
PN_i	Positive-Negative (dictionary i)
PP	Positive financials, positive tweets
r	Return
R	Intraday return (also $Intraday$ in text)
R^2	Determination coefficient
$Sentiment$	Sentiment score
SC	Stock Capital
t	Counter (in a time context)
T_Sen	Tweet Sentiment
U	Utility
$Valence$	Valence score
Z	Z-score

Greek symbols

α	Drift
β	Regression coefficient
$\tilde{\beta}$	Standardized regression coefficient

δ	Partial derivative
ε	Residual
θ	Information set
σ_i	Standard deviation of i

Other symbols

Δ	Delta
€	Euro
#	Hashtag
∑	Sum
$	US-Dollar or Cashtag (in a social media context)
$\vec{V}$	Vector V

1 Introduction

'We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely.'

- Edward O. Wilson

With the advancement of digitalization, our lives have changed significantly, especially in terms of the availability and handling of information. Never before have so many people on this planet been able to both consume an infinitely appearing amount of information and produce their own content easily and for free. While about 16% of the world's population had access to the internet in 2005, by 2022 this figure had already risen to an estimated 66%, or 75% among those aged 15 to 24. In particular, in Europe as well as in North and South America, the user share is even higher, ranging between 80% and 90% (International Telecommunication Union (2022)). In 2022, 103.66 zettabytes of data were shared – one zettabyte being equivalent to one billion terabytes. By 2027, Sörries and Wissner (2023) project this data volume will nearly triple to 284.30 zettabytes.

In the sense of the US biologist Edward O. Wilson, this surplus of information does not automatically equate to increased wisdom. Rather, it is important to act as a 'synthesizer', selecting information based on its relevance and truthfulness, appropriately aggregating it, and using it thoughtfully as a basis for decision-making. This is easier said than done, especially against the backdrop of the ever-increasing importance of social media platforms both in the media landscape and in everyday life, where anyone can largely share information unchecked, thereby consciously or unconsciously spreading misinformation. With the help of 'bots' even individuals or groups can attempt to assign greater relevance to their viewpoints. The dangers of bots on Twitter prominently came into focus for the first time during the 2016 US election (Orabi et al. (2020), Bessi and Ferrara (2016)).

Against this background, the question arises as to how social media information affects financial markets. The power of social media was recently highlighted in the highly publicized case of the GameStop stock, but investors also exchange information daily

outside of such extreme events, for example, by using cashtags like '$AAPL' for the stock of Apple Inc. (Long et al. (2023), Umar et al. (2021)). In economic research, the term 'investor sentiment' has become established, which can be extracted from social media data. This work aims to examine whether and how investor sentiment can impact individual investment decisions on a micro level and, consequently, financial markets on a macro level.

1.1 On the role of information and investor sentiment on capital markets

When examining how information impacts financial markets, an engagement with the Efficient Market Hypothesis (EMH) by Fama (1970) is unavoidable. Fama (1970) defines a market as fully informationally efficient if all available information is included in the information set and correctly interpreted in relation to a market price P. Consequently, the expected price $E(P_{i,t+1})$ of an asset i at time $t+1$ must correspond to the expectation of this price at time t given the information set θ_t available at that time:

$$E\big(P_{i,t+1} - E(P_{i,t+1}|\theta_t)\big) = 0 \tag{1}$$

If we assume the existence of private insider information, insiders can use this information to generate systematic excess returns. In this case, the market is defined as semi-strong efficient and equation (1) becomes an inequality. If the information set contains only historical price data, such excess returns can also be achieved using public fundamental information (and, of course, insider information).

Fama (1970) himself, as well as other researchers, have addressed the (empirical) determination of the degree of informational efficiency. A problematic aspect of this examination is that a definitive verification of the hypothesis is always subject to the 'joint hypothesis problem' which states that rejecting a hypothesis always involves assuming another necessary hypothesis. In the case of testing informational efficiency, this would always be the assumption of a correct asset pricing model. Nonetheless, there is a wide range of evidence in financial research for markets that are not (fully) informationally efficient due to excess volatility (Shiller, Fischer, and Friedman (1984)) or various calendar or fundamental anomalies as the January effect or value anomaly.[1]

[1] A good overview is given by Latif et al. (2011).

Particularly, the assumption associated with the Efficient Market Hypothesis that investors are rational, which is necessary for the correct processing of all information, has come under scrutiny from its critics. This assumption is not compatible with the practical observation of historical events such as the Tulip Mania of 1637, the South Sea Bubble of 1720, or more recent events like the Dotcom Bubble of the early 2000s (Aliber and Kindleberger (2015), Baker and Wurgler (2006), Malkiel (2003), Mackay (1869)).

Building on the concept of 'noise' mainly introduced by Kyle (1985) and Black (1986), Long et al. (1990) establish the Noise Trader Theory, in which asset prices can deviate from their fundamental value in the short to medium term. According to the Efficient Market Hypothesis, irrational noise traders would be driven out of the market by rational arbitrageurs in the long run. However, Long et al. (1990) argue that limits of arbitrage exist, ensuring that irrational noise traders can persist in the market. Since the behavior of noise traders and their reaction to subsequent information is unpredictable and therefore a longer-term mispricing is possible, arbitrageurs are increasingly exposed to noise trader risk the more risk-averse they are and the shorter their investment horizon is. This effect is amplified by the presence of transaction costs in the market and the empirical observation that (especially individual) investors tend to avoid taking short positions (Stambaugh, Yu, and Yuan (2012), Miller (1977)). Consequently, Long et al. (1990) argue that investor sentiment should be considered in asset pricing models. Regardless of asset pricing, markets with many noise traders appear more volatile. Investor sentiment (or noise trading) can also explain the frequently observed fact that markets are often more volatile than estimated by models not utilizing investor sentiment (Giglio and Kelly (2018), Shiller (1981), Miller (1977)).

The assumption of zero information costs, put forth by Fama (1970), conflicts with real-world conditions. This assumption, especially the absence of information costs, has sparked debates in financial literature: while Fama defines this assumption as merely 'sufficient', Grossman and Stiglitz (1980) theoretically model that this condition is 'necessary'. Essentially, the authors argue that under the assumption of costs C for information θ, the price P of an asset in a competitive market can only contain information if market participants are compensated with at least a small premium for gathering the information.

In a (nearly) fully informationally efficient market, as per Fama, there can be initially incentives for information gathering. However, the costly information obtained by informed investors also becomes observable to uninformed investors through the price. To save on costs C while maintaining the same profit, informed investors might cease information gathering. Consequently, this configuration leads to no market equilibrium, a situation later known as the 'information paradox' (see Figure 1).

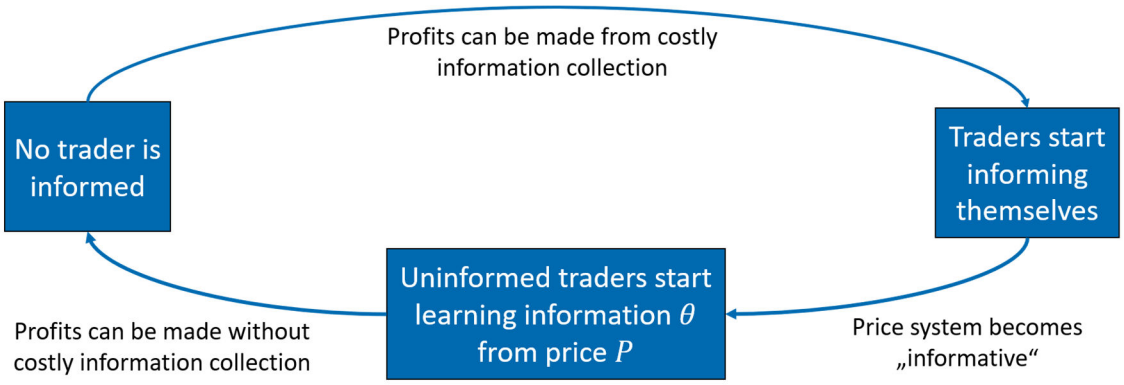


Figure 1: The information paradox following Grossman and Stiglitz (1980)

If the market is not fully informationally efficient, the price only partially reflects the information, allowing uninformed investors to continue free-riding. At the same time, monetary incentives remain for informed investors to gather information, enabling a balance where both groups can achieve equal expected utilities. This equilibrium can only occur in a situation where the market is not completely efficient, thus maintaining incentives for informed investors to continue their information gathering activities.

Apart from the model of Grossman and Stiglitz (1980), one can simplify the determination θ of by assuming that a rational market participant might choose the size of the information set θ in the following manner, depending on a concavely shaped utility function $U(\theta)$ and convex information costs $C(\theta)$:

$$\frac{\partial U}{\partial \theta} = \frac{\partial C}{\partial \theta}$$
$$with \ \frac{\partial U}{\partial \theta} > 0, \frac{\partial^2 U}{\partial \theta^2} < 0, \frac{\partial C}{\partial \theta} > 0 \ and \ \frac{\partial^2 C}{\partial} > 0 \quad (2)$$

In this sense, it is desirable to design information gathering in such a way that the ratio between the information content and the costs of collection is maximized, making the

utility function of information gathering $U(\theta)$ as steep and the cost function $C(\theta)$ as flat as possible. This fundamental idea aligns with the results of the market model by Grossman and Stiglitz (1980), in which an equilibrium situation is particularly possible for low-cost and/or precise information. Jing, Wu, and Wang (2021) conclude that social media could gradually deprofessionalize the job of financial analysts, who previously could save investors information costs through their specialization.

From a theoretical perspective, the role of social media information can impact investors in terms of investor sentiment in the following ways:

(1) Social media as an information cost-saving institution that provides a wide range of (free) information for investors while also pre-selecting and prioritizing the collection and processing of information through the work of communities and algorithms.
(2) Social media information as a pure proxy for measuring investor sentiment, to which noise traders react and further propagate on the platforms.

1.2 Investor Sentiment

1.2.1 Definition

Before addressing the question of appropriate measurement of investor sentiment, the term should first be defined and delineated. The definition of investor sentiment in economic literature is not clear-cut. For instance, Aggarwal (2022) criticizes that past research on investor sentiment has focused too much on measurement without a unified definition. This thesis is supported by the observation that individual sentiment measures (especially in early sentiment research) are sometimes independent of each other or even contradictory (Qiu and Welch (2004)).

The most fundamental definition is provided by Shleifer and Summers (1990), who define investor sentiment as the ‘overall investor attitude towards financial markets’. Investor sentiment thus encompasses any kind of emotions, moods, beliefs, or future expectations prevailing in the market. Similarly, Barberis, Shleifer, and Vishny (1998) offer a similar definition, specifying the proposal of Shleifer and Summers (1990) by stating that these beliefs occur in ‘formations’ that lead to over- and underreactions in capital markets. Baker and Wurgler (2006) build on the Noise Trader Theory outlined in the previous subsection, as proposed by Long et al. (1990), defining it as the propensity to speculate

about the future based on noise. In a later work, they define this propensity as ‘optimism’ or ‘pessimism’ towards financial markets (Baker and Wurgler (2007)).

The most common specification focuses on the fundamental value of an asset.[2] While Brown and Cliff (2004) initially define investor sentiment as market expectations relative to a ‘norm’, still quite broadly formulated, others such as Zhang (2008) and Shefrin and Belotti (2008) utilize the concept of fundamental value, viewing investor sentiment as a (subjective) belief of investors regarding this value. The understanding of investor sentiment aims to encompass all these thoughts and beliefs of investors (deviating from the fundamental value), triggered by irrational behavior or cognitive biases. Following the distinction made by Bormann (2013) between short-term and long-term investor sentiment, this work will primarily focus on the short-term component. Bormann (2013) highlights that short-term sentiment deals more with rapidly changing but stronger emotions. Emotions differ from the long-term component, mood, in that they directly relate to a person or object – in this case, an asset – while mood represents a longer-term persistent attitude of the investor.

Along with the insights from Section 1.1, the measurement of this conceptual construct will now be discussed, or as Baker and Wurgler (2007) put it: ‘Now, the question is no longer, as it was a few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and quantify its effects.’ This is a task, as Aggarwal (2022) rightly points out, that is far from being completed, and thus, particularly ‘more studies [...] by using computationally intensive sentiment analysis techniques for prediction of financial markets’ are needed.

1.2.2 Measurement

1.2.2.1 Market-based

Market-based measures, along with survey-based measures, represent the oldest research strand in this field. Aggarwal (2022) further divides this area into ‘related market measures’, ‘direct equity market activity’, and ‘market performance measures’.

One of the most well-known measures in the first group is the Closed-End Fund Discount (CEFD), first introduced as a sentiment measure by Lee, Shleifer, and Thaler (1991).

[2] At this point, it should also be noted once again that the definitional appeal to the fundamental value of an asset leads again to the ‘joint hypothesis problem’ with the assumption of a correct asset pricing model.

Investor sentiment, in this context, is seen as the difference between the Net Asset Value (NAV) of the security holdings of the fund and its market value, which is predominantly held by individual investors (Lee, Jiang, and Indro (2002)). A positive (negative) change in the discount indicates pessimism (optimism) among investors, as the change in market value relative to NAV is more negative (less positive) or more positive (less negative), respectively. Similarly, the observation of Net Mutual Fund Redemptions (NMFD) follows a similar direction. While Malkiel (1977) provides rational explanations for these relationships, they also acknowledge that a significant portion of discounts and redemptions can only be explained by irrational behavior.[3] In recent years, Ben-Rephael, Kandel, and Wohl (2012) have proposed using net exchanges instead of net redemptions and demonstrated a stronger correlation with 'noise' in the American financial market.

Also popular as an investor sentiment measure are Initial Public Offering (IPO) data. Ritter (1991) and Cornelli, Goldreich, and Ljungqvist (2006) use the success of IPOs in the form of first-day returns as a sentiment measure. Baker and Wurgler (2007) take a step back and introduce the number and volume of IPOs as a sentiment measure, assuming that companies time their IPOs according to market sentiment. Additionally, Baker and Wurgler (2004) had previously introduced that dividend premiums (the difference between the average market-to-book ratio of dividend payers and non-payers) have an inverse relationship with investor sentiment. Lastly, within the first group of market-based measures, insider trading is also included, for which Seyhun (2000) and Jeng, Metrick, and Zeckhauser (1999) show that insiders can still earn abnormal returns even within the bounds of legality.

In the area of 'direct equity market activity' Baker and Stein (2004) use market liquidity as a proxy for investor sentiment. They assume that irrational noise traders are subject to a short sale constraint, whereby higher market liquidity automatically indicates a more positive sentiment and vice versa. Similar to the number of IPOs mentioned earlier is the use of the ratio between equity issues to total (equity & debt) issues. Baker and Wurgler (2000) empirically show that a high ratio is usually followed by lower returns, indicating an inverse relationship with investor sentiment. Since irrational behavior is primarily

[3] Baker and Wurgler (2007), Frazzini and Lamont (2007), Brown et al. (2003) and Neal and Wheatley (1998) give various explanatory approaches.

attributed to smaller investors, odd lot sales[4] were long used as a measure of investor sentiment (Barber (1994), Brodie (1940)). However, Neal and Wheatley (1998) demonstrated that the influence on stock returns is minimal. More recent studies have also introduced the put-call ratio (Wang, Keswani, and Taylor (2006)) and the buy-sell imbalance (Kumar and Lee (2006)).

'Market performance measures' encompass the widest range of variations. One of the earliest variations was the Bearish Sentiment Index, which indicates the proportion of bearish investment advisors relative to all investment advisors. Solt and Statman (1988) show that this measure is not capable of predicting asset returns. Further, the ARMS index is defined as the ratio between the ratio of advancing and declining stocks and the ratio between advancing and declining volume – just to name one of the technical indicators. An index value of one implies a neutral market, a value below one implies a bullish market, and a value above one implies a bearish market, as this value suggests more (less) volume in the average advancing stock and less (more) volume in the declining stock (Wang, Keswani, and Taylor (2006)).

Baker and Wurgler (2006) emphasize the use of multiple measures simultaneously by employing principal component analysis (PCA). In their widely used Baker and Wurgler Sentiment Index, they include the closed-end fund discount, first-day IPO returns, IPO volume, dividend premium (value-weighted), and the ratio between equity issues to total issues. Empirical applications have already demonstrated significant relationships in stock markets (Yu and Yuan (2011)), bond markets (Nayak (2010)), and external corporate finance decision-making (McLean and Zhao (2014)).

Nevertheless, market-based measures also face significant criticism. Foremost among these is the issue of endogeneity, where changes in other market factors can influence the assessment of investor sentiment. For example, Baker and Wurgler (2006) initially included the turnover ratio of the NYSE as the sixth indicator in their sentiment index. However, they had to remove this indicator due to the misleading impact of the substantial increase in high-frequency trading. Additionally, Qiu and Welch (2004) demonstrate that

[4] An odd lot trade is an order that is not divisible by 100 without remainder. These order sizes are typically chosen by individual traders.

market-based measures, such as the CEFD, do not align with the next group of sentiment measures – survey-based sentiments – that directly query sentiment.

1.2.2.2 Survey-based

Survey-based investor sentiment measures have already been collected extensively even before the discussion about market efficiency and noise trading began. To this day, representative (target) groups are regularly asked the same questions about their (economic) expectations at regular intervals. Similar to the 'market performance measures" there is also an enormous variety of different indices within survey-based measures. The following will present those that have garnered the most attention in economic research. Broadly, the collected surveys can be divided into those that measure the sentiment of professionals and those that measure the sentiment of individuals. In line with the assumption that noise trading is predominantly driven by individuals and that individuals are easier to reach for surveys, it is not surprising that the latter group enjoys significantly more popularity in research.

Investors Intelligence (II) describes a measure from the first group as advisor sentiment. Weekly market newsletters issued by (semi-)professional investors are analyzed. Although the collected data is used to fill out a survey, the evaluation of the newsletters resembles early methods of text-based measurement of investor sentiment, which will be discussed in the following chapter. The index of the American Association of Individual Investors (AAII) also surveys individual investors weekly about their expectations for the stock market over the next six months. Fisher and Statman (2000) show that both measures, despite different investor groups, exhibit a significant correlation of 0.47 and only a weak negative relationship with future returns. Later analyses confirm the observation of a contrarian relationship with significant results (Verma and Soydemir (2009), Verma and Verma (2008), Zwergel and Klein (2006), Brown and Cliff (2004)).[5]

Also noteworthy are measures that, while not explicitly focusing on investor sentiment, instead examine consumer sentiment. The University of Michigan's Index, based on telephone interviews with at least 50 questions, sets the sentiment with a number of participants in the mid-three-digit range. On the other hand, The Conference Board Consumer Confidence Index relies on five questions about business, income, and

[5] Data for both available at https://www.investorsintelligence.com/x/default.html (Investors Intelligence) and https://www.aaii.com/sentiment-survey (AAII).

employment, administered to approximately 5000 households.[6] Otoo (2000) demonstrates that both measures nonetheless strongly correlate with each other and exhibit a significant relationship with stock prices.

The most well-known weekly survey published in the German-speaking region is Sentix, which predominantly collects German data. Sentix provides a clear distinction between individual and professional investors, as well as differentiates between short- and mid-term perspectives. These specifications have made this survey measure attractive for recent studies once again. Schmeling (2007) utilizes Sentix data as a proxy for 'smart money' and noise trader risk. Heiden, Klein, and Zwergel (2013) demonstrate medium-term relationships between professional sentiment and exchange rates, which are significant but highly fluctuating. Bormann (2013) utilizes the division into short- and middle-term perspectives to align with the distinction made in Section 1.2.1 between short-term emotions and medium- to long-term mood.[7]

In a comparative study involving six surveys, including all the surveys mentioned here except Sentix, Greenwood and Shleifer (2014) find that survey sentiments are highly correlated with each other and also exhibit relationships with past market returns and mutual fund inflows. 'These results suggest that survey measures of investor expectations are not meaningless noise but are rather reflections of widely shared beliefs about future market returns, which tend to be extrapolative in nature' (Greenwood and Shleifer (2014)).

Survey data also present challenges. Apparently, the data are available at too short frequencies. This is particularly problematic in light of recent findings in which intraday relationships between investor sentiment and assets are demonstrated (e.g., Gao and Liu (2020), Behrendt and Schmidt (2018), Renault (2017), Sun, Najand, and Shen (2016)). Furthermore, some researchers question the complete credibility of both the representativeness of the sample population and the responses themselves. Incentives may exist to provide distorted information in order to influence the decisions of others, including political institutions, for one's own benefit (Zhou (2018), Singer (2002)).

[6] Data for both available at http://www.sca.isr.umich.edu/ (University of Michigan) and https://www.conference-board.org/topics/consumer-confidence (Conference Board).
[7] Data available at https://www.sentix.de/.

1.2.2.3 Text-based

Text-based measures can largely circumvent the problems associated with market-based and survey-based measures. They widely exhibit exogeneity and provide unbiased data since the observed individuals do not feel they are being observed, which imparts a 'living lab' characteristic. Additionally, as noted in Section 1, the availability of such data is continually growing. Initially, the limitations that led to this area being largely ignored in early sentiment research were primarily due to limited computational capacity and the lack of advanced algorithms for evaluation. Today, these limitations have been largely addressed, though there is always room for further improvement.

In a seminal contribution, Antweiler and Frank (2004) analyzed 1.5 million messages on internet stock message boards from Yahoo! Finance and Raging Bull, finding initial links to stock market volatility. In most studies, investor sentiment is always represented as a ratio of news classified as 'positive' and 'negative'. Closely related to text-based measures are media-based measures, where sentiment is derived from search behavior or media coverage on specific topics. While media-based measures will not be further explored here, a comprehensive overview can be found in works like Aggarwal (2022).

Textual analysis, in contrast, can be categorized into 'dictionaries", 'statistical methods", 'word & sentence encoders' and 'transformers'. Unlike the methods discussed in previous chapters, these forms are not competitive but rather represent successive stages of development, with each building upon the last, despite minor inconsistencies (Mishev et al. (2020)). The following Section provides a brief overview of the fundamental technical developments of each stage.

Dictionaries

In the dictionary approach, texts are classified solely based on the words used, using a word list or dictionary that assigns classifications to words. In its basic form, this method completely ignores the syntactic properties of the text and assumes an independence between the words in a sentence (known as the 'bag-of-words' approach, Loughran and McDonald (2016)). For example, a tweet would be classified into a category (e.g., 'positive') if the majority of the classifiable words in the tweet belong to that category. [8]

[8] Other thresholds, deviating from the majority, can of course also be defined.

The first dictionaries emerged much earlier in sociological and psychological contexts, with the Harvard IV-4 dictionaries as part of the Harvard General Inquirer (GI) in the 1960s being their most well-known representative. Loughran and McDonald (2011) explain in their work, using the word 'risk' as an example, that words in the context of finance and accounting can have different connotations than in general language, as illustrated by the Harvard GI. Consequently, they themselves, as well as other researchers like Henry (2008), developed various dictionaries with different focuses around finance and accounting. Although Renault (2017) later emphasized that highly specialized, field-specific dictionaries can have advantages over more advanced algorithms, research has increasingly focused on these algorithms due to their enormous successes, making dictionaries less critical for text classification today.

Statistical methods

The simplest form is the count vectorizer, where a text is translated into a vector that documents the occurrence of each word. Consequently, the most frequently occurring words can be analyzed for different classifications. Count vectorizers also neglect semantics and overrepresent irrelevant filler words – a problem that can be addressed using the algorithm 'term frequency – inverse document frequency' (TF-IDF, Mishev et al. (2020)). Following the basic idea of Sparck Jones (1972), TF-IDF introduces a penalizing term, the 'inverse document frequency' (IDF), in addition to the 'term frequency' (TF) known from the count vectorizer, if a word frequently appears in multiple documents (e.g., sentences) within the entire corpus (e.g., a text) being analyzed.

Word- & Sentence-Encoders

Word-encoders are the first to assign a crucial role to context. The idea is to assign similar vectors, known as 'embeddings', to words that appear in a similar context. Over time, various word-encoder designs have been developed. In 'Word2Vec' introduced by Mikolov et al. (2013), vectors can be created unsupervised using the 'Continuous-Bag-of-Words' (CBOW) and 'Skip-Gram' (SG) models. In the CBOW model, the aim is to predict a single target word based on its context, while in the SG model, the context is evaluated based on the word. CBOW works faster and handles common words and phrases better than SG (Mikolov et al. (2013)). Later advancements like Global Vectors for Word Representation (GloVe, Kelton and Pennington (2020)) and FastText

(Bojanowski et al. (2017)) address weaknesses of 'Word2Vec' such as handling unknown words (Mishev et al. (2020)).

Sentence-encoders differ in that embeddings are assigned not at the word level but at the sentence or at least the phrase level. Using a pre-trained neural network, variably long sentences are translated into a fixed-size numerical representation. Among the most well-known sentence-encoders are Doc2Vec (Le and Mikolov (2014)), Universal Sentence Encoder (USE, Cer et al. (2018)), and Language-Agnostic Sentence Representations (LASER, Artetxe and Schwenk (2019)).

Transformers

In their paper 'Attention is All You Need' Vaswani et al. (2017) introduce the 'Transformers' model architecture, which revolutionized natural language processing and enabled newer developments such as BERT or ChatGPT. Transformers use an encoder-decoder structure (Figure 2, left side for encoder, right side for decoder) with two crucial innovations that allow them to outperform previous models like Recurrent Neural Networks (RNN) or their advancements like Long Short-Term Memory (LSTM): positional encoding and (self-)attention.

Positional encoding means that both the input embeddings and the output embeddings are augmented with an additional component that signals their position. Information about the position of a word in a sentence or text is thus no longer part of the model structure as before but rather part of the model's data.

The concept of attention is already known in RNNs (and thus also in LSTMs) and describes the weighting of words based on their importance. The crucial advantage of the (Self-)Attention introduced by Vaswani et al. (2017) is that the attention weights can be computed simultaneously in this method, allowing for parallelization of the process.

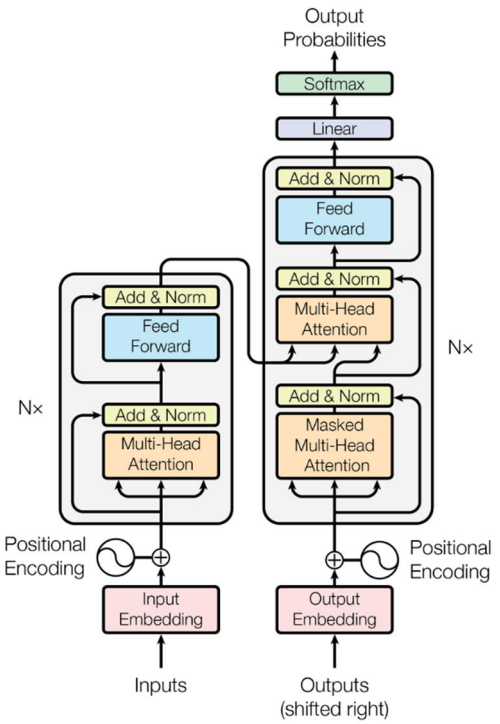


Figure 2: The Transformer – model architecture by Vaswani et al. (2017)

Additionally, the reference window in which words can be related to each other is no longer limited, enabling Transformers to have a significantly longer memory (Mishev et al. (2020)). Compared to word encoders, the embeddings of a word can take on multiple values, allowing, for example, the word ‘Apple’ to be recognized in the respective context as a fruit or a company (Vaswani et al. (2017)).

Building on this, in 2018, Devlin et al. (2018) introduced a language representation model called ‘Bidirectional Encoder Representations from Transformers’ (BERT). Unlike previous Transformers, BERT considers context bidirectionally for the first time – that is, both left and right. This is achieved in pre-training through a Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM randomly masks 15% of the words (tokens) to learn to predict them from the context. For NSP, segment embeddings are added to the word and position embeddings to include information about the sentence structure. BERT is thus able to recognize the relationship between sentences, which is particularly important for tasks such as question answering (Devlin et al. (2018)).

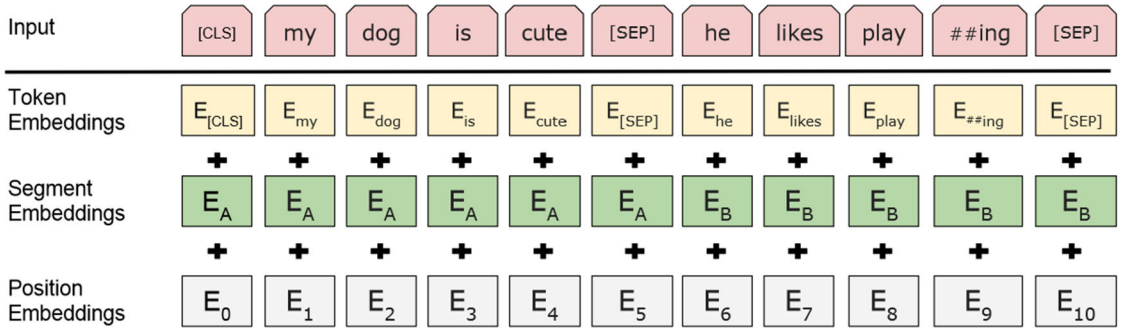


Figure 3: Example for a BERT input representation by Devlin et al. (2018)

In an initial version, the authors create BERT in a ‘Base’ and a ‘Large’ version, which differ in the number of layers used, the hidden size, self-attention heads, and the number of parameters. Later, other authors introduce further specifications of BERT, such as RoBERTa (Liu et al. (2019)) or DistilBERT (Sanh et al. (2020)), aiming to achieve even better results or significantly reduce computational overhead while maintaining comparable performance. In the financial context Araci (2019) fine-tuned BERT for the first time and demonstrated that their BERT version, called ‘FinBERT’ outperformed previous approaches such as LSTM or ELMo. Due to the relatively easy accessibility of BERT models, for example through libraries like transformers in Python, nearly 10,000

text classification BERT models are available on platforms like *Huggingface.co* up to now.

1.3 Research gap & contribution

As has already been demonstrated and will be further illustrated in the following chapters, many studies have already focused on the development and application of various investor sentiment measures. Therefore, this work will primarily focus on a more detailed experimental investigation of the mechanisms through which investor sentiment operates, as well as on the technical and, above all, ideological advancement of text-based sentiment measures.

In their work, Aggarwal (2022) and Bormann (2013) rightly criticize the overly rigid focus on the purely technical development of new measures beyond a fundamental understanding of the causal relationships and psychology of investors. Often, only individual findings from the field of behavioral finance are used for explanation without placing them in a broader framework, such as the Prospect Theory proposed by Kahneman and Tversky (1979). An essential component here is the experimental research in the laboratory to demonstrate whether and how investor sentiment (derived from text contributions) arises and through which channels it affects individual investment decisions.

Previous experimental research contributions have focused more on the ‘what?’ rather than the more necessary ‘how?’ for comprehensive understanding. Hales, Kuang, and Venkatarman (2011) show in their study that investors are more influenced by vivid language, especially when the observed text contradicts their own beliefs. These observations are complemented by Tan, Ying Wang, and Zhou (2014) and Rennekamp and Witz (2021), who investigate the impact of readability. Finally, Boulu-Reshef et al. (2023) examine the influence of emojis and find a significant yet marginal effect on investment decisions.

In Section 2, the focus will be on experimentally examining the impact of investor sentiment – specifically on social media – on individual investment decisions. Diverging from previous studies that primarily concentrated on the properties of the text, this section will utilize a mediator analysis to scrutinize the channels through which these effects operate. To achieve this, a decision-making scenario will be created on an information

platform, where participants can access both historical and fundamental data as well as social media data about a fictional company. Participants will then make an investment decision based on this information. In addition to the investment decision itself, the study will also observe participants' perceptions of financial and social media sentiment.

As demonstrated in the previous chapters, technical advancements in the field of textual analysis are indispensable and – as shown by Mishev et al. (2020) – continually evolving positively. BERT models, in particular, have gained significant popularity in recent years due to their superior accuracy rates compared to other models and their relatively simple application and development possibilities. However, in the financial context, Text-Based Emotion Detection (TBED) has been comparatively neglected compared to other disciplines, as indicated by Zad et al. (2021) in a review study. The works in Section 3 and 4 will aim to accomplish two tasks:

In a first step, tweets will be analyzed for application in the economic context using both the EmoLex dictionary by Mohammad and Turney (2013) and the two-dimensional positive-negative dictionaries introduced in Section 1.2.2.3. Contrary to most current works, this study does not aim to set new accuracy records but rather to make a conceptual contribution by emphasizing the use of multi-dimensional emotion classifications in future. Building on this, in Section 4, a BERT model will be fine-tuned to classify tweets based on their emotions. Given that a similar model has already been created by Vamossy and Skog (2020), this third study will sharpen the focus in the spirit of Renault (2017). Renault (2017) emphasizes in the dictionary context (and the work in Section 3 confirms this again) that proper evaluation necessitates alignment with the underlying text type. Therefore, the question arises as to how homogeneous texts on social media platforms are. By categorizing traders into the groups 'Novice', 'Intermediate', and 'Professional' linguistic differences will be highlighted, and three BERT models will be fine-tuned and evaluated specifically for these groups, based on the model by Vamossy and Skog (2020).

2 The relevance and influence of social media posts on investment decisions – an experimental approach based on Tweets

2.1 Abstract

We conducted an experiment to examine the role of positive and negative tweets (generated by AI) on investment behavior, comparing them with provided historical and fundamental financials. Through mediator analysis, we discovered that positive tweets have a significantly positive mediating effect on investment amounts, while negative tweets have a negative impact. Importantly, we found that this effect is not primarily driven by the perception of the tweets; rather, positive tweets influence individuals' perception of a company's financials which is the most influencing factor in individuals' investment decision. In this manner our study contributes to the existing literature by (1) proving evidence for a causal effect of social media investor sentiment on investment behavior on capital markets and especially (2) focussing how the influence channels are built.

2.2 Introduction

Predominantly starting with Kyle (1985) and Black (1986) the influence of noise in financial markets has aroused the interest of many researchers in the field of behavioral finance. In financial research the role of noise traders has been widely discussed as noise trading is supposed to explain why stock prices could differ from their fundamental value. This idea contradicts the idea of information-efficient markets stated in the Efficient Market Hypothesis (EMH) by Fama (1970). Fama (1965) himself argues that irrational noise traders would meet rational traders on financial markets who trade against them. This should result in systematic losses for noise traders who will leave the market because of the behavior of rational arbitrageurs. Long et al. (1990) oppose that there are limits to arbitrage due to risk aversion and short time horizons allowing noise traders to temporarily diverge prices from the fundamental value. Consequently, the development, identification (and prediction) of noise has become a main interest of research in financial research.

Market or investor sentiment defined as market's general, psychological environment is believed to wield considerable influence over noise trading, thereby anticipated to impact

stock prices. Given the non-trivial nature of observing investor sentiment, the debate on its influence within financial markets pivots on identifying the most appropriate measure. Over time, three main distinct measurement approaches have emerged: market-based, survey-based, and text-based methodologies.[9]

The approach last mentioned, which has gained and continues to enjoy widespread popularity, aligns with the ascent of social media platforms like Twitter, Facebook, and Instagram. Their expanding user bases, coupled with increasingly accessible textual analysis tools such as Google BERT with nearly 9,000 trained models on *Huggingface.co*, have propelled this approach. Consequently, researchers have probed the potential impact of a platform's content on stock market performance. Given investors' limited attention spans, their investment decisions often exhibit biases toward assets that consciously or subconsciously grab their attention – such as through framing techniques (Barber and Odean (2008)). As a result, social media platforms may indeed sway individual investment choices (Liu (2020)). Johnson and Tversky (1983) previously noted that sentiment has the power to influence investors' risk perceptions. Kaplanski et al. (2015) corroborate this observation, even detecting the effects of investors' personal happiness on their investment behavior. Additionally, Baker and Wurgler (2007) conclude that the debate no longer revolves around whether sentiment influences market participants but rather focuses on the intensity of its impact and how best to measure it.

Despite empirical findings predominantly suggesting relationships, discussions surrounding causality, particularly the causal direction and channels, have surfaced. This area has been experimentally explored across various papers in economic literature. Hales, Kuang, and Venkatarman (2011) contribute to linguistic analysis in financial accounting research (e.g. Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Feldman et al. (2010)) by demonstrating that investors are more susceptible to the influence of vivid language compared to dull language of the same sentiment in financial reporting. This effect is especially pronounced when the underlying information is preference inconsistent. Studies by Tan, Ying Wang, and Zhou (2014) and Rennekamp and Witz (2021) echo these findings, suggesting that text can significantly impact investors' judgments, particularly when the readability of the text is low or when the

[9] A good overview about the three measurements is given for example in Aggarwal (2022).

language used is informal. Moreover, Miller (2010) finds that lengthy and less readable filings lead to reduced trading, prompting small investors to halt trading activities. The chosen information channel also plays a role. Kelton and Pennington (2020) note that investors tend to identify more with a CEO when communication occurs through Twitter compared to the company's website. A recent and comparable study by Boulu-Reshef et al. (2023) specifically examines the influence of emojis in social media posts (tweets) on financial professionals. Their research indicates a significant yet marginal impact of these messages on investment decisions.

Despite the specific experimental findings, there remains a limited understanding of the intricate mechanisms underlying these effects. A deeper examination of the influential channels could significantly enhance our comprehension of individuals' investment behavior. Thus, we aim to contribute to the aforementioned literature by investigating individuals' investment choices and their perceptions of financial and social media sentiment within an experimental setting encompassing various financial and social media information sources.

Through the application of mediation analysis, our study seeks to scrutinize whether and through which channels these distinct information sources exert an influence on perceived sentiment. Subsequently, we aim to explore how these perceptions, in turn, impact investment decisions. We go in line with prior findings, but also find using mediator analysis that the tweets do not have significant influence on investment decision directly as well as over the mediator perceived tweet sentiment. Moreover, the tweets influence the perceived financial sentiment which has a large and significant influence on the investment decision.

The remainder of this paper is structured as follows: Section 2.3 provides a detailed description of the methods utilized to gather financial and social media data within the experimental framework, aiming for authenticity. It further delves into the implementation process, concluding with the formulation of hypotheses based on the established setting in Section 2.4. Section 2.5 offers a concise overview of the collected data, leading into the presentation of our findings. This includes a mediation analysis elucidating the impact on investment decisions. Finally, Section 2.6 serves as the

conclusion, where we summarize our observations in light of previous literature, and highlight potential avenues for future research.

2.3 Experimental design

Our experimental design aims to assess the impact of social media posts, specifically tweets on the platform 'X' (formerly 'Twitter'), on the investment behavior of individuals. Taking into consideration aspects of loss aversion following prospect theory by Kahneman and Tversky (1979), we are also interested in observing this behavior with positive and negative versions of provided financials and tweets. To achieve this, we divided our test subjects into six different groups, as outlined in Table 1.

Group	Tag	Financials	Twitter
1	PP	Positive	Positive
2	PN	Positive	Negative
3	NP	Negative	Positive
4	NN	Negative	Negative
5	P	Positive	*none*
6	N	Negative	*none*

Table 1: Grouping

In the following subchapters, we describe the specified investment setting along with the design of positive and negative financials and tweets. We conclude our introduction to the experimental design by detailing the incentive system. Subsequently, we derive our hypotheses based on our key findings in the introduction and our experimental design.

2.3.1 Investment setting

Test participants were instructed to gather information about the fictional company 'Glubon AG'[10] of which they already owned 100 stocks, each valued at 10€ (resulting in a total stock capital of 1000€). Based on a brief company description (refer to Figure 10 in Section 2.7.1.1), stock charts, financial metrics (see Section 2.3.2) and (for groups 1 to 4) posts on the platform Twitter[11] ('Tweets', see Section 2.3.3), participants had to decide whether to sell or buy stocks at a rate of 10€ each. Each participant also possessed 1000€

[10] AG is the German abbreviation for 'Aktiengesellschaft,' which translates to 'stock company'.
[11] Before the conclusion of our experiment, 'Twitter' had unexpectedly been rebranded to 'X'. We chose to keep using the name Twitter, as most participants might not be familiar with the new branding and the name 'Twitter' has been used to provide information to the participants.

of free capital, and the decision was limited to holding between zero stocks and 2000€ of free capital or holding 200 stocks and 0€ of free capital at the end of the experiment. After all participants made their decisions, a new stock price per group would be calculated, as explained in Section 2.3.4. This calculation also affected the total capital (and consequently, the number of lottery tickets) of the participants. Therefore, the experimental setting is limited to one period and each participant makes only one decision.

All information was presented on a self-designed, Visual Basic-based information and trading platform, exemplified by the opened (negative) *Financials* tab in Figure 4. On this platform, our participants could freely navigate between three tabs: *Company description*, *Financials*, and *Social Media*, to gather information for the final decision in the investment decision tab. Thanks to the autonomous coding of the platform, we were also able to track all transitions between tabs and monitor the time spent within each tab.

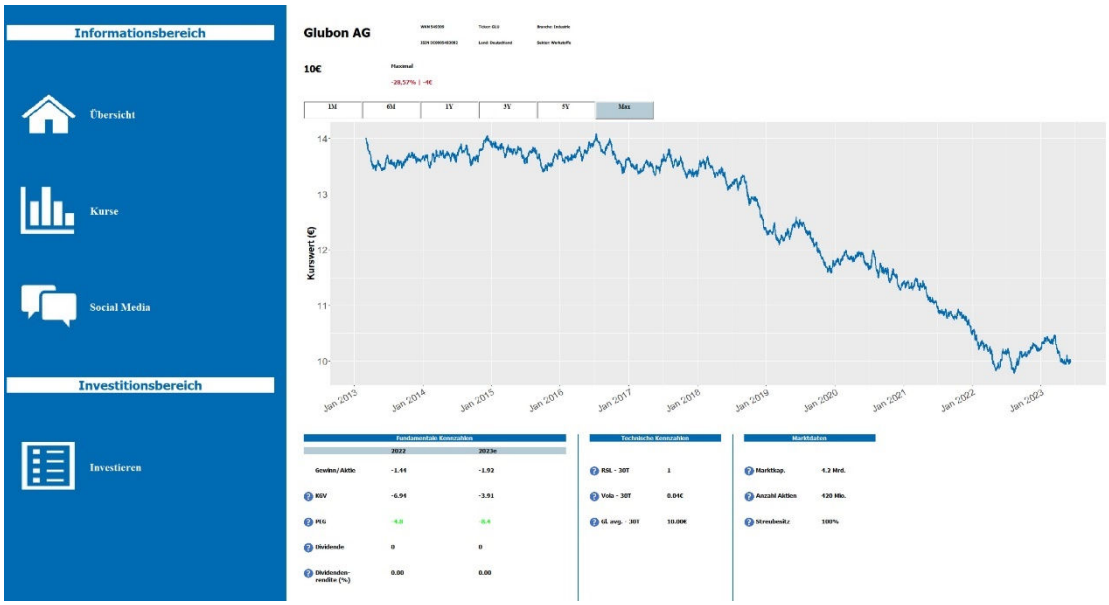


Figure 4: Platform's interface (*Financials* tab opened, negative version)

2.3.2 Financials

The structure of the financials tab is modeled after financial websites such as Yahoo! Finance, presenting charts for different time horizons along with financial figures. The positive and negative cases can be found in Figure 11 and Figure 12 in the appendix 2.7.1.3.

The stock price development was simulated using a random walk with drift, as described in formulas (3) and (4). To enhance the authenticity of the development, a new drift α_i was drawn from a normal distribution with a positive mean for the positive case every 30 days, as detailed in formula (5).

$$P_t = P_{t-1} + \alpha_i + \varepsilon_t \tag{3}$$

with

$$\varepsilon_t \sim N(0,1) \tag{4}$$

and an every 30 days t changing α_i

$$\alpha_i \sim N(1,25) \tag{5}$$

For the negative case, daily returns were reversed, and both stock price developments were scaled to a price of 10€ on the last day.

Additionally, participants could find financial figures below the charts, designed to appeal to economically educated participants who assumed the market, following Fama (1970), to be semi information-efficient. Even less economically educated participants could benefit from this information, as each figure was explained by clicking the '?' buttons next to the figure. The provided positive (negative) financial figures included positive (negative) profits per share, positive (no) dividends/dividend returns, positive (negative) price-earnings ratios for the previous year as well as expected for the current year. Furthermore, figures for low (high) volatilities, relative strength, 30 days moving average, as well as information about the market capitalization, free float, and number of shares, were presented.

Consequently, we are aware of possible biases in the perception of the financials of Glubon as 'positive' and 'negative', especially for the charts, due to prior findings in behavioral finance (in this case, especially the disposition effect empirically introduced by Shefrin and Statman (1985)). Therefore, we ask the participants about their perception as well as their judgment regarding plausibility and trustworthiness of the given financials after the investment decision.

2.3.3 Tweets

Tweets were presented as the result of a search for the cashtag '$GLU' of the imaginary Glubon AG on the platform Twitter. The content of the tweets was generated using OpenAI's ChatGPT queries mentioned in Table 9 in the appendix 2.7.1.2. Due to different queries, positive, negative, and neutral tweets were created by the AI using varying maximum lengths (20, 70, or 140 characters) as well as in colloquial and non-colloquial language. From the created database of 180 tweets, we sampled 40 tweets each for groups 1 & 3 and groups 2 & 4, as stated in Table 2. The tweets provided on the platform for group 1 & 3 not only contain positive tweets but also a minor number of neutral and negative tweets for authenticity reasons. The same holds true vice versa for the tweets provided to group 2 & 4. To ensure that this does not affect the treatment, participants were asked for their perception of the tweets after the investment decision.

Query specification			**Occurences per group**		
sentiment	colloquial	max character	1 & 3	2 & 4	5 & 6
Positive		20	5		0
Positive		70	5		0
Positive		140	5	randomly	0
Positive	X	20	5	picked 3	0
Positive	X	70	5		0
Positive	X	140	5		0
Neutral		20			0
Neutral		70			0
Neutral		140	randomly	randomly	0
Neutral	X	20	picked 7	picked 7	0
Neutral	X	70			0
Neutral	X	140			0
Negative		20		5	0
Negative		70		5	0
Negative		140	randomly	5	0
Negative	X	20	picked 3	5	0
Negative	X	70		5	0
Negative	X	140		5	0
		$\sum$	40	40	0

Table 2: Queries and presence of tweet type per group

To enhance authenticity further, we added ChatGPT-generated German usernames as well as randomly picked profile pictures from the academic dataset delivered by the

company ‘Generated Photos’. The picture dataset, including estimators for gender, race, and the emotion shown in the picture, allowed us to pick a diverse spectrum of mostly happy profile pictures. While we randomly ordered the sampled tweets per group, the order of profile names and pictures is the same in every group. Ultimately, replies, retweets, likes and impressions were drawn from a normal distribution with a higher mean if the tweet sentiment fits the group’s social media treatment than for tweets of another sentiment as those factors can also influence investors’ perception following Cade (2018) or Rennekamp and Witz (2021). All these operations lead to a social media tab as exemplified in Figure 5.[12]

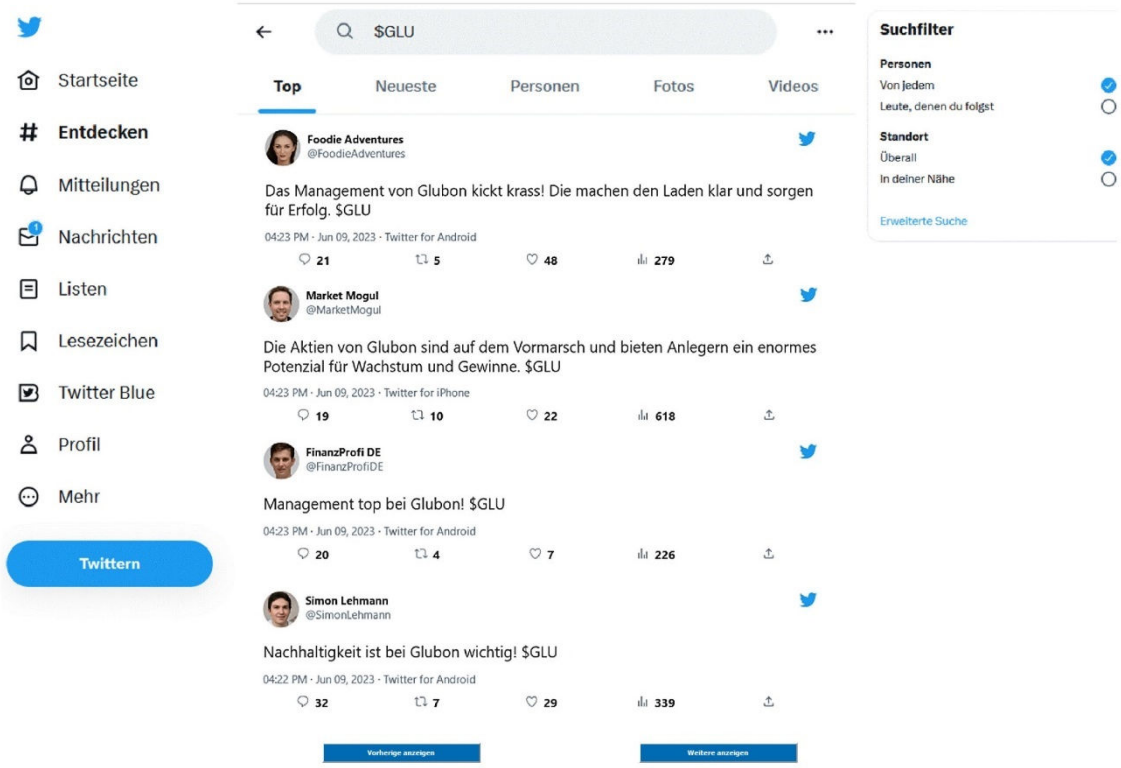


Figure 5: *Social media* tab, site 1 of 10 opened, positive version

Consequently, this operationalization does not mimic a potential ‘timeline’ of the users and can be more accurately compared to a search for the company’s cashtag ($) in the Twitter feed. We assume that potential effects reported in Section 2.5 would be more pronounced if tweets had been posted by users our test participants would have decided

[12] A translated example for a tweet of every query type mentioned in Table 2 can be found in Table 9 in the appendix 2.7.1.2.

to follow in real life, which would not have been possible to mimic reliably in an experiment. Additionally, the AI-generated content could possibly be recognized by the users. Therefore, we asked the participants for their assessment of the trustworthiness of the tweets.

2.3.4 Implementation

The experiment took place in a lab at the Heinrich-Heine-University Düsseldorf in July and August 2023 with an open registration for everyone speaking German fluently. Over time we collected data from 300 participants mainly containing economic students but also professionals and students from other disciplines. From the 300 participants we use 259 responses for our dataset excluding 41 participants who failed at answering at least 3 of 4 control questions regarding the given setting and incentive system correctly.

In addition to fixed compensation, participants were incentivized by a lottery which ensures conscientious behavior by the participants (Holt and Laury (2002)). Each participant started the experiment with a total capital of 2000€ (1000€ stock capital, 1000€ free capital), which translated into 2000 tickets for the lottery (1€ equals 1 ticket). Depending on the decisions made within each reference group, a new stock price was calculated, affecting the stock capital and total capital of each participant based on their decision. Figure 6 illustrates how the decision to buy or sell 50 stocks affects the total capital, and consequently, the number of lottery tickets, if the stock price increases to 15€ (blue situation) or decreases to 5€ (green situation).

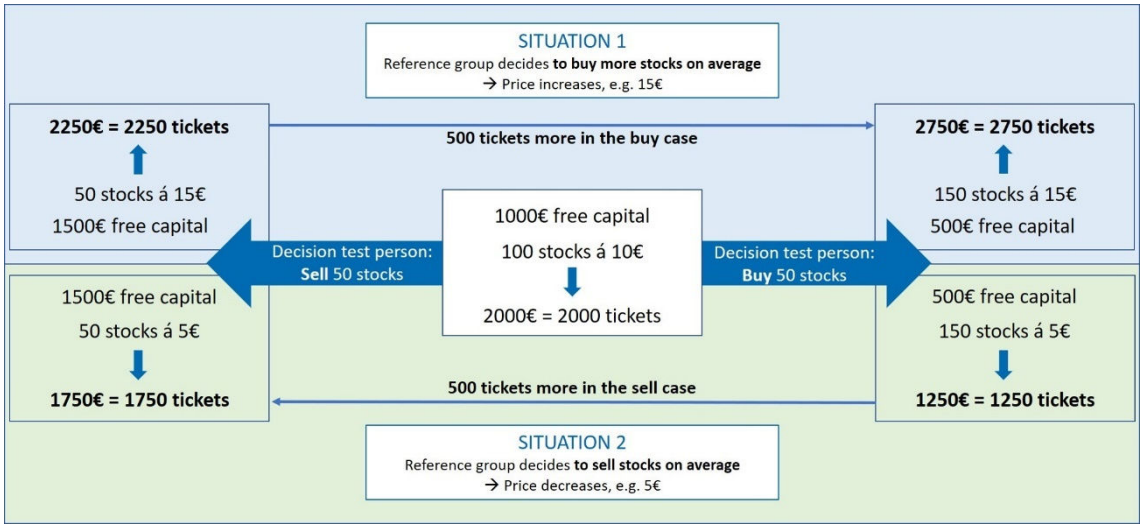


Figure 6: Ticket outcomes under different situations and decisions

For the calculation of the new stock price, P_1, in each group i with N_i participants, we use a simplified stock pricing formula that interprets the return of the stock, r_i, as the ratio between the change in cumulated stock capital in t_1, $SC_{i,1}$, and the cumulated stock capital in t_0, $SC_{i,0}$:

$$r_i = \frac{SC_{i,1} - SC_{i,0}}{SC_{i,0}} \tag{6}$$

Consequently, the new price per group i ($P_{i,1}$) is calculated as

$$P_{i,1} = P_0 * (1 + r_i) \tag{7}$$

which is limited between

$$\lim_{SC_{i,1} \to 0} P_1 = 0 \tag{8}$$

and

$$\lim_{SC_{i,1} \to 2000*N_i} P_1 = 20 \tag{9}$$

Further, we collected variables for controlling purposes regarding participants' demographics (as gender, age, income & risk tolerance following Holt and Laury (2002)), financial experience and social media usage.

2.4 Hypotheses

In the context of the Efficient Market Hypothesis (Fama (1970)), it can be assumed that economic agents process information provided to them appropriately, thereby adjusting their actions to the existing information environment. As indicated by the relevant literature and various economic studies, both social media (see i.a. Antweiler and Frank (2004), Baker and Wurgler (2006), Da, Engelberg, and Gao (2015), Das and Chen (2007), Renault (2017), Sun, Najand, and Shen (2016), Tetlock (2007)) and financial indicators influence the investment calculus of individuals.

However, Tversky and Kahneman (1974) in their highly regarded study considered the starting point of Behavioral Finance, demonstrated that due to behavioral biases, the

available information is inadequately processed using experience and heuristics (Ritter (2003)). In this context, differences may arise in the consideration of various information sources and their interpretation leading to departures from rational decision-making calculations, as exemplified by phenomena such as noise trading. Thus, it can be assumed that different economic agents may consider various information sources differently based on their experiences and perceptions.

In our specific case, economic agents have access to social media posts in the form of tweets and financials (historical and fundamental) for their investment decisions. The goal of this study is to examine whether the provided information has an impact on individuals' investment decisions. However, in the context of the presented behavioral biases, it is also necessary to investigate how the tweets and financial information were perceived by each participant (sentiment) and whether this sentiment also influences the investment decision. To address this question, a mediation analysis will be employed, aiming to answer the following main hypotheses:

H1: *There is a mediating effect of Financial Sentiment on the investment decisions of individuals.*

H2: *There is a mediating effect of Tweet Sentiment on the investment decisions of individuals.*

In our analysis, we draw insights from Baron and Kenny (1986) and Zhao, Lynch, and Chen (2010) to elucidate the intricate mechanism by which provided information and the associated sentiment shape investment decisions. Our approach involves examining both the direct impact of tweets and financials on investment decisions and their indirect effects mediated by two factors: *Tweet Sentiment and Financial Sentiment*. Furthermore, we also examine the influence of tweets on Financial Sentiment and the influence of financials on Tweet Sentiment to account for a potential deviation from rational decision-making in the context of Behavioral Finance.

Hence, the following sub-hypotheses arise:

H1.1: *There is an indirect effect of tweets via the mediator Tweet Sentiment on the investment decisions of individuals.*

H1.2: *There is an indirect effect of tweets via the mediator Financial Sentiment on the investment decisions of individuals.*

H1.3: *There is a direct effect of tweets on the investment decisions of individuals.*

H2.1: *There is an indirect effect of financials via the mediator Financial Sentiment on the investment decisions of individuals.*

H2.2: *There is an indirect effect of financials via the mediator Tweet Sentiment on the investment decisions of individuals.*

H2.3: *There is a direct effect of financials on the investment decisions of individuals.*

2.5 Results

Before proceeding with the analysis of the data from the conducted experiment in the next chapter, we will first delve into the collected information of the participants. To do this, the data is divided into three categories, with the last category further subdivided into three more categories. All information discussed below can be found in Table 3.

The 'Participants' behavior' category encompasses the 'Stocks held' by participants at the end of the experiment, thus reflecting their investment decision. By definition, the values in this category can only be integers in the interval $[0, 200]$, where 0 represents the sale of all initially (100) held stocks, and 200 represents the maximum purchase of 100 additional stocks within the available budget. This interval was utilized, as evident from the maximum and minimum values, with participants acquiring, on median, an additional 10 stocks, while, on average, only 1.6 additional stocks were acquired by a standard deviation of 61.73 stocks.

The second category, 'participants' sentiment', includes the sentiment of the participants regarding the given tweets and financials. After making their investment decisions, participants were tasked with using a Likert scale ranging from 1 to 5 to assess how they perceived the given tweets and financials.

	Min	Max	Med	Mean	Sd	Dummy
Participants' behavior						
Stocks held	0	200	110	101.60	61.73	
Participants' sentiment						
Tweets	1	5	2	2.63	1.59	
Financial	1	5	3	3.01	1.43	
Participants' characteristics						
Demographic						
Age	17	62	23	24.93	7.25	
Male	0	1	1	0.60	0.49	X
Risk	0	10	5	4.86	1.79	
Income	0	10	1	1.77	2.00	
Financial						
Economic	0	1	1	0.61	0.49	X
Cap market	0	1	1	0.60	0.49	X
Social Media						
Usage	0	14	2	2.48	1.74	
Twitter	0	1	0	0.26	0.44	X

Table 3: Participants' information

In this context, a value of 1 corresponds to a very negative sentiment, 3 to a neutral one, and 5 to a very positive sentiment. These pieces of information serve in the further development of the work both to validate whether the given treatment was perceived by the participants according to its intention and to highlight whether perception, rather than the actual information, has an impact on investment decisions. The entire possible interval of [1,5] was also utilized by the participants for both Social Media and Financial Sentiment, with the Social Media Sentiment being more negative on both average and median compared to the Financial Sentiment.

The last category, 'Participants' characteristics', includes characteristics of the participants regarding their demographic information, financial experience, and social media usage. The category of 'Demographics' includes the age, gender, risk attitude and income of the participating individuals. The youngest participant was 17 years old, and the oldest person was 62 years old. Based on the median (23) and the average age (24.93), it can be observed that, as expected, it is a relatively young participant group since this study was conducted at an university.

The variable 'Male' is a dummy variable, which takes the value 1 for participants who identify as male. To account for the three different gender specifications of the participants and considering that only one observation is labeled as gender-diverse, a dummy variable is used. As indicated by the median and the mean, there is a slight majority of male participants in the present dataset.

The 'Risk' variable measures the risk tolerance of each participant with values ranging from [0,10], which was determined using the Holt-Laury test.[13] A value of 0 indicates a high-risk appetite, while a value of 10 reflects a pronounced risk aversion. In the present dataset, the majority of participants are therefore more risk-averse.

Furthermore, participants were asked about their income, which could be indicated in increments of 500. Thus, the number 0 represents an income of 0-500€, and the number 10 (the maximum in this dataset) represents an income of more than 5000€. Hence, we observe a relatively low-income level of 1.77 on average, which again, is to be expected since the experiment was conducted at an university.

Aside from demographic information, additional data was collected on participants' financial background and social media usage to consider their effects in the further analysis. In terms of economic characteristics, there is a dummy variable indicating whether a participant has an economic-related background in form of an university degree or an apprenticeship. The variable 'Cap market' indicates whether a participant has been active in a capital market. In terms of social media characteristics, the dummy variable 'Twitter' differentiates whether a participant uses or has used the social media platform Twitter, as this study focuses primarily on this platform for social media posts. Additionally, the variable 'Usage' indicates how many hours per day a participant uses social media channels.

Overall, the majority of participants have been active in the capital market and are currently or have previously pursued a study with an economic background. However, most participants do not use the social media platform Twitter. Furthermore, participants

[13] Holt and Laury measure individuals' risk aversion by presenting two lotteries. Participants are asked to choose between a less risky and a riskier but potentially more profitable lottery in 10 different scenarios, with the probability of the higher payoff increasing in each iteration. The degree of risk aversion is determined by the switching point from the less risky to the riskier lottery, with the rational switch based on expected value occurring after the fourth iteration. Therefore, values above 4 indicate increased risk aversion. For a more detailed overview, see Holt and Laury (2002).

spend an average of 2.48 hours (2 hours in median) per day on social media channels. However, it is important to note that one participant with a daily usage of 14 hours is a clear outlier, which needs to be critically considered in the subsequent ANOVA analysis.

The collection of the data described above allows, on one hand, drawing conclusions about the characteristics of the participating individuals to assess the generalizability of the results of the present study. On the other hand, these variables serve as control variables in a later chapter to check the robustness of the results.

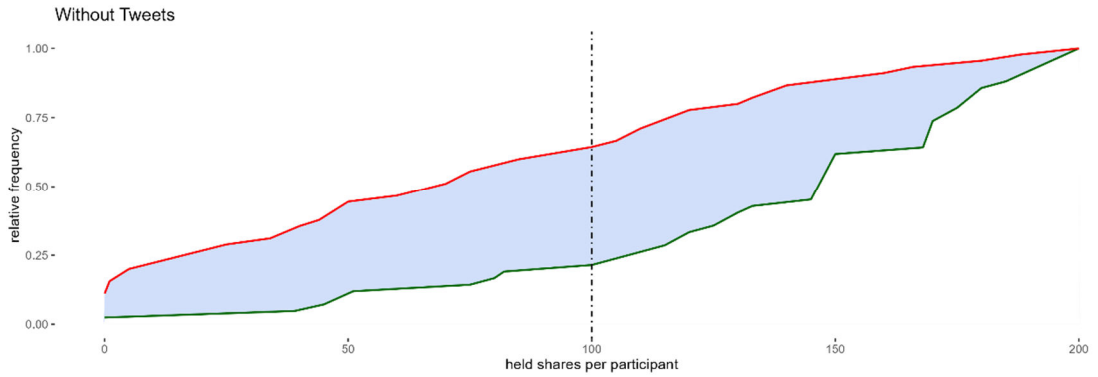


Figure 7: Cumulative relative frequency of Stocks held (without tweets)

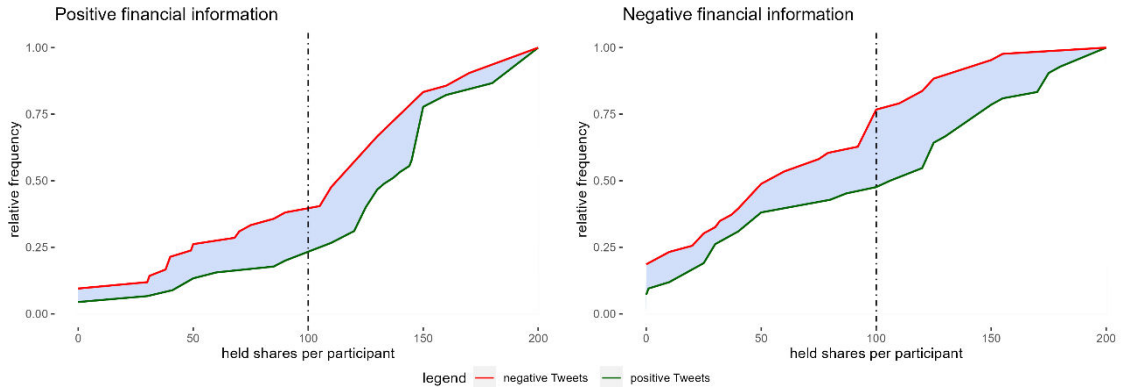


Figure 8: Cumulative relative frequency of Stocks held (with tweets)

After examining participants' behavior, sentiment, and characteristics, the next step is to take a closer look at these factors for each group. Since this study aims to contribute to the explanation of individuals' investment behavior, Figure 7 and Figure 8 are used to

provide an overview of the differences in investment behavior between the individual groups.[14]

Firstly, the cumulative relative frequency of stocks held for the groups without tweets is examined (Figure 7). The two groups only differ in the provided financials. It can be seen that the group with positive financials (P), represented in green, holds more stocks throughout the entire distribution compared to the comparison group with negative financials (N). Looking at the density distribution of the other groups (Figure 8), which were provided with tweets, a similar pattern emerges. The compared groups always differ in the provided tweets, while the financials do not differ in the individual comparisons. It becomes clear that both in the case of positive and negative financials, there is a difference in the held stocks. In both cases, participants who were provided with positive tweets (PP, NP) hold more stocks throughout the entire distribution compared to the groups with negatively connotated tweets (PN, NN).

2.5.1 Analysis of Variance & Post-hoc Test

Based on these observations, an Analysis of Variance (ANOVA) will be conducted subsequently to examine whether the held stocks differ significantly among the individual groups. In addition to differences in participant behavior, an examination will also be conducted to determine whether there are differences in participants' sentiment and characteristics among the individual groups. The ANOVA results and the means for every aspect analyzed are depicted in Table 4.

The F-statistic of the ANOVA clearly indicates that there are significant differences between individual groups regarding the average number of stocks held at the end of the experiment. On average, groups with positive financials hold more stocks than those with given negative financials. In particular, the control group with positive indicators without social media posts (P) holds the most stocks on average. Furthermore, a difference can be observed between the groups with positive financials and positive or negative social media treatment (PN & PP). Participants in the group with positive social media posts (PP) hold, on average, about 23 more stocks compared to participants with negative posts (PN), which might hint towards an influence of the given social media treatment. A similar pattern emerges when examining the groups with negative financials and different

[14] For an overview of the different groups see Table 1.

social media treatments (NN & NP). Participants in the group with positive social media posts (NP) hold, on average, about 30 more stocks than the comparison group with negative posts (NN). The NN group also holds the lowest number of stocks on average, even when compared to the control group with negative financials and no social media posts (N).

	PP	**PN**	**NP**	**NN**	**P**	**N**	**F-Stat**
			Participants' behavior				
Stocks held	128.77	105.97	97.35	66.86	136.52	75.15	10.53***
			Participants' sentiment				
Tweets	3.91	1.38	3.80	1.37	NA	NA	127.99***
Financial	4.15	4.04	2.23	1.72	4.23	1.68	93.54***
			Participants' characteristics				
Demographic							
Age	26.68	25.33	23.28	25.79	23.71	24.62	1.37
Male	0.511	0.38	0.42	0.34	0.35	0.4	0.63
Risk	5.17	4.61	4.73	4.88	4.61	5.06	0.75
Income	2.07	1.93	1.64	1.91	1.36	1.69	1.72
Financial							
Economic	0.60	0.62	0.62	0.49	0.79	0.53	0.00
Cap market	0.53	0.57	0.69	0.58	0.67	0.53	0.07
Social Media							
Usage	2.29	2.21	3.28	2.24	2.57	2.32	0.00
Twitter	0.20	0.21	0.33	0.33	0.26	0.20	0.06

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: ANOVA between the different groups

Moreover, the results of the ANOVA regarding participants' perceptions reveal that the given treatments (social media posts) were perceived by the participants in accordance with their intended sentiment. There are significant differences in the perception of the sentiment of social media posts among the individual groups, as measured on a Likert scale. The groups with positive tweets (PP & NP) perceive these posts significantly more

positively on average (deviation of approximately 2.5 units) compared to the groups with negatively formulated tweets. A different perception also exists regarding the financials. The groups with positive financials (PP, PN & P) perceive them on average significantly more positively than the groups with given negative financials (NP, NN & N). These results suggest that the treatments were perceived according to their intended purpose.

Finally, ANOVA was used to compare participants' characteristics across the individual groups (for characteristics where such a method is meaningful). The results indicate that there are no significant differences in terms of the participants' characteristics, suggesting a balanced distribution of participants.[15]

Group comparison	**Stocks held**		**Tweet Sentiment**		**Financial Sentiment**	
	diff	p adj.	diff	p adj.	diff	p adj.
PN - PP	-22.802	0.420	-2.530	0.000	-0.108	0.992
NP - PP	-31.420	0.105	-0.102	0.991	-1.917	0.000
NN - PP	-61.917	0.000	-2.539	0.000	-2.435	0.000
P - PP	7.746	0.988			0.083	0.998
N - PP	-53.622	0.000			-2.467	0.000
NP - PN	-8.619	0.982	2.429	0.000	-1.810	0.000
NN - PN	-39.116	0.020	-0.009	1.000	-2.327	0.000
P - PN	30.548	0.137			0.190	0.911
N - PN	-30.821	0.118			-2.359	0.000
NN - NP	-30.497	0.134	-2.437	0.000	-0.517	0.063
P - NP	39.167	0.021			2.000	0.000
N - NP	-22.202	0.451			-0.549	0.036
P - NN	69.663	0.000			2.5171	0.000
N - NN	8.295	0.983			-0.032	1.000
N - P	-61.368	0.000			-2.549	0.000

Table 5: Post-hoc test between each group

Although the results of the ANOVA indicate significant differences between the means of the six groups in terms of participant behavior and perception, such an analysis does

[15] As previously noted, there is an outlier with 14 hours of social media usage. The effect of this outlier is evident in the elevated mean of social media usage for the NP group. However, in this context, this outlier should not pose a problem, as even when considering this outlier, there is no significant difference between the individual groups. Moreover, if the outlier were to be excluded, the average of this group should align even more closely with the lower average of the other groups.

not provide insight into the specific nature of these differences. Therefore, a post-hoc test, specifically the Tukey post-hoc test, is employed (Tukey (1992)). This test allows for detailed comparisons between each group and the others, enabling a pairwise comparison across all groups. The results of the post-hoc test can be found in Table 5.

The group-wise comparison of participant behavior (stocks held) reveals that groups with opposing financials significantly differ in their purchasing behavior ($NN - PP$, $N - PP$, $NN - PN$, $P - NP$, $P - NN$, $N - P$), with groups having negative financials, as expected, holding fewer stocks. Furthermore, the results from the preceding ANOVA analysis is confirmed in the sense that the treatments of sentiment and financials were perceived by the participants according to their intended purpose. Thus, the groups with divergent sentiment in social media posts consistently differ statistically highly significantly in their perception of tweets.

The same applies to the treatment of financials. The metrics are perceived as intended by the authors. However, two group comparisons stand out. Although groups NP, NN, N were each provided with the same financial information, these pieces of information were perceived statistically significantly differently. In the $N - NP$ comparison, this difference is significant at a 5% level, and in the $NN - NP$ group comparison, it is still significant at a 10% level.

Since the respective groups all received the same financial information, they differ only in the sentiment of the provided social media posts. In both group comparisons ($NN - NP$ and $N - NP$), participants received positively connotated tweets. Thus, it can be presumed that the sentiment, especially if the tweets contain positive sentiment, of the given tweets has an influence on individuals' perception of financial information, which in turn might influence an individuals' investment decision. To test this hypothesis, a statistical analysis using a mediation analysis will be conducted subsequently.

2.5.2 Mediation Analysis

Mediation analysis (Baron and Kenny (1986)) is used to measure the effect of (an) independent variable(s) on a dependent variable. For this purpose, both the direct influence of the independent variable(s) on the dependent variable and the indirect effect of the independent variable through a mediator are estimated.

In the present analysis, due to the identified group differences, there is reason to believe that the provided tweets and financials have a direct impact on the investment decisions of the participants (*H1.3* & *H2.3*). Thus, these variables are chosen as independent variables to assess their direct influence on the investment decision made. Furthermore, the results of the preceding chapter provide grounds to assume that the actual manifestations of tweets and financials influence how these variations are perceived by the participants, and in turn, this sentiment has an impact on the investment decision (*H1.1* & *H2.1*). First evidence that tweets (financials) can also frame perceived Financial (Tweet) sentiment (*H1.2* & *H2.2*) can be seen in Table 5 as the perceived sentiment of financials was significantly more positive when the tweets were of a positive nature. Hence, through the mediation analysis, the model illustrated in Figure 9 is estimated.

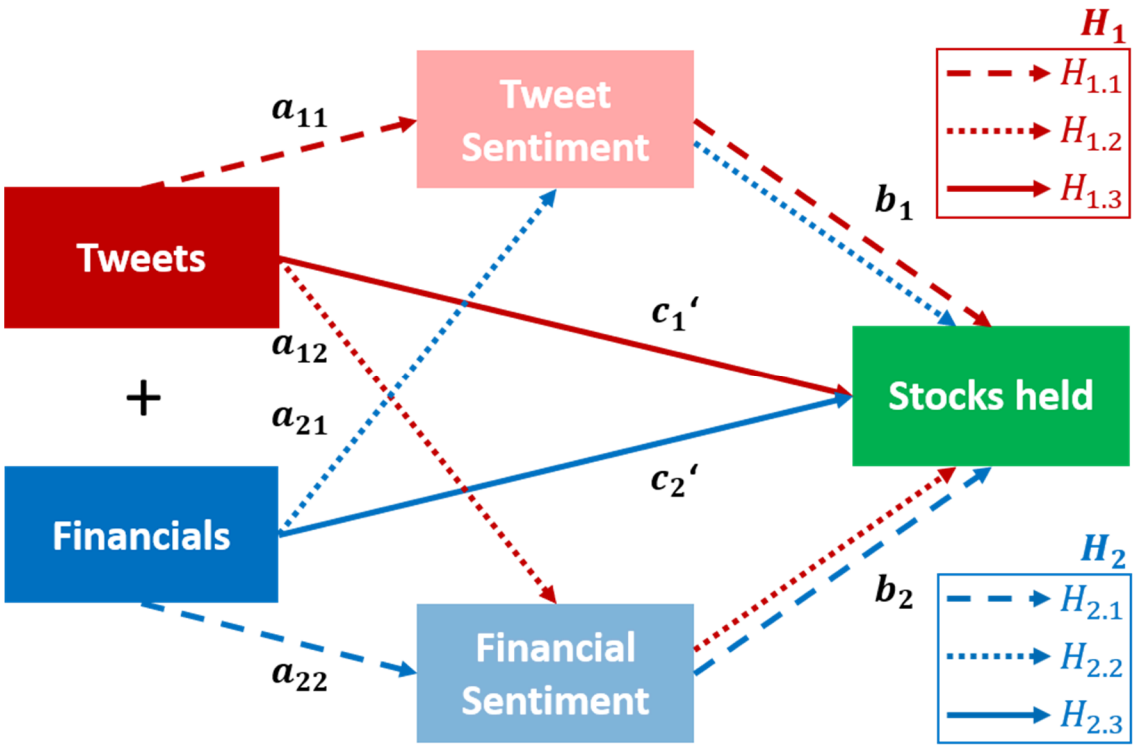


Figure 9: Mediation analysis

This model uses the provided tweets and financials as dependent variables and the perception of their sentiment as mediators to explain the stocks held by the participants and test our hypotheses. In the presented base model (A) of a two-mediator model, a total of 4 different regressions need to be estimated to determine the direct and indirect effects of each regressor and takes the following form:

$$Stocks\_held = i_1 + c_1 * Tweets + c_2 * Financials + \varepsilon_1 \quad (10)$$

$$Stocks\_held = i_2 + c'_1 * Tweets + c'_2 * Financials + b_1 * Tweet\_Sentiment + b_2 * Financial\_Sentiment + \varepsilon_2 \quad (11)$$

$$Tweet\_Sentiment = i_3 + a_{11} * Tweets + a_{21} * Financials + \varepsilon_3 \quad (12)$$

$$Financial\_Sentiment = i_4 + a_{12} * Tweets + a_{22} * Financials + \varepsilon_4 \quad (13)$$

To check the robustness of the results of this base model, additional control variables are subsequently added to the estimation. Model (B) includes the demographic information about the participants already presented earlier. In contrast, model (C) has been expanded to include financial and social media characteristics, while model (D) contains both demographic information and financial and social media characteristics.

Please be aware that for assessing the influence of Tweet Sentiment, it is imperative to exclusively consider the groups provided with tweets, given that participants in groups *P* and N were not exposed to any tweets, thus rendering them incapable of developing any Tweet Sentiment. Consequently, the models are estimated with $N = 172$ observations. The results of these estimation models can be found in Table 6.

The results of model (A) show that the given tweets do not have a direct impact on stocks held. However, as expected, the given tweets have a strong and highly significant influence (a_{11}) on the first mediator, the Tweet Sentiment (T_Sen). However, this mediator does not have a statistically significant impact (b_1) on stocks held either, so in this case, we can neither assume a mediating or direct effect, contradicting *H1.1* & *H1.3*. This is also confirmed by the statistically insignificant indirect effect $a_{11} * b_1$. The given financials do not have a statistically significant direct influence on stocks held, which rejects *H2.3*.

	Effect type		(A)	(B)	(C)	(D)
Stocks held	**Direct**					
	Tweets	(c_1')	0.108	0.108	0.087	0.087
			(0.115)	(0.113)	(0.116)	(0.114)
	Financials	(c_2')	-0.142	-0.153	-0.153	-0.157
			(0.106)	(0.107)	(0.099)	(0.100)
	T_Sen	(b_1)	0.059	0.076	0.086	0.096
			(0.115)	(0.111)	(0.113)	(0.110)
	F_Sen	(b_2)	0.560***	0.583***	0.575***	0.590***
			(0.115)	(0.114)	(0.111)	(0.111)
	Age			0.066		0.053
				(0.081)		(0.080)
	Male			0.098		0.046
				(0.064)		(0.067)
	Income			-0.105		-0.120
				(0.089)		(0.084)
	Risk			-0.057		-0.054
				(0.059)		(0.060)
	Economic				-0.045	-0.047
					(0.063)	(0.063)
	Cap Market				0.146*	0.150*
					(0.064)	(0.067)
	Usage				-0.026	-0.035
					(0.075)	(0.080)
	Twitter				-0.070	-0.051
					(0.065)	(0.066)
	Indirect					
	$a_{11} \rightarrow b_1$	($a_{11} * b_1$)	0.046	0.059	0.068	0.075
			(0.090)	(0.087)	(0.089)	(0.086)
	$a_{21} \rightarrow b_1$	($a_{21} * b_1$)	0.001	0.001	0.002	0.002
			(0.004)	(0.004)	(0.005)	(0.005)
	$a_{12} \rightarrow b_2$	($a_{12} * b_2$)	0.063*	0.065*	0.065*	0.066*
			(0.027)	(0.028)	(0.028)	(0.028)
	$a_{22} \rightarrow b_2$	($a_{22} * b_2$)	0.429***	0.447***	0.441***	0.452***
			(0.096)	(0.095)	(0.093)	(0.093)
T_Sen	**Direct**					
	Tweets	(a_{11})	0.784***	0.784***	0.784***	0.784***
			(0.047)	(0.047)	(0.047)	(0.047)
	Financials	(a_{21})	0.018	0.018	0.018	0.018
			(0.048)	(0.048)	(0.048)	(0.048)
F_Sen	**Direct**					
	Tweets	(a_{12})	0.112*	0.112*	0.112*	0.112*
			(0.048)	(0.048)	(0.048)	(0.048)
	Financials	(a_{22})	0.767***	0.767***	0.767***	0.767***
			(0.048)	(0.048)	(0.048)	(0.048)
R^2	Stocks held		0.261	0.285	0.293	0.311
	T_Sen		0.615	0.615	0.615	0.615
	F_Sen		0.605	0.605	0.605	0.605

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Regressions estimate equations (10) to (13) for the models (A) to (D) with gradual inclusion of controls and the respective sample size $N = 172$ for each model.

Table 6: Results Mediation Analysis models (A)-(D)

Although the financials' direct effect does not exert a statistically significant influence (c_2'), there is an indirect impact of the financials on Stocks held through the mediator Financial Sentiment. Stocks held are primarily influenced by the Financial Sentiment and therefore by the perception of the nature of the financial information provided. This indirect effect ($a_{22} * b_2$) is statistically highly significant and substantial, thereby confirming *H2.1*. In this case, we can speak of full mediation (Baron and Kenny (1986), Zhao, Lynch, and Chen (2010)).

As suspected from the results of the previous chapter, the mediator Financial Sentiment is also influenced by the tweets at a 5% significance level (a_{12}). Thus, Financial Sentiment serves as a mediator for both the financials and tweets to explain Stocks held.

The indirect effect of tweets on Stocks held through Financial Sentiment ($a_{12} * b_2$) is relatively smaller than the indirect effect $a_{22} * b_2$ – however, it is significant and thus provides a first explanation for the group differences with the same financials ($NN - NP$, $N - NP$) from Table 5 and confirms *H1.2*. However, *H2.2* must be rejected, as financials do not exert a significant influence on the perception of tweets.

These effects remain significant even with the gradual inclusion of control variables concerning the participants' demographics, their financial background and social media usage (models (B) to (D)). The direct effect of tweets on Stocks does not exert a significant on Stocks held in any model leading to the continued rejection of hypothesis *H1.3*. The strength of the significant direct and indirect effects on stocks held ($a_{12} * b_2$ and $a_{22} * b_2$) in model (A) is slightly increased in models (B) to (D), while most control variables do not exert a significant influence on Stocks held. When all control variables are included in model (D), only the previous experience in capital markets at a 5% significance level has an impact on the stocks held. In case of existing experience in capital markets more stocks are held by participants. According to the respective R^2 values for the two mediators, the presented models explain above 60% of the total variance of the perceived Financial Sentiment and the perceived Tweet Sentiment. Also, the investment decision of held stocks can be explained with an R^2 of over 30%.

The measured effect $a_{12} * b_2$ provides an explanation for the group differences in stocks held, as depicted in Figure 9, when the financials are the same. However, especially Table 5 provides grounds to assume that tweets primarily affect Financial Sentiment when the

financials are of a negative nature ($NN - NP$, $N - NP$), as in these cases, there are significant differences in perception at a 10% level for $NN - NP$ and a 5% level for $N - NP$ respectively, which is why a more in-depth analysis of this observation is needed.

Therefore, in the next step, we divide our overall dataset into participants who received positive financials and participants who were given negative financials for their investment decision. Subsequently, we estimate further separate mediator models for both groups. The base models for positive and negative financials (E) and (F) without control variables take the following form:

$$Stocks\_held = i_1 + c_1 * Tweets + \varepsilon_1 \tag{14}$$

$$\begin{aligned} Stocks\_held = i_2 + c'_1 * Tweets \\ + b_1 * Tweet\_Sentiment \\ + b_2 * Financial\_Sentiment + \varepsilon_2 \end{aligned} \tag{15}$$

$$Tweet\_Sentiment = i_3 + a_{11} * Tweets + \varepsilon_3 \tag{16}$$

$$Financial\_Sentiment = i_4 + a_{12} * Tweets + \varepsilon_4 \tag{17}$$

Both basic models are consequently expanded with the demographic, financial, and social media characteristics to check the robustness of the estimations. The results of the estimation of these models (G) and (H) are depicted in Table 7.

The results of the estimations (E) and (F) confirm, on the one hand, the highly significant direct effect of Financial Sentiment on Stocks held (b_2) and, as expected, the highly significant influence of tweets on Tweet Sentiment. However, on the other hand by dividing the overall dataset, differences in the impact of positive and negative tweets become evident. In the case of positive financials (E), unlike the estimation with negative financials (F) and the previously estimated models (A) and (B), tweets do not exert a significant influence on the Financial Sentiment (a_{12}) and, consequently, exert no indirect effect ($a_{12} * b_2$) on the Stocks held, either.

	Effect type		(E)	(F)	(G)	(H)
Stocks held	**Direct**					
	Tweets	(c_1')	0.021	0.165	-0.013	0.119
			(0.214)	(0.133)	(0.205)	(0.138)
	T_Sen	(b_1)	0.200	-0.039	0.272	-0.007
			(0.209)	(0.136)	(0.194)	(0.135)
	F_Sen	(b_2)	0.290***	0.442***	0.305***	0.475***
			(0.097)	(0.116)	(0.101)	(0.111)
	Age				0.001	-0.060
					(0.117)	(0.082)
	Male				0.095	-0.009
					(0.105)	(0.097)
	Income				-0.075	-0.178
					(0.114)	(0.150)
	Risk				-0.036	-0.070
					(0.098)	(0.086)
	Economic				-0.111	0.020
					(0.093)	(0.093)
	Cap Market				0.166	0.150
					(0.097)	(0.099)
	Usage				-0.099	-0.028
					(0.116)	(0.114)
	Twitter				-0.121	-0.024
					(0.093)	(0.092)
	Indirect					
	$a_{11} \rightarrow b_1$	($a_{11} * b_1$)	0.163	-0.029	0.221	-0.005
			(0.170)	(0.103)	(0.157)	(0.102)
	$a_{12} \rightarrow b_2$	($a_{12} * b_2$)	0.021	0.114*	0.022	0.123*
			(0.030)	(0.046)	(0.032)	(0.049)
T_Sen	**Direct**					
	Tweets	(a_{11})	0.813***	0.756***	0.813***	0.756***
			(0.062)	(0.071)	(0.062)	(0.071)
F_Sen	**Direct**					
	Tweets	(a_{12})	0.073	0.257*	0.073	0.257*
			(0.106)	(0.105)	(0.106)	(0.105)
R^2	Stocks held		0.142	0.244	0.224	0.301
	T_Sen		0.660	0.572	0.660	0.572
	F_Sen		0.005	0.066	0.005	0.066
Group	N		87	85	87	85
	Financials		Positive	Negative	Positive	Negative

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Regressions estimate equations (14) to (17) for the models (E) to (H) with gradual inclusion of controls and the respective sample size N for each model.

Table 7: Results Mediation Analysis models (E)-(H)

Therefore, the observable variance of Financial Sentiment, which has the dominant influence on Stocks held, can be explained to a significantly lesser extent in the model with positive financials (E) in comparison to the model with negative financials (F) since the nature of the given financials does exert an influence on the investment decision of individuals. As a result, the financial sentiment can be explained to a slightly but higher extent in model (F) than in model (E).

All results remain robust for both models even when control variables are included, where model (G) represents the model with control variables and positive financials, and model (H) includes control variables and negative financials.

Overall, our observations align with the initial assumptions and indicate that the Financial Sentiment is particularly influenced when the available financials are negative, and the tweets contradict them in their statements. In addition, it can be seen that individuals tend to have a loss aversion as b_2 is considerable higher for negative (models (F) and (H)) than positive (models (E) and (G)) financials.

Transferring this idea of loss aversion to the given tweets we also divide the dataset by the nature of tweets in Table 8 estimating the following equations:

$$Stocks\_held = i_1 + c_2 * Financials + \varepsilon_1 \tag{18}$$

$$\begin{aligned} Stocks\_held = i_2 + c'_2 * Financials \\ + b_1 * Tweet\_Sentiment \\ + b_2 * Financial\_Sentiment + \varepsilon_2 \end{aligned} \tag{19}$$

$$Tweet\_Sentiment = i_3 + a_{21} * Financials + \varepsilon_3 \tag{20}$$

$$Financial\_Sentiment = i_4 + a_{22} * Financials + \varepsilon_4 \tag{21}$$

	Effect type		(I)	(J)	(K)	(L)
Stocks held	**Direct**					
	Financials	(c_2')	-0.063	-0.413*	-0.100	-0.425**
			(0.132)	(0.162)	(0.125)	(0.155)
	T_Sen	(b_1)	0.006	0.129	0.066	0.159
			(0.101)	(0.094)	(0.098)	(0.091)
	F_Sen	(b_2)	0.479***	0.847***	0.551***	0.849***
			(0.145)	(0.181)	(0.141)	(0.175)
	Age				0.003	0.062
					(0.111)	(0.119)
	Male				0.079	0.032
					(0.101)	(0.109)
	Income				-0.168	-0.079
					(0.101)	(0.138)
	Risk				-0.001	-0.117
					(0.017)	(0.089)
	Economic				-0.110	-0.023
					(0.083)	(0.099)
	Cap Market				0.195	0.091
					(0.101)	(0.090)
	Usage				-0.100	0.064
					(0.116)	(0.132)
	Twitter				-0.085	-0.066
					(0.096)	(0.101)
	Indirect					
	$a_{21} \rightarrow b_1$	($a_{21} * b_1$)	0.000	0.001	0.003	0.001
			(0.014)	(0.014)	(0.008)	(0.018)
	$a_{22} \rightarrow b_2$	($a_{22} * b_2$)	0.325***	0.743***	0.374***	0.744****
			(0.110)	(0.161)	(0.109)	(0.154)
T_Sen	**Direct**					
	Financials	(a_{21})	0.042	0.007	0.042	0.007
			(0.699)	(0.109)	(0.699)	(0.109)
F_Sen	**Direct**					
	Financials	(a_{22})	0.679***	0.877***	0.679***	0.877***
			(0.080)	(0.052)	(0.080)	(0.052)
R^2	Stocks held		0.193	0.296	0.293	0.340
	T_Sen		0.002	0.000	0.002	0.000
	F_Sen		0.461	0.769	0.461	0.769
Group	Observations		87	85	87	85
	Tweets		Positive	Negative	Positive	Negative

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Regressions estimate equations (18) to (21) for the models (I) to (L) with gradual inclusion of controls and the respective sample size N for each model.

Table 8: Results Mediation Analysis models (I)-(L)

In the case of negative tweets, the effect of perceived Tweet Sentiment (b_1) remains insignificant. Nevertheless, it might be noteworthy that the b_1 coefficients in the negative

models (J) and (L) of 0.129 and 0.167 are higher than in the positive models (I) and (K) and also show smaller standard errors leading to p-values decreasing from 95% to 17%, respectively from 57% to 7%. This observation could give justification for not rejecting *H1.1* but should not be overvalued as the effect is negligible aligning with the observations of Boulu-Reshef et al. (2023). In contrast, no significant differences for the given financials between positive and negative tweets can be found.[16]

2.6 Conclusion

The objective of this study is to illuminate the causal pathway of available information on the investment decisions of economic agents. Specifically, the focus is on a detailed examination of the impact of social media posts and their perception. To achieve this goal, a laboratory experiment was conducted, providing participants with various pieces of information in the form of financial data and tweets to inform an investment decision. The aim is to draw conclusions about the causal channels of the provided information based on the investment decisions made by the participants at the end of the experiment.

Following their investment decisions, participants were surveyed regarding their perception of the financials and tweets using a Likert scale. This allows for an examination of whether participants perceived the information in line with the author's intentions. As significant differences in participants' perceptions between the individual groups were expected, it can be inferred that the information was perceived as intended. Furthermore, the financial and tweet sentiment provide an opportunity for a more in-depth analysis of the causal pathway of these two pieces of information.

To address this, the method of mediation analysis was employed to separate the influence of the given information into direct and indirect effects. It was revealed that particularly the perception of information has a significant effect on the investment decisions of economic agents. While the sentiment of tweets does not directly influence investment decisions (or just with a legible impact when tweets are negative), the tweets do impact the perception of financials, which in turn significantly influences investment decisions. This result is in line with existing literature in two different ways. On the one hand we

[16] Following Zhao, Lynch, and Chen (2010) we can observe a competitive mediation with $a_{22} * b_2 * c_2' < 0$ in the models (J) and (L) leading to a summed whole effect of the Financials which is nearly the same as of the positive pendants (I) and (K).

show that social media sentiment does influence the investment decisions of individuals, which has previously also been shown by i.a. Antweiler and Frank (2004), Baker and Wurgler (2006), Da, Engelberg, and Gao (2015), Das and Chen (2007), Renault (2017), Sun, Najand, and Shen (2016) and Tetlock (2007). On the other hand, our results align with the findings of Behavioral Finance. Contrary to the participants' self-reported statements, their investment decisions are subconsciously influenced by the provided tweets, indicating the existence of biases in the information processing process.

In this specific case, the behavior of the participants suggests the presence of the anchoring effect, as presented by Tversky and Kahneman (1974). According to this effect, the tweets, with their content, act as a mental anchor that distorts the interpretation of the financial information. Additionally, we observe a differential impact of tweets on Financial Sentiment when the financials are positive or negative. Our results suggest that an influence exists when negative financial information is present, and the tweets contradict it, i.e., they are positively framed. This could be rooted in the prospect theory, wherein, in the case of losses expressed through negative financials, participants, due to their risk aversion, behave differently than in the case of positive financials. In this scenario they may be more susceptible to information from tweets that deviate from the financials. The results of our study provide three starting points for further research and the practical application of sentiment analysis regarding the precise direction of the impact of social media sentiment we presented.

Firstly, the models discussed could be expanded to include moderators that could serve as catalysts for the strength of the effect of social media sentiment. This could provide insights into relevant factors influencing the susceptibility of economic agents to social media sentiment. However, such an analysis would require a broader participant base and, consequently, a higher number of observations per study group than was the case in this study. Secondly, the influence of bot-generated tweets on our participants suggests that despite the automated generation of these tweets, an impact on economic agents occurs. It seems possible to influence the assessment of a company's financial situation using computer-generated social media content. For an accurate measurement of this approach compared to the use of human-generated tweets, appears necessary in light of the advancing development of AI.

Finally, our results indicate that the influence of social media sentiment on investor decisions is of an indirect nature. Therefore, it seems advisable to take this into greater consideration in future analyses. To the best of our knowledge, this is the first experimental study that dissects the causal pathway of social media sentiment through a mediation analysis into direct and indirect effects, aiming to gain a deeper understanding of its impact on the investment behavior of economic agents.

2.7 Appendix

2.7.1 Platform's interface and content

2.7.1.1 Company description

Figure 10: Company description interface (German language, translation below)

Translation: We are Glubon - Glubon improves the everyday life with intelligent solutions for multiple generations. For 125 years we are driven by our vision every day improving our all and future generation's life with our innovative and sustainable products and technologies. At our company everything is dedicated to our guiding principle: 'grow responsible'.

With over 120,000 employees in over 50 different countries we belong the worldwide leading suppliers of industry and consuming goods. To our innovation and product range count multiple intelligent solutions in the Sections plastics, carbon, metal and glass.

2.7.1.2 Tweets

German ChatGPT query: Generiere mir 10 <*colloquial*> deutsche <*sentiment*> Tweets über die imaginäre Firma Glubon bezüglich Ihrer Aktien, Finanzen, Strategie, Nachhaltigkeit oder Ihres Managements mit maximal <*max length*> Zeichen und dem Cashtag $GLU sowie keinen Emojis.

Translated ChatGPT query: Generate 10 <*colloquial*> German <*sentiment*> Tweets about the imaginary company Glubon regarding their stocks, financials, strategy,

sustainability or management with maximal <*max length*> characters and the cashtag $GLU as well as no emojis for me.

Variations:

<*colloquial*> = {'colloquial', ' '}

<*sentiment*> = {'positive', 'negative', 'neutral'}

<*max length*> = {20, 70, 140}

Original Tweet	**Translation**	**Max length**	**sentiment**	**colloquial**
Glubon zeigt beeindruckende Finanzergebnisse und beweist erneut, warum sie ein solider Wert für langfristige Investitionen sind. $GLU	Glubon shows impressive financial results and proves again why they are a solid value for long-term investments. $GLU	140	positive	no
Glubon-Aktien performen hervorragend und bieten Anlegern eine solide Rendite. $GLU	Glubon-stocks perform excellently and deliver investors a solid return. $GLU	70	positive	no
Top-Finanzen bei Glubon! $GLU	Top-Financials at Glubon! $GLU	20	positive	no
Die Aktien von Glubon sind der Hammer, Leute! Die machen richtig Knete und lassen uns alle mitverdienen. $GLU	The Glubon stocks are amazing, folks! They're making serious dough and letting all of us earn a share. $GLU	140	positive	yes
Glubon-Aktien ballern richtig! Hier gibt's fette Gewinne, Brudi! $GLU	The Glubon stocks are really booming! There are fat profits here, bro! $GLU	70	positive	yes
Glubon-Aktien abgefahren! $GLU	Glubon stocks are off the charts! $GLU	20	positive	yes

Glubon berücksichtigt Nachhaltigkeitsaspekte in ihrem Geschäft und strebt einen verantwortungsbewussten Umgang mit Ressourcen an. $GLU	Glubon considers sustainability aspects in their business and aims for responsible resource management. $GLU	140	neutral	no
Glubon legt Wert auf Nachhaltigkeit und Ressourcenschonung. $GLU	Glubon emphasizes sustainability and resource conservation. $GLU	70	neutral	no
Strategie solide. $GLU	Strategy is solid. $GLU	20	neutral	no
Die Aktien von Glubon sind ganz okay, nichts Weltbewegendes, aber auch keine Totalausfälle. Mal sehen, wie's weitergeht. $GLU	The Glubon stocks are just okay, nothing groundbreaking, but not total disappointments either. Let's see how it goes. $GLU	140	neutral	yes
Finanzen bei Glubon okay, nix Besonderes, aber auch nicht im Keller. So mittel halt. $GLU	Finances at Glubon are okay, nothing special, but not at rock bottom either. Just average. $GLU	70	neutral	yes
Management ganz okay. $GLU	Management is quite okay. $GLU	20	neutral	yes
Die Strategie von Glubon ist zum Scheitern verurteilt, kein Wunder, dass sie den Markt nicht dominieren können. $GLU	Glubon's strategy is doomed to fail; no wonder they can't dominate the market. $GLU	140	negative	no

Finanzen bei Glubon katastrophal, rote Zahlen ohne Ende. Keine gute Wahl für Anleger. $GLU	Finances at Glubon are catastrophic, endless red figures. Not a good choice for investors. $GLU	70	negative	no
Strategie bei Glubon schwach. $GLU	Strategy at Glubon is weak. $GLU	20	negative	no
Ey, die Aktien von Glubon sind voll der Reinfall, voll im Keller! Wer da investiert, hat echt 'nen Schaden. Finger weg! $GLU	Hey, Glubon stocks are a complete flop, way down in the dumps! Investing there is a real mistake. Stay away! $GLU	140	negative	yes
Finanziell geht's bei Glubon den Bach runter, die sind pleite! $GLU	Financially, Glubon is going downhill, they're bankrupt! $GLU	70	negative	yes
Nachhaltigkeit Fehlanzeige. $GLU	No sustainability in sight. $GLU	20	negative	yes

Table 9: Tweet examples per ChatGPT query

2.7.1.3 Financials

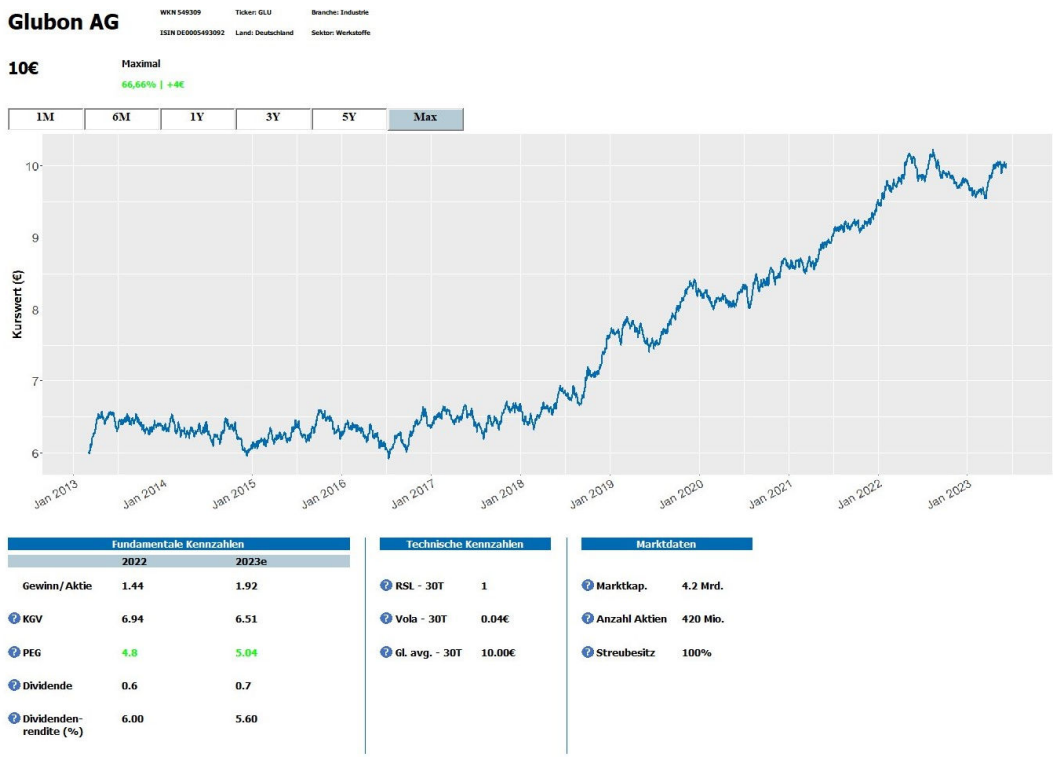


Figure 11: *Financials* tab, max chart opened (positive version)

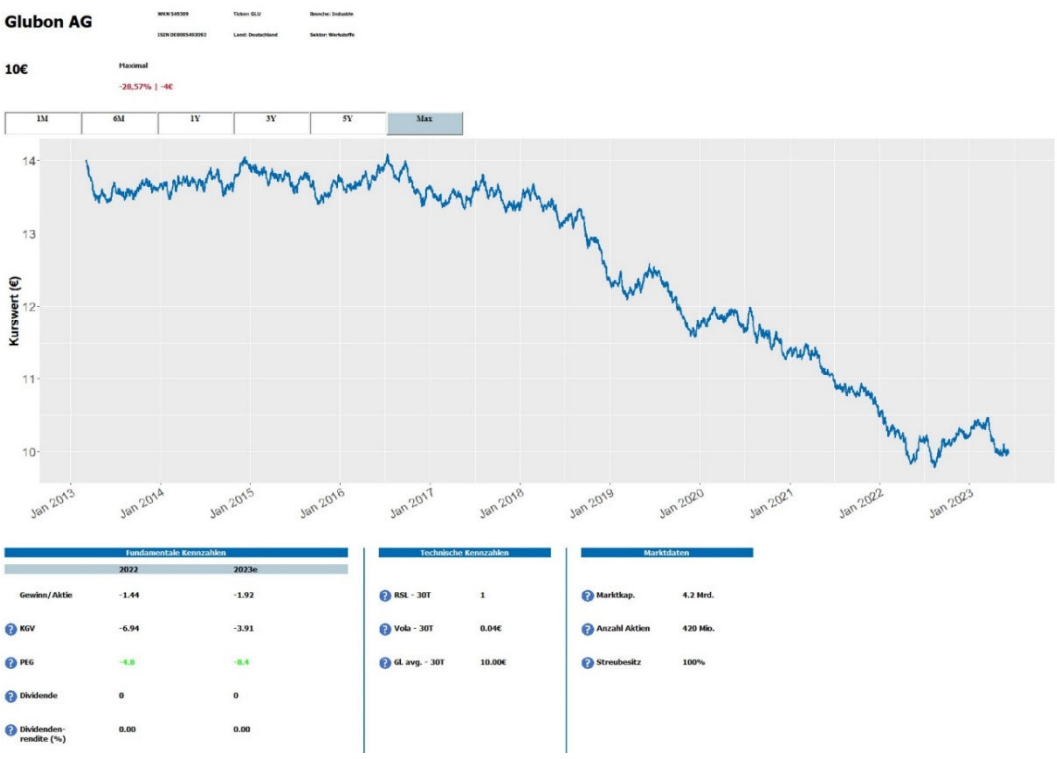


Figure 12: *Financials* tab, max chart opened (negative version)

2.7.2 Robustness checks

2.7.2.1 Removal of slowest and fastest participants

	Effect type		(M)	(N)	(O)
Stocks held	**Direct**				
	Tweets	(c_1')	0.096	0.101	0.111
			(0.118)	(0.118)	(0.122)
	Financials	(c_2')	-0.138	-0.180	-0.161
			(0.117)	(0.099)	(0.117)
	T_Sen	(b_1)	0.088	0.073	0.063
			(0.115)	(0.113)	(0.117)
	F_Sen	(b_2)	0.571***	0.617***	0.599***
			(0.115)	(0.111)	(0.130)
	Age		0.057	0.083	0.086
			(0.082)	(0.081)	(0.085)
	Male		0.061	0.058	0.073
			(0.069)	(0.067)	(0.069)
	Income		-0.140	-0.138	-0.160
			(0.088)	(0.087)	(0.092)
	Risk		-0.043	-0.066	-0.058
			(0.064)	(0.061)	(0.066)
	Economic		-0.043	-0.040	-0.034
			(0.064)	(0.064)	(0.065)
	Cap Market		0.148*	0.165*	0.164*
			(0.068)	(0.068)	(0.069)
	Usage		-0.037	-0.021	-0.023
			(0.081)	(0.080)	(0.080)
	Twitter		-0.055	-0.031	-0.035
			0.068	(0.064)	(0.066)
	Indirect				
	$a_{11} \rightarrow b_1$	($a_{11} * b_1$)	0.069	0.057	0.049
			(0.090)	(0.088)	(0.092)
	$a_{21} \rightarrow b_1$	($a_{21} * b_1$)	0.002	0.001	0.002
			(0.005)	(0.004)	(0.005)
	$a_{12} \rightarrow b_2$	($a_{12} * b_2$)	0.065*	0.067*	0.066*
			(0.028)	(0.030)	(0.030)
	$a_{22} \rightarrow b_2$	($a_{22} * b_2$)	0.457***	0.447***	0.484***
			(0.110)	(0.093)	(0.112)
T_Sen	**Direct**				
	Tweets	(a_{11})	0.786***	0.784***	0.787***
			(0.049)	(0.049)	(0.050)
	Financials	(a_{21})	0.024	0.018	0.024
			(0.049)	(0.049)	(0.050)
F_Sen	**Direct**				
	Tweets	(a_{12})	0.113*	0.109*	0.110*
			(0.046)	(0.049)	(0.047)
	Financials	(a_{22})	0.801***	0.773***	0.809***
			(0.046)	(0.050)	(0.047)
R^2	Stocks held		0.302	0.319	0.309
	T_Sen		0.620	0.616	0.671
	F_Sen		0.659	0.612	0.621
N	Observations		163	163	154

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Regressions estimate equations (10) to (13) for the models (M) to (O) with gradual exclusion of the 5% fastest (M)/slowest (N)/fastest and slowest participants(O))

Table 10: Results Mediation Analysis models (M)-(O)

2.7.2.2 Results per check questions correctly answered

	Effect type		(P)	(Q)	(R)
Stocks held	**Direct**				
	Tweets	(c_1')	0.049	0.061	0.087
			(0.114)	(0.113)	(0.114)
	Financials	(c_2')	-0.160	-0.146	-0.157
			(0.096)	(0.097)	(0.100)
	T_Sen	(b_1)	0.137	0.136	0.096
			(0.109)	(0.109)	(0.110)
	F_Sen	(b_2)	0.564***	0.562***	0.590***
			(0.106)	(0.107)	(0.111)
	Age		0.023	0.019	0.053
			(0.070)	(0.070)	(0.080)
	Male		0.029	0.039	0.046
			(0.067)	(0.067)	(0.067)
	Income		-0.125	-0.122	-0.120
			(0.076)	(0.076)	(0.084)
	Risk		-0.076	-0.084	-0.054
			(0.059)	(0.059)	(0.060)
	Economic		-0.040	-0.039	-0.047
			(0.062)	(0.062)	(0.063)
	Cap Market		0.141*	0.123	0.150*
			(0.066)	(0.065)	(0.067)
	Usage		-0.064	-0.041	-0.035
			(0.081)	(0.079)	(0.080)
	Twitter		-0.000	-0.012	-0.051
			(0.065)	(0.065)	(0.066)
	Indirect				
	$a_{11} \rightarrow b_1$	($a_{11} * b_1$)	0.108	0.107	0.075
			(0.086)	(0.086)	(0.086)
	$a_{21} \rightarrow b_1$	($a_{21} * b_1$)	0.003	0.003	0.002
			(0.007)	(0.007)	(0.005)
	$a_{12} \rightarrow b_2$	($a_{12} * b_2$)	0.071**	0.070**	0.066*
			(0.027)	(0.026)	(0.028)
	$a_{22} \rightarrow b_2$	($a_{22} * b_2$)	0.430***	0.428***	0.452***
			(0.088)	(0.089)	(0.093)
T_Sen	**Direct**				
	Tweets	(a_{11})	0.787***	0.785***	0.784***
			(0.045)	(0.046)	(0.047)
	Financials	(a_{21})	0.022	0.022	0.018
			(0.047)	(0.047)	(0.048)
F_Sen	**Direct**				
	Tweets	(a_{12})	0.126**	0.125**	0.112*
			(0.046)	(0.047)	(0.048)
	Financials	(a_{22})	0.763***	0.762***	0.767***
			(0.048)	(0.048)	(0.048)
R^2	Stocks held		0.287	0.297	0.311
	T_Sen		0.621	0.618	0.615
	F_Sen		0.608	0.604	0.605
N	Observations		182	180	120

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Regressions estimate equations (10) to (13) for the models (P) to (R) including all particpants (P) and participants who correctly answered at least 2(Q) or 4(R) control questions.)

Table 11: Results Mediation Analysis models (P)-(R)

2.8 Declaration of (co-)authors and record of accomplishments

Title: The relevance and influence of Social Media posts on investment decisions – an experimental approach based on tweets

Author(s): Lars M. Kürzinger (Heinrich Heine University Düsseldorf)
Philipp Stangor (Heinrich Heine University Düsseldorf)

Conferences: Participation and presentation at 'Forschungskolloquium Finanzmärkte', 25th January 2023, Düsseldorf, Germany

Participation and presentation at 'HVB Doctoral Colloquium', 2nd – 3rd February 2024, Münster, Germany

Publication: SSRN published. Submitted to the 'Journal of Behavioral and Experimental Finance', double-blind peer-reviewed journal. Current status: Under review.

Share of contributions:

Contributions	Lars M. Kürzinger	Philipp Stangor
Research Design	**65%**	**35%**
Development of Research Question	50%	50%
Method development	80%	20%
Research performance & analysis	**50%**	**50%**
Literature Review and framework development	50%	50%
Data collection, preparation and analysis	20%	80%
Analysis and discussion of results	80%	20%
Derivation of implications and conclusions	50%	50%
Manuscript preparation	**45%**	**55%**
Final draft	50%	50%
Finalization	40%	60%
Overall contribution	**60%**	**40%**

31.05.2024,

Date, Philipp Stangor

31.05.2024,

Date, Lars M. Kürzinger

3 Measuring investor sentiment from Social Media Data – an emotional approach

3.1 Abstract

We employ a multi-dimensional approach extracting investor sentiment from social media data using the NRC-Emotion Association Lexicon. Considering a vast number of short text messages from the financial microblogging platform StockTwits, we analyze different emotions contained in each message. Subsequently, we classify these posts as bullish or bearish signals on basis of their emotional profile using machine learning techniques to develop aggregated investor sentiment. This classification outperforms comparable classifications based on non-economic or two-dimensional dictionaries in terms of accuracy and data efficiency. Consequently, we are able to predict intraday returns for the S&P 500 and NASDAQ 100.

3.2 Introduction

With the rise of social media platforms such as Twitter, Facebook and Instagram and their growing popularity, many researchers have investigated the potential influence of a platform's content on the performance of stock markets. As investor's attention is found to be limited, their investment behavior tends to be biased towards investments that consciously or unconsciously attract their attention (Barber and Odean (2008)). In this case, social media platforms might affect an individual's investment decision (Liu (2020)). Johnson and Tversky (1983) already observed that sentiment is able to affect investors' perception of risk. Kaplanski et al. (2015) confirm this finding, even going so far as to detect the effects of investors' personal happiness on their investment behavior. Furthermore, Baker and Wurgler (2007) conclude that what is in question is no longer whether sentiment influences market participants but rather how strong its effect may be and how its measured.

In this regard, we extract aggregated investor sentiment by analyzing a vast number of social media posts and examine the sentiment's influence on market movement.

Figure 13 outlines the progress made in economic text analysis. Linguistic text analysis initially classified single words by matching them with their predefined connotation in a linguistic dictionary that was originally derived from psychological analysis. Hence, a

word's connotation is usually distinguished between 'positive' and 'negative' (see, i.a., Antweiler and Frank (2004), Baker and Wurgler (2007), Gao and Yang (2017), Kim and Kim (2014) and Sun, Najand, and Shen (2016)). However, when conducting this analysis in an economic setting, one faces the question of whether a word's meaning in a psychological context might differ from its meaning in an economic context. Thus, the word 'risk' might be connotated (very) negatively in the first setting, while this might not be the case in an economic analysis. Therefore, starting with Henry (2008) and Loughran and McDonald (2011), sentiment dictionaries intentionally designed for economic uses of language have been created and used for a more economically specific analysis. Nevertheless, further linguistic challenges such as punctuation, slang, irony or emoticons were not considered, which led to the rise of rule-based models to further evaluate social media data. One of most prominent dictionaries and sentiment analysis tools in this context is the so-called *Valence Aware Dictionary and sEntiment Reasoner* (VADER), which is able to consider the abovementioned linguistic components, making it possible to estimate the degree to which a microblogging text contains positive or negative sentiment (Hutto and Gilbert (2014)).

Apart from those dictionary based approaches newer techniques such as Natural Language Processing (NLP) Transformers like i.a. BERT (Devlin et al. (2018)), XLNet (Yang et al. (2019)) and XLM (Lample and Conneau (2019)) recently emerged and have been optimized continuously. These NLP Transformers make use of different kinds of machine learning techniques to achieve high accuracies in text classification tasks.[17]

However, in this work we take a step back to answer the question of whether a multi-dimensional analysis might present a better starting point for sentiment analysis than a two-dimensional approach. We find that a multi-dimensional approach using emotions outperforms comparable classifications based on non-economic or two-dimensional dictionaries in terms of accuracy and data efficiency. When using the NRC-Emotion Association Lexicon created by Mohammad and Turney (2013) (also known as 'EmoLex'), we do not match positive or negative connotations with a given microblogging text but rather with up to eight different emotions associated with each word. As EmoLex is a dictionary without an economic background (Figure 13: A1) a

[17] For a more extensive overview concerning different approaches of sentiment analysis (including dictionary based approaches and NLP Transformers) see Mishev et al. (2020).

suitable benchmark is a positive-negative dictionary without an economic context (Figure 13: B1). To further validate our results, we compare the accuracy of our approach with other benchmark dictionaries that are already widely used in (economics) literature and practice and possess an economic background, as well (Figure 13: B2).

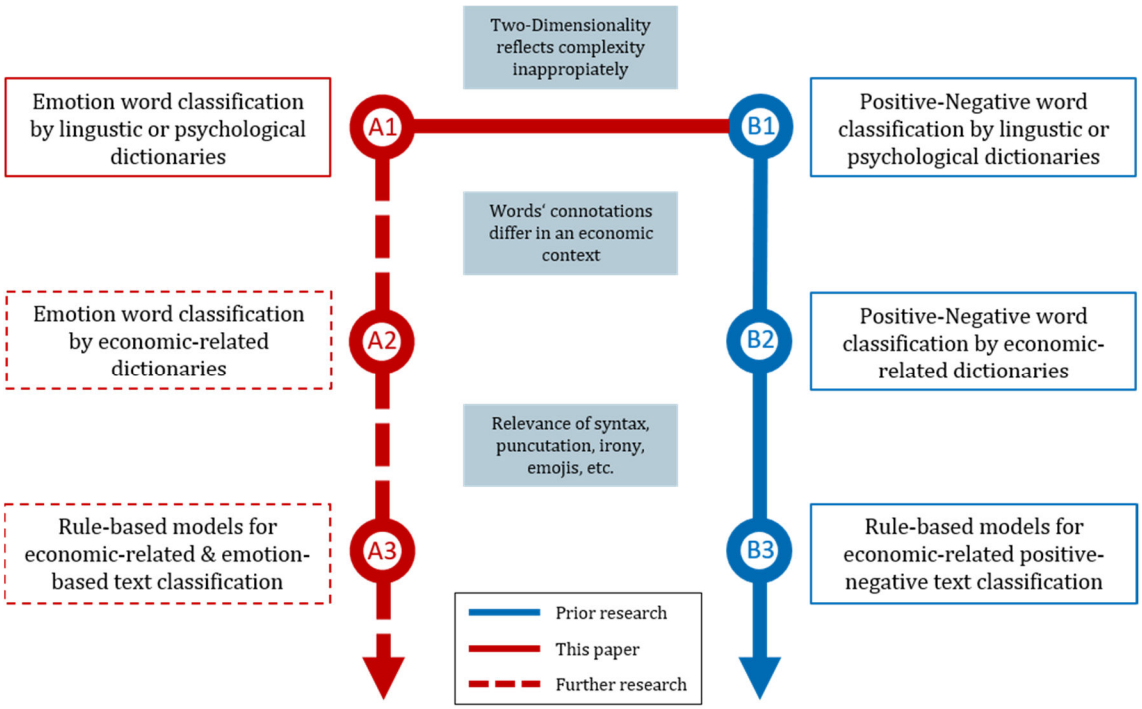


Figure 13: Progress of economic-related text-analysis research

Our results emphasize the need for (more specific) emotion-based *and* economic-related dictionaries. To the best of our knowledge, we are unaware of other studies using this explicit technique in the same way. With our results, we encourage economic research in textual sentiment analysis to focus more on multi-dimensional emotional approaches than on two-dimensional approaches as the most prominent positive-negative approaches used in the majority of related research. We expand the existing literature by outlining three main factors determining the success of a field-specific sentiment analysis dictionary: multi-dimensional scoring (for example emotions), economic word connotation and type of text. Our dictionary-based results from the beginnings of sentiment analysis (Figure 13: A1) also give implications for more sophisticated approaches of sentiment analysis. When considering our findings, one could expect the results of other approaches like more advanced dictionaries (Figure 13: A3) or NLP Transformers to profit from a more

dimensional analysis further improving classification results. Future research should take this hypothesis into consideration and validate our basic findings.

The remainder of this paper is structured as follows: Section 3.3 gives a short overview of the related literature regarding sentiment analysis. Section 3.4 presents our data, namely, the ideas from the social media platform StockTwits and the chosen stock market data for proving the economic relevance of our results. In Section 3.5, we describe our method, which leads to our results presented in Section 3.6. Section 3.7 concludes the paper, relates our observations with prior results found in the literature and provides an outlook on possible future research topics.

3.3 Literature review

Beginning with the work of Antweiler and Frank (2004), internet stock messages have been investigated for to their suitability to measure market sentiment and thus to predict the movement of markets. Antweiler and Frank (2004), among other studies, (see, i.a., Das and Chen (2007), Kim and Kim (2014)) do not find a significant relationship between sentiment and market returns but reveal a correlation among social media activity, trading volume and return volatility. Although Kim and Kim (2014) do not find any relationship of the abovementioned kind, other studies do find significant relationships between intraday sentiment and intraday returns (Sun, Najand, and Shen (2016), Gao and Yang (2017)) or overnight returns (Renault (2017)). One possible explanation for the differing results might lie in the changing composition of social media users and their behavior over time, as Renault (2017) argues.

Following Grossman and Stiglitz (1980), however, we assume that market participants are able to obtain small excess returns as compensation for continuous information gathering, contradicting market information efficiency (Jensen (1978)). These additional returns can be viewed as a reward for monitoring and analyzing market information that compensate market participants for the costs associated with monitoring and maintaining the market's signals. In a competitive market setting, however, small excess returns are assumed to be short-lived since professional investors will exploit any value-relevant information to gain an information advantage over their competitors (Renault (2017)).

Therefore, individuals will make use of any institution that reduces information costs by centralizing, selecting and verifying information, which explains the emergence of

information service providers such as Reuters or Bloomberg. Usually, fees for using these services exceed small investors' capabilities. Social media platforms may represent one means of filling this gap, making it easier to obtain potential value-relevant information. This finding is in line with Baker and Wurgler (2007), for example, who argue that sentiment effects hold especially for 'small-capitalization, younger, unprofitable, high-volatility, non-dividend-paying, growth companies or stocks of firms in financial distress' since they might be more difficult to value due to increasing information costs.

Furthermore, in a behavioral finance context, stock prices may differ from their fundamental value due to possibly irrational investor behavior. Bullish or bearish expectations among noise traders might therefore be able to move stock prices (Long et al. (1990)). For example, individuals tend to overvalue a conversation partner's opinion (DeMarzo, Vayanos, and Zwiebel (2003)) or may be more willing to invest in certain assets because they have aroused their attention consciously or unconsciously (Barber and Odean (2008)). Behavioral biases such as these might be one of the reasons that social media sentiment analysis appears tempting in a financial setting, as it may provide an explanation for individuals' noisy behavior in the sense of Black (1986) and simultaneously provide an explanation for why people participate in social media platforms such as StockTwits and publish their beliefs. In the setting described by DeMarzo, Vayanos, and Zwiebel (2003) and Giannini, Irvine, and Shu (2018), it might even be rational for institutional investors, who are often assumed to be less susceptible to biases, to follow opinion leaders since they are able to move markets or even become influential themselves. Furthermore, communication between market participants appears to be suitable to convince hesitant market participants to invest in certain assets, as they learn of other individuals who share a similar opinion about an investment possibility (Cao, Coval, and Hirshleifer (2002), Antweiler and Frank (2004)). Knowledge of these ways of behavior might even provide incentives to individuals to deliberately spread rumors about assets in an attempt to profit from the expected reactions their followers might take (van Bommel (2003)), thereby explaining questions concerning the motivation of informed investors to publish their information (see Xiong et al. (2019)). Bullish or bearish expectations among noise traders are therefore able to move stock prices (Long et al. (1990), Black (1986)).

For this purpose, we define sentiment as a market's general, psychological environment. Currently, three different methods to obtain a market's sentiment can be found in the literature. The first alternative resembles the analysis of market-based data such as trading volumes, IPO returns or IPO volumes using high-frequency data (e.g., Lee, Shleifer, and Thaler (1991) or Baker and Wurgler (2006)). However, Qiu and Welch (2004) and Da, Engelberg, and Gao (2015) argue that these types of studies suffer from the vast number of potential variables at hand and their interdependencies. Second, surveys such as the Consumer Sentiment Index represent another method of measuring investor sentiment (i.a. Brown and Cliff (2005)) but are only frequently conducted and therefore suffer from low frequency, making them unsuitable for analyzing short-lived excess returns. Additionally, little incentive exists to truthfully answer survey questions, resulting in potentially biased survey results (Singer (2002)). As a consequence, we employ the third alternative in the form of a textual-based analysis with reference to Tetlock (2007) and Renault (2017), using a linguistic approach to evaluate text data from the microblogging platform *StockTwits*. This approach enables us to make use of high-frequency text data created by the platform's users and the data's living lab properties, negating the abovementioned issues.

3.4 Data

3.4.1 StockTwits

In this study, we use data from the microblogging platform StockTwits as formerly done in, for example, Renault (2017), Giannini, Irvine, and Shu (2018) and Cookson and Niessner (2020). Ranked by the website analytics tool Alexa as the 768th most popular website in the USA as of April 2022, the platform addresses individuals, professionals and institutions who want to share their opinions, thoughts and ideas about financial topics. Sprenger et al. (2014) correctly note how many (early) results in the field of financial textual sentiment research lack statistical significance because of using un- or inadequately filtered data. In this manner, the platform's concept addresses those emerging problems in a way not done by other microblogging platforms (e.g., Twitter) without losing the advantage of generating a considerable amount of real-time data. Another noteworthy benefit of StockTwits data is the user's ability to flag their ideas as 'bullish' or 'bearish', thereby eliminating the need for researchers to manually classify ideas into each category. In prior research, this issue has often led to the problem of

misclassification due to subjective classification. An additional noteworthy feature of our data is the possibility for users to reveal information about themselves within the shown categories in Table 12:

Category	**Possible Expressions**
Trading Experience	Novice, Intermediate, Professional
Holding Period	Day Trader, Position Trader, Swing Trader, Long Term Investor
Trading Approach	Technical, Momentum, Growth, Fundamental, Global Macro, Value
Trading Asset	Equities, Options, Forex, Futures, Bonds, Private Companies

Table 12: User categories and possible expressions

As in every self-classification task, there is obvious potential for misclassification by the users, especially due to the possible benefits of over- or underestimating themselves. In our case, we do not expect systematic problems to occur for the last three categories of Table 12, since the categories are well distinguished from one another and understandable for the users interested in participating on such a platform and there is no incentive for misclassification. However, there might exist an incentive for users to overestimate their trading experience before the community such that self-classification could suffer from bias. Nevertheless, differences between the expressions ('Novice', 'Intermediate' and 'Professional') can be interpreted in the following.

At the end of 2020, StockTwits had traffic of over 40,000 active users[18] sharing nearly 300,000 ideas on average per day. Figure 14 illustrates that (1) both numbers have strongly increased in recent years and (2), for this reason, the latter results are constantly in need of updates and improvements (e.g., Renault (2017) also states himself).

[18] As active users per day, we define the number of users who published at least one idea on the platform on that day.

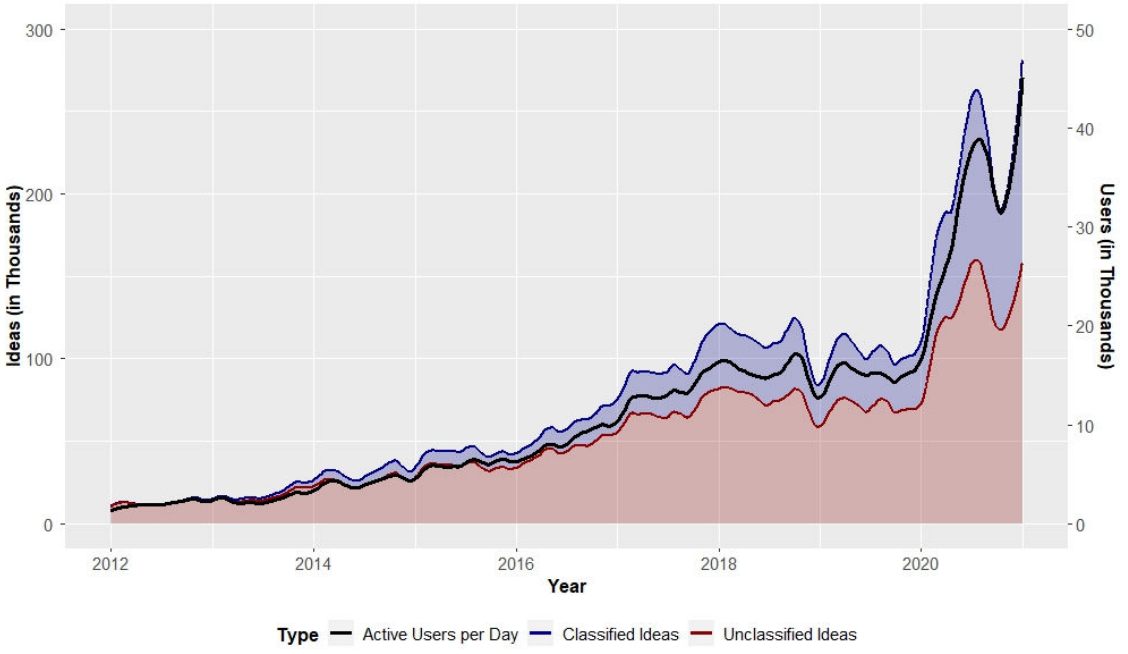


Figure 14: Number of shared ideas and active users per day (loess-smoothed)

To update the latter research results, we use all ideas published on the platform from January 2012 until the end of December 2020 by accessing the StockTwits Developer API. After clearing the data of ideas that are not suitable for the measurement of textual sentiment, for example, ideas that only contain 'cashtags' as identifiers for several stocks ($), pictures or hyperlinks, 250,321,511 ideas remain in the chosen time horizon; 75,414,994 (30.13%) have been classified by the StockTwits community - 62,826,233 (25.10%) as 'bullish' (N_{Bu}) and 12,588,761 (5.03%) as 'bearish' (N_{Be}) To the best of our knowledge, this is one of the largest StockTwits data samples used in published relevant research to date.

The higher rate of bullish ideas can be explained by the predominantly bullish market conditions in the chosen time horizon and the fact that individuals generally tend to share positive rather than negative news. The major share of unclassified ideas illustrates how important a suitable classification with the help of emotion scores is. On average, 34,833 ideas were classified on StockTwits per day in 2019. Assuming an equal distribution over time, approximately 1,451 ideas per hour or only 24 ideas per minute were published on average in 2019. Considering the discussed economic theory, enlarging the dataset by classifying all published ideas improves prediction quality and allows for a more detailed analysis (e.g., for single stocks).

3.4.2 Stock Data

In addition to our main research topic of measuring investor sentiment, we want to emphasize the economic relevance of this generated sentiment by attempting to forecast intraday returns. We do so by observing the development of derived investor sentiment shortly before stock market closing on the previous day, $t-1$, and after the opening on the next day t. The stock data we use to analyze the predictive power of investor sentiment are retrieved from *Thomson Reuters Eikon*. As depicted in Figure 15, the timeframe with the highest average activity on StockTwits coincides with the opening hours of the American stock markets in contrast to the European and Asian markets. This observation most likely derives from the fact that according to Alexa around 54% of all visitors to the platform StockTwits originate from the United States (49.2%) and Canada (5.2%) as of April 2022.

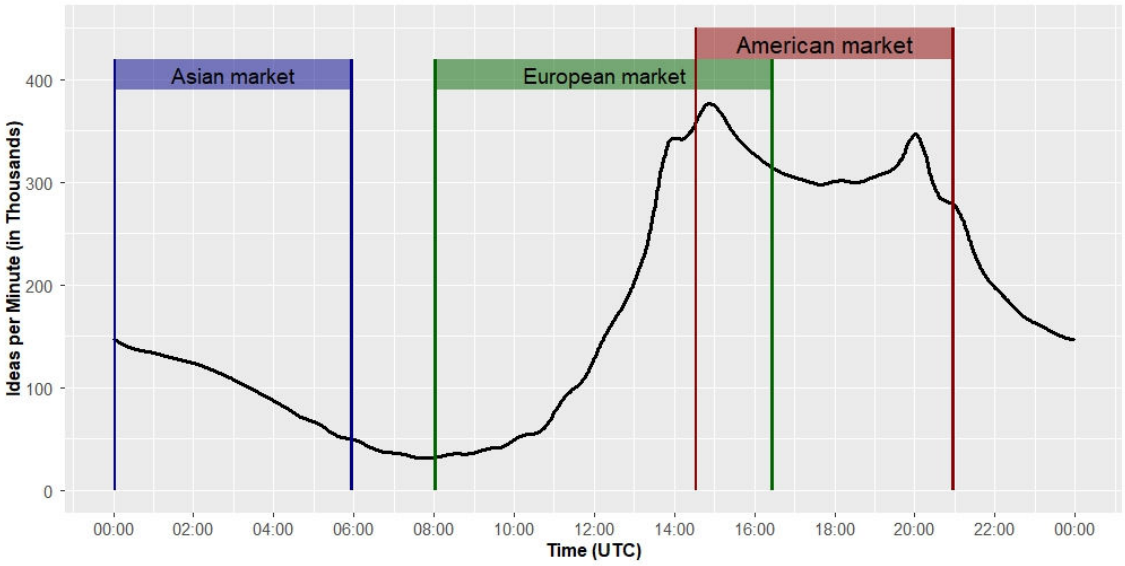


Figure 15: Creation time of shared ideas on StockTwits

As we expect most conversation to be held about topics concerning the US stock market and we find, in accordance to Cookson and Niessner (2020), the platform's users to have an affinity for technology companies, we obtain besides the S&P 500 data for the NASDAQ 100. The NASDAQ 100 appears to be suitable to appropriately represent the North American financial market in general but is also more focused on technology stocks, thereby addressing user affinity. The examined time period corresponds to the

time interval selected for our StockTwits data spanning from 01/2012 to 12/2020 ($T = 2263$).

Due to their statistically desirable characteristics, we use logarithmic returns as a steady measure of performance. We compute the intraday return of the S&P 500 and NASDAQ 100 on a given trading day t ($Intraday_t$) as formula (22)

$$Intraday_t = \ln\left(\frac{Closing_t}{Opening_t}\right) \tag{22}$$

depicts, where $Opening_t$ denotes the opening price on the given day t and $Closing_t$ the closing price on the same day.

3.5 Methodology

3.5.1 Converting Text to Emotion Scores

As previously defined, our aim is to improve previous research results, which we seek to accomplish by using the NRC Word-Emotion Association Lexicon (also known as 'EmoLex') introduced by Mohammad and Turney (2013) as a multi-dimensional text classification approach. In contrast to the predominantly used sentiments 'positive' and 'negative' in related literature, Mohammad and Turney (2013) created EmoLex containing the basic emotions 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise' and 'trust' proposed by Plutchik (1984). Using the R package 'syuzhet' developed by Jockers (2015), we are able to access the collected word list from Mohammad and Turney (2013) containing 14,182 unigrams and 25,000 word senses. The word list consists of the most frequently used unigrams and bigrams measured by the Google n-gram corpus, which are part of the *Macquarie Thesaurus* dictionary of words from the WordNet Affection Lexicon, and at most word-sense pairs from the General Inquirer, which have at least two or three senses. The authors split the classification task into independently solvable 'human intelligence tasks' (HITs), which are solved by users (so-called 'turkers') on the Amazon platform 'Mechanical Turk'. Thus, emotion scores can be extracted from the individual classification by turkers (Mohammad and Turney (2013)).

A further problem that often emerges while working with word lexicons to identify sentiment scores - regardless of whether positive-negative polarity or emotions are

examined - is that word lexicons do not include all possible formats a word might take. For example, a matching algorithm would miss the word 'lovers' if only the root 'love' is part of the lexicon. With *stemming* and *lemmatization*, linguistics proposes two possible solutions for this issue. While stemming algorithms attempt to determine a word's root by detecting and removing suffixes ('lovers' to 'lover' and 'loves' to 'love'), lemmatization attempts to group inflected forms into a single group ('lovers' and 'loves' to 'love'). For our analysis, we use the lemmatization list (41,531 words) created by Mechura (2016), which we access via the R package 'textstem' from Rinker (2018).

Table 13 illustrates how three representative ideas had been edited before we matched them with EmoLex to extract their emotion scores. In addition to the lemmatization of the strings, we remove whitespaces, stopwords, hyperlinks, hash- (#) or cashtags ($) and punctuation.

Idea		**Emotion scores**							
Origin	Edited	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Costco should report stronger than expected December comps, says Stifel Nicolaus - $COST - http://stks.co/1k8b	Costco report strong expect December comps say Stifel Nicolaus	0	1	0	0	0	0	1	1
What word would you use to describe your feelings about $USDCAD since 1.0655? hatred..anger.. boredom..frustration.. #forex #cad	What word use describe feeling since hatred anger boredom frustration forex cad	5	1	3	2	1	3	1	2
Finally $TZA I am in green. I am off to enjoy my weekend. Signing off early. Lot of stress and anxiety. Need a break. Good luck to all.	Finally I green I enjoy weekend sign early Lot stress anxiety Need break Good luck	1	5	1	1	5	1	4	4

Table 13: Conversion from origin ideas to edited ideas and resulting emotion scores

Furthermore, Table 14 depicts the mean emotion scores within the dataset of classified ideas grouped by the classification of the users. Bearish ideas tend to be loaded with

words associated with anger, disgust, fear and sadness, while bullish ideas tend to be loaded with words associated with anticipation, joy, surprise and trust. Except for the emotions anticipation and surprise, which do not appear to be clearly assignable to one of the classifications, all results match intuition. Using a Welch two-sample t-test, we check whether the difference in means is different from zero. For all types of emotions, we can reject the null hypothesis that groups' mean scores do not differ with a significance level below 0.01%.

Emotion	**Classification**		**Welch Two Sample t-test**
	Bullish	Bearish	
Anger	0.2388	**0.3161**	-407.47 ***
Anticipation	**0.6055**	0.5236	310.52 ***
Disgust	0.1121	**0.2015**	-620.27 ***
Fear	0.2449	**0.3282**	-428.09 ***
Joy	**0.3745**	0.2975	400.79 ***
Sadness	0.1682	**0.2635**	-562.90 ***
Surprise	**0.2891**	0.2880	5.86 ***
Trust	**0.5599**	0.5049	205.64 ***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$
Sample: Classified ideas between 01/2012 and 12/2020 ($N_{Bu} + N_{Be} = 75,414,994$)

Table 14: Mean emotion scores of classified ideas per group ('bullish'/'bearish')

With the generated emotion scores, we train a machine learning algorithm with the aim of classifying the 69.87% unclassified ideas as 'bullish' or 'bearish' using their emotion scores. As our dataset is strongly unbalanced with many 'bullish' ideas and less 'bearish' ones we first balance it by randomly picking ideas from each group with the following sample size:

$$N_{Sample} = \frac{\min(N_{Bu}, N_{Be})}{2} = 6{,}294{,}380 \tag{23}$$

Furthermore, we divide the sample ($N_{Bu,Sample} + N_{Be,Sample} = 2 * N_{Sample} = 12{,}588{,}760$) into a training and a test dataset with a proportion of 80 to 20. Subsequently, we divide the training dataset with the same proportion into two further datasets that the algorithm uses for training and validation. The model used contains three dense layers, the first two layers deliver 64 units using a relu activation function while the last layer delivers one unit using a sigmoid activation which is the probability that an idea is

classified as 'bullish'.[19] In this manner, we classify ideas with a probability above or equal to 0.5 as 'bullish' and below 0.5 as 'bearish' using the test dataset in a first step. Thus, we define the accuracy of our model as the percentage of its correct classification within the test dataset.

3.5.2 Benchmarks

Consequently, we need to choose qualified benchmarks to compare the accuracy results of ideas classified by EmoLex with the classification results of other dictionaries to evaluate the performance of our approach. Therefore, we conduct the same classification task as described in Section 3.5.1 with other dictionaries commonly used in the economic literature. As mentioned at the beginning of this work, our aim is to underline the need to create an emotion-based *and* economic-related dictionary. For this purpose, we separately analyze the benefits of both dictionary types, defining the two following hypotheses:

H1: *The classification accuracy and economic relevance of emotion-based dictionaries are higher than the accuracy of positive-negative-based dictionaries in text with an economic background.*

As noted in the introduction, we expect that an emotion-based dictionary such as the EmoLex dictionary with its eight dimensions (emotions) is more suitable to capture the complexity of (everyday) language. The compared dictionaries need to be created for the same type of language because words differ in connotation across contexts. This aspect brings us to our second hypothesis (*H2*), in which we expect that field-specific dictionaries are more capable of classifying words correctly from a text originating in this specific field.

H2: *The classification accuracy and economic relevance of economic-related dictionaries are higher than the accuracy of non-economic-related dictionaries in text with an economic background.*

Table 15 presents an overview of the properties of the chosen benchmark dictionaries. To examine the first hypothesis (*H1*) we use the accuracy rates of a positive-negative

[19] Beforehand, we've also tried out linear regression and logit/probit models, which have already been outperformed by a simple neural network with three dense layers in terms of accuracy of classification. As the classification problem is not that complex, more complex networks only delivered small increases in accuracy and the use of them has been rejected by the authors with respect to proportionality.

dictionary that is not economically related. The simplest approach is the use of the positive and negative scores of EmoLex (PN_{EL}), which we will also check for their accuracy contribution but which we are also questioning with respect to their independence from EmoLex emotion scores. Hence, we implement the positive and negative scores from the Harvard General Inquirer (PN_{GI}), since they have been used in many economic studies since the beginning of textual analysis research. Eventhough, this dictionary is not economic related (i.a. Da, Engelberg, and Gao (2015), Engelberg, Reed, and Ringgenberg (2012) or Tetlock (2007)).

Dictionary		**Score type**		**Economic-related**	**No. of words**
Name	Symbol	Emotions	Pos.-Neg.		
EmoLex	EM_{EL}	X			14,154
	PN_{EL}		X		
H GI	PN_{GI}		X		3642
LM	PN_{LM}		X	X	2709
Henry	PN_{HE}		X	X	190

Table 15: Properties of commonly used dictionaries in economic literature

For the examination of the second hypothesis (*H2*), we need an economic-related positive-negative dictionary. Most prominent in this context is the work of Loughran and McDonald (2011), who created such a dictionary for evaluating the text tone of financial reports (PN_{LM}). Despite the broad use of this dictionary in economic research (i.a. Da, Engelberg, and Gao (2015), Chen et al. (2014), Kearney and Liu (2014), Engelberg, Reed, and Ringgenberg (2012), Dougal et al. (2012)) we also consider the dictionary by Henry (2008), as it is one of the first economic-related dictionaries that focuses on the influence of earnings press releases' tone on investor decision-making (PN_{HE}).[20]

3.5.3 Deriving Investor Sentiment

As the next and last step, we use the received classification to measure investor sentiment. This task is only needed for the investigation of the relevance of our main results – the accuracy of the different dictionaries. Intuitively, we expect times of high investor sentiment on the platform to be characterized by a high number of bullish ideas N_{Bu} relative to the number of bearish ideas N_{Be} and vice versa. In Figure 15, we show that the

[20] Please note, that none of the dictionaries considered in this work have been developed to catch the tone of language used in social media as discussed in Section 3.2.

number of ideas is increasing over time, and thus we need to correct the bullish-bearish spread with the number of classified ideas in total. Following Antweiler and Frank (2004), we define investor sentiment on a given day t derived from a specific dictionary $i = \{EM_{EL}, PN_{EL}, PN_{GI}, PN_{LM}, PN_{HE}\}$ as

$$Sentiment_{i,t} = \frac{N_{Bu,i,t} - N_{Be,i,t}}{N_{Bu,i,t} + N_{Be,i,t}} \tag{24}$$

where $N_{Bu,t}$ and $N_{Be,t}$ are the numbers of bullish and bearish ideas, respectively, on a given day t. The resulting measure is bounded in the interval $[-1,1]$, where a value of 1 denotes the best possible investor sentiment and one of -1 the worst.

3.6 Results

3.6.1 Classification Accuracy

Before we use the derived measure for intraday return prediction, we compare the accuracy of all scored ideas within the analyzed dictionaries. Table 16 shows various descriptive statistics of the summed generated scores for the four dictionaries.

The mean number of scores per idea of EM_{EL}, which is 2.54, is nearly 40% higher as the next highest score of PN_{GI}, which takes a value of 1.83. As the emotion scores contain eight different emotions, it was predictable that EmoLex emotions EM_{EL} would exhibit the highest statistics per idea on average. The emotions 'anticipation' and 'trust' from EM_{EL} exhibit the highest scores on average. As the majority of ideas in our dataset have been classified as bullish, this result was to be expected as well. As already highlighted in Table 14, both of these emotions have a strong bullish connotation and occur especially in bullish ideas.

Nevertheless, the difference in scores between the economic-related and the non-economic-related positive-negative dictionaries attracts our attention. Both non-economic-related dictionaries PN_{EL} and PN_{GI} possess considerably higher scores than PN_{LM} and PN_{HE}, which are economically related. This finding hints at two possible conclusions. On the one hand, the mean score of PN_{HE} might suffer from its short amount of words (see Table 15) relative to PN_{LM}. On the other hand, as PN_{LM} and PN_{HE} possess a mainly economic background, it seems that the language used on social media platforms, in our case StockTwits, is distinct from language in economic texts as for

example financial reports. This finding emphasizes that in addition to the claims 'emotional-based' and 'economic-related', a perfectly designed field-specific dictionary for social media text analysis should focus on the text type used on such platforms.

	Dictionary	Mean	Median	Sd	Min	Max	Unique expressions
EM_{EL}	**All**	**2.54**	**1**		**0**	**112**	**638,712**
	anger	0.23	0	0.54	0	24	19
	anticipation	0.56	0	0.88	0	19	20
	disgust	0.13	0	0.39	0	21	19
	fear	0.25	0	0.56	0	25	22
	joy	0.33	0	0.64	0	16	17
	sadness	0.19	0	0.48	0	24	20
	surprise	0.27	0	0.56	0	12	13
	trust	0.57	0	0.90	0	21	21
PN_{EL}	**All**	**1.23**	**1**		**0**	**59**	**484**
	positive	0.77	0	1.15	0	27	28
	negative	0.46	0	0.82	0	48	31
PN_{GI}	**All**	**1.83**	**1**		**0**	**245**	**1,062**
	positive	1.01	1	1.53	0	245	75
	negative	0.82	0	1.25	0	198	74
PN_{LM}	**All**	**0.36**	**0**		**0**	**107**	**265**
	positive	0.17	0	0.49	0	107	40
	negative	0.19	0	0.50	0	62	33
PN_{HE}	**All**	**0.22**	**0**		**0**	**77**	**208**
	positive	0.15	0	0.45	0	47	33
	negative	0.07	0	0.28	0	76	27

Sample: 01/2012 - 12/2020 ($N = 250,321,511$)

Table 16: Descriptive statistics of generated scores from textual analysis

This conclusion is further strengthened when considering the median score of 0 for PN_{LM} and PN_{HE}, indicating that less than 50% of all ideas contain at least one word that can be classified as positive or negative. Another noteworthy feature in Table 16 is the number of unique expressions of word scores found in the data when classifying ideas with different dictionaries. It becomes apparent that due to its higher dimensionality and mean score, the most unique combinations of scores by far occur when using EM_{EL} (638,712), confirming the proposed ability to capture the underlying complexity of (social media) text in greater detail than two-dimensional approaches do.

To further compare the different dictionaries and their performance, we analyze their respective classification accuracies for the full classified dataset. Table 17 illustrates that when including all 75,414,994 ideas previously classified as `bullish' or `bearish' by the users, EM_{EL} scores highest with an accuracy of 55.73%, followed by both non-economic-related dictionaries PN_{EL} and PN_{GI} with accuracies of 55.37% and 54.66%, respectively. Again, PN_{LM} and PN_{HE} surprisingly obtain the lowest accuracies, with 53.32% and 52.38% at first glance. As explained above, the relatively low accuracy rate of PN_{LM} and PN_{HE} derives from the low classification rate of the words contained in the analyzed ideas.

Dictionary	All Ideas (in %)			Scored Ideas (in %)			
	All	Bullish	Bearish	All	Bullish	Bearish	Loss
EM_{EL}	55.73	53.88	61.12	58.17	56.63	60.64	36.51
PN_{EL}	55.37	54.05	57.94	57.69	57.77	57.62	40.26
PN_{GI}	54.66	53.89	55.79	56.14	57.08	55.41	29.53
PN_{LM}	53.32	51.97	60.30	60.36	60.61	60.11	73.13
PN_{HE}	52.38	60.97	51.33	57.74	62.26	55.66	82.12

Sample: 01/2012 - 12/2020 ($N_{Bu} + N_{Be} = 75,414,994$)
Results differ max. $\pm 0.05\%$ using only for the test datset for validation.

Table 17: Accuracy of scoring by different dictionaries

This is why we further compute the accuracy rate for all ideas that contain at least one scored word in each dictionary. When only considering ideas containing at least one score, all dictionaries experience a growth in accuracy. In particular, the accuracy rates of PN_{LM} and PN_{HE} disproportionately increase to 60.36% and 57.74%, exceeding all other growth rates. Nevertheless, this increase comes with the loss of approximately 73.13% and 82.12%, respectively, of all potentially available ideas. Apparently, a tradeoff between accuracy and data loss exists and needs to be considered when using either dictionary. Therefore, exclusively observing the accuracy rate might not be adequate. Apart from the two economic-related dictionaries PN_{LM} and PN_{HE}, EM_{EL} provides the highest accuracy rate (58.17%) with the second lowest percentage of ideas lost (36.51%).

On this basis, Figure 16 shows the relationship between the share of excluded data and the accuracy of our classification. Gradually, we exclude data points whose prediction values from the trained algorithm are most uncertain by moving simultaneously from 0.5 to 0 (bearish predicts) and 0.5 to 1 (bullish predicts). In general, all dictionaries profit in

accuracy from this operation. Nevertheless, some dictionaries profit more than others. At a data loss level of 95%, the accuracy of the economic-related positive-negative dictionaries reaches around 67% for PN_{LM} or nearly 70% for PN_{HE}, while the non-economic-related positive-negative dictionaries only reach an accuracy of approximately 62% (PN_{GI}) and 66% (PN_{EL}). Furthermore, the emotion-based and non-economic-related dictionary EmoLex (EM_{EL}) performs even stronger than the dictionary by Henry (2008) with around 73% accuracy at a degree of data excluded slightly above 95%. The last mentioned EmoLex dominates all other dictionaries without exception at every degree of excluded data. It is clear that this dominance grows with the number of excluded, most uncertain predicted ideas.

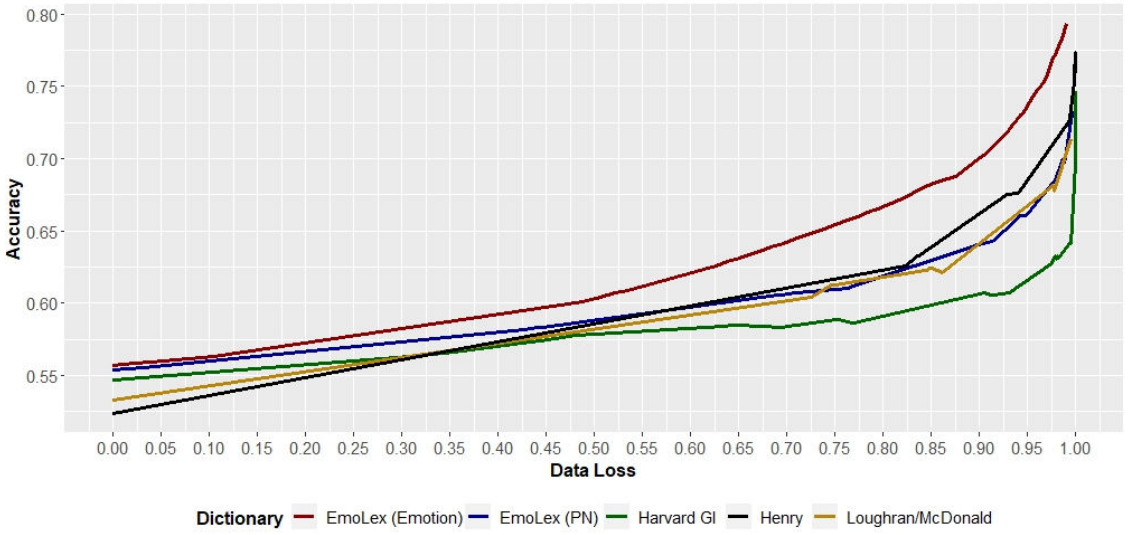


Figure 16: Relationship between the data loss and classification accuracy of different dictionaries

Plotting the histograms of the resulting prediction values for each dictionary suggests the abovementioned observations. As Figure 17 illustrates, all histograms show a high density around a value of 0.5, which is mainly caused by ideas without any score. Consequently, both economic-related dictionaries PN_{LM} and PN_{HE} with the highest rate of unscored ideas show the highest density at approximately 0.5, downgrading their accuracy when using the full dataset. Nevertheless, the prediction values of these dictionaries and EmoLex possess a considerably higher kurtosis than the positive-negative dictionaries (PN_{EM} and PN_{GI}), illustrating their power to classify economic text as 'bullish' or 'bearish' in a more certain way. Observing the tails of the prediction

distributions by calculating the 2.5% and 97.5% quantiles shows that EM_{EL} and PN_{HE} make the safest predictions, leading to the highest accuracy rates when more than 95% of the data are excluded.

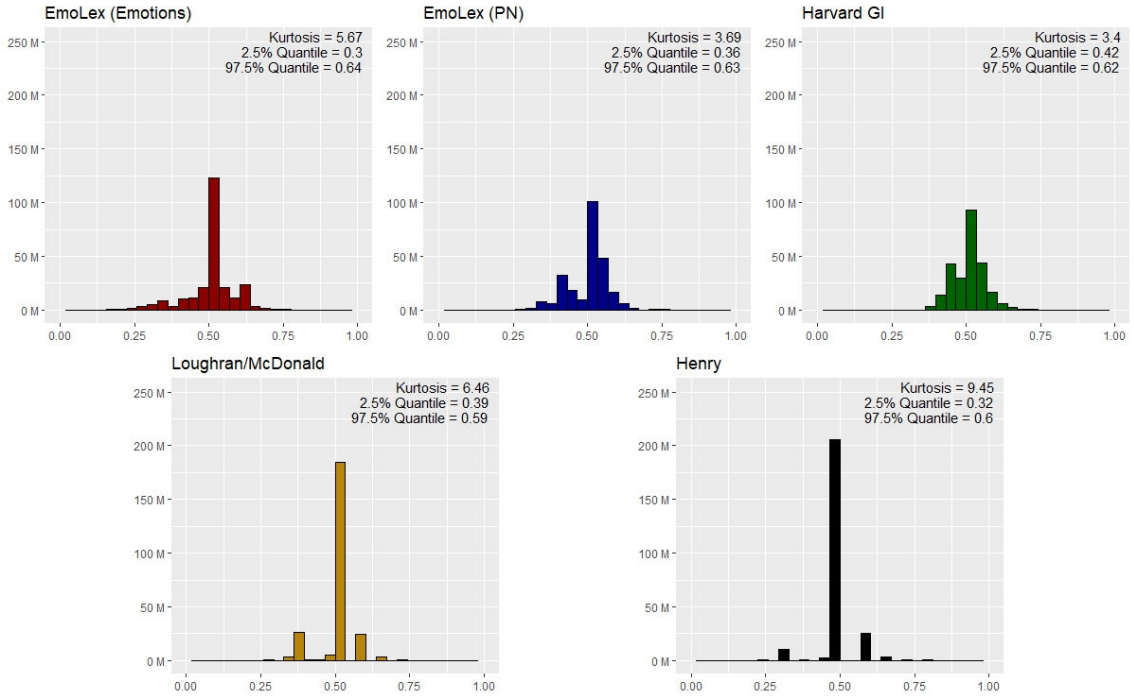


Figure 17: Histograms of prediction values of different dictionaries ($N = 250{,}321{,}511$)

Overall, the data show that the emotion-based dictionary performs slightly better than positive-negative dictionaries using full data. Nevertheless, the feature of multi-dimensionality leads to safer predictions in the tails of the prediction distribution. The same is true for the economic-related dictionary because of the use of accurate connotations. Subletting the data into the three in Table 12 mentioned self-classified trading experience groups (Novice, Intermediate, Professional) and observing the kurtosis of the extracted prediction values within each group gives a hint that the type of the observed text is important for accurate predictions, as well. This finding strengthens prior assumptions found in related literature by e.g. Giannini, Irvine, and Shu (2018).

Table 18 shows that despite the kurtosis of all dictionaries differ, most dictionaries also show the same positive tendency in kurtosis moving to professional text. Assuming that the language used between all groups differs, the need for wordlists addressing the language of other author groups is illustrated as the economic relevance of their posts cannot be ignored.

	EM_{EL}	PN_{EL}	PN_{GI}	PN_{LM}	PN_{HE}
All	**5.67**	**3.69**	**3.40**	**6.46**	**9.45**
Novice	5.55	3.48	3.07	5.56	9.34
Intermediate	5.70	3.53	3.13	5.54	9.06
Professional	6.18	3.56	3.09	5.97	9.23

Table 18: Kurtosis of prediction values of different dictionaries

Consequently, it remains to illustrate the hypothesized economic relevance of our findings. For this purpose, we use the generated scores from all dictionaries to measure investor sentiment and compare their explanatory power for intraday stock returns in the following.

3.6.2 Economic Relevance

As according to Figure 15 most ideas are published around the opening of the US stock market, and we attempt to use this amount of information to predict the intraday return of a specific trading day t by using the shift in sentiment between one hour before market opening of that trading day and the last market hour of the previous trading day $t-1$.[21] In detail, we therefore calculate investor sentiment ($Sentiment_{i,t,m}$) for each dictionary i on trading day t and with a time indicator m subsetting the classified ideas used.

$$Sentiment_{i,t,m} = \begin{cases} m = 1 \; for \; 01.30 \; p.m. \; to \; 02.30 \; p.m. \\ m = 2 \; for \; 08.00 \; p.m. \; to \; 09.00 \; p.m. \end{cases} \quad (25)$$

Hence, we define the shift in investor sentiment ($\Delta Sentiment_{i,t}$) derived by dictionary i on trading day t as:

$$\Delta Sentiment_{i,t} = Sentiment_{i,t,1} - Sentiment_{i,t-1,2} \quad (26)$$

We use this as an explanatory variable explaining the intraday return of the S&P 500 and the NASDAQ 100 on trading day t ($Intraday_t$), as defined in formula (22). As many studies offer the critique that identified relationships between investor sentiment and stock returns might simply be driven by autocorrelation, we introduce the intraday return on the previous trading day $t-1$ as a second explanatory variable to control for this

[21] By doing so, we assume ideas to be published shortly before stock market closing on day $t-1$ mostly contain a summary of that day, while users/investors focus primarily on upcoming events in the hour before market opening of the next trading day t.

effect in our regression model (i.e., Xiong et al. (2019)). Thus, we estimate the following linear model.

$$Intraday_t = \beta_0 + \beta_1 * Intraday_{t-1} + \beta_i * \Delta Sentiment_{i,t} + \varepsilon_t \qquad (27)$$

Table 19 illustrates the results of the regression for both indices for each dictionary. We calculate standardized coefficients ($\tilde{\beta}_1$ and $\tilde{\beta}_i$) because all derived investor sentiment measures have different statistical properties and the size of coefficients between the estimated models would not be comparable.

		$\tilde{\beta}_1$	$\tilde{\beta}_i$	R^2	$Adj.R^2$
S&P 500	EM_{EL}	-0.1031*	0.0646**	0.0151	0.0142
		(-2.3398)	(3.1252)		
	PN_{EL}	-0.1058*	0.0573**	0.0142	0.0134
		(-2.4094)	(2.6168)		
	PN_{GI}	-0.1046**	0.0375	0.0123	0.0115
		(-2.3812)	(1.8901)		
	PN_{LM}	-0.1042*	0.0656**	0.0152	0.0144
		(-2.3839)	(3.1109)		
	PN_{HE}	-0.1069*	0.0833***	0.0179	0.0170
		(-2.4458)	(3.4206)		
NASDAQ 100	EM_{EL}	-0.1523***	0.0641**	0.0273	0.0264
		(-4.5966)	(3.2067)		
	PN_{EL}	-0.1531***	0.0533*	0.0260	0.0252
		(-4.6213)	(2.3474)		
	PN_{GI}	-0.1522***	0.0214	0.0236	0.0228
		(-4.594)	(1.1112)		
	PN_{LM}	-0.1518***	0.0462*	0.0253	0.0245
		(-4.5953)	(2.4250)		
	PN_{HE}	-0.1528***	0.0188	0.0235	0.0227
		(-4.6135)	(0.8673)		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (We use robust White standard errors.)
Regressions estimate equation (28) with the sample: 01/2012 - 12/2020 ($T = 2263$)

Table 19: Intraday return predictability using different sentiment measures

Starting with the full dataset, differences in sentiment for all dictionaries except Harvard GI (PN_{GI}) have a significantly positive influence on the intraday return of the S&P 500 while the influence on NASDAQ 100 returns is for all dictionaries lower and no longer significant for the Henry dictionary (PN_{HE}). Surprisingly, observing the standardized coefficients and the goodness of fit measured by adjusted R^2 shows that (1) the sentiments derived from the EmoLex dictionary by Mohammad and Turney (2013), EM_{EL}, possess

the highest influence on intraday returns with standardized coefficients of 0.0641 for the NASDAQ 100, but (2) the influence for the economic-related dictionaries predicting S&P 500 return is even higher with (highly) significant coefficients of 0.0833 and 0.0656.

As these results are predominantly in line with the accuracy results in Table 17, we also calculate the same regressions while gradually excluding ideas with the most uncertain prediction values. Economically, such indifferent ideas could be interpreted as a 'hold' signal by the publisher. The results of this operation on the standardized coefficients ($\tilde{\beta}_i$) can be observed in Figure 18.

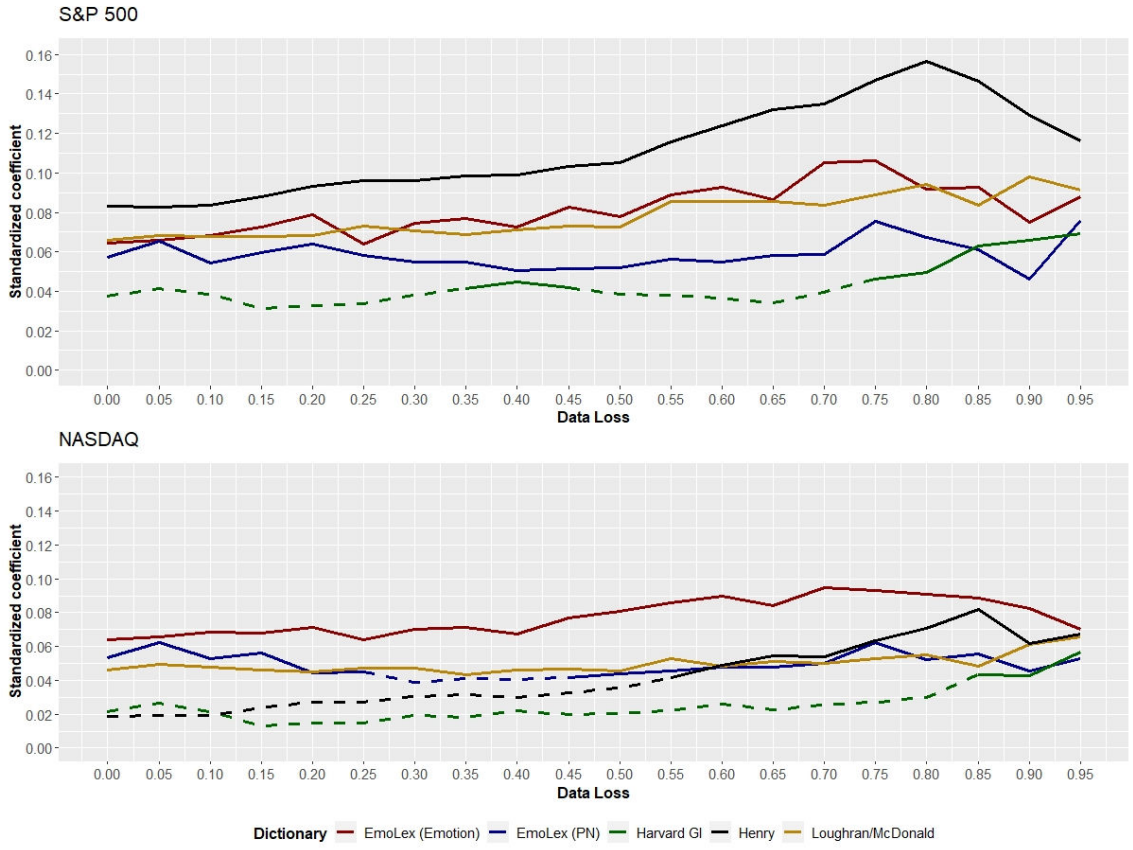


Figure 18: Development of the standardized coefficients (dashed if $p > 0.05$)

First, all dictionaries show a significant positive influence on intraday returns at data loss degrees above 55% despite of the Harvard GI dictionary. Nevertheless, investor sentiment derived from EmoLex or Henry (2008) (and at a high degree of uncertain predictions excluded also from Loughran and McDonald (2011)) dominates the non-economic-related positive-negative dictionaries in predictive power, with standardized coefficients

reaching maximum values of approximately 0.16 and an explained variance of up to 3% measured by adjusted R^2.[22]

Regarding our stated hypotheses (*H1* and *H2*) these findings can only be validated reliably by testing for differences between the coefficients. Therefore, following Clogg, Petkova, and Haritou (1995) and Paternoster et al. (1998) we calculate the Z-scores using the formula:

$$Z = \frac{\widetilde{\beta}_\iota - \widetilde{\beta}_J}{\sqrt{\sigma_{\widetilde{\beta}_\iota} + \sigma_{\widetilde{\beta}_J}}} \tag{28}$$

Proving *H1* we observe the difference between EM_{EL} and PN_{EL}/PN_{GI} in Figure 19. For the S&P 500 as well as the NASDAQ 100 differences for the Harvard GI are positive and highly significant, especially for higher degrees of excluded data. Differences with the related positive-negative version of EmoLex are positive but only become significant for higher levels of data excluded.

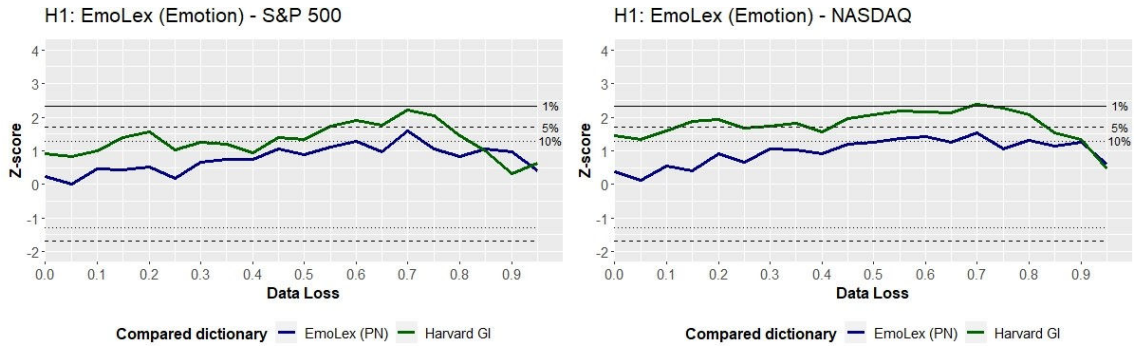


Figure 19: Development of Z-scores proving for *H1*

Proving *H2* we observe more mixed results while testing for differences between PN_{LM}/PN_{HE} and PN_{EL}/PN_{GI}. While for the S&P 500 both economic-related dictionaries overperform the Harvard GI (and Henry also the positive-negative EmoLex version), the Z-scores for the NASDAQ 100 are mainly insignificant around zero.

[22] More detailed regression results can be found in the appendix in Table 21 and Table 22.

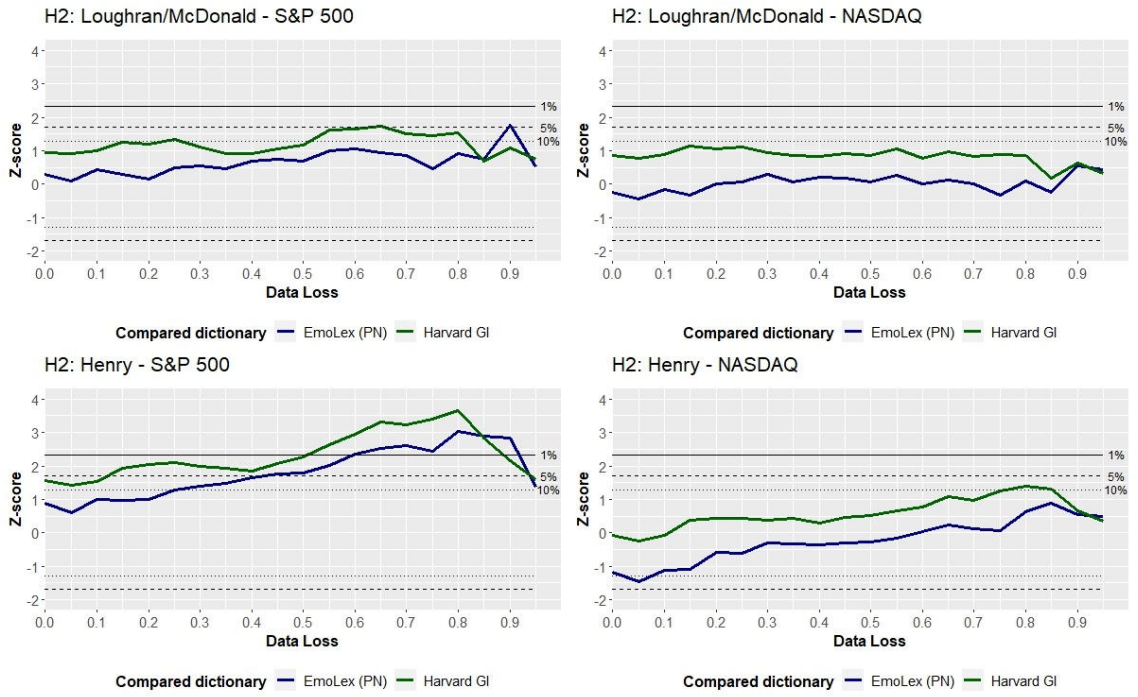


Figure 20: Development of Z-scores proving for *H2*

Overall, also bearing the accuracy results from Section 3.6.1 in mind, our first hypothesis (*H1*) cannot be rejected, as EmoLex emotion scores reach a higher accuracy in classifying 'bullish' or 'bearish' signals. Furthermore, the shift in investor sentiment has greater predictive power than the shift in investor sentiment derived from the non-economic-related positive-negative dictionaries, PN_{EL} and EM_{GI}, at all levels of excluded data.

In line with prior research (for example by Renault (2017)) the same holds true for the second hypothesis (*H2*) as the accuracy and economic relevance of investor sentiment derived with the help of economic-related dictionaries are higher than those of dictionaries without economic relations – even though Z-scores only indicated a significant difference between the coefficients for the S&P 500. If we consider this hypothesis in a more precise way, we also find that within economic-related dictionaries, accuracy and economic relevance can differ. Hence, the results of the dictionary created by Henry (2008), which is based on (financial) earnings press releases, are stronger than those of the dictionary created by Loughran and McDonald (2011), which originated in an accounting background, for our field-specific application.

		SP500		NASDAQ	
		$\tilde{\beta}_i$	$Adj.R^2$	$\tilde{\beta}_i$	$Adj.R^2$
Novice	EM_{EL}	0.0681***	0.0147	0.0428*	0.0250
		(3.6878)		(2.3803)	
	PN_{EL}	0.0646***	0.1420	0.0528**	0.0260
		(3.3921)		(2.8138)	
	PN_{GI}	0.0763***	0.0159	0.0479**	0.0246
		(4.0782)		(2.6742)	
	PN_{LM}	0.0764***	0.0159	0.0518**	0.0250
		(3.8073)		(2.6368)	
	PN_{HE}	-0.0302	0.0110	0.0074	0.0224
		(1.5879)		(0.4191)	
Intermediate	EM_{EL}	0.0488*	0.0125	0.0346	0.0235
		(2.5574)		(1.8814)	
	PN_{EL}	0.0510**	0.0127	0.0337	0.0235
		(2.6280)		(1.6675)	
	PN_{GI}	-0.0018	0.0101	-0.0128	0.0225
		(-0.0949)		(-0.7026)	
	PN_{LM}	-0.0437*	0.0120	0.0290	0.0232
		(2.2568)		(1.5514)	
	PN_{HE}	0.0709***	0.0151	0.0600**	0.0259
		(3.7072)		(3.1817)	
Professional	EM_{EL}	0.0476*	0.0123	0.0599**	0.0259
		(2.423)		(3.0924)	
	PN_{EL}	0.0463*	0.0122	0.0662**	0.0267
		(2.1606)		(3.2627)	
	PN_{GI}	0.0431*	0.0119	0.0407*	0.0240
		(2.2132)		(2.1496)	
	PN_{LM}	0.0364	0.0114	0.0185	0.0227
		(1.4446)		(0.8339)	
	PN_{HE}	0.0961***	0.0193	0.0255	0.0230
		(4.3384)		(1.2016)	
No Group	EM_{EL}	0.0409*	0.1170	0.0410*	0.0240
		(2.0418)		(2.1417)	
	PN_{EL}	0.0157	0.0103	0.0068	0.0224
		(0.7327)		(0.3277)	
	PN_{GI}	0.0186	0.0104	-0.0039	0.0223
		(0.8642)		(-0.2038)	
	PN_{LM}	0.0359	0.0114	0.0271	0.0230
		(1.6339)		(1.4423)	
	PN_{HE}	0.0349	0.0113	-0.0073	0.0224
		(1.7575)		(-0.4032)	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (We use robust White standard errors.)
Regressions estimate equation (28) with the sample: 01/2012 - 12/2020 ($T = 2263$)
Subgroup ideas: Novice 16,172,895 (6.46%) / Intermediate 42,480,800 (16.97%) / Professional 31,464,736 (12.57%) / No Group 160,203,326 (64.00%)

Table 20: Intraday return predictability using different sentiment measures by trader group

Further, our prior analysis indicates that in addition to the economic word connotation and the emotional scoring as a third factor, the type of text plays an important role in deriving investor sentiment from text. By dividing the published ideas into the three self-classified trading groups – 'Novice', 'Intermediate' and 'Professional' – and repeating the regression defined in formula (28) for each of those groups, as is clear from Table 20, the predictive power of most dictionaries reported in the full regression (Table 19) highly differs between user types, this is especially true for the two economic-related dictionaries.

We assume the language used by user who classify themselves as 'Professional' to be that specific that it only fits well for the financial earnings-related dictionary PN_{HE}. For other dictionaries the ideas differ too much from their origin texts which vice versa perform better for text written by users who classify themselves as 'Novice' or 'Intermediate'.[23] Despite the disadvantage of not being economic-related, shifts in investor sentiment from EmoLex emotion scores exhibit their slightly weaker predictive power for intraday returns by each of the three (or rather four) groups, which makes it more applicable for analyzing text from which the trader group of the publisher is unknown, as is common in most social media text data.

3.7 Conclusion

The field of social media sentiment analysis is fast moving due to the rapid growth of data published on platforms such as Twitter, Facebook or, in our case, StockTwits. This makes it necessary to regularly reevaluate preexisting research, adjust former methodologies and propose new methodologies.

The first part of this paper addressed the economic background of sentiment analysis and why it might be desirable to consider sentiment analysis when attempting to predict stock market movements. We do so by reviewing the preceding related literature. In the following, we present our obtained dataset, which we generated from the microblogging platform StockTwits. Furthermore, we provide detailed insight into the way we extract eight emotions from ideas published on StockTwits by using the NRC Word-Emotion Association Lexicon (EmoLex), enabling us to correlate these underlying emotions with

[23] For different degrees of data excluded no further findings could be made. Nevertheless, all relevant figures are reported in the appendix in Figure 21.

an individual's self-revealed bullish or bearish sentiment by using machine learning algorithms.

Consequently, this allows us to classify further ideas that have not been classified into bullish or bearish sentiment categories, thereby enriching our database. We make use of this extended database, which comprises approximately 250 million classified ideas, to correlate our sentiment findings with US stock market performance. We find that investor sentiment classified by emotion scores can be used to predict stock market movements weakly but more accurately than positive-negative scores derived from non-economic-related dictionaries. Although we are able to classify more ideas than the analyzed two-dimensional dictionaries, many ideas still cannot be scored by our approach due to missing occurrences of ideas' words in dictionaries' wordlists. This weakness is in line with prior research results and illustrates the need for further improvement.

In detail, our results define three main factors that determine the success of deriving investor sentiment with the help of textual sentiment in an economic context: multi-dimensional scoring (for example emotions), economic word connotation and type of text. By using supervised machine learning algorithms without taking common benchmark dictionaries into account, many researchers address those three factors with mostly noteworthy results. Nevertheless, for example, Renault (2017) comes to the same conclusion as Kearney and Liu (2014) and shows that field-specific dictionaries are more applicable than rough benchmark dictionaries as well as machine learning algorithms. Since the classification of a text by its publisher for training purposes, as in StockTwits data, is a rare feature of text data and self-classification of text by researchers often leads to misclassification, there remains a need to create multiple dictionaries addressing those three factors.

In accordance to our results it can be expected that more advanced dictionaries (Figure 13, A3) or NLP Transformers might also profit from the factors outlined above - especially multi-dimensional scoring. These considerations therefore give a reasoning for the recent emergence of field-specific and emotion-based NLP transformers (as for example specifications of RoBERTa/DistilRoBERTa).[24]

[24] An overview of various specification can be found on the model hub for NLP: Hugging Face.

3.8 Appendix

		Data Loss				
		10%	30%	50%	70%	90%
EM_{EL}	$\tilde{\beta}_1$	-0.1033*	-0.1025*	-0.1029*	-0.1032*	-0.1045*
		(-2.3451)	(-2.3353)	(-2.3465)	(-2.3627)	(-2.3862)
	$\tilde{\beta}_i$	0.0680**	0.0745***	0.0798***	0.1053***	0.0751***
		(3.2626)	(3.7124)	(3.7523)	(5.1316)	(3.8235)
	R^2	0.0156	0.0165	0.0170	0.0220	0.0166
	$Adj.R^2$	0.0147	0.0156	0.0161	0.0212	0.0157
PN_{EL}	$\tilde{\beta}_1$	-0.1057*	-0.1059*	-0.1059*	-0.1050*	-0.1045*
		(-2.4058)	(-2.4228)	(-2.4170)	(-2.4018)	(-2.3838)
	$\tilde{\beta}_i$	0.0544*	0.0546*	0.0521**	0.0585**	0.0462*
		(2.5038)	(2.5574)	(2.6970)	(3.1564)	(2.4923)
	R^2	0.0139	0.0139	0.0137	0.0144	0.0131
	$Adj.R^2$	0.0130	0.0131	0.0128	0.0135	0.0122
PN_{GI}	$\tilde{\beta}_1$	-0.1047*	-0.1046*	-0.1044*	-0.1049*	-0.1052*
		(-2.3848)	(-2.3836)	(-2.3782)	(-2.3918)	(-2.3998)
	$\tilde{\beta}_i$	0.0384	0.0378*	0.0385*	0.0396*	0.0658**
		(1.9375)	(1.9798)	(2.0029)	(2.1778)	(3.5362)
	R^2	0.0124	0.0124	0.0124	0.0125	0.0152
	$Adj.R^2$	0.0115	0.0115	0.0116	0.0116	0.0144
PN_{LM}	$\tilde{\beta}_1$	-0.1043*	-0.1042*	-0.1043*	-0.1042*	-0.1042*
		(-2.3870)	(-2.3878)	(-2.3883)	(-2.3910)	(-2.3812)
	$\tilde{\beta}_i$	0.0676**	0.0707**	0.0726**	0.0838***	0.0979***
		(3.1862)	(3.2686)	(3.4762)	(3.9877)	(4.9409)
	R^2	0.0155	0.0159	0.0162	0.0180	0.0205
	$Adj.R^2$	0.0146	0.0151	0.0153	0.0171	0.0197
PN_{HE}	$\tilde{\beta}_1$	-0.1070*	-0.1079*	-0.1079*	-0.1091*	-0.1073*
		(-2.4458)	(-2.4696)	(-2.4721)	(-2.5154)	(-2.4513)
	$\tilde{\beta}_i$	0.0836***	0.0961***	0.1052***	0.1349***	0.1294***
		(3.4497)	(4.2358)	(4.9281)	(6.3827)	(6.3361)
	R^2	0.0179	0.0202	0.0220	0.0291	0.0277
	$Adj.R^2$	0.0171	0.0193	0.0211	0.0283	0.02687

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (We use robust White standard errors.)
Regressions estimate equation (28) with the sample: 01/2012 - 12/2020 ($T = 2263$)

Table 21: S&P 500 intraday return predictability using different sentiment measures at different degrees of excluded uncertain predictions

		Data Loss				
		10%	30%	50%	70%	90%
EM_{EL}	$\tilde{\beta}_1$	-0.1525***	-0.1522***	-0.1524***	-0.1529***	-0.1527***
		(-4.6052)	(-4.5947)	(-4.6054)	(-4.6259)	(-4.6232)
	$\tilde{\beta}_i$	0.0683**	0.0700***	0.0807***	0.0948***	0.0823***
		(3.3783)	(3.6775)	(3.9760)	(4.7052)	(4.3509)
	R^2	0.0279	0.0281	0.0297	0.0322	0.0300
	$Adj.R^2$	0.0270	0.0272	0.0288	0.0313	0.0291
PN_{EL}	$\tilde{\beta}_1$	-0.1531***	-0.1530***	-0.1532***	-0.1527***	-0.1526***
		(-4.6213)	(-4.6176)	(-4.6236)	(-4.6109)	(-4.6113)
	$\tilde{\beta}_i$	0.0526*	0.0387	0.0438*	0.0499**	0.0454*
		(2.3474)	(1.8911)	(2.3668)	(2.7391)	(2.5473)
	R^2	0.0259	0.0247	0.0251	0.0257	0.0252
	$Adj.R^2$	0.0251	0.0238	0.0242	0.0248	0.0244
PN_{GI}	$\tilde{\beta}_1$	-0.1522***	-0.1523***	-0.1522***	-0.1541***	-0.1526***
		(-4.5947)	(-4.5999)	(-4.5966)	(-4.6035)	(-4.6075)
	$\tilde{\beta}_i$	0.0216	0.0191	0.0203	0.0253	0.0428*
		(1.1067)	(1.0505)	(1.221)	(1.4427)	(2.3948)
	R^2	0.0236	0.0235	0.0236	0.0238	0.0250
	$Adj.R^2$	0.0228	0.0227	0.0227	0.0230	0.0241
PN_{LM}	$\tilde{\beta}_1$	-0.1519***	-0.1520***	-0.1520***	-0.1517***	-0.1514***
		(-4.5971)	(-4.6048)	(-4.6062)	(-4.5944)	(-4.5763)
	$\tilde{\beta}_i$	0.0477*	0.0471*	0.0455*	0.0498*	0.0614**
		(2.4708)	(2.4640)	(2.4888)	(2.5340)	(3.1260)
	R^2	0.0255	0.0254	0.0252	0.0257	0.0267
	$Adj.R^2$	0.0246	0.0245	0.0244	0.0248	0.0261
PN_{HE}	$\tilde{\beta}_1$	-0.1529***	-0.1532***	-0.1533***	-0.1541***	-0.1531***
		(-4.6145)	(-4.6260)	(-4.6282)	(-4.6487)	(-4.6120)
	$\tilde{\beta}_i$	0.0193	0.0302	0.0356	0.0538**	0.0618**
		(0.8788)	(1.3968)	(1.7201)	(2.6888)	(3.042)
	R^2	0.0236	0.0241	0.0244	0.0261	0.0270
	$Adj.R^2$	0.0227	0.0232	0.0236	0.0252	0.0261

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (We use robust White standard errors.)
Regressions estimate equation (28) with the sample: 01/2012 - 12/2020 ($T = 2263$)

Table 22: NASDAQ 100 intraday return predictability using different sentiment measures at different degrees of excluded uncertain predictions

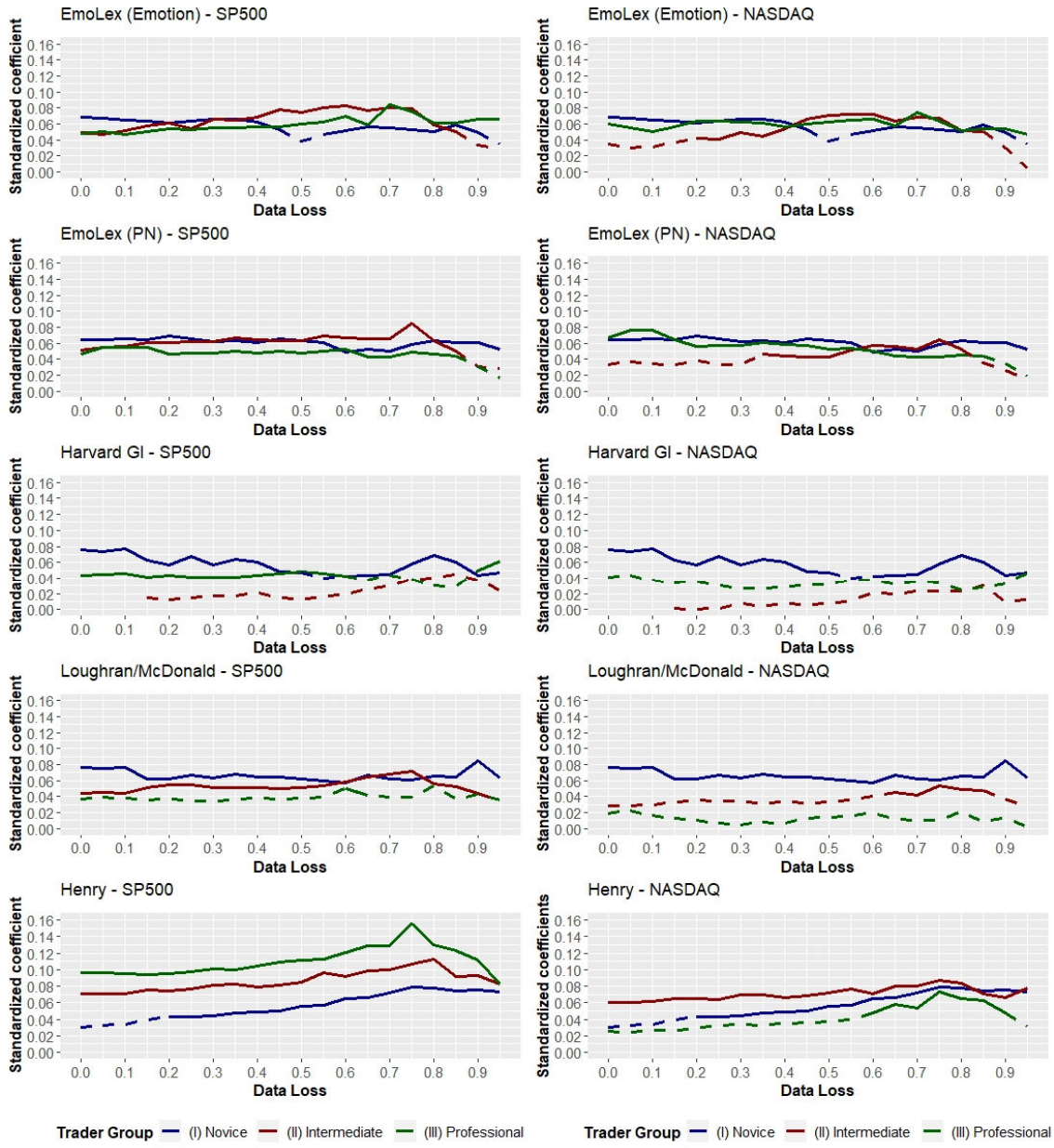


Figure 21: Development of the standardized coefficients (dashed if $p > 0.05$)

3.9 Declaration of (co-)authors and record of accomplishments

Title: Measuring investor sentiment from Social Media Data – an emotional approach

Author(s): **Philipp Stangor** (Heinrich Heine University Düsseldorf)
Lars M. Kürzinger (Heinrich Heine University Düsseldorf)

Conferences: Participation and presentation at 'HVB Doctoral Colloquium', 15th – 16th January 2021, online hosted by University of Paderborn, Germany

Participation and presentation at '58th Annual Meeting of the Eastern Finance Association (EFA)', 6th – 9th April 2022, Washington D.C. United States of America

Participation and presentation at 'HeiCAD Lightning Talks', 24th June 2022, Düsseldorf, Germany

Publication: SSRN published. Submitted to 'The Quarterly Review of Economics and Finance', double-blind peer-reviewed journal. Current status: Under review.

Share of contributions:

Contributions	**Philipp Stangor**	**Lars M. Kürzinger**
Research Design	**85%**	**15%**
Development of Research Question	90%	10%
Method development	80%	20%
Research performance & analysis	**70%**	**30%**
Literature Review and framework development	20%	80%
Data collection, preparation and analysis	90%	10%
Analysis and discussion of results	90%	10%
Derivation of implications and conclusions	80%	20%
Manuscript preparation	**100%**	**0%**
Final draft	100%	0%
Finalization	100%	0%
Overall contribution	**85%**	**15%**

31.05.2024,

Date, Philipp Stangor

31.05.2024,

Date, Lars M. Kürzinger

4 How do you talk finance on social media? – Extracting and identifying emotions from different trader groups

4.1 Abstract

This paper contributes to the research of economic textual analysis sentiment with an advanced approach for extracting emotions from social media utilizing Google BERT. Additionally, it will be shown that used language in an economic context on a social media platform, as in this case StockTwits, differs between the defined trader groups 'Novice', 'Intermediate' and 'Professional'. The paper empathizes dividing the training of neural networks for the purpose of defining emotions and sentiment with textual analysis on social media platforms utilizing Amazon MTurk and shows that higher classification accuracies can be reached by doing so. Additionally, in an economic context achieved sentiment measures profit from this split into trader groups on social media predicting intraday returns.

4.2 Introduction

One of the most important challenges in understanding noise trader behavior following Kyle (1985), Black (1986) and Long et al. (1990) became extracting investor or market sentiment more and more. In economic and financial research different methods have been developed to measure investor sentiment roughly subdivided into market-based, survey-based and text-/media-based methods.[25]

Especially investor sentiment derived from text became increasingly popular in research with the increasing availability and accuracy of developed methods and availability of (real-time) data. Starting with Antweiler and Frank (2004) predicting market volatility and trading volume from Internet stock messages many measures and procedures have been introduced in financial research. Tetlock (2007) constructs a measure of media pessism by simple counting negative words in the Wallstreet Journal column 'Abreast of the Market' with which he is able to predict stock market movements. In this manner, multiple word-dictionaries have been developed analyzing different types of text as for

[25] Zhou (2018) and Aggarwal (2022) give great overviews over the advances made in economic research.

example 10-K SEC filings (Loughran and McDonald (2011)), social media platforms (Renault (2017)) or earnings press releases (Henry (2008)).

Mishev et al. (2020) show in a chronological study the whole development from those lexicon-based methods over statistical methods (as count vectors (CV) & term frequency – inverse document frequency (TF-IDF)), word-encoders (as Word2Vec (Mikolov et al. (2013)) or ELMo (Peters et al. (2018))) as well as sentence-encoders (as Doc2Vec (Le and Mikolov (2014)) or LASER (Artetxe and Schwenk (2019))) to NLP transformers. They come to the conclusion that utilizing and fine-tuning NLP transformers as the many accessible versions of BERT have been the most promising approach in terms of accurately classifying (financial) text data. Consequentially, many researchers have developed extensions as DisitilBERT, FinBERT, etc. leading to nearly 10,000 text classification BERT models available on *Huggingface.co*.

The fine-tuned models often differ in the type of text used for training as well as in the output being two- (positive/negative) or multi-dimensional (e.g. emotional scoring). Section 3 has shown that multi-dimensional/multi-class scoring can be beneficial in terms of classification accuracy and economic relevance of derived investor sentiment what is going in line with the increasing development of multi-class models as for example findings in the field of Text-Based Emotion Detection (TBED, see Zad et al. (2021) for an overview) and especially 'EmTract' introduced by Vamossy and Skog (2020) as the first one applied in a financial context. Further, Section 3 has shown that intraday return predictability differs between the three defined trader groups 'Novice', 'Intermediate' and 'Professional' concluding that there is a need to get more into detail in the variety of social media language in economic context as it is unanswered whether different abilities to predict returns are determined by the quality of extracting sentiment or a different influence of information produced by those three trader groups.

With this in mind, the following work aims at developing a fine-tuned model for extracting emotions from social media text of different trader groups. I utilize the StockTwits ideas trained BERT model 'EmTract' by Vamossy and Skog (2020) and create three new models by finetuning only using 'Novice', 'Intermediate' or 'Professional' ideas which have been classified by workers on Amazon's platform MTurk. With this a part of the critique by Aggarwal (2022) that 'the scope of examining

IS [Investor Sentiment] needs to be enlarged with respect to the markets and the types of investors being studied' will be addressed.

The remainder of this paper is structured as follows: Section 4.3 presents the data and data collection, namely, the ideas from the social media platform StockTwits and the derived tweet classifications from workers on MTurk. Further, an introduction how the models will be fine-tuned concludes this Section. In Section 4.4 the fit of the three fine-tuned models will be evaluated and an analysis of the economic relevance will be done by trying to predict intraday returns of S&P 500 und NASDAQ 100. Section 4.5 concludes the paper, relates the observations made with prior results found in the literature and provides an outlook on possible future research topics.

4.3 Data & Methodology

4.3.1 Data & tweet characteristics

The data basis consists of all tweets[26] published on the microblogging platform StockTwits between January 2012 and December 2020. StockTwits is comparable to Twitter but is specifically tailored for users interested in discussing financial market topics. This ensures that the shared information and resulting sentiment are closely related to the financial markets. During this period, there were 250,321,511 tweets[27] available, with a tendency for more tweets in recent years due to platform growth.

The key advantage of this platform in the context of this work is that additional trading-related information about the users and the tweets they post is available. Users are free to classify themselves into three groups – 'Novice', 'Intermediate' and 'Professional' – based on their trading experience. Since this is a self-assessment, biases (e.g., due to over- or underconfidence) cannot be ruled out. However, it is assumed that outliers among the 40,000 daily active users in 2020 do not significantly influence the observations, and differences in trading experience between the groups can be considered as:

$$Professional > Intermediate > Novice$$

[26] In the StockTwits context, the posts are actually called 'ideas'. However, since these are equivalent to tweets on Twitter and the term 'tweet' is more commonly known in general language usage, I will also refer to the posts as 'tweets' in this work.

[27] After clearing the data of ideas that are not suitable for the measurement of textual sentiment, for example, ideas that only contain 'cashtags' as identifiers for several stocks ($), pictures or hyperlinks.

The second helpful characteristic of the data is that users can also tag their tweets as 'bullish' or 'bearish' signals, which will facilitate the creation of a more balanced sample for training the language models. Since the disclosure of both pieces of information is voluntary and not maintained by all users, the dataset is reduced to 23,324,809 tweets that contain information about the poster's trading experience as well as a 'bullish' or 'bearish' tag. Table 23 shows the number of tweets per group as well as 'bullish' or 'bearish' sentiment and provides hints on possible differences in syntax between the groups based on tweet length.

	N	**Length (Words)**				
		Min	Max	Mean	Median	Sd
Novice	**4,464,121**	**1**	**1058**	**8.17**	**6**	**8.31**
Bullish	3,889,215	1	1058	8.22	6	8.43
Bearish	574,906	1	739	7.89	6	7.46
Intermediate	**12,404,616**	**1**	**1450**	**8.59**	**7**	**9.28**
Bullish	10,335,627	1	1450	8.65	7	9.50
Bearish	2,068,989	1	1143	8.27	6	8.11
Professional	**6,456,072**	**1**	**1561**	**8.62**	**7**	**9.08**
Bullish	4,756,418	1	1561	8.70	7	9.57
Bearish	1,699,654	1	978	8.39	7	7.54

Table 23: Tweet characteristics per trader group

It can be observed that the group of 'Intermediate' traders accounts for the most tweets, followed by 'Professional' and 'Novice' traders. It is also worth noting that across all groups, significantly more 'bullish' tweets have been written than 'bearish' ones. This is attributed to both the predominantly bullish markets during the observation period and the tendency for people to share positive news more often. However, this imbalance must be considered later during the training of the language models, as otherwise, the model could be biased towards the bullish direction.[28]

[28] For example, the BERT model 'EmTract' by Vamossy and Skog (2020) shows a slight tendency to classify tweets as 'happy' that I will try to tackle by balancing the training data with equally 'bullish' and 'bearish' tweets again.

Regarding the length of tweets, there is a slight tendency towards longer tweets for more experienced trader groups, both in terms of the mean and the median. Figure 22 illustrates the differences between the groups more clearly using empirical distribution functions. The ‘Novice’ and ‘Intermediate’ groups exhibit a similar density function, while tweets from the ‘Professionals’ show a distinctly different trend towards longer tweets. Tweets with lengths exceeding 50 words are exceptions in this regard.

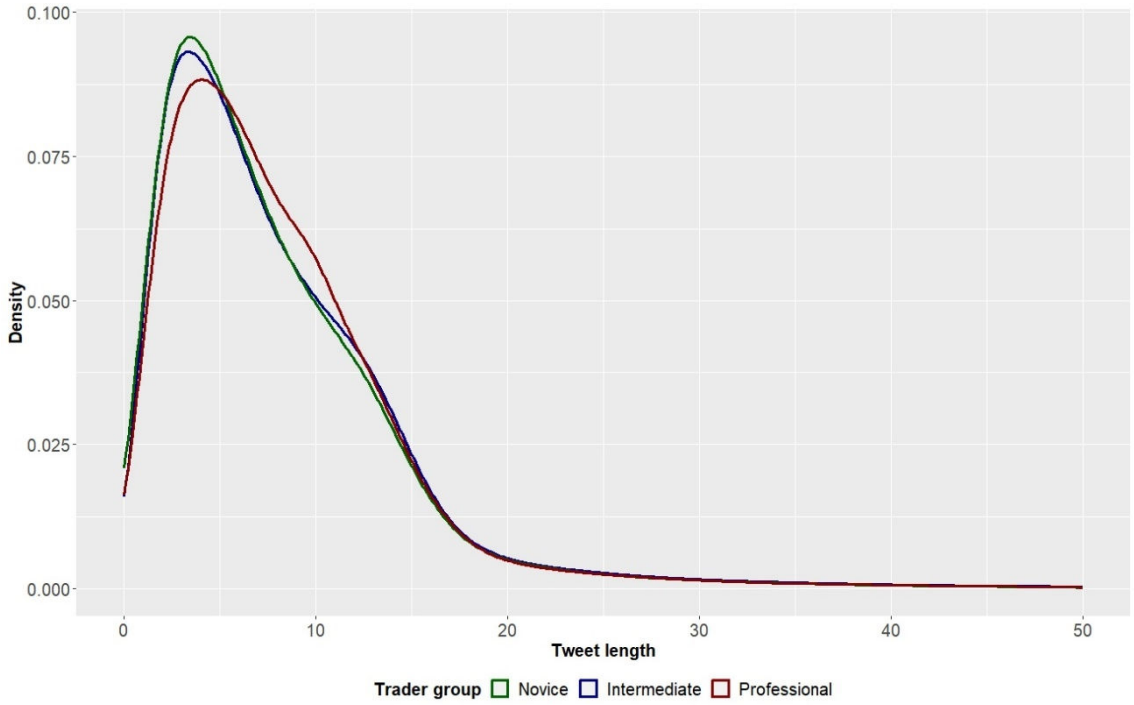


Figure 22: Tweet lengths per trader group

Besides this rather syntactically oriented observation, there are also substantive or linguistic differences between the groups. Figure 23 shows the most frequently used words after the tweets have been lemmatized[29] and cleaned of stop words. At first glance, only marginal differences are apparent, which is why it is worth taking a closer look at the words with the greatest differences in frequency of use between the groups, as depicted in Figure 24.

[29] Lemmatizing words in a string attempts to group inflected forms into a single group (’lovers’ and ’loves’ to ’love’). In this work, I apply the lemmatization list by Mechura (2016) which can be accessed in the R package 'textstem' by Rinker (2018).

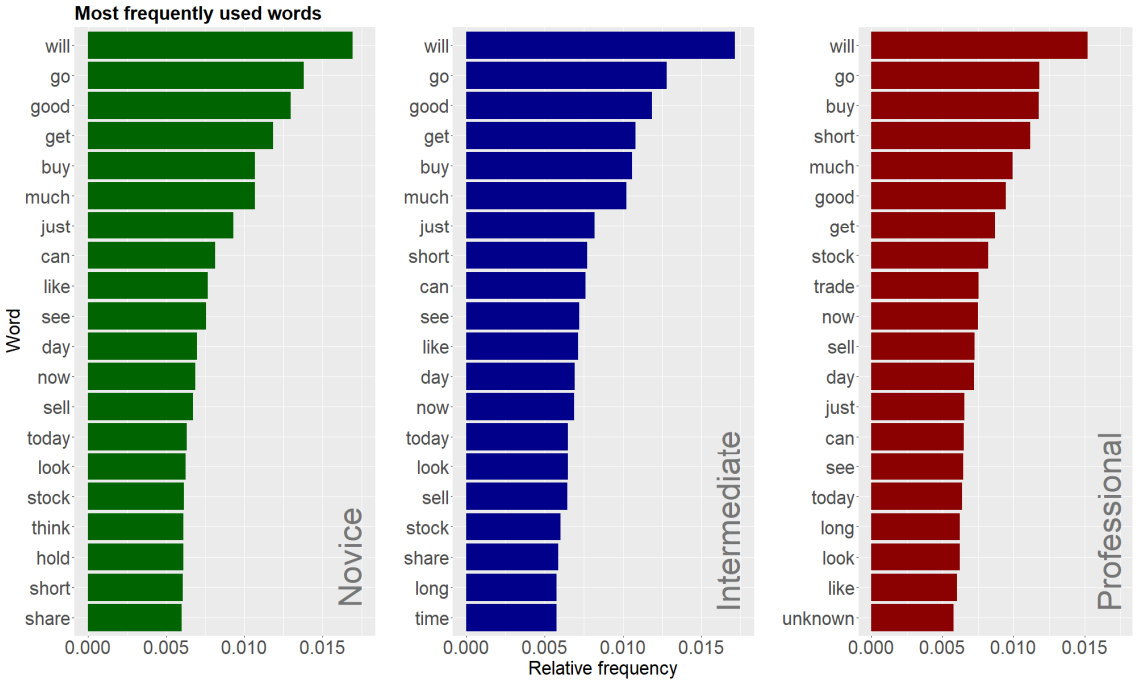


Figure 23: Words with highest frequency of usage between trader groups

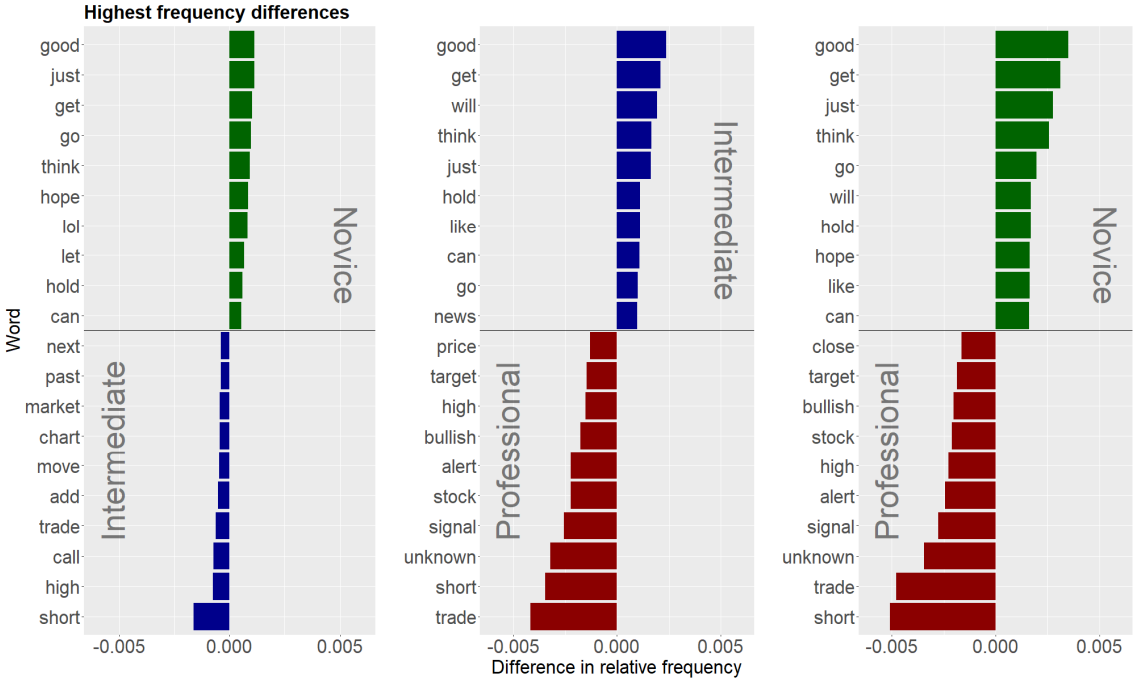


Figure 24: Words with highest difference in frequency of usage between trader groups

In this content analysis as well, it is noticeable that the differences between ‘Novice’ and ‘Intermediate’ tweets are smaller, and only the ‘Professionals’ clearly distinguish

themselves from the other groups by their more frequent use of 'trading terms'. One way to further objectify these observations is to assess language similarity using a Vector Space Model such as Cosine Similarity (CS) (see i.e. Li and Han (2013)). Cosine Similarity refers to the cosine of the angle between two vectors (here $\overrightarrow{Tweet1}$ and $\overrightarrow{Tweet2}$) using the standard dot product and Euclidean norms in an N-dimensional vector space as follows:

$$CS\left(\overrightarrow{Tweet1}, \overrightarrow{Tweet2}\right) = \frac{\overrightarrow{Tweet1} * \overrightarrow{Tweet2}}{\left|\overrightarrow{Tweet1}\right| \left|\overrightarrow{Tweet2}\right|} = \frac{\sum_{i=1}^{N} Tweet1_i * Tweet2_i}{\sqrt{\sum_{i=1}^{N} Tweet1_i^2} * \sqrt{\sum_{i=1}^{N} Tweet2_i^2}} \quad (29)$$

The length N of the vectors is determined by the number of different words used in both vectors (tweets). In the text context, cosine similarity measures, on a scale between 0 (no correlation) and 1 (complete correlation), the similarity between two texts in terms of their language usage.

	Mean	Sd	Max
Novice → Novice	0.0663	0.0985	0.9799
Novice → Intermediate	0.0671	0.0998	1.0000
Novice → Professional	0.0609	0.0938	0.8901
Intermediate → Intermediate	0.0663	0.0966	1.0000
Intermediate → Professional	0.0625	0.0943	0.9701
Professional → Professional	0.0583	0.917	1.0000

	Novice	Intermediate	Professional
Novice		-0.0007	0.0054***
		(-0.8184)	(6.1775)
Intermediate	-0.0008		0.0038***
	(-0.9271)		(4.3579)
Professional	-0.0026**	-0.0042***	
	(-3.1108)	(-4.9614)	

Table 24: Cosine similarity descriptives and differences between trader groups

For verification, two tweets per group were randomly selected, and the similarity between the first tweet between groups, as well as within a group with the second tweet, was computed. On average, the similarity between two randomly selected tweets across all comparisons was relatively low, ranging from 0.0583 to 0.0671 (see Table 24). However, it's also worth noting that, while Novice and Intermediate tweets exhibit similar cosine similarities (even among themselves), the group of Professionals shows a more distinct difference. Applying t-statistics to the data comparing mean values states the same picture as earlier mentioned in Figure 24 draws the same picture. Differences between the groups 'Novice' and 'Intermediate' are not significant while Novice and Intermediate tweets are more similar within those groups compared to Professional's tweets. Nevertheless, Professional tweets seem to be so diverse that they have the lowest cosine similarity with themselves.

All these implications suggest that language on social media platforms is not uniform and therefore should be considered in segmented fashion, as discussed in Section 4.2. Despite to the expected similarity of the groups 'Novice' and 'Intermediate', the following Section aims to train all three language models for the trader groups 'Novice', 'Intermediate' and 'Professional' using supervised machine learning, aiming to better capture the diversity of language used on the platform for research purposes.

4.3.2 Crowdsourcing emotion labels

The centerpiece of the training is the classification of tweets for each trader group to process them in a (supervised) machine learning model. In such a framework, to avoid unnecessary effort and typically enormous costs associated with collecting these classifications, *transfer learning* can be employed. In transfer learning, existing, similar language models can be fine-tuned with fewer new data compared to training a completely new model.

The classifications are collected through multiple participants on the MTurk platform (formerly i.e. used by Mohammad and Turney (2013) for emotion labeling) operated by Amazon. On MTurk, requesters can post small classification tasks, known as Human Intelligence Tasks (HITs), which are then completed by workers[30], who get paid for their

[30] Following Difallah, Filatova, and Ipeirotis (2018), workers are mostly from the US and India, balanced in terms of gender, and born after 1980 (60%), leading to a younger worker population compared to the US population.

work. One advantage of this platform is that classifications can be assigned to workers, often specialized in such tasks, and are typically processed quickly due to the large pool of available workers.[31] In this regard, Shank (2016) particularly recommends MTurk for pilot studies to assess general effectiveness before conducting a more costly, larger study. Various research contributions have already addressed ensuring the reliability of the collected data and developed best practices (e.g., Behrend et al. (2011), Buhrmester, Talaifar, and Gosling (2018), Landers and Behrend (2015)). In general, studies advise offering fair working conditions (clear instructions, transparent rejection policies, appropriate hourly wage, providing contact information, short approval time), filtering for reliable workers (HIT approval rate >95%, significant HIT history, and if necessary: specified demographics and characteristics), and ensuring reliable work through the task (e.g., including attention questions, checks for bot-like behavior).

This research aims to provide an exploratory first training and to stimulate further research in this area with the resulting outcomes in Section 4.4. Consequently, initially, 1,500 tweets per trader group (totaling 4,500 tweets) will be classified by the workers. Each tweet will be classified by five different workers to prevent both random misclassifications (e.g., with only one classification) and too many unclear results (e.g., with two different classifications after two assessments). Referring to the assumed superiority of multi-dimensional/multi-class classification mentioned in Section 3, the 'basic emotions' proposed by Ekman (1992) will be utilized, which play a crucial role in text-based emotional detection (TBED). By increasing the classification options from two or three ('positive', 'negative', 'neutral') to seven ('anger', 'disgust', 'fear', 'neutral', 'happy', 'sad', 'surprise'), the importance of multiple classifications of a tweet becomes evident, albeit at the expense of the quantity of classified tweets in the initial analysis.

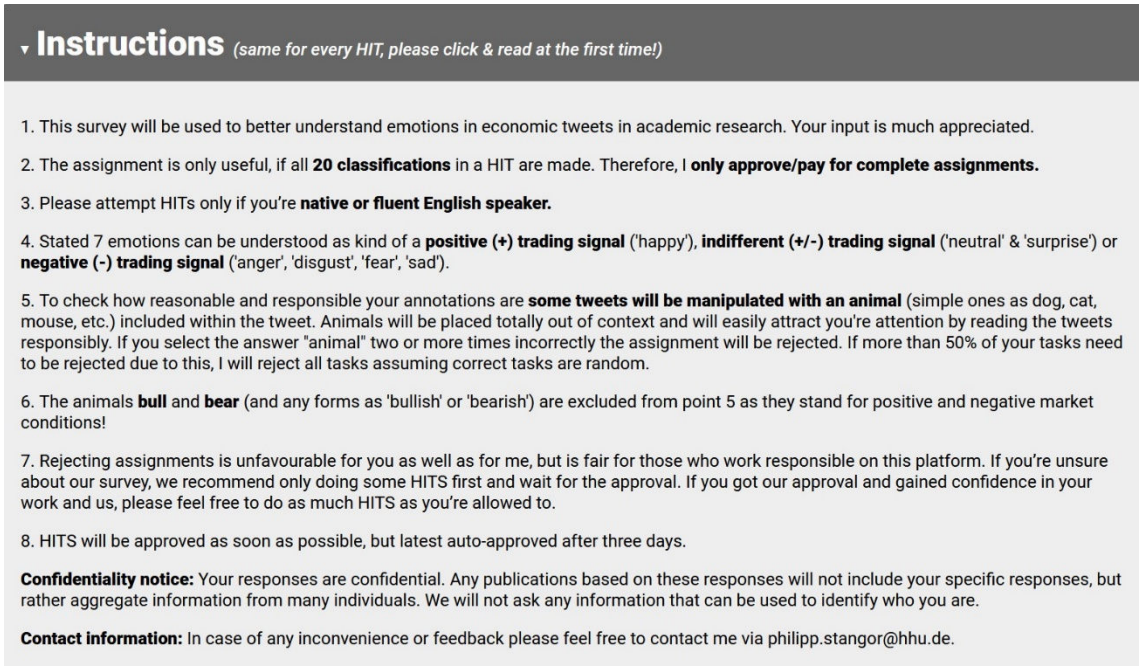
Since the quantity of tweets to be classified is limited, the first step is to identify tweets that particularly correspond to the language of each group. For this purpose, a score is calculated based on the mean of the relative word frequencies (see Figure 23 again for words with the highest frequency of usage per group) of all words i in the set N in each

[31] A good overview of the benefits and criticisms of using MTurk for research can be found in the literature, as discussed i.e. by Young and Young (2019).

tweet. This score reflects the increasing conformity of the tweet to the group with a higher value:

$$Score = \frac{1}{N}\sum_{i=1}^{N} WordFrequency_i \tag{30}$$

To avoid overly biased word choices or even 'spamming' of words, only tweets that use at least three different words are considered for analysis. Furthermore, since bullish markets have predominated in the sample period and users tend to share positive signals more often, a balanced sample (50:50) is created for each trader group based on the integrated 'bullish'/'bearish' classification in StockTwits.



Instructions *(same for every HIT, please click & read at the first time!)*

1. This survey will be used to better understand emotions in economic tweets in academic research. Your input is much appreciated.

2. The assignment is only useful, if all **20 classifications** in a HIT are made. Therefore, I **only approve/pay for complete assignments.**

3. Please attempt HITs only if you're **native or fluent English speaker.**

4. Stated 7 emotions can be understood as kind of a **positive (+) trading signal** ('happy'), **indifferent (+/-) trading signal** ('neutral' & 'surprise') or **negative (-) trading signal** ('anger', 'disgust', 'fear', 'sad').

5. To check how reasonable and responsible your annotations are **some tweets will be manipulated with an animal** (simple ones as dog, cat, mouse, etc.) included within the tweet. Animals will be placed totally out of context and will easily attract you're attention by reading the tweets responsibly. If you select the answer "animal" two or more times incorrectly the assignment will be rejected. If more than 50% of your tasks need to be rejected due to this, I will reject all tasks assuming correct tasks are random.

6. The animals **bull** and **bear** (and any forms as 'bullish' or 'bearish') are excluded from point 5 as they stand for positive and negative market conditions!

7. Rejecting assignments is unfavourable for you as well as for me, but is fair for those who work responsible on this platform. If you're unsure about our survey, we recommend only doing some HITS first and wait for the approval. If you got our approval and gained confidence in your work and us, please feel free to do as much HITS as you're allowed to.

8. HITS will be approved as soon as possible, but latest auto-approved after three days.

Confidentiality notice: Your responses are confidential. Any publications based on these responses will not include your specific responses, but rather aggregate information from many individuals. We will not ask any information that can be used to identify who you are.

Contact information: In case of any inconvenience or feedback please feel free to contact me via philipp.stangor@hhu.de.

Figure 25: MTurk instructions

Thus, with 4,500 tweets to be classified and 5 classifications per tweet, there are 22,500 classifications that need to be carried out on MTurk. An additional 2,500 classifications are allocated for incorporating 'attention/language checks.' In these checks, further tweets are randomly manipulated with animals (i.e., 'dog', 'cat', 'chicken'). Workers are expected to identify these tweets, ensuring adequate attention and a minimum level of English language understanding. A task (HIT) then contains 20 classifications to be processed, with 18 belonging to the dataset and 2 additional ones serving as attention

checks, manipulated with an animal. In addition to the attention/language checks[32], data quality is ensured through a HIT approval rate of >98% and a minimum of 5,000 previously approved HITs by the workers. Workers were paid $0.70 per HIT, resulting in an hourly wage of $12.60 under the assumption of 6 classifications per minute. Figure 25 and Figure 26 exemplify a provided HIT including the exact instructions given to the workers. As it also can be easily seen usernames, company ticker, numbers and company names have been replaced to prevent workers from being biased.

Tasks

With what emotion do you associate the **20** given tweets below most **in a trading context**? Select "Animal" only if there is an animal **(except bull and bear!)** in the tweet (see Instruction No. 5 & 6).

Usernames, stock ticker as $AAPL for Apple, company names and numbers have been replaced with {user}, {ticker}, {company} and {number} to ensure you're only focussed on text and not biased because of these.

	+	+/-		-				
Tweet	**Happy**	**Neutral**	**Surprise**	**Anger**	**Disgust**	**Fear**	**Sad**	**Animal**
{ticker} {ticker} will go down {number} - {number} - sell	○	○	○	○	○	○	○	○
buy {ticker} it will do like {ticker}	○	○	○	○	○	○	○	○
{user} i will go with {number} as well chicken	○	○	○	○	○	○	○	○
{ticker} will it go to $ {number} today ? !	○	○	○	○	○	○	○	○
{user} i will buy more and more and more and more !	○	○	○	○	○	○	○	○
{ticker} sold all out here cow , going short	○	○	○	○	○	○	○	○
{user} {user} will this go up from here ? i just got in at {number} . .	○	○	○	○	○	○	○	○
{user} well , you will lose {number} when it go down to {number}	○	○	○	○	○	○	○	○
{ticker} a + {number} % day will be good :-)	○	○	○	○	○	○	○	○
{ticker} i think it will go down to {number} s	○	○	○	○	○	○	○	○
{user} buy more and it will go up	○	○	○	○	○	○	○	○
{user} i am with you there . at {number} , i will go all in on shorts and puts .	○	○	○	○	○	○	○	○
they will get more {user}	○	○	○	○	○	○	○	○
{ticker} you will get more below $ {number} :)	○	○	○	○	○	○	○	○
{user} looks will go to {number} before {number}	○	○	○	○	○	○	○	○
{user} better get going !	○	○	○	○	○	○	○	○
{user} {user} it will go to $ {number} during the next r / s	○	○	○	○	○	○	○	○
{ticker} shorted since $ {number} {ticker} going short again on this	○	○	○	○	○	○	○	○
{ticker} do good and good will come to you .	○	○	○	○	○	○	○	○
{ticker} it will go up , it will go down . . but moreso it will go up	○	○	○	○	○	○	○	○
Tweet	**Happy**	**Neutral**	**Surprise**	**Anger**	**Disgust**	**Fear**	**Sad**	**Animal**
	+	+/-		-				

Submit

Figure 26: Exemplified MTurk HIT

[32] The HIT of a worker had been accepted if the classification 'animal' had been chosen less than two times incorrectly.

Overall, all classifications were collected between the end of March and the beginning of April 2024, with 995 different workers participating. A tweet was assigned an emotion based on the classification results if three out of five workers chose that emotion, or if only two out of the five classifications belonged to emotions that fall within the categorization of the valence measure formerly introduced by Breaban and Noussair (2018) used for economic analysis in Section 4.4.2:

$$Positive = \{happy\}$$
$$Negative = \{anger, disgust, fear, sad\}$$
$$Neutral = \{neutral, surprise\}$$

Table 25 displays the results of the unambiguous classifications by the MTurk workers. Overall, out of the 1,500 tweets per group, 1116 (74.44%) Novice, 1102 Intermediate (73.47%), and 1030 Professional (68.67%) tweets were classified. The tweets were predominantly classified as neutral (42.21%) and happy (30.97%). Aggregating the emotions into the valence categories ‘positive’, ‘negative’ and ‘neutral’ provides clearer insights into the core results, as depicted in Table 26. Here, neutral and positive categories dominate with approximately 50% and 30%, respectively. While the perceived emotion of the workers is subjective, it is expected that tweets tagged as ‘bullish’ (‘bearish’) would be more associated with positive (negative) emotions in valid classifications: A comparison between ‘bullish’ and ‘bearish’ tweets reveals this dependency without exceptions in all groups. Slightly noticeable is the higher rate of neutrally classified bearish tweets, which tends to come at the expense of a clearer negative classification.

	Happy (+)	Neutral (+/−)	Surprise (+/−)	Anger (−)	Disgust (−)	Fear (−)	Sad (−)
Novice	**342**	**454**	**100**	**75**	**39**	**47**	**59**
Bullish	252	206	55	27	18	13	12
Bearish	90	248	45	48	21	34	47
Intermediate	**328**	**456**	**113**	**73**	**33**	**48**	**51**
Bullish	227	211	58	21	11	14	7
Bearish	101	245	55	52	22	34	44
Professional	**336**	**461**	**108**	**61**	**21**	**34**	**24**
Bullish	242	196	51	28	7	12	4
Bearish	94	265	57	33	14	22	20
$\sum$	**1006**	**1371**	**321**	**211**	**93**	**129**	**134**
Bullish	721	613	164	76	36	39	23
Bearish	285	758	157	133	57	90	111

Table 25: Results of MTurk classification per group (emotions)

	Absolute (N)			**Relative (%)**		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Novice	**342**	**554**	**220**	**30.64**	**49.64**	**19.71**
Bullish	252	261	70	43.22	44.77	12.00
Bearish	90	293	150	16.88	54.97	28.14
Intermediate	**328**	**569**	**205**	**29.76**	**51.63**	**18.60**
Bullish	227	269	53	41.35	49.00	9.65
Bearish	101	300	152	18.26	54.25	27.49
Professional	**336**	**569**	**140**	**32.15**	**54.45**	**13.40**
Bullish	242	247	51	44.81	45.74	9.44
Bearish	94	322	89	18.61	63.76	17.62
$\sum$	**1006**	**1692**	**565**	**30.83**	**51.85**	**17.31**
Bullish	721	777	174	43.12	46.47	10.41
Bearish	285	915	391	17.91	57.51	24.57

Table 26: Results of MTurk classification per group (valence-categories)

4.3.3 Training BERT

To determine the emotions contained in all tweets for each trader group, a model to estimate emotion scores will now be developed. This will be based on the pretrained model 'EmTract' by Vamossy and Skog (2020), enabling fine-tuning with relatively few data points. Since Vamossy and Skog (2020) have already expanded the number of tokens to the language used on StockTwits, there is no need to add additional tokens. The pretrained model is publicly available on *Huggingface* and can be accessed using the 'transformers' library in Python (Wolf et al. (2019)).

Table 25 and Table 26, despite efforts to create a balanced panel, have shown that both emotions and valence categories occur differentially in the dataset, with negative emotions being particularly underrepresented. To counteract this effect, the loss function is weighted based on the occurrence of valence categories in the respective dataset. To validate the suitable model, a stratified five-fold cross-validation procedure is applied, where the dataset is divided into five parts, with each part being used for training and validating in turn, representing the distribution of emotions in the entire dataset. All models are evaluated based on accuracy, balanced accuracy, and F1-score. The results of this procedure for the optimal model are presented in Table 27 and briefly discussed in the following chapter.

4.4 Results

4.4.1 Model fit

After fine-tuning, the models 'NovEm' for Novice tweets, 'IntEm' for Intermediate tweets, and 'ProEm' for Professional tweets were estimated. Table 27 displays relevant fit measures for the models compared to the performance of the original EmTract model. In terms of overall accuracy, all fine-tuned models show a slightly improved performance in tweet classification, ranging from 43.9% to 45.6%. When aggregating emotions into the three valence categories, values up to 53.4% can be achieved with slight gains compared to the EmTract model. If we exclude the neutral category from the confusion matrix, clearer differences are observed for Novice tweets, with accuracy rates reaching up to 83.3%, which underscores a low rate of misclassifying positive tweets as negative, and vice versa. Similar (slightly weaker) rates are achieved for the other two groups.

		Novice		Intermediate		Professional	
		NovEm	Emtract	IntEm	Emtract	ProEm	Emtract
Emotions	**Accuracy**	**0.456**	**0.415**	**0.439**	**0.418**	**0.446**	**0.445**
	Bal. Acc.	**0.566**	**0.545**	**0.573**	**0.555**	**0.555**	**0.551**
	Happy	0.711	0.678	0.692	0.675	0.669	0.657
	Anger	0.557	0.515	0.541	0.504	0.535	0.503
	Disgust	0.497	0.499	0.494	0.499	0.508	0.537
	Fear	0.534	0.545	0.587	0.594	0.497	0.535
	Sad	0.567	0.519	0.585	0.533	0.567	0.508
	Neutral	0.594	0.558	0.596	0.573	0.590	0.600
	Surprise	0.504	0.500	0.519	0.507	0.518	0.512
	F1-Score	**0.234**	**0.192**	**0.247**	**0.200**	**0.220**	**0.204**
	Happy	0.568	0.525	0.541	0.518	0.519	0.502
	Anger	0.189	0.067	0.146	0.021	0.146	0.019
	Disgust	0.000	0.000	0.000	0.000	0.038	0.128
	Fear	0.113	0.115	0.187	0.158	0.025	0.091
	Sad	0.179	0.078	0.216	0.118	0.172	0.040
	Neutral	0.558	0.516	0.547	0.526	0.559	0.584
	Surprise	0.030	0.039	0.091	0.056	0.079	0.065
Valence	**Accuracy**	**0.534**	**0.511**	**0.533**	**0.522**	**0.522**	**0.519**
	Bal. Acc.	**0.642**	**0.612**	**0.644**	**0.625**	**0.620**	**0.618**
	Positive	0.710	0.681	0.695	0.679	0.671	0.665
	Negative	0.627	0.581	0.637	0.610	0.599	0.612
	Neutral	0.588	0.574	0.600	0.585	0.589	0.578
	F1-Score	**0.519**	**0.480**	**0.513**	**0.492**	**0.484**	**0.485**
	Positive	0.569	0.529	0.545	0.524	0.521	0.511
	Negative	0.418	0.340	0.409	0.370	0.340	0.358
	Neutral	0.569	0.571	0.584	0.582	0.590	0.586
P-N	**Accuracy**	**0.833**	**0.785**	**0.807**	**0.800**	**0.782**	**0.799**
	Bal. Acc.	**0.823**	**0.756**	**0.806**	**0.783**	**0.758**	**0.785**
	F1-Score	**0.862**	**0.834**	**0.839**	**0.844**	**0.835**	**0.842**
	%-Obs.	**31.24**	**28.87**	**30.56**	**28.67**	**28.28**	**27.16**

As emotions and valence are multi-class classifications overall F1-Score is calculated using macro-averaging taking all classes as equally weighted.

Table 27: Fit measures of 5-fold-crossvalidation using argmax-predictions

Since it's an unbalanced sample, the model accuracy alone can provide a distorted picture of the model fit. Therefore, it's worthwhile to consider both the balanced accuracy and the F1-score. The balanced accuracy assigns equal weight to each label, while the balanced accuracy per label is the average of sensitivity (true positive rate) and specificity (true negative rate) for each label. While the balanced accuracy shows a mixed picture, especially for negative emotions, both the mean balanced accuracy values and when grouped into valence categories show increases. The same trend is also observed for the F1-scores, which express the harmonic mean of sensitivity (true positive rate) and precision (positive predictive value).

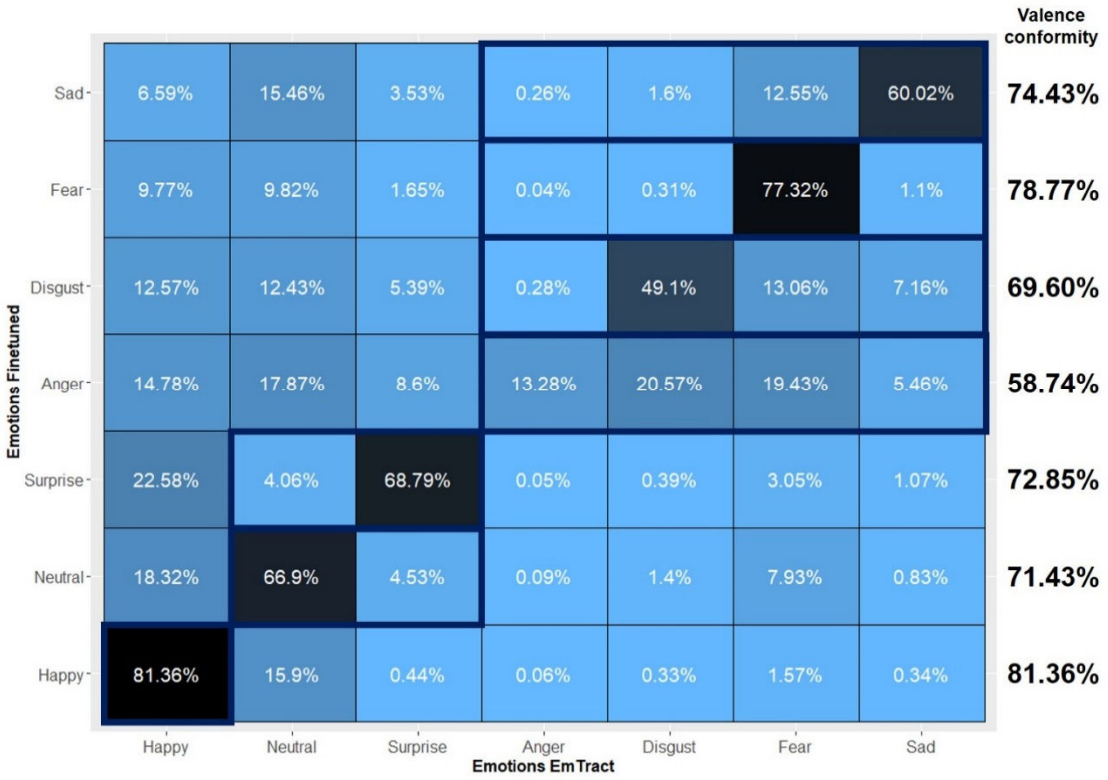


Figure 27: Confusion matrix fine-tuned and EmTract emotion labels (all tweets)

Applied to all 23,324,809 tweets (see Table 23), the distribution of emotion labels from both models is depicted in Figure 27. In the diagonal, it is evident that there continues to be a significant association between the classifications of both models even after fine-tuning. Nevertheless, differences are observed, particularly for negative emotions such as 'anger' and 'disgust' (only 13.28% and 49.10% of fine-tuned labels are the same as in EmTract). On one hand, these differences largely occur within their respective valence categories, resulting in conformity of valence mostly reaching values above 70%. On the

other hand, these differences also stem from the slight inclination towards classifying tweets as 'happy' and 'neutral' (see two first columns) by EmTract, which was partially offset by fine-tuning with the weighted loss function.

Overall, the three fine-tuned models show slight advantages. However, in addition to model fit, the economic relevance of the models adjusted by the classifications of workers on MTurk is also crucial, which will also be examined in the following Section.

4.4.2 Economic relevance

The influence of sentiment or the ability to predict stock returns using sentiment has been extensively studied. Most studies show significant predictive power, but it tends to be relatively low and applicable only for shorter time horizons. Following the approach practiced in 3.6.2, the intraday return[33] of a day t will be predicted using the change in sentiment (in the form of valence) from the close of trading on the previous day $t-1$ to the pre-market opening of the respective trading day t. By doing so, it is assumed that ideas published shortly before stock market closing on day $t-1$ mostly contain information relevant to that day, while investors focus more on the upcoming trading day t before market opening. Following Breaban and Noussair (2018) and Vamossy and Skog (2020), valence extracted by model i is defined as the difference between positive and negative emotions:

$$Valence_i = \underbrace{Happy_i}_{positive} - \underbrace{(Anger_i + Disgust_i + Fear_i + Sad_i)}_{negative} \tag{31}$$

In detail, valence ($Valence_{i,t,m}$) for each model $i = \{Finetuned, EmTract\}$ is calculated on trading day t and with a time indicator m subsetting the classified ideas used.

$$Valence_{i,t,m} = \begin{cases} m = 1 \; for \; 01.30 \; p.m. \; to \; 02.30 \; p.m. \\ m = 2 \; for \; 08.00 \; p.m. \; to \; 09.00 \; p.m. \end{cases} \tag{32}$$

Within the period m at day t the average score of each emotion of model i is calculated using them in equation (31) to calculate $Valence_{i,t,m}$. Hence, the change in valence ($\Delta Valence_{i,t}$) derived by model i on trading day t is defined as:

[33] Log-returns between opening and closing price at day t delivered by Bloomberg are calculated.

$$\Delta Valence_{i,t} = Valence_{i,t,1} - Valence_{i,t-1,2} \quad (33)$$

These shifts in valence are used as an explanatory variable explaining the intraday return of the S&P 500 and the NASDAQ 100 on trading day t ($Intraday_t$). Due to significantly lower data amouts before 2016 on hourly basis this analysis can only reliably be done for the period from 01/2016 to 12/2020 ($T = 1237$).[34] As many studies criticize that identified relationships between investor sentiment and stock returns might simply be driven by autocorrelation (which can be found for both indices in the first lag), I introduce the intraday return on the previous trading day $t - 1$ as a second explanatory variable to control for this effect (i.e. Xiong et al. (2019)). Consequently, following linear model will be estimated:

$$Intraday_t = \beta_0 + \beta_1 * \Delta Valence_{i,t} + \beta_2 * Intraday_{t-1} + \varepsilon_t \quad (34)$$

Table 28 presents the regression results for both indices and various models as well as for full data and subdivided by trader groups. Using the data from all groups (full data) it can be seen that the shift in valence significantly predicts the intraday return for both indices as well as with both models.[35] Nevertheless, it appears that (1) S&P 500 intraday returns can be predicted a little bit better than NASDAQ returns and (2) fine-tuned models deliver better results in predicting the intraday returns. Furthermore, differences within the trader groups are apparent, with Novice and Professional tweets exhibiting lower predictive power compared to tweets from Intermediate traders. However, all groups fail to surpass the predictive power of the entire dataset, implying that besides trader group membership, the data volume also plays a crucial role in the predictability of returns.

[34] Figure 15 in Section 3.4.2 shows that the number of ideas posted exponentially increases after 2016.
[35] Note that in this case Novice/Intermediate/Professional tweets are scored with the Novice/Intermediate/Professional fine-tuned models - NovEm, IntEm, ProEm - as well.

		SP500		NASDAQ	
		Fine-tuned	EmTract	Fine-tuned	EmTract
Full Data	*Intercept*	-0.0000	0.0000	0.0003	0.0004
		(-0.1170)	(0.3525)	(0.9708)	(1.3928)
	$\triangle Valence$	0.0126***	0.0118**	0.0111**	0.0095*
		(2.7492)	(2.3427)	(2.1718)	(1.6521)
	R_{t-1}	-0.1823***	-0.1823***	-0.1866***	-0.1868***
		(-2.9125)	(-2.8978)	(-4.2178)	(-4.2380)
	R^2	0.0406	0.0397	0.0391	0.0381
	F-Statistic	26.13***	25.51***	25.11***	24.44***
Novice	*Intercept*	0.0000	0.0001	0.0004	0.0004
		(0.3629)	(0.5634)	(1.4719)	(1.5548)
	$\triangle Valence$	0.0043*	0.0038	0.0020	0.0015
		(1.7199)	(1.4485)	(0.6893)	(0.4793)
	R_{t-1}	-0.1866***	-0.1868***	-0.1890***	-0.1890***
		(-3.0068)	(-2.9344)	(-4.3458)	(-4.2875)
	R^2	0.0369	0.0365	0.0362	0.0360
	F-Statistic	23.65***	23.38***	23.14***	23.05***
Intermediate	*Intercept*	0.0001	0.0001	0.0004	0.0004
		(0.5451)	(0.4834)	(1.5022)	(1.4385)
	$\triangle Valence$	0.0079**	0.0074*	0.0090**	0.0083*
		(2.3025)	(2.0059)	(2.2727)	(1.9134)
	R_{t-1}	-0.1851***	-0.1848***	-0.1882***	-0.1883***
		(-2.9652)	(-2.9744)	(-4.3014)	(-4.3001)
	R^2	0.0381	0.0378	0.0389	0.0384
	F-Statistic	24.47***	24.22***	24.95***	24.67***
Professional	*Intercept*	0.0000	0.0001	0.0003	0.0004
		(0.0909)	(0.4724)	(1.0659)	(1.4464)
	$\triangle Valence$	0.0051*	0.0049*	0.0054*	0.0045
		(1.7776)	(1.7406)	(1.7331)	(1.3229)
	R_{t-1}	-0.1802***	-0.1828***	-0.1868***	-0.1871***
		(-2.8617)	(-2.9388)	(-4.2628)	(-4.1960)
	R^2	0.0375	0.0371	0.0380	0.0371
	F-Statistic	24.05***	23.76***	24.37***	23.80***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ (HAC standard errors are used.)

Regressions estimate equation (35) with the sample: 01/2016 - 12/2020 ($T = 1237$)

Table 28: Intraday return predictability using (fine-tuned) EmTract model by trader group (1-hour-window)

Nonetheless, it must be noted that all measured significant relationships are relatively weak compared to previous research findings. Particularly when considering the regression results in the appendix in Table 29 without lagged returns, despite the significance of the effects, the low contribution to the explanatory power of the model is evident. Similarly, the differences resulting from the previous findings are also minimal, such that, for example, when comparing them using the Z-scores proposed by Paternoster et al. (1998), one would obtain insignificant results. However, the consistent systematic pattern remains striking, providing a clear implication that fine-tuning was beneficial. In line with the noise trader theory, these effects are only well measurable for short periods: Table 30 in the appendix shows that these observations have nearly disappeared by increasing the time window of observed tweets up to two hours.

4.5 Conclusion

This work has demonstrated that there are linguistic differences between various trader groups on both syntactic and content levels. The more professional the individual behind a tweet, the more likely the tweet is to be longer and to use more specialized financial terminology. This observation has largely been overlooked in the creation of general language models thus far. Therefore, with the help of tweet classifications on MTurk, the three models ‘NovEm’, ‘IntEm’, and ‘ProEm’ were estimated, which systematically show improvements in both fit measures and the economic relevance of the evaluated emotion scores.

This article implies several avenues for further research: Firstly, it seems worthwhile to train corresponding models with more data. As this work shows in a first step, that fit measures and economic relevance have slightly improved – even with a comparatively smaller dataset that has not fully exploited the potential of models fine-tuned according to trader groups. Secondly, the availability of a large amount of data appears to be a crucial factor, highlighting the need to expand the analysis to other platforms with even more traffic, such as Twitter. However, these platforms do not provide trader group classifications, necessitating models for classifying trader groups based on their written text. Finally, the economic relevance of the collected emotion scores can be scrutinized in many other ways. For example, Vamossy (2024) shows that direct use of emotion scores compared to the valence measure introduced by Breaban and Noussair (2018) can

be advantageous. Furthermore, with a sufficient amount of data at the sector or firm level, stronger correlations could be expected.

4.6 Appendix

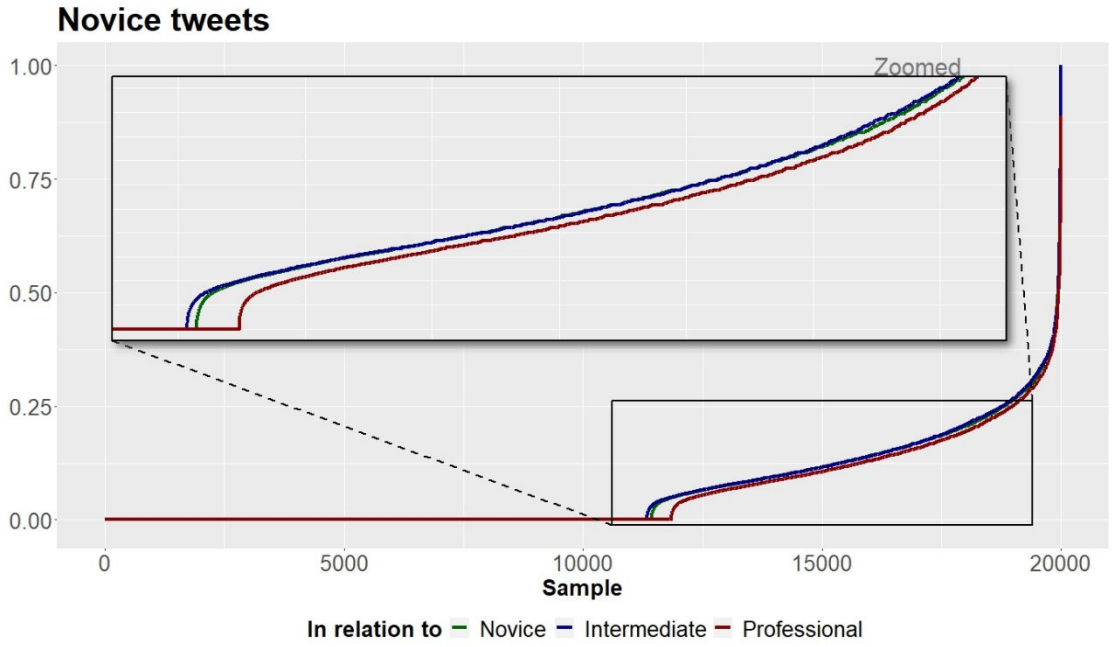


Figure 28: Distribution of sorted cosine similarities between trader groups (Novice)

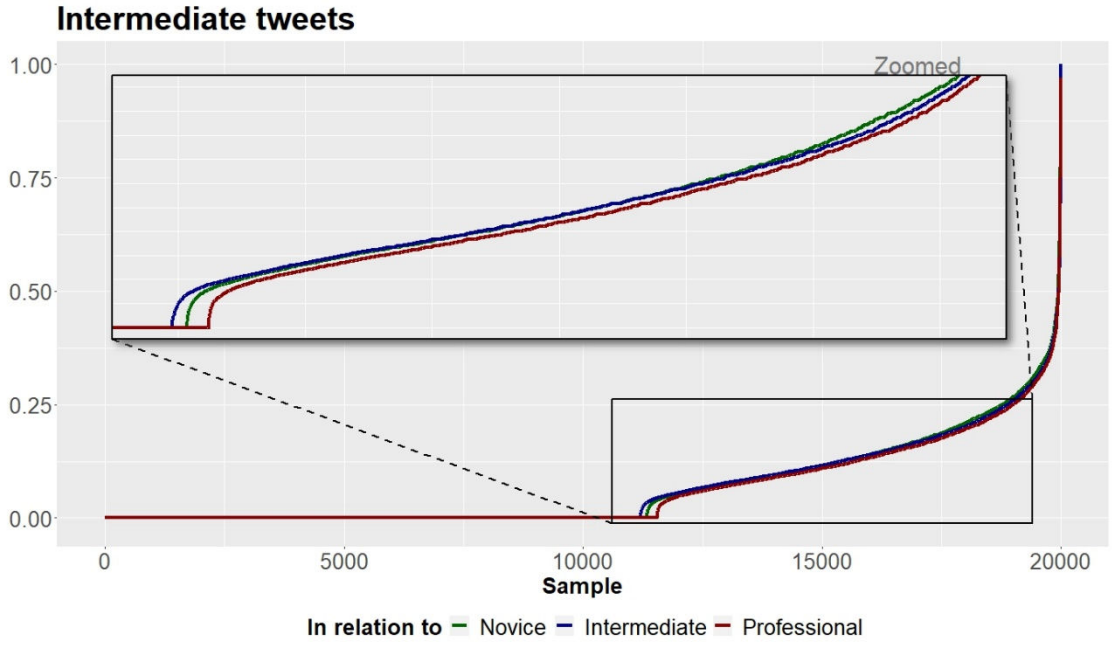


Figure 29: Distribution of sorted cosine similarities between trader groups (Intermediate)

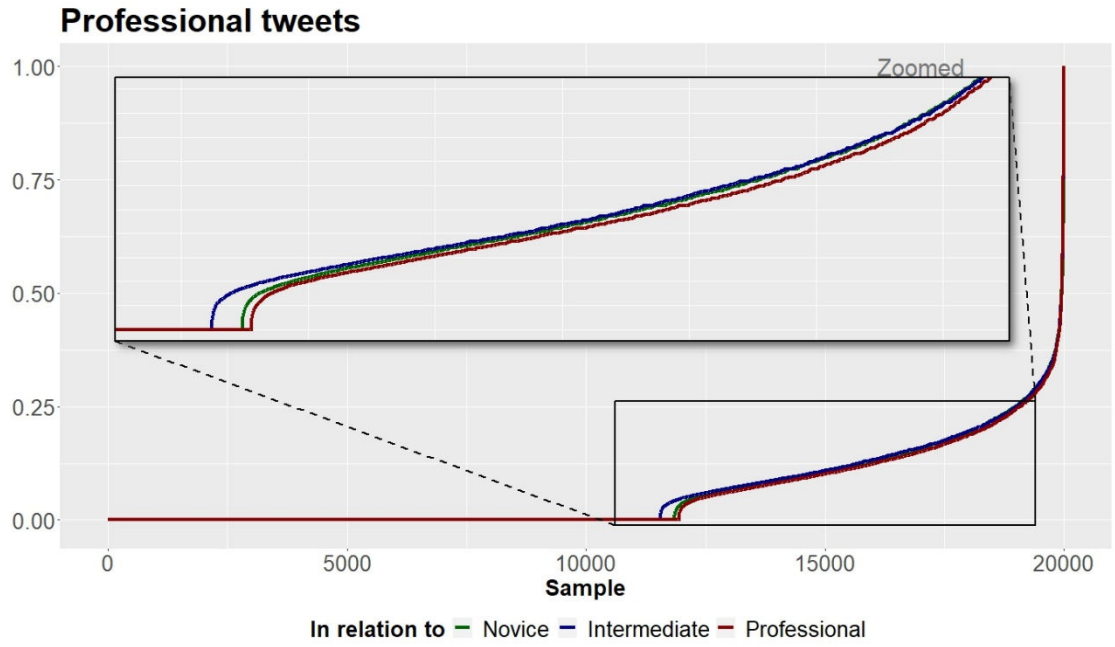


Figure 30: Distribution of sorted cosine similarities between trader groups (Professional)

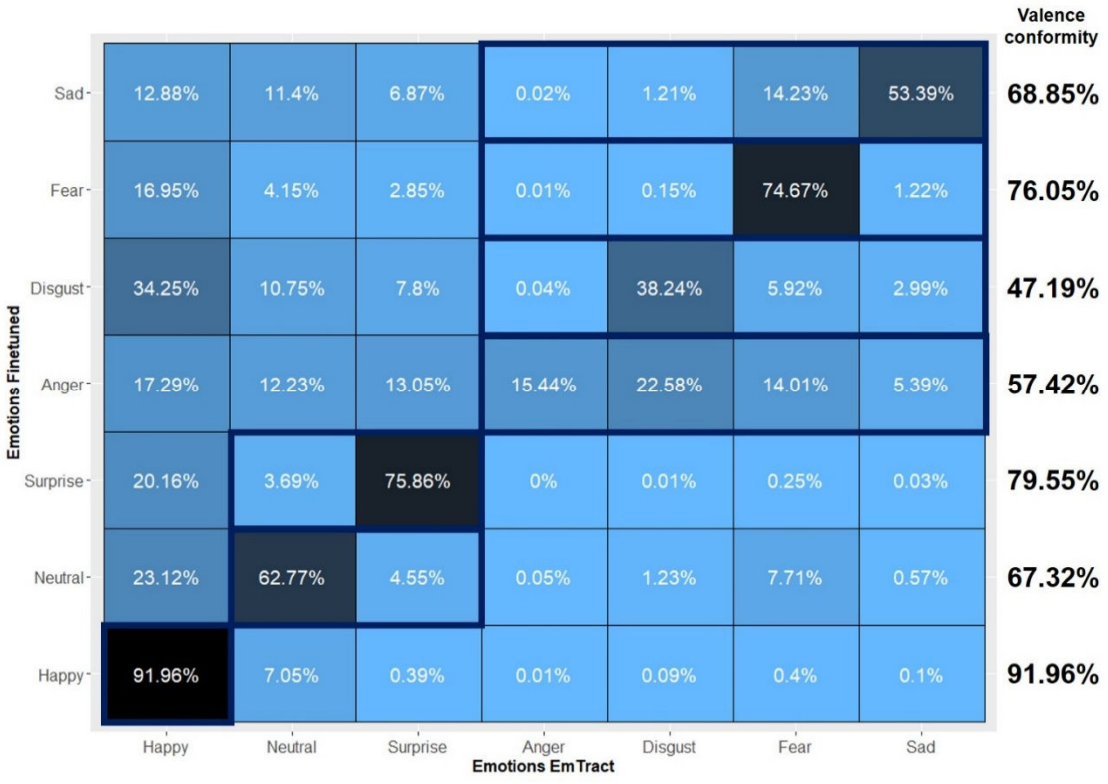


Figure 31: Confusion matrix fine-tuned and EmTract emotion labels (Novice tweets)

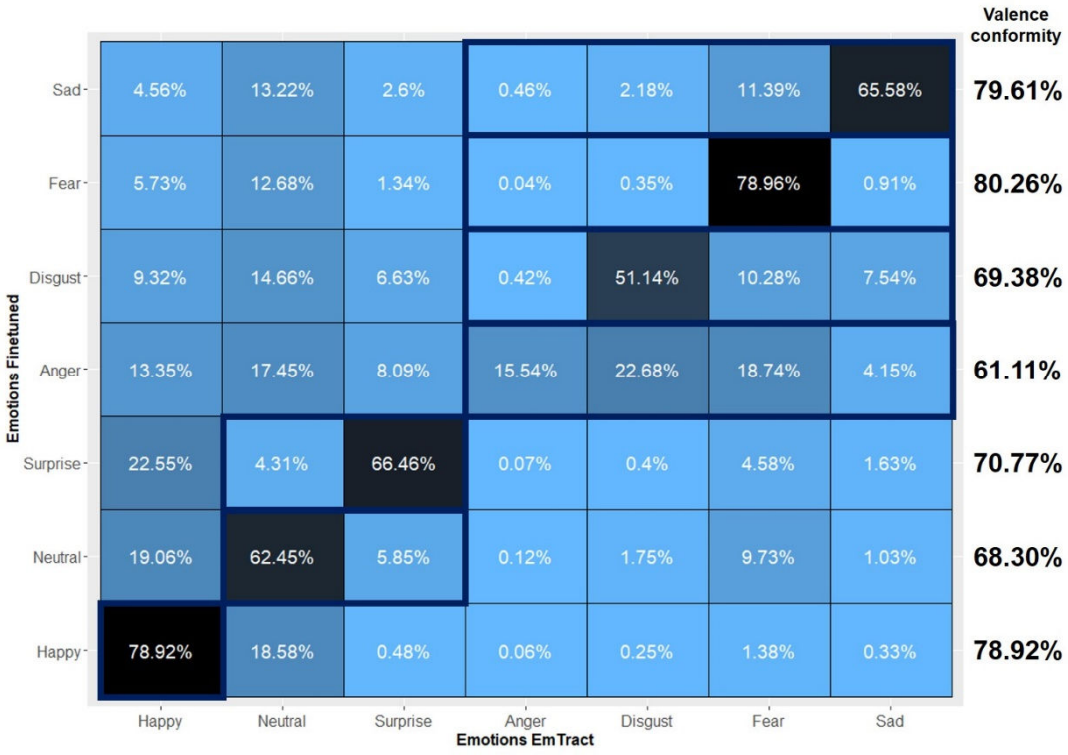


Figure 32: Confusion matrix fine-tuned and EmTract emotion labels (Intermediate tweets)

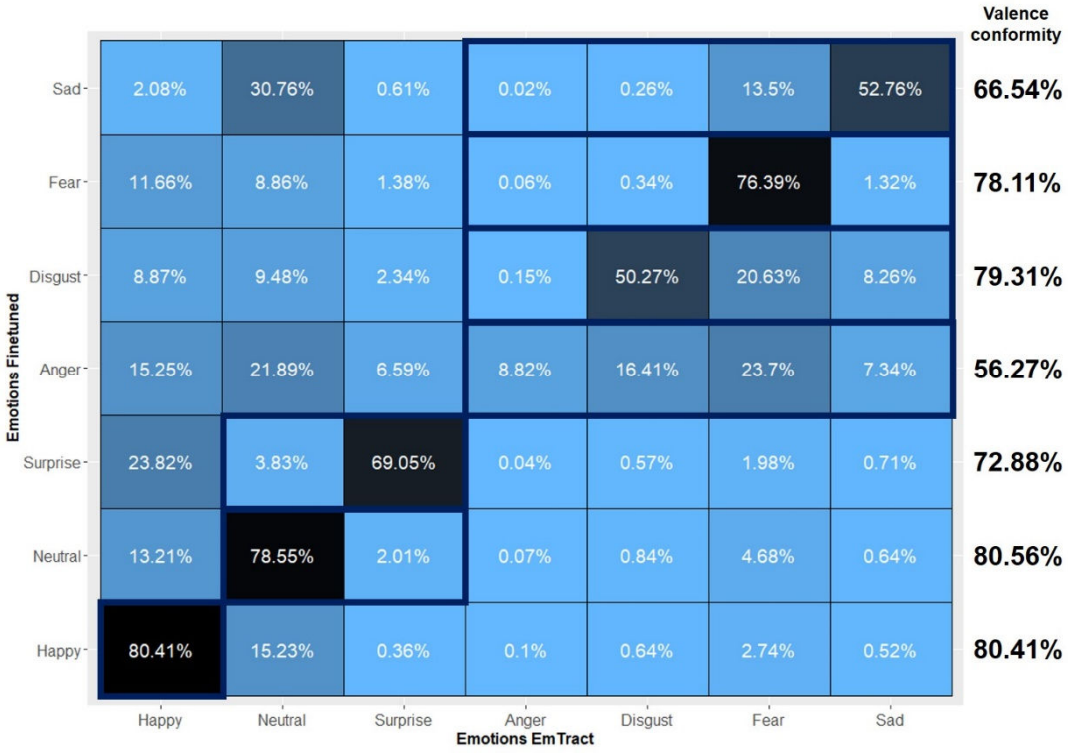


Figure 33: Confusion matrix fine-tuned and EmTract emotion labels (Professional tweets)

Regression without lagged return:

$$Intraday_t = \beta_0 + \beta_1 * \Delta Valence_{i,t} + \beta_2 * Intraday_{t-1} + \varepsilon_t \quad (35)$$

		SP500		NASDAQ	
		Fine-tuned	EmTract	Fine-tuned	EmTract
Full Data	*Intercept*	0.0000	0.0001	0.0003	0.0004
		(0.0242)	(0.5492)	(1.1642)	(1.4057)
	$\triangle Valence$	0.0143***	0.0117*	0.0098	0.0030
		(2.6737)	(1.8473)	(1.5466)	(0.4079)
	R^2	0.0054	0.0030	0.019	0.0001
	F-Statistic	6.75***	3.73*	2.35	0.18
Novice	*Intercept*	0.0000	0.0001	0.0004	0.0004
		(0.2778)	(0.5502)	(1.5311)	(1.5644)
	$\triangle Valence$	0.0049	0.0030	-0.0003	-0.0017
		(1.2378)	(0.7604)	(-0.0623)	(-0.3913)
	R^2	0.0013	0.0005	0.0000	0.0001
	F-Statistic	1.642	0.6526	0.00	0.14
Intermediate	*Intercept*	0.0001	0.0001	0.0004	0.0004
		(0.6111)	(0.6611)	(1.5041)	(1.5573)
	$\triangle Valence$	0.0110***	0.0075	0.0104**	0.0049
		(3.0578)	(1.6170)	(2.2418)	(0.8536)
	R^2	0.0044	0.0016	0.0029	0.0005
	F-Statistic	5.55**	2.05	3.592*	0.62
Professional	*Intercept*	0.0000	0.0001	0.0003	0.0004
		(0.2185)	(0.5072)	(1.2731)	(1.2773)
	$\triangle Valence$	0.0058*	0.0067*	0.0043	0.0025
		(1.6470)	(1.7024)	(1.1181)	(0.5569)
	R^2	0.0022	0.0025	0.0009	0.0002
	F-Statistic	2.71*	3.08*	1.09	0.30

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ (HAC standard errors are used.)

Regressions estimate equation (36) with the sample: 01/2016 - 12/2020 ($T = 1237$)

Table 29: Intraday return predictability using (fine-tuned) EmTract model by trader group (1-hour window)

		SP500		NASDAQ	
		Fine-tuned	EmTract	Fine-tuned	EmTract
Full Data	$Intercept$	0.0001	0.0001	0.0004	0.0005
		(0.2725)	(0.6031)	(1.4468)	(1.6370)
	$\triangle Valence$	0.0101*	0.0070	0.0052	-0.0019
		(1.7771)	(1.0796)	(0.8263)	(-0.2530)
	R_{t-1}	-0.1761***	-0.1784***	-0.1825***	-0.1859***
		(-2.7838)	(-2.7302)	(-4.2320)	(-4.1511)
	R^2	0.0360	0.0344	0.0348	0.0343
	F-Statistic	23.04***	21.98***	22.25***	21.94***
Novice	$Intercept$	0.0001	0.0001	0.0004*	0.0005*
		(0.4116)	(0.6036)	(1.6957)	(1.6623)
	$\triangle Valence$	0.0035	0.0016	-0.0022	-0.0036
		(0.8738)	(0.3955)	(-0.4862)	(-0.8364)
	R_{t-1}	-0.1856***	-0.1863***	-0.1902***	-0.1909***
		(-2.8934)	(-2.9353)	(-4.4787)	(-4.3674)
	R^2	0.0357	0.0351	0.0360	0.0364
	F-Statistic	22.82***	22.47***	23.07***	23.31***
Intermediate	$Intercept$	0.0002	0.0002	0.0005	0.0005
		(0.6348)	(0.6766)	(1.4724)	(1.6230)
	$\triangle Valence$	0.0078**	0.0041	0.0069	0.0016
		(2.0837)	(0.8237)	(1.4724)	(0.2805)
	R_{t-1}	-0.1772***	0.1802***	-0.1815***	-0.1845***
		(-2.7091)	(-2.8664)	(-4.0521)	(-4.1361)
	R^2	0.0355	0.0338	0.0355	0.0343
	F-Statistic	22.73***	21.60***	22.75***	21.94***
Professional	$Intercept$	0.0001	0.0001	0.0004	0.0004
		(0.3868)	(0.5777)	(1.4552)	(1.5564)
	$\triangle Valence$	0.0039	0.0043	0.0025	0.0003
		(1.0474)	(1.0742)	(0.6294)	(0.0600)
	R_{t-1}	-0.1804***	0.1805***	-0.1864***	-0.1874***
		(-2.8321)	(-2.9065)	(-4.2693)	(-4.2157)
	R^2	0.0348	0.0349	0.0355	0.0352
	F-Statistic	22.26***	22.28***	22.73***	22.53***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ (HAC standard errors are used.)

Regressions estimate equation (35) with the sample: 01/2016 - 12/2020 ($T = 1237$)

Table 30: Intraday return predictability using (fine-tuned) EmTract model by trader group (2-hour-window)

4.7 Declaration of (co-)authors and record of accomplishments

Title: How do you finance talk on Social Media? Extracting and identifying emotions from different trader groups

Author(s): **Philipp Stangor** (Heinrich Heine University Düsseldorf)

Conferences: -

Publication: SSRN published. Submitted to 'The Journal of Finance and Data Science', double-blind peer-reviewed journal. Current status: Under review.

Share of contributions:

Contributions	Philipp Stangor
Research Design	**100%**
Development of Research Question	100%
Method development	100%
Research performance & analysis	**100%**
Literature Review and framework development	100%
Data collection, preparation and analysis	100%
Analysis and discussion of results	100%
Derivation of implications and conclusions	100%
Manuscript preparation	**100%**
Final draft	100%
Finalization	100%
Overall contribution	**100%**

31.05.2024,

Date, Philipp Stangor

5 Concluding remarks

The purpose of this work was to meaningfully expand the existing literature on the relevance and impact of investor sentiment (derived from social media) on capital markets. Following a theoretical introduction to the role of information in capital markets and the construct of 'investor sentiment' along with its various definitions and measurement methods in the first section of the work, the relationship has been approached experimentally first.

The experiment aligns with existing literature by demonstrating that social media posts, in this case, tweets with a certain sentiment, influence individuals' investment decisions. A key new finding is that social media sentiment does not directly affect investment decisions but instead manipulates the fundamental financial perception of the company, which primarily determines the investment. The channel of perceived social media sentiment remains irrelevant to the decision, and the influence occurs subconsciously. It is also noteworthy that all content was generated by AI, highlighting the potential susceptibility to manipulation by bots. This presents an initial point for further research to critically examine this vulnerability in more detail.

Regarding the measurement of investor sentiment from social media data, several insights can be noted. Firstly, the multi-dimensional emotion classification of financial texts appears to have untapped potential that can be realized with the current available tools. Regardless, a deeper focus on the underlying text is opportune. Using social media data as an example, it becomes evident that simply categorizing texts based on their source is insufficient. Instead, it's crucial to consider the different groups with varying language usage at each exchange venue. This also poses the challenge that, if the text is not explicitly linked to a specific group of individuals – which is usually the case – this classification must be performed by another model first. The potential for further research in this area seems virtually limitless, given the numerous exchange venues and user groups, the continuously increasing volume of available data, and the ongoing advancement of new technological capabilities.

In summary, both the reviewed literature and the empirical findings of this study confirm the notion that 'money talks' in the context of this research. The influence of investor sentiment on aspects such as asset pricing is relatively minor. However, it remains unclear

whether this minor impact is due to its economic nature or the result of insufficient definitions and measurements of investor sentiment. Future research will definitely benefit from to expected advances in the field of Natural Language Processing. Nevertheless, it will be crucial not to neglect the foundational pillars of an unified definition and experimental exploration of the (psychological) mechanisms while pursuing continuous technical perfection.

6 References

Aggarwal, Divya, 2022, Defining and measuring market sentiments: a review of the literature, *Qualitative Research in Financial Markets* 14, 270–288.

Aliber, Robert Z., and Charles P. Kindleberger, 2015, *Manias, Panics, and Crashes* (Palgrave Macmillan UK, London).

Antweiler, Werner, and Murray Z. Frank, 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *The Journal of Finance* 59, 1259–1294.

Araci, Dogu, 2019, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, *arXiv preprint*.

Artetxe, Mikel, and Holger Schwenk, 2019, Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond, *Transactions of the Association for Computational Linguistics* 7, 597–610.

Baker, Malcolm, and Jeremy C. Stein, 2004, Market liquidity as a sentiment indicator, *Journal of Financial Markets* 7, 271–299.

Baker, Malcolm, and Jeffrey Wurgler, 2000, The Equity Share in New Issues and Aggregate Stock Returns, *The Journal of Finance* 55, 2219–2257.

Baker, Malcolm, and Jeffrey Wurgler, 2004, A Catering Theory of Dividends, *The Journal of Finance* 59, 1125–1165.

Baker, Malcolm, and Jeffrey Wurgler, 2006, Investor Sentiment and the Cross-Section of Stock Returns, *The Journal of Finance* 61, 1645–1680.

Baker, Malcolm, and Jeffrey Wurgler, 2007, Investor Sentiment in the Stock Market, *Journal of Economic Perspectives* 21, 129–151.

Barber, Brad M., 1994, Noise Trader Risk, Odd-Lot Trading, and Security Returns, *NBER working paper*.

Barber, Brad M., and Terrance Odean, 2008, All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors, *Review of Financial Studies* 21, 785–818.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

Baron, Reuben M., and David A. Kenny, 1986, The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* 51, 1173–1182.

Behrend, Tara S., David J. Sharek, Adam W. Meade, and Eric N. Wiebe, 2011, The viability of crowdsourcing for survey research, *Behavior research methods* 43, 800–813.

Behrendt, Simon, and Alexander Schmidt, 2018, The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility, *Journal of Banking & Finance* 96, 355–367.

Ben-Rephael, Azi, Shmuel Kandel, and Avi Wohl, 2012, Measuring investor sentiment with mutual fund flows, *Journal of Financial Economics* 104, 363–382.

Bessi, Alessandro, and Emilio Ferrara, 2016, Social Bots Distort the 2016 US Presidential Election Online Discussion, *First Monday* 21, 1–14.

Black, Fischer, 1986, Noise, *The Journal of Finance* 41, 528–543.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017, Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics* 5, 135–146.

Bormann, Sven-Kristjan, 2013, *Sentiment indices on financial markets: What do they measure?*, Kiel: Kiel Institute for the World Economy (IfW) .

Boulu-Reshef, Béatrice, Catherine Bruneau, Maxime Nicolas, and Thomas Renault, 2023, An Experimental Analysis of Investor Sentiment, in David Bourghelle, Pascal

Grandin, Fredj Jawadi, and Philippe Rozin, eds.: *Behavioral Finance and Asset Prices* (Springer International Publishing, Cham).

Breaban, Adriana, and Charles N. Noussair, 2018, Emotional State and Market Behavior, *Review of Finance* 22, 279–309.

Brodie, Henry, 1940, Odd-Lot Trading on the New York Stock Exchange and Financial Decentralization, *Southern Economic Journal* 6, 488.

Brown, Gregory W., and Michael T. Cliff, 2004, Investor sentiment and the near-term stock market, *Journal of Empirical Finance* 11, 1–27.

Brown, Gregory W., and Michael T. Cliff, 2005, Investor Sentiment and Asset Valuation, *The Journal of Business* 78, 405–440.

Brown, Stephen, William Goetzmann, Takato Hiraki, Noriyoshi Shirishi, and Masahiro Watanabe, 2003, Investor Sentiment in Japanese and U.S. Daily Mutual Fund Flows, *National Bureau of Economic Research*.

Buhrmester, Michael D., Sanaz Talaifar, and Samuel D. Gosling, 2018, An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use, *Perspectives on psychological science: a journal of the Association for Psychological Science* 13, 149–154.

Cade, Nicole L., 2018, Corporate social media: How two-way disclosure channels influence investors, *Accounting, Organizations and Society* 68-69, 63–79.

Cao, H. H., Joshua D. Coval, and David Hirshleifer, 2002, Sidelined Investors, Trading-Generated News, and Security Returns, *Review of Financial Studies* 15, 615–648.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, 2018, Universal Sentence Encoder, *arXiv preprint*.

Chen, Hailiang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media, *Review of Financial Studies* 27, 1367–1403.

Clogg, Clifford C., Eva Petkova, and Adamantios Haritou, 1995, Statistical Methods for Comparing Regression Coefficients Between Models, *American Journal of Sociology* 100, 1261–1293.

Cookson, J. A., and Marina Niessner, 2020, Why Don't We Agree? Evidence from a Social Network of Investors, *The Journal of Finance* 75, 173–228.

Cornelli, Francesca, David Goldreich, and Alexander Ljungqvist, 2006, Investor Sentiment and Pre-IPO Markets, *The Journal of Finance* 61, 1187–1216.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2015, The Sum of All FEARS Investor Sentiment and Asset Prices, *Review of Financial Studies* 28, 1–32.

Das, Sanjiv R., and Mike Y. Chen, 2007, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, *Management Science* 53, 1375–1388.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel, 2003, Persuasion Bias, Social Influence, and Unidimensional Opinions, *The Quarterly Journal of Economics* 118, 909–968.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint*.

Difallah, Djellel, Elena Filatova, and Panos Ipeirotis, 2018, Demographics and Dynamics of Mechanical Turk Workers, in Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, eds.: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (ACM, New York, NY, USA).

Dougal, Casey, Joseph Engelberg, Diego García, and Christopher A. Parsons, 2012, Journalists and the Stock Market, *Review of Financial Studies* 25, 639–679.

Ekman, Paul, 1992, Are there basic emotions?, *Psychological review* 99, 550–553.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg, 2012, How are shorts informed?, *Journal of Financial Economics* 105, 260–278.

Fama, Eugene F., 1965, The Behavior of Stock-Market Prices, *The Journal of Business* 38, 34–105.

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *The Journal of Finance* 25, 383.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, 2010, Management's tone change, post earnings announcement drift and accruals, *Review of Accounting Studies* 15, 915–953.

Fisher, Kenneth L., and Meir Statman, 2000, Investor Sentiment and Stock Returns, *Financial Analysts Journal* 56, 16–23.

Frazzini, Andrea, and Owen A. Lamont, 2007, The Earnings Announcement Premium and Trading Volume, *NBER working paper*.

Gao, Bin, and Xihua Liu, 2020, Intraday sentiment and market returns, *International Review of Economics & Finance* 69, 48–62.

Gao, Bin, and Chunpeng Yang, 2017, Forecasting stock index futures returns with mixed-frequency sentiment, *International Review of Economics & Finance* 49, 69–83.

Giannini, Robert, Paul Irvine, and Tao Shu, 2018, Nonlocal Disadvantage: An Examination of Social Media Sentiment, *The Review of Asset Pricing Studies* 8, 293–336.

Giglio, Stefano, and Bryan Kelly, 2018, Excess Volatility: Beyond Discount Rates*, *The Quarterly Journal of Economics* 133, 71–127.

Greenwood, Robin, and Andrei Shleifer, 2014, Expectations of Returns and Expected Returns, *The Review of Financial Studies* 27, 714–746.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Hales, Jeffrey, Xi J. Kuang, and Shankar Venkatarman, 2011, Who Believes the Hype? An Experimental Examination of How Language Affects Investor Judgments, *Journal of Accounting Research* 49, 223–255.

Heiden, Sebastian, Christian Klein, and Bernhard Zwergel, 2013, Beyond Fundamentals: Investor Sentiment and Exchange Rate Forecasting, *European Financial Management* 19, 558–578.

Henry, E., 2008, Are Investors Influenced By How Earnings Press Releases Are Written?, *Journal of Business Communication* 45, 363–407.

Holt, Charles A., and Susan K. Laury, 2002, Risk Aversion and Incentive Effects, *American Economic Review* 92, 1644–1655.

Hutto, C., and Eric Gilbert, 2014, VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, *Proceedings of the International AAAI Conference on Web and Social Media* 8.

International Telecommunication Union, 2022, Measuring digital development, *Facts and Figures 2022*.

Jeng, Leslie, Andrew Metrick, and Richard Zeckhauser, 1999, The Profits to Insider Trading: A Performance-Evaluation Perspective, *National Bureau of Economic Research*.

Jensen, Michael C., 1978, Some anomalous evidence regarding market efficiency, *Journal of Financial Economics* 6, 95–101.

Jing, Nan, Zhao Wu, and Hefei Wang, 2021, A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction, *Expert Systems with Applications* 178, 115019.

Jockers, M. L., 2015, Syuzhet. Extract Sentiment and Plot Arcs from Text.

Johnson, Eric J., and Amos Tversky, 1983, Affect, generalization, and the perception of risk, *Journal of Personality and Social Psychology* 45, 20–31.

Kahneman, Daniel, and Amos Tversky, 1979, Prospect Theory: An Analysis of Decision under Risk, *Econometrica* 47, 263.

Kaplanski, Guy, Haim Levy, Chris Veld, and Yulia Veld-Merkoulova, 2015, Do Happy People Make Optimistic Investors?, *The Journal of Financial and Quantitative Analysis* 50, 145–168.

Kearney, Colm, and Sha Liu, 2014, Textual sentiment in finance: A survey of methods and models, *International Review of Financial Analysis* 33, 171–185.

Kelton, Andrea S., and Robin R. Pennington, 2020, If You Tweet, They Will Follow: CEO Tweets, Social Capital, and Investor Say-on-Pay Judgments, *Journal of Information Systems* 34, 105–122.

Kim, Soon-Ho, and Dongcheol Kim, 2014, Investor sentiment from internet message postings and the predictability of stock returns, *Journal of Economic Behavior & Organization* 107, 708–729.

Kumar, Alok, and Charles M. Lee, 2006, Retail Investor Sentiment and Return Comovements, *The Journal of Finance* 61, 2451–2486.

Kyle, Albert S., 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315.

Lample, Guillaume, and Alexis Conneau, 2019, Cross-lingual Language Model Pretraining.

Landers, Richard N., and Tara S. Behrend, 2015, An Inconvenient Truth: Arbitrary Distinctions Between Organizational, Mechanical Turk, and Other Convenience Samples, *Industrial and Organizational Psychology* 8, 142–164.

Latif, Madiha, Shanza Arshad, Mariam Fatima, and Farooq, 2011, Market efficiency, market anomalies, causes, evidences, and some behavioral aspects of market anomalies, *Research Journal of Finance and Accounting* 2, 1–13.

Le, Quoc V., and Tomas Mikolov, 2014, Distributed Representations of Sentences and Documents, *arXiv preprint*.

Lee, Charles M. C., Andrei Shleifer, and Richard H. Thaler, 1991, Investor Sentiment and the Closed-End Fund Puzzle, *The Journal of Finance* 46, 75–109.

Lee, Wayne Y., Christine X. Jiang, and Daniel C. Indro, 2002, Stock market volatility, excess returns, and the role of investor sentiment, *Journal of Banking & Finance* 26, 2277–2299.

Li, Baoli, and Liping Han, 2013, Distance Weighted Cosine Similarity Measure for Text Classification, in David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao, eds.: *Intelligent Data Engineering and Automated Learning – IDEAL 2013* (Springer Berlin Heidelberg, Berlin, Heidelberg).

Liu, Bing, 2020, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (Cambridge University Press).

Liu, Yinhan, Myle Ott, Naman Goyal, Du Jingfei, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 2019, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint*.

Long, J. B. de, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Noise Trader Risk in Financial Markets, *Journal of Political Economy* 98, 703–738.

Long, Suwan, Brian Lucey, Ying Xie, and Larisa Yarovaya, 2023, "I just like the stock": The role of Reddit sentiment in the GameStop share rally, *Financial Review* 58, 19–37.

Loughran, Tim, and Bill McDonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill McDonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

Mackay, Charles, 1869, *Memoirs of Extraordinary Popular Dilusions* .

Malkiel, Burton G., 1977, The valuation of closed-end investment-company shares, *The Journal of Finance* 32, 847–859.

Malkiel, Burton G., 2003, The Efficient Market Hypothesis and Its Critics, *Journal of Economic Perspectives* 17, 59–82.

McLean, David R., and Mengxin Zhao, 2014, The Business Cycle, Investor Sentiment, and Costly External Finance, *The Journal of Finance* 69, 1377–1409.

Mechura, M. B., 2016, Lemmatization list - English (en).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Distributed Representations of Words and Phrases and their Compositionality, *arXiv preprint*.

Miller, Brian P., 2010, The Effects of Reporting Complexity on Small and Large Investor Trading, *The Accounting Review* 85, 2107–2143.

Miller, Edward M., 1977, Risk, Uncertainty, and Divergence of Opinion, *The Journal of Finance* 32, 1151.

Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov, 2020, Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers, *IEEE Access* 8, 131662–131682.

Mohammad, Saif M., and Peter D. Turney, 2013, Crowdsourcing a word-emotion association lexicon, *Computational Intelligence* 29, 436–465.

Nayak, Subhankar, 2010, Investor Sentiment and Corporate Bond Yield Spreads, *Review of Behavioral Finance* 2, 59–80.

Neal, Robert, and Simon M. Wheatley, 1998, Do Measures of Investor Sentiment Predict Returns?, *The Journal of Financial and Quantitative Analysis* 33, 523.

Orabi, Mariam, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel, 2020, Detection of Bots in Social Media: A Systematic Review, *Information Processing & Management* 57, 102250.

Otoo, Maria W., 2000, Consumer Sentiment and the Stock Market, *SSRN Electronic Journal*.

Paternoster, Raymond, Robert Brame, Paul Mazerolle, and Alex Piquero, 1998, Using the Correct Statistical Test for Equality of Regression Coefficients, *Criminology* 36, 859–866.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 2018, Deep contextualized word representations, *arXiv preprint*.

Plutchik, Robert, 1984, A general psychoevolutionary theory of emotion, in Robert Plutchik, and Henry Kellerman, eds.: *Theories of Emotion* (Acad. Press, Orlando).

Qiu, Lily, and Ivo Welch, 2004, Investor Sentiment Measures, *National Bureau of Economic Research*.

Renault, Thomas, 2017, Intraday online investor sentiment and return patterns in the U.S. stock market, *Journal of Banking & Finance* 84, 25–40.

Rennekamp, Kristina M., and Patrick D. Witz, 2021, Linguistic Formality and Audience Engagement: Investors' Reactions to Characteristics of Social Media Disclosures*, *Contemporary Accounting Research* 38, 1748–1781.

Rinker, T. W., 2018, Textstem. Tools for stemming and lemmatizing text.

Ritter, Jay R., 1991, The Long-Run Performance of initial Public Offerings, *The Journal of Finance* 46, 3–27.

Ritter, Jay R., 2003, Behavioral finance, *Pacific-Basin Finance Journal* 11, 429–437.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, 2020, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint*.

Schmeling, Maik, 2007, Institutional and individual sentiment: Smart money and noise trader risk?, *International Journal of Forecasting* 23, 127–145.

Seyhun, H. N., 2000, *Investment intelligence from insider trading* (MIT Press, Cambridge, Mass.).

Shank, Daniel B., 2016, Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk, *The American Sociologist* 47, 47–55.

Shefrin, Hersh, and M. L. Belotti, 2008, Risk and return in behavioral SDF-based asset pricing models, *Journal of Investment Management* 6, 1–18.

Shefrin, Hersh, and Meir Statman, 1985, The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence, *The Journal of Finance* 40, 777.

Shiller, Robert J., 1981, Alternative tests of rational expectations models: The case of the term structure, *Journal of Econometrics* 16, 71–87.

Shiller, Robert J., Stanley Fischer, and Benjamin M. Friedman, 1984, Stock Prices and Social Dynamics, *Brookings Papers on Economic Activity* 1984, 457.

Shleifer, Andrei, and Lawrence H. Summers, 1990, The Noise Trader Approach to Finance, *Journal of Economic Perspectives* 4, 19–33.

Singer, Eleanor, 2002, The use of incentives to reduce nonresponse in household surveys, in R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, eds.: *Survey Nonresponse* (Wiley).

Solt, Michael E., and Meir Statman, 1988, How Useful is the Sentiment Index?, *Financial Analysts Journal* 44, 45–55.

Sörries, Bernd, and Matthias Wissner, 2023, *IKT in den Stromverteilernetzen: Aktueller Stand und Ausblick vor dem Hintergrund einer sektoralen Datenökonomie*, Bad Honnef: WIK Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste .

Sparck Jones, Karen, 1972, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation* 28, 11–21.

Sprenger, Timm O., Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe, 2014, Tweets and Trades: the Information Content of Stock Microblogs, *European Financial Management* 20, 926–957.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2012, The short of it: Investor sentiment and anomalies, *Journal of Financial Economics* 104, 288–302.

Sun, Licheng, Mohammad Najand, and Jiancheng Shen, 2016, Stock return predictability and investor sentiment: A high-frequency perspective, *Journal of Banking & Finance* 73, 147–164.

Tan, Hun-Tong, Elaine Ying Wang, and B. O. Zhou, 2014, When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments, *Journal of Accounting Research* 52, 273–302.

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *The Journal of Finance* 63, 1437–1467.

Tukey, John W., 1992, *Exploratory data analysis* (Addison-Wesley, Reading, Mass. u.a.).

Tversky, A., and D. Kahneman, 1974, Judgment under Uncertainty: Heuristics and Biases, *Science* 185, 1124–1131.

Umar, Zaghum, Mariya Gubareva, Imran Yousaf, and Shoaib Ali, 2021, A tale of company fundamentals vs sentiment driven pricing: The case of GameStop, *Journal of Behavioral and Experimental Finance* 30, 100501.

Vamossy, Domonkos F., 2024, Social Media Emotions and Market Behavior, *arXiv preprint*.

Vamossy, Domonkos F., and Rolf Skog, 2020, EmTract: Investor Emotions and Market Behavior, *SSRN Electronic Journal*.

van Bommel, Jos, 2003, Rumors, *The Journal of Finance* 58, 1499–1520.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is All you Need, *Advances in Neural Information Processing Systems* 30.

Verma, Rahul, and Gökçe Soydemir, 2009, The impact of individual and institutional investor sentiment on the market price of risk, *The Quarterly Review of Economics and Finance* 49, 1129–1145.

Verma, Rahul, and Priti Verma, 2008, Are survey forecasts of individual and institutional investor sentiments rational?, *International Review of Financial Analysis* 17, 1139–1155.

Wang, Yaw-Huei, Aneel Keswani, and Stephen J. Taylor, 2006, The relationships between sentiment, returns and volatility, *International Journal of Forecasting* 22, 109–123.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush, 2019, HuggingFace's Transformers: State-of-the-art Natural Language Processing, *arXiv preprint*.

Xiong, Xiong, Chunchun Luo, Ye Zhang, and Shen Lin, 2019, Do stock bulletin board systems (BBS) contain useful information? A viewpoint of interaction between BBS quality and predicting ability, *Accounting & Finance* 58, 1385–1411.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, 2019, XLNet: Generalized Autoregressive Pretraining for Language Understanding.

Young, Jacob, and Kristie M. Young, 2019, Don’t Get Lost in the Crowd: Best Practices for Using Amazon’s Mechanical Turk in Behavioral Research, *Journal of the Midwest Association for Information Systems*, 7–34.

Yu, Jianfeng, and Yu Yuan, 2011, Investor sentiment and the mean–variance relation☆, *Journal of Financial Economics* 100, 367–381.

Zad, Samira, Maryam Heidari, James H. Jones, JR, and Ozlem Uzuner, 2021, Emotion Detection of Textual Data: An Interdisciplinary Survey, in : *2021 IEEE World AI IoT Congress (AIIoT)* (IEEE).

Zhang, Cathy, 2008, Defining, modeling, and measuring investor sentiment, *University of California, Berkeley, Department of Economics (Thesis)*.

Zhao, Xinshu, John G. Lynch, and Qimei Chen, 2010, Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis, *Journal of Consumer Research* 37, 197–206.

Zhou, Guofu, 2018, Measuring Investor Sentiment, *Annual Review of Financial Economics* 10, 239–259.

Zwergel, Bernhard, and Christian Klein, 2006, On the Predictive Power of Sentiment - Why Institutional Investors are Worth Their Pay, *SSRN Electronic Journal*.

Statutory declaration

I, Philipp Stangor, swear that I am writing this dissertation independently and without inadmissible outside help, taking into account the ‘principles for ensuring good scientific practice at the Heinrich Heine University Düsseldorf’.

Düsseldorf, 3rd of June, 2024 (Philipp Stangor)