MEASURING SENTIMENT

THIS IS THE SUBTITLE

STEFAN WINTER

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2067606

COMMITTEE

dr. Peter Hendrix
prof. dr. The Second Reader

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

November 23, 2021

ACKNOWLEDGMENTS

Some room for acknowledgements.

MEASURING SENTIMENT

THIS IS THE SUBTITLE

STEFAN WINTER

Abstract

This is where the abstract goes. Don't forget to change the variables in `main.tex` to change all general placeholders shown in this document. The `frontmatter.tex` file should be left alone.

1 INTRODUCTION

Modern society has been able to access information, communicate ideas, and become part of a community due to the advent of the internet. Online discussion boards are playing a critical role by providing a platform where people can do so. Those discussion boards are also used by a variety of people to talk about the stock market and discuss trading strategies. Recently, the Reddit forum WallStreetBets has become one of the most well-known and influential investing online-forums.
Even though the Reddit subforum was created in 2012 already, it received the majority of its media exposure in 2021 as a result of a short-squeeze of the GameStop (GME) stock, which drove the stock price up hundreds of percentage points. However, it was not the rapid price appreciation that amazed market participants. Instead, it was the unprecedented decentralized and coordinated buying of Gamestop shares by members of the WallStreetBets community that attracted attention (Abhinav & Jalaj, 2021). Organizing the mass-coordinated buying of stock, however, requires that enough participants share the same sentiment. According to several studies, social media sentiment has a particularly strong impact on uninformed traders (Danbolt Jo & Evangelos, 2015).
Interestingly, finance scholars did not consider Reddit as a platform capable of having such a significant impact on the financial markets. As a result, the site has been neglected in their research (Long Cheng & Larisa, 2021). Hence, this thesis will try to answer the following Research Question:

> *How can sentiment analysis best be performed on the WallStreetBets Reddit-forum?*

To begin, it must be determined how the discussions about the Gamestop stock on WallStreetBets should be handled to serve as suggestive input features for sentiment analysis. One of the challenges, is the heavy use of peculiar terminology and domain-specific phrases on the WallStreetBets forum, as well as many novel words (Abhinav & Jalaj, 2021). According to recent research, sentiment lexicons and corpora with a focus on a certain domain produce superior sentiment analysis results compared to a general-purpose sentiment lexicon or corpora (Park Sungrae & Il-Chul, 2015). Furthermore, the text data needs to be cleaned and pre-processed in order to be accurately processed by a machine learning algorithm (Jemai Fatma & Sahbi, 2021). As a result, the following sub-research question was formed:

RQ1 *How can the domain-specific language of the Reddit forum WallStreetBets best be incorporated into sentiment analysis?*

Or, format it as you desire (tip: you can nest `itemize` as well). You can alternate *emph* and **textbf** however you wish. This should cover most of the things required for the introduction.

Subsequently, the machine learning models can be trained to perform sentiment analysis. However, each machine learning algorithm has its own idiosyncrasies and assumptions, and no single classifier works optimally in all possible scenarios. Hence, it is a good idea to evaluate the results and performance of different machine learning algorithms. As a result, the best model with a given set of hyperparameters can be selected to solve a particular problem (Sebastian & Vahid, 2019, p. 53).

This thesis will explore traditional machine learning methods such as Naive Bayes (NB) and Support Vector Machines (SVMs), as well as deep learning methods like Long Short Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). Due to the high dimensionality of textual data, deep learning methods have shown to outperform traditional machine learning techniques in recent research. That can be explained by the ability of deep learning methods to automatically learn the most important features, whereas traditional methods may suffer from the curse of dimensionality (Fu Xianghua, Min, & Huihui, 2018). *(Note: Here I should have 5 authors. However, Latex throws an error when I add all names -> fix!!)* As was mentioned earlier, however, no classifier works best on all scenarios which is why the next research question needs to be answered:

RQ2 *Which sentiment analysis approach performs best on predefined key performance indicators?*

2 RELATED WORK

Copy paste BibTeX code[1] and put it in `references.bib`. After, you can cite some work – using `\citep`. You can refer to the author of e.g. Minsky (1961) directly like using `\cite` (this does not work when using bracket-citation). If you use bracket-style,[2] you might want use `\citeauthor` when citing, like: see Ananny and Crawford Ananny and Crawford (2018). If you want to add pages you can use brackets in `\citep[][p. 5]{mackay2003information}`, which looks like: (MacKay & Mac Kay, 2003, p. 5). The first brackets can be used for things like *see*, and *e.g.*. If you want to cite multiple authors, simply comma-separate them (`\citep{minsky1961steps,mackay2003information}`) and it will aggregate them automatically (MacKay & Mac Kay, 2003; Minsky, 1961).

Gauging sentiment of online forums to predict movements in stock prices has been a research subject for many years now. (R. & Y., 2007) did a study on the Yahoo! message board, which was amongst the first ones on the internet for investors to exchange ideas.

also showed that as the discussion volume on WallStreetBets increased, the volatility of certain stocks got amplified. (Umar Zaghum & Shoaib, 2021) also found that sentiment of investors on WallStreetBets affected the returns of the Gamestop stock.

However, they also show that other features such as the put-call ratio and the short-sale volume had a strong impact on the stock price. (Long Cheng & Larisa, 2021) tried to uncover the impact of specific emotions such as "Angry, Fear, Happy, Sad and Surprise" from the comments on

WallStreetBets discussions on intraday changes of the stock price of the affected stock. While they conclude that the tone as well as the number of comments have an impact on the stock price, they show that the number of comments is not directly related to sentiment. Additionally, they argue it is the number of comments that is posted within an hour that has the biggest effect on one minute changes in the stock price. Furthermore, the paper shows that the emotions Sad, Anger and Surprise have a significant impact on the gamestop 1-minute stock price. The Happy sentiment does not show a significant impact on 1-minute price changes, however, a causality test showed a link between the Happy sentiment and intraday returns of the GME stock. In addition, the paper shows, that sentiment only impacts intraday returns if a thread has more than 2000 comments. Hence, the authors confirm that Reddit sentiment has an impact on the stock market. They also argue that any asset that is targeted by a large

[1] Using e.g. the quote icon in GScholar, then BibTeX at the bottom.
[2] Find the `natbib` part in the `main.tex` LaTeX script.

crowd from wallstreetbets can become a subject of excessive volatility, without being driven by any fundamental reasons. However, since the WallStreetBets 'meme-stock movement' is a relatively recent phenomenon, there is very little research on the impact of WallStreetBets on individual stocks, especially with regards to sentiment analysis. Additionally, of all the published research none account for the domain-specific language used on the forum. Because of the frequent usage of terminology that is specific to WallStreetBets, this can lead to incorrect conclusions.

Of course, this also applies to research in other fields, which usually also use a general-purpose sentiment lexicon, because of the cost associated with building a domain-specific one. However, it has been demonstrated that using a domain-specific knowledge base results in more accurate sentiment analysis (Sungrae Park & Moon, 2015). It is argued that there is no general-purpose sentiment lexicon that can be optimally applied on all domains. In different domains, some terms can have completely different meanings. A good example is the word "unpredictable", which would have negative sentiment for electronics but can be a positive label for movies. It has been demonstrated that by adapting sentiment lexicons to a certain domain performance for sentiment classification can be enhanced (Lu Yue & ChengXiang, 2011). This adapted lexicon can then be searched to find and score the sentiment of a specific word (Muhammad, 2014). While lexicon-based methods have found widespread adoption, mainly due to their simplicity, more advanced machine learning methods have also shown strong performance (Wang Yanyan & Marco, 2020). For this reason other research deviates from the aforementioned lexicon-based approaches. Instead, they examine how deep learning methods can be used to automatically detect and identify domain-specific words from sentences. By doing so it is assumed that the algorithm can not only detect whether domain-specific words are used (sentence-level detection), but also to identify the exact position of the term in the sentence (token-level identification). Hence, it is possible to detect new meanings of words in an already existing corpus. In addition, this approach also allows to classify novel words, that do not yet exist in a dictionary. This can be achieved by having models that formulate domain-specific word detection as a sequence-labelling task. Furthermore, novel domain-specific words can be learned by understanding the contextual structure of a sentence (Pei Zhengqi & Yang, 2019). Those out-of-vocabulary tokens can be learned in the hidden layers of LSTMs (Sepp & Jürgen, 1997). To further optimize performance, models can be improved, by applying a character-based convolutional neural network to encode the spelling of words (Pei Zhengqi & Yang, 2019). Even though the literature suggests many innovative ways to enhance model performance by a few percentage points, the

biggest benefits seem to come from high quality input data in the form of domain-specific labeled data. Creating a domain-specific annotated corpus to train machine learning models, however, is not without its own challenges. For example, working with multiple human annotators can lead to discrepancies in the annotation results (Kim Jin-Dong & Junichi, 2008). Additionally, it is hard to estimate the total annotation cost and can depend on whether the annotator is capable of understanding the language for the task at hand (Arora Shilpa & Carolyn, 2009). Additionally, labelling an entire dataset incurs extremely high costs, which can be avoided. With the support of an Active Learner, a complete domain-specific corpus with its respective labels can be created using only partial annotations (Park Sungrae & Il-Chul, 2015). One of the key concepts of Active Learners is that if a machine learning algorithm is allowed to choose the data from which it learns, it will achieve higher accuracy with less training data. If a considerable amount of the data is unlabeled, this is especially desirable. As a result, the total cost of annotation can be reduced drastically. Research shows that the total number of manual annotations can be reduced by 80% when using an Active Learner instead of randomly selecting data to label (Jason & Miles, 2004). If data is manually annotated at random, the annotator will invest a lot of time into labeling irrelevant instances. This may incur costs which could be avoided with an Active Learner. It is argued that Passive Learning, or randomly selecting instances to be labeled by an annotator, is especially costly if the class distribution of the data is imbalanced or if there are many very similar documents. For example, if a specific feature set appears on only 1% of instances, the annotator would have to label 1000 documents to cover the feature set on 10 relevant documents. When it comes to document similarity, large clusters of very similar documents might be identifiable. Because features may be barely distinctable, the annotator might spend a lot of effort labeling uninformative instances when selecting them random. An Active Learner, on the other hand, suggests which instances the annotator should label. Those instances can be determined on various quantitative metrics (Miller Blake & R, 2020).

3 METHOD

If you define any equations (`\begin{equation}...`), you probably might want to define everything using math operators (e.g., `$D$`) and cite the work (!). So for example, following , representing a document $d \in D$ as $\text{tf}(d)$, we define a probabilistic model $(d|Y = y)$ for all documents in class y, and select y most likely to generate d:

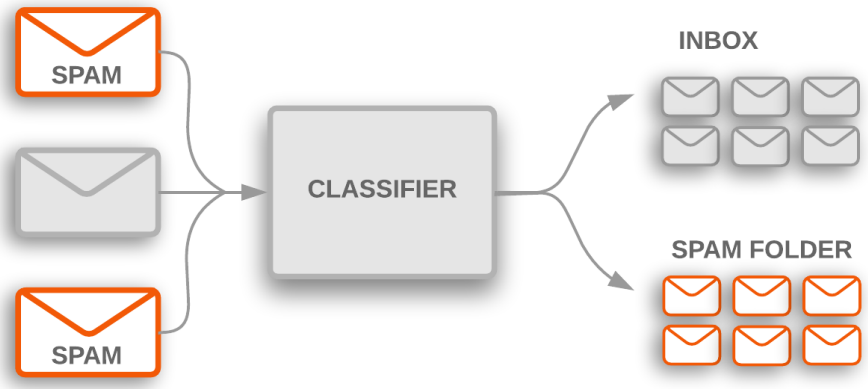

Figure 1: Spam classification example. Source: Google (CC BY 4.0).

Table 1: Best scoring models classifying bots, on Twitter and Facebook respectively. F_1 scores report positive (bot) class. Outline text left (l) and numbers right (r).

		F_1 score	
PCA	Models	Twitter	Facebook
300	Linear SVM ($C = 0.1$)	0.51	**0.91**
	Random Forest ($S = 5, F = 5$)	0.71	0.85
	Naive Bayes	0.61	0.73
500	Linear SVM ($C = 0.1$)	0.55	0.84
	Random Forest ($S = 5, F = 5$)	**0.76**	0.71
	Naive Bayes	0.41	0.64
	Majority	0.50	0.60

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(d|y) \cdot P(y) \tag{1}$$

With this we can detect spam (see Figure 1) or bots. Note that the figures (and tables for that matter) might not always be placed in this section (oh no)! LaTeX determines where to best put your objects, so don't worry about that. The reader will find them. **NOTE**: this Figure has a Creative Commons license; you cannot re-use other authors' figures without explicit permission or permissive licensing (as this would mean copyright infringement). You can refer to the equations as well (Equation 1)!

4 RESULTS

You have results and want to show them — probably in a table of some kind as you can see in Table 1. Highlight important scores with `\textbf{}`, use booktabs commands for structure: `\toprule \midrule \bottomrule`. APA does not allow vertical lines.

4.1 *Some Model*

If you have anything specific to talk about, use subsections, and refer to them as Section 4.1. Don't use paragraphs or subsubsections.

5 DISCUSSION

The 'to the moon' WallStreetBets movement had a tremendous impact on the lives of individuals, both to the positive and negative. Besides that, however, many investment funds have also been negatively impacted by the recent short-squeezes. While it might seem noble to root for individuals who try to force large funds out of their positions at big losses, it is easy to forget that many of those funds manage money for charitable endowments, pensions and others. Furthermore, such disruptions to the financial markets can harm its stability, thus causing spillover effects which can also negatively impact the lives of many people (Lyócsa Štefan & Tomáš, 2021). By being able to accurately measure and monitor the sentiment on WallStreetBets, market participants and regulators are able to preemptively take measures.
However, since the wallstreetbets subreddit has become very popular just recently, there is little academic research about the impact of the community on financial markets so far. Even though there is some research about sentiment analysis on wallstreetbets, that research does not use state of the art algorithms to perform sentiment analysis. This thesis not only tries to shine some light on those new and influential market participants, but also tries to put forward some methods that work best to perform sentiment analysis on the forum.
Not only did this thesis compare the performance of different models, but also proposed a highly efficient and reliable way to create a domain-specific annotated corpus, which can be used as the input to aforementioned models. To my knowledge, this thesis is the first research that creates a domain-specific corpus for the WallStreetBets forum. Researchers, such as Talamás (2021), specifically propose future work on "inclusion of features derived from alternative manipulation of the data like sentiment analysis could lead to new insights". I strongly believe that the methods proposed

in my thesis can lead to better sentiment classifiers, which can then be used in other scientific or industrial applications.

6 CONCLUSION

Done.

REFERENCES

Abhinav, A., & Jalaj, P. (2021). Wallstreetbets against wall street: The role of reddit in the gamestop short squeeze. *Indian Institute of Management Bangalore Research Paper Series*.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Arora Shilpa, N. E., & Carolyn, R. (2009, 01). Estimating annotation cost for active learning in a multi-annotator environment. *HLT-NAACL*. doi: 10.3115/1564131.1564136

Danbolt Jo, S. A., & Evangelos, V.-N. (2015). Investor sentiment and bidder announcement abnormal returns. *Journal of Corporate Finance*, 164-179.

Fu Xianghua, L. J., Yang Jingying, Min, F., & Huihui, W. (2018). Lexicon enhanced lstm with attention for general sentiment analysis. *IEEE Access*, 71884-71891.

Jason, B., & Miles, O. (2004, jul). Active learning and the total cost of annotation. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (p. 9-16). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-3202

Jemai Fatma, H. M., & Sahbi, B. (2021). Sentiment analysis using machine learning algorithms. *International Wireless Communications and Mobile Computing*, 775-779.

Kim Jin-Dong, O. T., & Junichi, T. (2008, 02). Corpus annotation for mining biomedical events from lterature. *BMC bioinformatics*, *9*, 10. doi: 10.1186/1471-2105-9-10

Long Cheng, L. B. M., & Larisa, Y. (2021). 'i just like the stock' versus 'fear and loathing on main street': The role of reddit sentiment in the gamestop short squeeze. *SSRN Electronic Journal*.

Lu Yue, D. U., Castellanos Malu, & ChengXiang, Z. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th international conference on world wide web* (p. 347–356). New York, NY, USA: Association for Computing

Machinery. Retrieved from https://doi-org.tilburguniversity.idm.oclc.org/10.1145/1963405.1963456 doi: 10.1145/1963405.1963456

Lyócsa Štefan, B. E., & Tomáš, V. (2021). Yolo trading: Riding with the herd during the gamestop episode. *Finance Research Letters*.

MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Miller Blake, L. F., & R, M. W. (2020). Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches. *Political Analysis*, *28*(4), 532–551. doi: 10.1017/pan.2020.4

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8-30.

Muhammad, A. (2014, 05). Detection and scoring of internet slangs for sentiment analysis using sentiwordnet. *Life Science Journal*, *11*, 66-72. doi: 10.6084/M9.FIGSHARE.1609621

Park Sungrae, L. W., & Il-Chul, M. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 38-44.

Pei Zhengqi, S. Z., & Yang, X. (2019, November). Slang detection and identification. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 881–889). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K19-1082 doi: 10.18653/v1/K19-1082

R., D. S., & Y., C. M. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 1375-1388.

Sebastian, R., & Vahid, M. (2019). *Python machine learning*. Packt Publishing.

Sepp, H., & Jürgen, S. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735

Sungrae Park, W. L., & Moon, I.-C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, *56*, 38-44.

Umar Zaghum, Y. I., Gubareva Mariya, & Shoaib, A. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of gamestop. *Journal of Behavioral and Experimental Finance*.

Wang Yanyan, L. J., Yin Fulian, & Marco, T. (2020, 08). Automatic construction of domain sentiment lexicon for semantic disambiguation. *Multimedia Tools and Applications*, *79*. doi: 10.1007/s11042-020-09030-1

APPENDIX A

If you have nothing to append: remove this. You can do a page referral for these, like: Appendix A (page 10).

APPENDIX B

And this!