TweetFinSent: A Dataset of Stock Sentiments on Twitter

Yulong Pei[1], **Amarachi Mbakwe**[2], **Akshat Gupta**[2], **Salwa Alamir**[1]
Hanxuan Lin[3], **Xiaomo Liu**[2], **Sameena Shah**[2]
[1]JP Morgan AI Research, London, UK
[2]JP Morgan AI Research, New York, USA
[3]JP Morgan, Shanghai, China
{yulong.pei,xiaomo.liu}@jpmchase.com

Abstract

Stock sentiment has strong correlations with the stock market but traditional sentiment analysis task classifies sentiment according to having feelings and emotions of good or bad. This definition of sentiment is not an accurate indicator of public opinion about specific stocks. To bridge this gap, we introduce a new task of stock sentiment analysis and present a new dataset for this task named TweetFinSent. In TweetFinSent, tweets are annotated based on if one gained or expected to gain positive or negative return from a stock. Experiments on TweetFinSent with several sentiment analysis models from lexicon-based to transformer-based have been conducted. Experimental results show that TweetFinSent dataset constitutes a challenging problem and there is ample room for improvement on the stock sentiment analysis task. TweetFinSent is available at https://github.com/jpmcair/tweetfinsent.

1 Introduction

Sentiment analysis, as a classical research problem in machine learning and natural language processing, aims to analyze peoples opinions, sentiments, and emotions towards entities such as products, services, organizations, individuals, and their attributes (Liu, 2012). A large amount of attention in industry and research community has been given to analysing sentiment of Twitter feeds. This has been done to analyse the effectiveness and predicting the result of election campaigns (Wang et al., 2012; Ramteke et al., 2016), analyse Twitter mood during the Covid-19 outbreak (Manguri et al., 2020; Dubey, 2020) and to analyse and predict the stock market. It has been repeatedly shown in literature that the Twitter sentiment has strong correlations with the stock market, with several works on predicting the stock market movement based on Twitter sentiment (Bollen and Mao, 2011; Bollen et al., 2011; Mittal and Goel, 2012). For instance, recent discussions of meme stocks on social media such as Twitter and Reddit have attracted significant attention and influenced the sentiment of investors especially young and inexperienced investors[1]. Therefore, it is of great value to analyse stock sentiment in both practice and research.

Despite the wide interest and importance, most existing research on sentiment analysis focused on distinguishing if the text contains or a user has feelings or emotions of good or bad. However, in the financial domain, we would like to analyse more specific and concrete sentiment, i.e., we aim to re-calibrate the definition of sentiment to include this desired property such as gaining or expecting to gain positive or negative return from a stock. Although traditional sentiment analysis of Twitter feeds correlates with the stock market dynamics to some extent, it is not an accurate indicator of public opinion about financial returns of specific stocks. In worst case, traditional sentiment analysis methods may classify tweets into controversy sentiment due to various factors such as finance-specific terms. Some representative examples are shown in Table 1. To bridge the gap, we introduce the concept of *stock sentiment*, where a positive sentiment indicates the opinion of a stock value increasing, a negative sentiment indicates the opinion of a stock value decreasing, and a neutral sentiment indicating that the given sentence does not make predictions for either. Stock sentiment is inherently related to the mention of a specific stock in the sentence. Based on the new definition of stock sentiment, we introduce the task of stock sentiment analysis, underlining the need for moving away from the traditional sentiment analysis definition.

[1]https://en.wikipedia.org/wiki/Meme_stock

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 37 - 47

Table 1: Some examples showing the differences between traditional sentiment and stock sentiment. For the traditional sentiment analysis, RoBERTa-base model trained on 124M tweets and fine-tuned for sentiment analysis with the TweetEval benchmark (Loureiro et al., 2022) is used.

Tweet	Target Ticker	Traditional Sentiment	Stock Sentiment
Bubbles burst an any given moment. Maybe $TSLA bubble will burst with the Bitcoin buy.	$TSLA	Neutral	Negative
$BABA is on yolo status and I almost sold $BIDU lol.	$BABA	Neutral	Positive
$SOFI Not touching it. I love the company though. We all know the rules, and know what happens during the lockup expiry	$SOFI	Positive	Negative
*Buy the f*cking dip! Hold the line! $AMC $GME $NOK*	$AMC	Negative	Positive

We then construct an expert-annotated dataset for stock sentiment analysis called TweetFinSent which will be made publicly available to the research community. We benchmark this dataset with various state-of-the-art baselines. Experimental results show that TweetFinSent dataset constitutes a challenging problem and there is ample room for improvement on the stock sentiment analysis task.

In summary, our main contributions are three-fold:

- We construct and release TweetFinSent, a new Twitter stock sentiment dataset. To the best of our knowledge, this is the first resource for stock sentiment analysis.
- We demonstrate the utility of the TweetFinSent dataset by evaluating different types of state-of-the-art sentiment analysis models on our dataset.
- We investigate the performance of different baselines and outline the challenge of the stock sentiment analysis task and future directions.

2 Related Work

The tremendous growth of unstructured text data has spurred research in NLP, especially in the area of sentiment analysis, which involves classifying and analyzing of people's opinions, emotions, and sentiments from textual data (Liu, 2012). In NLP, sentiment analysis plays a significant role in analyzing the emotions or feelings behind written texts which serve different purposes depending on the domain of its applications. Since sentiment analysis is an increasingly valuable tool for many organisations to enhances their decision-making, it has been extended to variety of use cases. However, we'd like to argue the use case of this study is unique in the sense that stock sentiment on Twitter is considerably different from traditional sentiment analysis. In the following, we review most relevant prior work and then highlight the value of our study and dataset.

Twitter sentiment analysis: Twitter sentiment analysis is an important area and has attracted much attention. It is considered a more challenging problem than general sentiment analysis on conventional texts because of the frequent use of slang, irregular words, informal words, and a vast number of tweets on various topics. Twitter sentiment analysis has applications in business management, public actions understanding, political analysis, and other domains. Previous works in Twitter sentiment analysis include sentiment analysis to assist stock prediction (Qasem et al., 2015; Pagolu et al., 2016), discovering brand perception (Arora et al., 2015; Gursoy et al., 2017), and analyzing and predicting election results (Xia et al., 2021; Budiharto and Meiliana, 2018). Researchers proposed different methods to solve this problem including lexicon-based (Elbagir and Yang, 2019), machine learning (Qasem et al., 2015), and hybrid methods (Kolchyna et al., 2015). Recent works (Bozanta et al., 2021; Mathew and Bindu, 2020) have applied transformers for sentiment analysis tasks.

Stock sentiment analysis: stock sentiment analysis significantly differs from general sentiment. It differs in terms of domain and purpose. The purpose behind stock sentiment analysis is usually to predict the stock markets reaction to the sentiments hidden in the text. Previous works have attempted to forecast stock prices using price history. Recent works have begun using textual data for predicting the stock markets reaction. For example,

stock market values were predicted using news articles (Kalyani et al., 2016), news headlines (Nemes and Kiss, 2021), and sentiments on social media (Qasem et al., 2015; Mittal and Goel, 2012). Apple Inc. companys news data were collected by (Kalyani et al., 2016) and performed sentiment analysis using supervised machine learning to understand the relationship between news and stock trend. Sentiment analysis of economic news headlines was used by (Nemes and Kiss, 2021) to predict the stock value changes for giant tech companies. (Xing et al., 2020) investigated the error patterns of some widely acknowledged sentiment analysis methods in the finance domain. There have been several sources of data for stock sentiment analysis. Popular sources of data include Financial PhraseBank (Araci, 2019), Yahoo Finance (Koukaras et al., 2022), Finviz (Nemes and Kiss, 2021), StockTwits Data (Araci, 2019), and SemEval (Cortis et al., 2017).

Twitter sentiment for stock analysis: Since Twitter provides a real-time information channel that can generate information about the market even before the leading newswires, it has been investigated for stock analysis. For example, (Souza et al., 2015) showed that social media can be a valuable source in the analysis of the financial dynamics in the retail sector. Also, the collective mood states (happy, calm) derived from large-scale Twitter feeds were correlated to the value of the Dow Jones industrial average over time (Bollen and Mao, 2011). Likewise, the rise and fall in stock prices and public sentiments in tweets were shown in (Pagolu et al., 2016; Smailović et al., 2013) to be strongly related. One of the challenges in Twitter sentiment analysis is lack of labeled data. Most recent works (Pagolu et al., 2016; Aattouchi et al., 2022; Nousi and Tjortjis, 2021) extracted tweets from the Twitter platform. Although some of these datasets are usually prepared by automatic sentiment detection of messages or manually determining the sentiments (Skuza and Romanowski, 2015), they are still in realm of traditional definition ("good" and "bad") of sentiments for stock movements. However, this study is more about retail investors' expected gain or loss from their investments as "stock sentiment" (please refer to Section 3.1 for the formal definition).

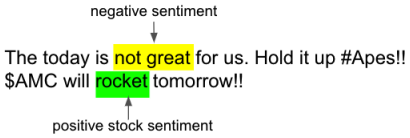


Figure 1: Sentiment vs Stock Sentiment

To the best of our knowledge, no labeled Twitter stock sentiment analysis dataset exists so far. In this paper, we construct and release an expert-annotated Twitter stock sentiment analysis dataset for the downstream stock analysis. This dataset is an essential step toward addressing the missing link of such a dataset in financial industry. The goal of releasing this dataset is to spur the development of more advanced algorithms and for the effective comparisons of these algorithms.

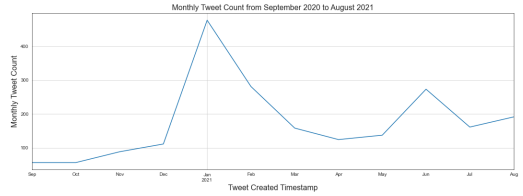
3 The TweetFinSent Dataset

3.1 Task Definition

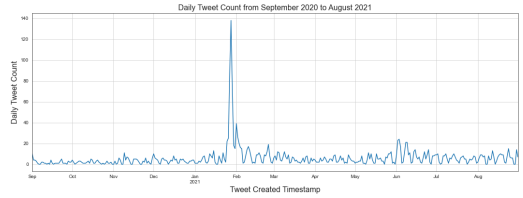
This study concentrates on a hypothetical use case that financial analysts need conduct equity analyses for a list of stocks and would like to take into account impact of online meme stock communities, in which these stocks may gain popularity on social media platforms like Twitter. Retail investors may rally on these platforms and have collective investment actions on them. Therefore, it can be important for financial analysts to understand the online stock sentiments which are defined as follows.

- **Positive**: Gained or expected to gain positive return from a stock
- **Negative**: Received or expected to receive negative return from a stock
- **Neutral**: Other situations

As one can observe, the stock sentiment in this study correlates but also differentiates from the ordinary sentiment which has been well studied in various scenarios such as product reviews and public opinions etc. These commonly discussed sentiments are more about feelings and emotions of good and bad (Liu, 2012). Nonetheless, the stock sentiment is more about price moving up and down. Stock sentiment and ordinary sentiment can certainly be the same thing. But they sometimes also can be completely unrelated. Figure 1 shows such an example where the indicators for different sentiments are highlighted. In this tweet, the ordinary sentiment to the market is negative,



(a) Number of tweets per month.



(b) Number of tweets per day.

Figure 2: Number of tweets in TweetFinSent during the time. The number of tweets spike correlates with the GameStop short squeeze in January 2021. The subreddit r/WallStreetBets posts, comments, and Twitter tweets by retail investors related to four meme stocks (GameStop, Nokia, AMC, and Blackberry) initiated the GameStop short squeeze in January, 2021 (tefan Lyócsa et al., 2022; Didier et al., 2022; Chohan, 2021).

but it also expects a specific stock $AMC to rise, which indicates positive stock sentiment. More examples can be found in Table 1.

In the context of social media, an online post such as a tweet P may contain the discussions of multiple stock tickers $G = \{g_1, g_2, ..., g_n\}$, we are interested in calculating the stock sentiment $S(g|G, P)$ towards a target ticker g within a post P. For example, given the following tweet:

> *@PhoShoBro I sold $1000 worth today of my $CLOV and threw it in my $FUBO position and some in $LGHL*

if the target ticker is $CLOV, the stock sentiment is *negative* because this user sold $CLOV. However, if the target ticker is $FUBO or $LGHL, the sentiment is *positive* because she bought $FUBO and $LGHL which indicates that she expected positive return from them. Note that in our TweetFinSent dataset, given a tweet, the target ticker is also provided.

3.2 Data Preparation

We collected 300 stock tickers of interests covering technology, consumer goods and energy etc. various sectors. We then used Twitter's standard search API[2] to retrieve recent 7 days' tweets containing one or multiple stock tickers of interests. Due to the rate limit of Twitter API, at most 17,280 tweets can be collected everyday. The data collection process was ongoing for 12 months from Sep., 2020 to Aug., 2021. Since this study only focuses on the English content, non-English tweets were filtered by the language tag in tweet metadata from API and also using some heuristics developed by authors. After that, a random sample of 2,113 tweets were selected for stock sentiment annotation to construct the TweetFinSent dataset. The volume of tweets per month and per day in TweetFinSent are shown in Figure 2. It is observed that there are two peaks in Figure 2a and 2b. This is consistent with the fact that retail investors initially gathered on r/wallstreetbets[3] and then on Twitter to start a short squeeze on GameStop, pushing their stock prices up significantly from January 22, 2021[4].

3.3 Annotation Procedure

The annotation procedure consists of three steps: (1) annotation guideline discussion to establish criteria of assigning sentiment labels; (2) pilot annotation exercise to resolve annotators' discrepancy (if there is any) of understanding annotation guideline; (3) and final annotation on the entire dataset.

Annotation guideline. Since stock sentiment is notably distinct from ordinary sentiment, a professional financial analyst who is an expert of equity research helped to establish the annotation guidelines on detailed rules of POSITIVE, NEGATIVE, NEUTRAL based on the definition of stock sentiment described earlier. 5 other domain experts were recruited to annotate the entire dataset. To guarantee they are on the same page, the annotators discussed the labeling rules in the guideline with the financial analyst. Through this process, we found some of labeling rules are not straightforward because of the complexity of the languages to express expectations of financial returns on social media. Some labeling rules and non-trivial examples are shown in Appendix.

Pilot annotation. Due to the challenges to be consistent with the labeling rules as shown above, we decided to incorporate an extra step

[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search

[3] https://www.reddit.com/r/wallstreetbets/

[4] https://en.wikipedia.org/wiki/R/wallstreetbets

Table 2: TweetFinSent inter-annotator agreement before and after conflict resolution.

	before	after
Positive	80.4%	90.0%
Neutral	77.8%	90.2%
Negative	67.8%	77.5%
Overall	77.5%	88.5%

for pilot annotation, which is unusual in other annotation tasks (Conforti et al., 2020; Orbach et al., 2020). Our financial analyst expert who created the guideline annotated 50 random samples by himself as the gold label set. They were assigned to every annotator as a pilot annotation exercise. The annotation disagreement (about 20%) with gold labels were discussed among annotators to align with the guideline and avoid potential ambiguity in the final annotation process.

Final annotation. During the final annotation process, 5 domain experts went through the pilot annotation and became the final annotators. 4 of them were assigned to annotate the whole dataset, in which each tweet was independently labeled by at least 2 annotators. The 5th annotator was used to resolve the conflicts in other 4 as a mean of controlling the data quality. If labels of 3 annotators are different, then that data point will be discarded.

3.4 Data Quality Assessment

In order to assess inter-annotator agreement, we calculate the pairwise Cohen's Kappa (κ). The average κ obtained was 0.67, which is substantial (Cohen, 1960) and interpreted as the moderate level of agreement (McHugh, 2012). To guarantee the data quality, we introduce an additional step to resolve the conflicts in annotations. Instead of adding new annotators with potential noise, we utilize an existing annotator. In practice, our conflict resolution step requires two annotators who have conflicted labels to discuss the annotations with a third annotator in order to achieve the agreement. We calculate the inter-annotator agreement ratio overall and at the class level before and after the conflict resolution. The results are presented in Table 2. In this comparison, it can be observed that with this conflict resolution step, we can achieve higher inter-annotator agreement as well as higher data quality. In fact, our overall agreement (88.5%) is higher than some previous sentiment analysis datasets; e.g., the

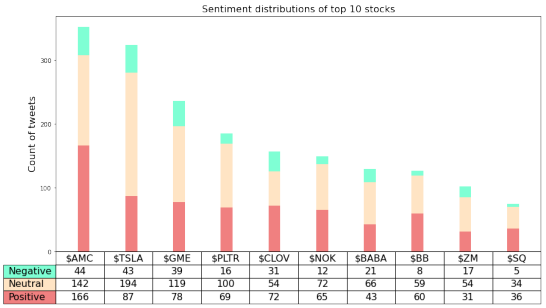


Figure 3: Sentiment distributions of top 10 stocks in TweetFinSent dataset.

inter-annotator agreement in Obama-McCain Debate dataset is 83.7% (Speriosu et al., 2011).

Moreover, in the cases where annotators disagree, we investigate the extent of the disagreement by measuring the distance between classes. If a Positive sentiment has value 1, Negative as -1 and Neutral as 0. Then we subtract the difference between the annotators and find that in 86.7% of the disagreements, it was with a difference of 1. In other words, it is more likely to differ on a Positive versus a Neutral sentiment than a Negative one, which happened to be the exact case for 67.9% of the disagreements. Another observation is that even after conflict resolution, the agreement in negative samples is still lower than that in positive and neutral samples. By investigating some cases, the possible reasons are: (1) the number of negative samples is smaller, so a small number of conflict can increase the disagreement, and (2) it is more difficult to determine if a tweet is negative due to various factors such as sarcasm, complicated emotions, and lack of context. For instance, given the tweet

> *Too many people drank the Kool aid. Telling you ....take your profits. Stack your cash. $tsla $zm $aapl*

the annotation conflict happens between *Positive* and *Negative*. This tweet contains complicated sentiments: being positive because the user gained positive return (with **profits**) while being negative because the user expected to gain negative return in the future (taking *cash* instead of buying stocks).

3.5 Data & Label Analysis

TweetFinSent dataset contains 2,113 tweets where the numbers of positive, neutral, and negative samples are 816, 1,030, and 267, re-

(a) Most frequent positive terms. (b) Most frequent negative terms. (c) Most frequent neutral terms.

Figure 4: Most frequent terms in TweetFinSent with different sentiment classes.

spectively. The distribution of different sentiment classes is quite imbalanced, i.e., there are much less negative samples. This imbalance may influence the performance of sentiment analysis methods and we will show more details in the experiments. We also show the sentiment distribution of 10 most discussed stocks in the dataset in Figure 3. One can observe that they are the meme stocks gaining most popularity among retail investors on social media during the period of data collection.

The most frequent terms in positive, negative, and neutral tweets in TweetFinSent dataset are shown in Figure 4. In positive samples, Twitter users talked more about 1) actions including to **buy** and **hold** stocks, 2) finance-specific expressions such as **to the moon**, **buy the dip** and **short squeeze** which was a hot topic during the period of data collection. All these discussions indicate positive (expected) return. In negative samples, more discussions are related to **sell** or **short** certain stocks and some stocks were **significantly overvalued**. They show the negative (expected) return. In neutral tweets, more tweets shared news or statistics about stock market, e.g., **pre-market stocks trend** and both **call** and **put** have been discussed.

4 Experimental Studies

4.1 Experimental Setup

We first preprocess the dataset by removing URLs and username (mentioning using @ notation)[5]. Hashtags are not processed because we observe that in financial domain some hashtags are indicators for special sentiment, e.g., *#YOLO* and *#WSB*[6]. Furthermore, it is common for a hashtag to refer to a particular stock ticker which represents the target for the sentiment analyzer. The data is split into training and test set with 1,113 and 1,000 tweets respectively. To make a fair comparison, we will keep the train-test split for all baselines.

4.2 Baselines

Since the task of stock sentiment analysis is different from traditional sentiment analysis and existing methods are not directly suitable for this task, we adopt several architectures that are commonly used in text classification and Twitter analysis for this problem. In details, three types of methods have been tested:

Lexicon-based methods In this experiment we adopt *Vader*[7] (Hutto and Gilbert, 2014) for our lexicon-based baseline because as the valence-based lexicon, *Vader* provides not just the binary polarity, but also the strength of the sentiment expressed in the given text. *Vader* is a rule-based sentiment analyzer that utilizes lexicons specifically trained on social media data. We were able to extract the lexicons list that contains a sentiment both English words and emoticons. Domain experts in Finance provided us a list of key words along with a sentiment class of 'Positive' or 'Negative'. We therefore modified the lexicon list we extracted based on the words provided, and gave a higher weighting to these relevant financial keywords. For example the sentiment scores in the lexicon file ranged from +3.4 to -3.9, and words like 'long' and 'short', were not present in the list as these words were classed as 'Neutral'. However in the financial context they would be 'Positive' and 'Negative' respectively. We enforce this by assigning a +5.0 score for pos-

[5]Note that there are more complicated preprocessing steps that could improve the performance especially in methods relying on feature engineering. We highlight our contributions on dataset construction and leave these preprocessing steps for future work.

[6]https://en.wikipedia.org/wiki/R/wallstreetbets

[7]https://github.com/cjhutto/vaderSentiment

itive keywords and -5.0 for negative ones. To be consistent with supervised methods, we use the lexicon-based methods only on the test set.

Pre-trained embedding. To conduct a comprehensive evaluation, both context-independent and context-dependent pre-trained word embeddings are compared. For each type of word embedding approach, we select different pre-trained embeddings that have been trained on general corpus and Twitter data. Specifically,

- For context-independent approaches, GloVe (Pennington et al., 2014) (including the original model *GloVe* pre-trained on general corpus like Wikipedia and the domain-specific model *GloVe-Twitter* pre-trained on Twitter) is selected.
- For context-dependent models, we use *DistilBERT* (Sanh et al., 2019), *FinBERT* (Araci, 2019), and RoBERTa (Liu et al., 2019) (including the original *RoBERTa* model pre-trained on general corpus and specific *RoBERTa-Twitter* model pre-trained on Twitter and fine-tuned for sentiment analysis task (Loureiro et al., 2022)).

After getting the embeddings, SVM and Gradient Boosted Decision Trees are employed to classify the sentiment using pre-trained embeddings as features.

Fine-tuned embedding models. Intuitively, due to the different patterns in our stock sentiment analysis task, general sentiment lexicons and pre-trained models may not perform well. Therefore, we fine-tune these pre-trained embedding models to verify the performance. Considering the advances of pre-trained language models, we only fine-tune these transformer models, i.e., DistilBERT, FinBERT, and RoBERTa. To make a fair comparison, we use the same train-test split, i.e., we use the training set to fine-tune the model and report the results on the test data.

4.3 Evaluation Metrics

The stock sentiment analysis is a typical multi-class classification task, so commonly used classification evaluation metrics can be easily adapted. Thus, following previous studies, in the experiments we utilize *Accuracy* and *F1* as the evaluation metrics. In particular, for *F1* scores, we report both macro average and weighted average versions.

It's worth noting that our constructed dataset contains more positive and neutral tweets than negative ones. To better understand the performance of different methods, we also calculate the F1 score for each class.

4.4 Benchmark Results

Benchmark results on these baselines are shown in Table 3. It can be observed that fine-tuned RoBERTa-Twitter achieved the best performance w.r.t all metrics. It makes sense because this model has been pre-trained on Twitter and fine-tuned for sentiment analysis task. By continuing to fine-tune on task-specific data, i.e., stock sentiment tweets in our experiments, the performance can be further improved.

Another observation is that in machine learning models, more advanced models generally achieve better performance which is consistent with other tasks. For example, context-dependent models are superior to context-independent models. One interesting and counter-intuitive result is that FinBERT performed worse than DistilBERT. This observation is consistent with previous study (Peng et al., 2021). A possible reason is that although FinBERT is trained for the financial domain, content from Twitter has different patterns from regular documents such as financial news texts and company press releases that FinBERT has been pre-trained on (Malo et al., 2014). However, fine-tuning cannot always guarantee better performance. After fine-tuning, although overall performance of DistilBERT and FinBERT has been improved, both F-1 scores for Negative tweets decreased.

It is also worth mentioning is that performance degradation can be observed for all models on negative tweets compared to positive and neutral ones. The major reason is that in the dataset, the size of negative samples is much smaller than that of positive and neutral ones. Such imbalance may make the models learn less representative information from the negative samples. Another reason is that there are different ways to express negative sentiment in financial domains including 1) using finance-specific terms, e.g., *put* and *short*, 2) using negation, and 3) using sarcasm or irony.

It is surprising that lexicon-based methods performed quite well compared to advanced deep learning models. In particular, finance

Table 3: Benchmark results of stock sentiment analysis using different baselines.

Methods	Overall performance			Per-class F-1		
	accuracy	macro avg F1	weighted avg F1	Positive	Neutral	Negative
Vader lexicon	0.4760	0.3592	0.3972	0.1840	0.6154	0.2781
Vader+Finance lexicon	0.5810	0.5269	0.5727	0.5342	0.6503	0.3962
GloVe+SVM	0.5340	0.4312	0.5157	0.4821	0.6275	0.1839
GloVe+GDBT	0.5420	0.4551	0.5335	0.4993	0.6397	0.2262
GloVe-Twitter+SVM	0.5140	0.3828	0.4872	0.5681	0.5215	0.0588
GloVe-Twitter+GDBT	0.5600	0.4823	0.5488	0.5248	0.6348	0.2872
DistilBERT+SVM	0.6020	0.5607	0.6017	0.5857	0.6557	0.4408
DistilBERT+GBDT	0.5920	0.5340	0.5871	0.5548	0.6667	0.3805
FinBERT+SVM	0.5750	0.5098	0.5694	0.5479	0.6465	0.3348
FinBERT+GBDT	0.5820	0.5262	0.5782	0.5537	0.6500	0.3750
RoBERTa-Twitter+SVM	0.5980	0.5594	0.5991	0.5982	0.6391	0.4409
RoBERTa-Twitter+GBDT	0.6320	0.5868	0.6306	0.6349	0.6701	0.4554
Fine-tuned DistilBERT	0.6180	0.5271	0.6095	0.6345	0.6838	0.2629
Fine-tuned FinBERT	0.6190	0.4923	0.5967	0.6390	0.6830	0.1548
Fine-tuned RoBERTa-Twitter	**0.7230**	**0.6785**	**0.7196**	**0.7436**	**0.7482**	**0.5439**

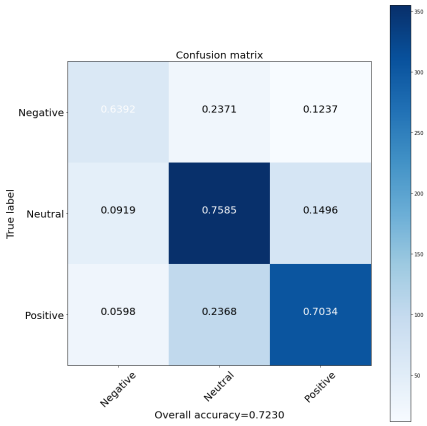


Figure 5: Confusion matrix of model output.

lexicons even outperformed GloVe including original one and GloVe pre-trained on Twitter data. Besides, Vader+finance lexicon performed better than general Vader lexicon. This comparison not only indicates the special characteristics of our constructed dataset and challenges of the stock sentiment analysis problem but also demonstrates the importance of prior knowledge in domain-specific tasks.

4.5 Discussions

To better understand the task of stock sentiment and TweetFinSent dataset, we select Fine-tuned RoBERTa-Twitter, the baseline achieving best performance, to further analyse. The confusion matrix of the prediction is shown in Figure 5. We can see that it performed poor on negative samples and achieved similar results on positive and neutral samples. Although Fine-tuned RoBERTa-Twitter outperformed other baselines with 0.72 accuracy, compared to existing Twitter sentiment analysis studies, the performance is acceptable but far from good. For example, different datasets and methods have been evaluated in (Saif et al., 2013) where the accuracy can reach to 0.8 even to 0.9 in some datasets. Therefore, on the one hand, this shows that TweetFinSent constitutes a challenging problem. On the other hand, there is ample room for improvement on the stock sentiment analysis task. Some research directions may be of interest for future work. From the data perspective, how to handle the data imbalance and improve the performance on negative data may improve the effectiveness of proposed models. From the methodological perspective, since finance lexicon showed its effectiveness, integrating prior knowledge of finance and stock into advanced machine learning models may boost the performance. Release of the TweetFinSent dataset enables researchers to further explore these directions.

5 Conclusions

We presented TweetFinSent, a new dataset for stock sentiment analysis and it contains 2,113 expert-annotated tweets covering different stocks. Different from existing sentiment analysis dataset, TweetFinSent defines sentiment based on whether a user gained or expected to gain positive or negative return from a stock rather than having feelings and emotions of good or bad. Our experiments with several sentiment analysis models indicated that there is a huge gap between machine learning models and human annotations. Thus, the TweetFinSent dataset constitutes a challenging problem and there is ample room for improvement on the stock sentiment analysis task.

References

Issam Aattouchi, Ait Mounir, Saida el Mendili, and Fatna Elmendili. 2022. Financial sentiment analysis of tweets based on deep learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, 25:1759–1770.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Deepali Arora, Kin Fun Li, and Stephen W Neville. 2015. Consumers' sentiment analysis of popular phone brands and operating system preference using twitter data: A feasibility study. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, pages 680–686. IEEE.

J. Bollen and H. Mao. 2011. Twitter mood as a stock market predictor. *Computer*, 44(10):91–94.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Aysun Bozanta, Sabrina Angco, Mucahit Cevik, and Ayse Basar. 2021. Sentiment analysis of stocktwits using transformer models. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1253–1258. IEEE.

Widodo Budiharto and Meiliana Meiliana. 2018. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data*, 5(1):1–10.

Usman W Chohan. 2021. Counter-hegemonic finance: The gamestop short squeeze. *Available at SSRN 3775127*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. Association for Computational Linguistics (ACL).

Sornette Didier, Sandro Lera, Jianhong Lin, and Ke Wu. 2022. Non-normal interactions create socio-economic bubbles. *arXiv preprint arXiv:2205.08661*.

Akash Dutt Dubey. 2020. Twitter sentiment analysis during covid-19 outbreak. *Available at SSRN 3572023*.

Shihab Elbagir and Jing Yang. 2019. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, page 16.

Umman Tugba Gursoy, Diren Bulut, and Cemil Yigit. 2017. Social media mining and sentiment analysis for brand management. *Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology*, 3(1):497–551.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Joshi Kalyani, Prof Bharathi, Prof Jyothi, et al. 2016. Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*.

Olga Kolchyna, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste. 2015. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

Paraskevas Koukaras, Christina Nousi, and Christos Tjortjis. 2022. Stock market prediction using microblogging sentiment analysis and machine learning. In *Telecom*, volume 3, pages 358–378. MDPI.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. 2020. Twitter sentiment analysis on worldwide covid-19 outbreaks.

Kurdistan Journal of Applied Research, pages 54–65.

Leeja Mathew and VR Bindu. 2020. A review of natural language processing techniques for sentiment analysis using pre-trained models. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 340–345. IEEE.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Anshul Mittal and Arpit Goel. 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15:2352.

László Nemes and Attila Kiss. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3):375–394.

Christina Nousi and Christos Tjortjis. 2021. A methodology for stock movement prediction using sentiment analysis on twitter and stocktwits data. In *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–7.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. *arXiv preprint arXiv:2012.14541*.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mohammed Qasem, Ruppa Thulasiram, and Parimala Thulasiram. 2015. Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 834–840. IEEE.

Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. Election result prediction using twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)*, volume 1, pages 1–5. IEEE.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis. *Emotion and Sentiment in Social and Expressive Media*, page 9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Michał Skuza and Andrzej Romanowski. 2015. Sentiment analysis of twitter data within big data distributed environment for stock prediction. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1349–1354. IEEE.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2013. Predictive sentiment analysis of tweets: A stock market application. In *International workshop on human-computer interaction and knowledge discovery in complex, unstructured, big data*, pages 77–88. Springer.

Thársis Souza, Olga Kolchyna, Philip Treleaven, and Tomaso Aste. 2015. Twitter sentiment analysis applied to finance: A case study in the retail industry.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63.

Hao Wang, Doğan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120.

Ethan Xia, Han Yue, and Hongfu Liu. 2021. Tweet sentiment analysis of the 2020 us presidential election. In *Companion Proceedings of the Web Conference 2021*, pages 367–371.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

tefan Lyócsa, Eduard Baumöhl, and Tomá Výrost. 2022. Yolo trading: Riding with the herd during the gamestop episode. *Finance Research Letters*, 46:102359.

A Appendix

A.1 Annotation Rules and Examples

1. [RULE]: Stock sentiment of a target ticker should be assessed only based on its own context. If there are multiple tickers in the same tweet, contexts of other tickers should have no impact to the target ticker.

   [EXAMPLE]: *"$AMC rocketed today! $BB $NOK $TSLA $GME."* The sentiment to $AMC is clearly POSITIVE. If the target is $GME though, then the sentiment should be NEUTRAL.

2. [RULE]: The assessment of sentiment should follow the subjective expectation. When both current and future returns are discussed, the focus should be on the future return.

   [EXAMPLE]: *"$TSLA revenue failed expectation, indicating a red day. However I will still buy at the dip"* should be POSITIVE. Because although the fact of $TSLA has negative return currently, the user still expects positive return in future and thus wants to keep buying.

3. [RULE]: Besides the normal buy or sell trades, other trade types like call vs put or long vs short can also reflect the expectation of positive or negative return.

   EXAMPLE: *"short $clov at this point"* is NEGATIVE. *"$ABIO ought $5 call options June 2021... easy buy, trading at book value."* is POSITIVE.

4. [RULE]: Besides the normal textual content, some slangs and hashtags indicating buy or sell, up or down are salient signals of stock sentiment and should contribute to the final sentiment assessment of the whole tweet.

   [EXAMPLE]: *Apes*, *to the moon*, *diamond hand* (risk tolerant, hold positions for long time), *#squeeze*, *#toMoon* are POSITIVE signals. Meanwhile *paper hand* (sell too early) is an example of NEGATIVE signals.

5. [RULE]: Some emojis in social media indicating "up"/"down" trend or expectation are salient signals of stock sentiment.

   [EXAMPLE]: 🚀 🔥 📈 are POSITIVE signals and 📉 is a NEGATIVE signal.

6. [RULE]: The received or expected return should be directional, i.e. either up or down. Ambiguous direction should be considered as NEUTRAL.

   [EXAMPLE]: *"$AMC cannot stop!"* or *"Looks like $tsla having its typical Tuesday."* are NEUTRAL since the content in the tweet is not enough to tell the direction.

A.2 Implementation Details

We use spaCy[8] to extract pre-trained *GloVe* embedding and obtain *GloVe-Twitter* embedding from the original paper[9] (Pennington et al., 2014). For classifiers, we use the implementations of linear SVM[10] and Gradient Boosting classifier[11] in scikit-learn. We use PyTorch and Hugging Face to obtain and fine-tune pretrained transformers including *DistilBERT*[12], *FinBERT*[13] and *RoBERTa*[14]. The settings of major hyper-parameters for transformers are: batch size is 16, max training epochs is 5, and max sequence length is 256. We use Adam as the optimizer with learning rate 2e-5 and the dropout rate is 0.1. The other hyperparameters are set by default. e.g., hidden size is 768 and number of attention heads is 12.

[8] https://spacy.io/usage/embeddings-transformers

[9] https://nlp.stanford.edu/projects/glove/

[10] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[11] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[12] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

[13] https://huggingface.co/ProsusAI/finbert

[14] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest