

Enhancing Financial Market Sentiment Analysis Using Dataset Augmentation

Dante Contella^{1*}, Johnson Kinyua² and Charles Mutigwe³

¹College of Information, Sciences, and Technology, The Pennsylvania State University, State College, PA, USA (dantecontella@gmail.com)

²College of Information, Sciences, and Technology, The Pennsylvania State University, State College, PA, USA (jdk450@psu.edu)

³College of Business, Western New England University, Springfield, MA, USA (charles.mutigwe@wne.edu)

*corresponding author

Abstract – We developed a series of financial social media sentiment analysis models starting with PyFin-Sentiment, a recent state-of-the-art domain-specific large language model for the financial domain. We focused on label accuracy and improvements in the quality and quantity of the dataset by collecting more data. We were able to improve model accuracy from 0.70 to 0.93, supporting a well-known maxim that models are just as good as the data they are trained on. This created a consistently accurate model that outperformed other existing financial sentiment models. We also used nuanced approaches such as separating social media data into short and long comments which drastically improved the accuracy of our model.

Keywords – sentiment analysis, machine learning, financial market, social media, StockTwits

I. INTRODUCTION

Investment decisions often rely on information from external sources to accurately make decisions. This information is a key aspect in helping analysts and investors identify trends, associations and other relationships in order to make predictions about where a stock's price or the market is

headed. The outlets to acquire such information are vast, especially in the age of the Internet. Important financial reports and organizational news are easily accessible through a quick internet search. More often than not, most of this news is reported in real time, by major news outlets, giving investors access to information in the most time efficient manner. Information such as earnings reports, upcoming conferences, mergers, hirings and departures of key figures, major sales contracts or partnerships, and internal financial documents are the most prevalent of information types. This information is mostly catered towards institutional investors. However, after the COVID pandemic, there has been a noticeable uptrend of individuals engaging in investment opportunities [1]. The uptrend of retail investors results in the typical investment information influencing a larger investor pool. This new investor pool also provides new sources of information on investor trends.

Social media platforms such as Reddit and X (formally Twitter) have their own sub-communities dedicated to retail investors. The interest for retail investors to communicate about investment opportunities has risen to the point where there are dedicated websites, such as StockTwits, which are solely focused on retail investors discussing potential investments. Included with these retail investors, CEOs, public figures, and financial advisors share their opinions on these various social media platforms [2]. The presence of such prominent figures, paired with the slew of retail investors giving financial opinions in real time leaves social media as a valuable body of knowledge to analyze the latest investment information. Paired with this investment information on social media platforms comes the need to properly analyze all of these opinions. The most common way that these opinions are

displayed across social media platforms are by comments. Short form comments, ranging from a few words to 2-3 sentences, are the most common. Short form opinions and information vary from the typical investment information inputs which primarily come in the form of news articles or financial statements that can be many paragraphs to pages long [4]. Short form content, such as social media comments, provide sentiment about an investment, which is the goal for most financial information. The sentiment provides positive, negative, or neutral opinions regarding certain investments. In this paper we will construct an AI model to classify the sentiment of financial social media comments. The ability to capture investor sentiment of a specific investment in a time and resource effective manner can be a useful tool to make decisions regarding investments.

Researchers have used a variety of techniques for sentiment analysis in the financial domain including machine learning, dictionary-based approaches, deep learning and large language models [5]. Sentiment analysis in the financial domain requires domain-specific models because this domain uses its own vocabulary, and general-purpose models such as (BERT) [6], ULMFit [7], ELMo [8], XLNet[9], GPT [10]; and LLama [11]; and are likely to underperform. For that reason, several models for financial sentiment analysis have been developed including FinBERT [12, 13, 14], NTSU-Fin [15], PyFin-Sentiment [16], FinSoSent [5], etc.

One major challenge when developing social sentiment analysis models for the financial domain is lack of datasets on which these models can be trained. There are some datasets that have been used for model development in this area such as SemEval- 2017 Task 5 [17], and Fin-SoMe [18], SSIX [19], Fin-Lin [20], Sanders [21], and Tarborda [22]. These datasets vary in size and context and there are issues related to multiple sentiments being present in a document. Some of them are relatively small to enable models to learn. For example, SemEval-2017 Task 5 contains a subtask (subtask 1) which consists of 2510 labeled messages from StockTwits and Twitter; and SSIX consists of only 2886 financial messages

from StockTwits and Twitter with opinion targets. The *FinSoMe* data set consists of 10,000 social media posts from StockTwits, and authors labeled every post with a market sentiment.

PyFinSentiment [16] is a recent state-of-the-art model developed recently which can be applied to social media posts in the financial domain. The developers of this model, Wilksch and Abramova, have reported that their PyFin-Sentiment model was able to outperform current sentiment analysis, and specifically financial sentiment models in accuracy. Models tested in their work included FinBERT, VADER, NTUSD-Fin, and Twitter RoBERTa [23]. PyFin-Sentiment, it is a logistic regression model trained off of data from StockTwits as well as other social media comments. The developers used the scikit-learn Python library to create the model. After testing the performance of a number of models including, but not limited to Random Forest, Decision Trees, and SVC models, the developers came to the conclusion that the Logistic Regression model consistently performed the best. Logistic Regression is commonly used in classification tasks, such as sentiment analysis, in which it is sensible that logistic regression has great performance. The predictions of the PyFinSentiment model had labels of positive, negative, and neutral. Comments with a positive sentiment were labeled with a 1, neutral comments were labeled with a 2, and comments with a negative sentiment were labeled with a 3. We picked this model as a starting point for enhancing performance in financial sentiment analysis as discussed in the materials and methods section.

This paper is organized as follows: The analysis of related work is introduced in Section 2. Section 3 will describe the resources, tools, methods used, and steps

taken to conduct this research. The results of the research will be presented in Section 4. Section 5 will be dedicated to final discussions, conclusions, and possible future work.

II. MATERIALS AND METHOD

We first decided to confirm the results reported by developers of PyFin-Sentiment by running the model against its own dataset. Considering it was trained on a dataset consisting of about 10,000 social media finance related social media comments, we expected the model to perform relatively well. These comments were hand labeled as well, based on this we expect the accuracy of the training data to be relatively high. When we ran the model against its own dataset, it yielded an accuracy score of about 75%. This result would be considered a good model in many instances. We decided to test the model against some of our own data. We collected 10,000 comments from the popular financial platform StockTwits. These comments were already labeled by users as bullish or bearish. We tested these against the PyFinSentiment model, which yielded us with promising results similar to the test of its own dataset. In order to demonstrate the repeatability of PyFin-Sentiment model performance results, we used the scikit-learn Python library to develop two models. One of which was trained on the PyFin-Sentiment data and the other trained on our 10,000 StockTwit comments. Note that these StockTwit comments do not include a neutral category, the importance of this will be discussed in later sections. After creating these models, we tested them against partial parts of the trained dataset and as well as the opposite. It is also important to note that for the 10,000 StockTwits model, we omitted the neutral category for the testing data. We were able

to garner results of around 71-75% accuracy. Considering our testing was similar to PyFin-Sentiments results, we decided to develop our models further.

As discussed in the above section, we found it most useful to use the logistic regression model, in alignment with results reported in PyFin-Sentiment.

Considering our proposed model only used around 10,000 comments as training data, it was first assumed that more training data was needed to accurately predict financial sentiment. As a result, we scraped more data from StockTwits, to the scale of around 130,000 comments. However, we suspected that the StockTwits data could be misrepresented. While the data contain the labels of bullish and bearish, we anticipated that there may be some people who incorrectly label their comments not only on purpose for sarcastic reasons, but also inadvertently. This issue was also noted in [5] that many financial sentiment datasets have meaningful inaccuracies as they are usually not labeled by financial professionals.

With this possibility of mislabeling, we needed the most accurate way to properly label the financial sentiment of the 130,000 StockTwit comments. We decided to automate the labeling process by using 4 different state-of-the-art sentiment analysis models on each comment in order to get four sentiment scores, and to then get the average sentiment score for each comment. The state-of-the-art sentiment analysis models we used were PyFin-Sentiment, FinBERT, Twitter RoBERTa, and VADER. These models all have reported high accuracies on financial sentiment analysis, while VADAR has consistently high rates in general sentiment analysis. It is an important factor to realize all 4 of these models classified sentiment in different ways. For example, PyFin-Sentiment graded sentiment with scores of 1-3, while models like FinBERT used a variable weighted scale to grade sentiment. It was of the utmost importance to standardize model outputs to the same scale in order to properly calculate averages. We choose the model for PyFin-Sentiment to be the standard with 1 as the label for positive sentiment, 2 as the label for neutral sentiment, and 3 as the label for negative sentiment.

After all of the labeling was done we took the averages of all 4 models, and used the averages as the new correct sentiment label for the comment. However, considering all of the options, the outputs were not always a whole number, but could be a decimal such as 1.75. After observing a large amount of the results, it was concluded that a fair final grading system would be as follows: 1-1.75 would be positive 2 and 2.25 would be neutral, and 2.5-3 would be negative. This resulted in the most accurate outcomes after reclassifying and double checking final results.

This dataset would yield very promising results, but technically they were not hand graded, but rather the average of top performing models. In order to negate the possible risk of relying too heavily on this data, we used other research of financial sentiment analysis. These datasets included the PyFin-Sentiment dataset, the Sanders dataset, and the FinSoMe dataset. All of these were labeled, in some cases by professionals of finance. While these also had slightly different labels associated with them, in a similar fashion to the way the previous models labeled differently, we were able to break down the datasets labels into 3 categories: 1 for positive, 2 for neutral, and 3 for negative. However, since classification typically uses the lower number as negative, we decided to switch the categories to 1 for negative, 2 for neutral, and 3 for

positive. This left us with 2 datasets that were relatively accurate in an attempt to train our models. The results will be discussed in the next section.

Figures and Tables

Table 1. All Models using 10k Comments and PyFin Sentiment Data

Model	Accuracy	Precision Macro	Recall Macro	F1 Score Macro
Logistic Regression	0.70894	1 - 0.72 2 - 0.64 3 - 0.75	0.71 0.61 0.79	0.72 0.62 0.77
ANN	0.71369	1 - 0.72 2 - 0.64 3 - 0.76	0.75 0.61 0.77	0.73 0.63 0.77
Random Forest	0.67891	1 - 0.72 2 - 0.64 3 - 0.68	0.67 0.56 0.79	0.69 0.60 0.63
Decision Tree	0.57022	1 - 0.58 2 - 0.51 3 - 0.62	0.57 0.55 0.58	0.58 0.53 0.60
SVC	0.70894	1 - 0.73 2 - 0.63 3 - 0.76	0.72 0.62 0.78	0.72 0.62 0.77

Short Comment Model Performance:
Accuracy: 0.9243697478991597
Classification Report:
precision recall f1-score support
1 0.93 0.68 0.78 2102
3 0.92 0.99 0.95 8251
average 0.92 10353
macro avg 0.93 0.83 0.87 10353
weighted avg 0.92 0.92 0.92 10353

Long Comment Model Performance:
Accuracy: 0.8832278015019879
Classification Report:
precision recall f1-score support
1 0.88 0.56 0.68 1531
3 0.88 0.98 0.93 5260
average 0.88 6791
macro avg 0.88 0.77 0.81 6791
weighted avg 0.88 0.88 0.87 6791

Fig. 1 Comment Length Accuracy

Accuracy on combined dataset: 0.9222548882355687
Classification Report on combined dataset:
precision recall f1-score support
1 0.92 0.99 0.95 67546
3 0.93 0.68 0.79 18170
accuracy 0.92 85716
macro avg 0.93 0.83 0.87 85716
weighted avg 0.92 0.92 0.92 85716

Fig. 2 Accuracy on Combined Dataset with Length

Accuracy on Test Set: 0.9198553429771349
Classification Report:
precision recall f1-score support
1 0.79 0.85 0.82 3634
3 0.96 0.94 0.95 13510
accuracy 0.92 17144
macro avg 0.87 0.90 0.88 17144
weighted avg 0.92 0.92 0.92 17144

Fig. 3 Accuracy on Combined Dataset including Length Separation with Different Vectorization Technique

Sampled Data with SMOTE Model Performance:
Accuracy: 0.9156910389514763
Classification Report:
precision recall f1-score support
1 0.77 0.86 0.82 2935
3 0.96 0.93 0.95 10646
accuracy 0.92 13581
macro avg 0.87 0.90 0.88 13581
weighted avg 0.92 0.92 0.92 13581

Accuracy on combined dataset: 0.9341079845069765
Classification Report on combined dataset:
precision recall f1-score support
1 0.81 0.90 0.85 18170
3 0.97 0.94 0.96 67546
accuracy 0.93 85716
macro avg 0.89 0.92 0.90 85716
weighted avg 0.94 0.93 0.94 85716

Fig. 4 Updated Comment Length Accuracy Scores with New Resampling

Hybrid Model Accuracy: 0.9060312645823612
Classification Report:
precision recall f1-score support
1 0.91 0.62 0.74 3634
3 0.91 0.98 0.94 13510
accuracy 0.91 17144
macro avg 0.91 0.80 0.84 17144
weighted avg 0.91 0.91 0.90 17144

Fig. 5 Updated Comment Hybrid Model with Different Vectorization Techniques

III. RESULTS

We first developed a Python script to extract data from Stocktwits, a popular social media platform for investors. Using this script, we created a dataset of 10,000 comments. These comments were then categorized as either bullish or bearish, allowing us to create an initial sentiment analysis model. When tested, this model showed good accuracy, 0.70, indicating that the data collected was highly indicative of market sentiment.

Encouraged by the initial results, we sought to enhance our model's robustness by incorporating another dataset, PyFin-Sentiment, which is widely recognized for its comprehensive financial sentiment annotations. We used this dataset to build a new model and achieved an accuracy rate of 75%. To validate the performance of the PyFin-Sentiment model, we tested it against its own dataset and observed good accuracy, 0.7524 or about 75%.

Understanding that data quality significantly affects model performance, we proceeded to refine our datasets hoping to improve the performance of the model. We employed various

data cleaning techniques, which involved removing noisy data and standardizing comment formats. Details of the cleaning formula are as follows: convert all text to lowercase, replace ticker symbols with “TICKER”, replace mentions of usernames with “@user”, replace digits with “9”, replace new line characters with spaces, replace URL links with “link”.

Despite these efforts, when we attempted to merge the PyFin-Sentiment data with our Stocktwits dataset, the combined model’s performance was barely impacted. We hypothesized that this result was due to differences in the formality of the comments across the two datasets; the PyFin-Sentiment comments appeared to be more formal compared to the more casual and colloquial nature of StockTwits comments.

To explore further improvements, we experimented with more machine learning models including Artificial Neural Networks (ANN), Random Forest, Decision Tree, and Support Vector Classification (SVC). Each model was tested against our combined dataset to evaluate performance variations. The results varied widely, but none significantly outperformed our baseline logistic regression model. The results are shown in Table 1 below.

Given these outcomes, we decided to continue using logistic regression as our primary model. This decision was also influenced by the fact that PyFin-Sentiment could be considered as a state-of-the-art tool in financial sentiment analysis, and recommended logistic regression as its preferred model.

Realizing that our models could benefit from a larger dataset, we expanded our Stocktwits data collection to 100,000 comments. With this significantly larger dataset, we conducted a thorough comparison to identify any differences in model performance. We noticed that many existing models were built on the assumption that comments could be neatly categorized into positive, negative, or neutral. To challenge this assumption, we employed four different financial sentiment analysis models—FinBERT, PyFin-Sentiment, VADER, and TwitterRoberta—to classify the comments.

Utilizing these models, we replaced manual comment classification with automated processes, which allowed us to enhance the accuracy of our sentiment classifications. We integrated the average score given by the four previously mentioned models of 100,000+ comments from Stocktwits into our model, resulting in a modest improvement in performance, 0.86305 or about 86%. However, after further testing of this model on proven datasets such as Sanders, FinSoMe, and PyFin-Sentiment, the true accuracy of the model did not reflect the results for comments pulled from sources other than StockTwits. After analyzing these results, we decided to create a model using only the datasets of Sanders, FinSoMe, PyFin-Sentiment as they seemed to be the most accurate. This model showed an accuracy of over 70%. To enhance this new model, we then decided to revert to the original two-class classifier approach, categorizing comments as either positive or negative. This adjustment led to a significant 10% boost in model accuracy, 0.81618. This model maintained a consistent accuracy score of over 80% for most datasets that it was tested on. This could lead to the possible example that having many hand classified comments by professionals varying in formality, can lead to excellent training data for financial sentiment analysis models. However we were limited to the around 30,000 comments that came from these data sources, so the next best test would be to

combine the averaged comments of over 100,000 with the 30,000 comments.

Further analysis of the hand-picked comments versus the 100,000 Stocktwits comments initially suggested that differences in formality were causing inconsistencies in our models. Upon deeper investigation, however, we discovered that the length of the comments played a more critical role in determining model performance. This realization prompted us to develop two new models: one focused exclusively on shorter comments and the other on longer comments. Both models demonstrated substantial improvements in accuracy. Results displayed in Figures 1 and 2 below.

Building on these findings, we created an additional model that applied different vectorization techniques based on comment length. We also attempted to apply SMOTE as a resampling technique. By tailoring vectorization and resampling processes to the specific characteristics of the text, these models also showed excellent results, further validating our approach to data-specific model optimization. In addition to these techniques, we changed the data labels in our dataset to 0 and 1 instead of 1 and 3, but the results showed no noticeable differences.

These optimal models were generated by combining the Sanders, FinSoMe, and PyFin-Sentiment dataset along with the 100,000+ comment dataset that was based on the averages. Different test/train data splits were used along with new vectorization techniques. Results are displayed in Figures 3-5.

IV. DISCUSSION

It is important to note that the models are only going to be as good as the data that they are trained on. Reiterating the point from [4] and [5], if you took the average person and looked at some of the listed comments it would be hard for one to decide if it is positive, negative, or neutral. This would be based on your understanding of financial knowledge, sarcasm, and language in general. Considering these factors, it is seemingly difficult to create an accurate machine learning model that can properly handle all types of financial information such as knowledge, people, and sarcasm base to name a few categories. Our best approach seemed to be getting a well-rounded model with equal data on different topics and from different sources. Based upon where the source of comments comes from can also have an effect on whether or not the model performs well. Our Model performed exceptionally well on StockTwits data. StockTwits, Reddit, and X (formally Twitter) all may have different types of information regarding formality, sarcasm, and financial knowledge. Having a well-rounded dataset is a crucial part in financial sentiment analysis. The other factor that we noticed is nuanced methods such as comment length. We hypothesize that with longer comments, there is more room to contradict previous statements which may cause confusion to models. Some overlooked features may still exist and could be further analyzed.

While we acknowledge that sentiment prediction can naturally lend itself to an ordinal framework (positive, neutral, negative), we opted to treat each sentiment class as a separate binary classification task due to both practical and performance considerations in initial model development. This approach allowed for greater flexibility in handling class imbalance and optimizing individual classification thresholds, which can be particularly useful given the sometimes

ambiguous nature of financial sentiment in social media contexts.

We recognize the potential benefits of employing models that explicitly account for the ordinal nature of sentiment, such as the ordered-logit model. These methods may better capture the structured relationship between sentiment levels, potentially leading to improved predictive performance. For reference, one machine-learning study by Jiang deals with ordered outcomes in the context of credit rating prediction (AAA, AA, A, B, etc) [3]. Incorporating such approaches represents a valuable direction for future research.

Recommendations for further development would include finalizing a model based off of recognizing neutral comments and combining that with the identification of positive and negative comments. It is important to note that through analysis of our collected data, the neutral comments primarily consisted of either unrelated comments or the comments had slight positive or negative connotations. Considering this, it may prove more effective to replace the neutral classifier with unrelated, but this we are unsure of. It may be possible that creating different models for different platforms, ex: one for StockTwits, one for Reddit, etc. would be more beneficial as language seems to differ by platform as well. It would also be recommended to create larger datasets of hand labeled financial comments to create a more accurate model. Further tuning of models using a similar strategy of the models created in this paper may help in the development of financial sentiment classification

V. CONCLUSION

Our study highlights the complexities involved in financial sentiment analysis of social media comments. While logistic regression proved to be a reliable model for this task, the research also underscored the importance of data characteristics, such as comment length and style, in influencing model performance. Moving forward, these insights will be crucial in developing more accurate sentiment analysis tools for financial markets, potentially enabling more informed decision-making based on social media sentiment.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

REFERENCES

- [1] R. Ortmann, M. Pelster, and S. T. Wengerek, "COVID-19 and investor behavior," **Finance Res. Lett.**, vol. 37, p. 101802, 2020. [Online]. Available: <https://doi.org/10.1016/j.frl.2020.101802><https://doi.org/10.1016/j.frl.2020.101802>
- [2] Ludwig, Zachary & Perkowski, Patryk. (2021). An Analysis of How Twitter Impacts Financial Markets. *Journal of Student Research*. 10. 10.47611/jsrhs.v10i3.2224.
- [3] Jiang, Y. (2023). *A Primer on Machine Learning Methods for Credit Rating Modeling*. IntechOpen. doi: 10.5772/intechopen.107317
- [4] Georgios Fatouros, John Soldatos, Kalliopi Kouroumalis, Georgios Makridis, Dimosthenis Kyriazis, Transforming sentiment analysis in the financial domain with ChatGPT, *Machine Learning with Applications*, Volume 14, 2023, 100508, ISSN 2666-8270, <https://doi.org/10.1016/j.mlwa.2023.100508>. (<https://www.sciencedirect.com/science/article/pii/S2666827023000610>)
- [5] Delgadillo, J.; Kinyua, J.; Mutigwe, C. FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models. *Big Data Cogn. Comput.* 2024, 8, 87. <https://doi.org/10.3390/bdcc8080087>
- [6] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers), pp. 4171–4186.
- [7] Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 15–20 July 2018; Volume 1: Long Papers, pp. 328–339.
- [8] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, 1–6 June 2018; Volume 1 (Long Papers), pp. 2227–2237.
- [9] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* 2019, 1, 8.
- [10] OpenAI, GPT-4.1. Available online: <https://platform.openai.com/docs/models/gpt-4.1>.
- [11] Meta, Llama 3.0. Available online: <https://ai.meta.com/blog/meta-llama-3/>.
- [12] Araci, D.T.; Zulkuf Genc, Z. FinBERT: Financial Sentiment Analysis with BERT. *Prosus AI Tech Blog*. 2020. Available online: <https://medium.com/prosus-ai-tech-blog/finbert-financial-sentiment-analysis-with-bert-b277a3607101>.
- [13] Desola, V.; Hanna, K.; and Nonis, P. FinBERT: Pretrained Model on SEC Filings for Financial Natural Language Tasks; Technical Report; University of California: Los Angeles, CA, USA, 2019.
- [14] Liu, Z.; Huang, D.; Huang, K.; Li, Z.; and Zhao, J. FinBERT: A Pretrained Financial Language Representation Model for Financial Text Mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, Virtual, 7–15 January 2021; pp. 4513–4519.
- [15] C. Chen, H. Huang, and H. Chen, "NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications," in *Proceedings of the LREC 2018 Workshop "The First Financial Narrative Processing Workshop (FNP 2018)"*.
- [16] M. Wilksch, and O. Abramova, "PyFin-sentiment: Towards a machine-learning-based model for deriving sentiment from financial tweets," *International Journal of Information Management Data Insights*, 3 (2023) 100171. <https://doi.org/10.1016/j.jjime.2023.100171>
- [17] Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., et al., (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Association for computational linguistics (ACL)* (pp. 519–535). Association for Computational Linguistics. 10.18653/v1/S17-2089.
- [18] Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2020). Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6106–6110)
- [19] Gaillat, T.; Zarrouk, M.; Freitas, A.; Davis, B. The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan; 7–12 May 2018; pp. 2671–2675.

- [20] Daudert, T. A Multi-Source Entity-Level Sentiment Corpus for the Financial Domain: The Fin-Lin Corpus. arXiv **2020**, arXiv:2003.04073. Available online: <http://arxiv.org/abs/2003.04073>
- [21] Saif, H.; Fernandez, M.; He, Y.; Alani, H. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), Turin, Italy, 3 December 2013.
- [22] Taborda, B.; de Almeida, A.; Dias, J.C.; Batista, F.; Ribeiro, R. Stock Market Tweets Data. IEEE Dataport **2021**.
- [23] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv **2019**, arXiv:1907.11692. Available online: <http://arxiv.org/abs/1907.11692>