Social Media Sentiment Analysis of Financial Communities

Ronnie Nop
nopronnie@gmail.com

Abstract

The application of sentiment analysis on the users of social media platforms can be a powerful tool to gauge the overall opinions of an online community towards a particular topic or event. Within the finance domain, this becomes extremely useful as we can compute the overall sentiment of an online community towards a particular stock, different sectors, and major indices. In this paper, we look to explore various methods and techniques to compute the sentiment of user-based comments within online communities focused on trading the stock market. We fine-tune a BERT transformer using a labeled dataset of 4,300 stock market-related tweets and self-annotated reddit comments which was found to outperform traditional machine learning techniques, popular rule-based models, and other pre-trained transformers.

1 Introduction

Social media can be thought of as a mirror placed in front of humanity that reflects on both the light and dark aspects of human nature and how we as humans learn, provide insights, and understand the world around us. The opinions and emotions we express on social media platforms provide us an invaluable corpus of text in which artificial intelligent systems can model and extract general sentiment among a community. This makes the task of sentiment analysis extremely useful as it allows us to automate the processing of large amounts of text and extract the general sentiment of a community toward domain-specific topics.

In the finance domain, sentiment analysis is an important tool for obtaining alternative data via social media platforms and extracting investor sentiment on individual stocks, sectors, and major indices. Sentiment can then be generated into alpha signals that are used to build profitable trading portfolios. However, one the challenges of performing sentiment tasks particularly for social media tasks is 1) the lack of a large corpora of labeled data, 2) inference and serving time, and 3) yielding satisfactory results that researchers can be confident in.

The recent advancements of deep learning and NLP tools for processing text and performing various language tasks attempts to offer a solution to the current problems facing researchers.

The following paper looks to build an efficient pipeline that scrapes, pre-processes, and extracts sentiment from social media comments from Reddit and Twitter. Social media sentiment will be computed using traditional machine learning models, a rule-based lexicon model, and lastly, pre-trained and fine-tuned transformer models.

The central hypothesis is that the performance of pre-trained transformers fine-tuned on our domain-specific data will yield better performance than traditional machine learning models, overcoming the need for large, labeled datasets, as well as reducing training time. We also expect our rule-based model to lack performance but yield competitive inference speeds.

While constructing pipelines for scraping, pre-processing, and extracting sentiment, we will look at efficiency in terms of both predictive power and inference speed as a major criterion for yielding satisfactory results.

2 Related Work

Currently, there are several papers that have proposed research methods for various language tasks in the financial domain. Researchers from each paper introduce key and important ideas to approach language tasks that this paper looks to utilize and compare performance to our domain-specific task, social media sentiment for stocks.

The introduction of attention mechanisms (Vaswani et al., 2017) have given rise to powerful transformer models such as BERT (Devlin, 2019) and its variants FinBERT (Araci et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh, et al., 2019). These transformers are trained on extremely large corpuses of text and have achieved state-of-the-art performance on various language tasks. The general architecture consists of encoder-decoder blocks containing special attention layers that allow the model to build contextual representations for language. With the help of Hugging Face, these transformers can be accessed through open-sourced techniques to be leveraged by researchers to help them solve their domain-specific tasks.

BERT (Bidirectional Encoder Representation from Transformers) builds contextual representations of words bidirectionally using a combination of a masked language modeling (MLM) objective and a Next Sentence Sequencing (NSP) task. The MLM objective randomly replaces a small percentage of the input tokens with a MASK token in which the training objective looks to predict the actual token that was masked. The NSP task looks to predict the next sentence in a sequence allowing the model to learn which sentences should occur sequentially.

RoBERTa looks to robustly optimize BERT by making changes to some of the training objectives in the original BERT architecture. The major difference is in the MLM objective where RoBERTa dynamically masks tokens during each training epoch contrary to BERT which statically assign masks that are repeated across all training epochs. Furthermore, RoBERTa completely removes the NSP objective. RoBERTa is also trained on substantially more text data than BERT.

FinBERT is a pre-trained NLP transformer built on top of BERT that is fine-tuned on text related to the financial domain. Their researchers found that fine-tuning the final layers of the original BERT architecture on text related to their domain-specific task yield a +15% increase in accuracy compared to current state-of-the-art methods at the time. One of the key insights arising from this paper is that even with significantly less training data, fine-tuning a transformer with domain-specific data can yield similar or even better performance than other language models.

DistilBERT looks to tackle one of the major disadvantages of BERT and its variants, which is model size. There are many key advantages to reducing model size which includes reduced computational costs, training time, and serving time. Researchers from Hugging Face proposed DistilBERT as a faster and cheaper alternative to BERT while achieving similar performance. DistilBERT reduces the original BERT architecture by 40% from 110 million parameters to 66 million parameters. The team found that DistilBERT retained 97% of BERT's performance while at the same time achieving faster inference speed by 60%.

Non-ML based approaches have also been developed for sentiment analysis. Researchers (Hutto, et. al., 2014) with the help of sociologists, psychologists, and computer scientists introduced VADER (Valence Aware Dictionary for sEntiment Reasoning), a rule-based model for social media sentiment analysis. VADER is a very complex sentiment lexicon that is able to handle sentiments related to punctuation, word-cases, contractions, slangs, and emojis. Researchers found that it has outperformed traditional machine learning models such as Naïve Bayes and Support Vector Machines while also being more computationally efficient in training and avoiding out-of-memory issues.

We look to utilize the insights from each paper when building the entire sentiment pipeline from scraping, pre-processing, and computing social media sentiment.

3 Data

We will use two datasets obtained from Reddit and Twitter that will be used to perform sentiment analysis and evaluate models for comparison. The combined dataset contains 4,300 comments that are labeled positive, neutral, and negative. The dataset contains slightly more positive texts, and a similar number of neutral and negative texts.

Data	Positive	Neutral	Negative
Reddit	1126	961	913
Twitter	528	424	348
Total	**1654**	**1337**	**1309**

Table 1: Sentiment labels for the datasets

3.1 r/WallStreetBets Daily Discussion

We utilize PRAW and PushShift Reddit APIs to scrape comments from the Daily Discussion Threads from the subreddit r/WallStreetBets for two years from January 1st, 2020 to December 31st, 2021. Initially, 440,000 comments were scraped and 3,000 were randomly sampled to be self-annotated with the labels positive, neutral, and negative. Sentiment was labeled based relative to the stock price. A positive sentiment was labeled if the user expressed the stock going up. A neutral sentiment was labeled if the user expressed the stock was trading sideways. Lastly, a negative sentiment was labeled if the user expressed the stock going down.

Sentiment	Reddit User-Comment
Positive	PLTR IS GOING UP BIG TIME!
Neutral	anyone heard of pltr?
Negative	PLTR sorry boys its crashing to 0

Table 2: Sentiment labels for r/WallStreetBets

3.2 Stock Market Tweets

The second dataset was obtained through a Kaggle search leading to an original dataset provided by Bruno Taborda, et al. 2021. The team built a Twitter API which scraped 943,672 tweets containing hashtags corresponding to the top 25 companies in the S&P500 from April 9th to July 16th, 2020. 1,300 tweets were manually annotated by the team with the sentiment labels positive, neutral, and negative.

4 Models

Models will range from traditional machine learning models, rule-based models, pre-trained transformers, and fine-tuned transformers which will be compared using two metrics: 1) macro average F1-score and 2) inference time in seconds.

4.1 Traditional Machine Learning Models

To establish baseline models for comparison, we will train logistic regression and naïve bayes classifiers using simple pre-processing schemes and optimization frameworks explained in the following Experiments section. We expect these models to act as baselines to be compared to our more complex transformers outlined below.

4.2 Rule-Based Models

Non-machine learning strategies will be explored using a lexicon-based model called VADER (Valence Aware Dictionary for Sentiment Reasoning). VADER is tuned specifically on social media data providing positive, neutral, and negative polarity scores. One major advantage is its ability to correctly classify sentiments of emoji and slang-terms within a community. Two variations of the VADER will be constructed. We expect the VADER models to perform similarly to our baseline models outline above as well as achieving much faster inference time for serving.

4.3 Transformers

We will use pre-trained BERT and RoBERTa transformers that were trained on large corpuses of text from Wikipedia and Twitter data to classify sentiment. We expect these models to perform better than our baselines and rule-based models based on F1 metrics with the tradeoff of having longer inference times.

Finally, we will fine-tune BERT and DistilBERT transformers, which we will call WSBERTs and WSBERTs-DistilBERT respectively, on 4,300 stock market-related tweets and self-annotated comments from the popular subreddit group, r/WallStreetBets. We expect the fine-tuned transformers to achieve the highest F1-scores as well as the longest inference times.

Transformers will be trained using the Hugging Face open-sourced libraries.

5 Experiments

The combined dataset of 4,300 comments consisting of stock market-related tweets and self-annotated Reddit comments were split into training, evaluation, and testing sets. The final performance metrics for all the models were computed using the same, unseen testing set. The primary performance metric used for evaluation was macro average F1-score and the secondary metric, inference speed, was computed in seconds using 1-month of subreddit data consisting of 18,500 user comments.

The logistic regression and naïve bayes classifiers were trained using the training set and their hyperparameters were tuned using grid search cross-validation. The data was pre-processed using 1) a count vectorizer, 2) stop word removal tool, and 3) a TF-IDF transformer. The final evaluation for their macro avg F1-score was computed using the testing set.

Two variations of the VADER model were constructed and compared against. The first variation is the original un-tuned VADER sentiment analyzer. The second variation is an updated VADER sentiment analyzer whose lexicon and polarity scores were updated and fine-tuned using new words, slang-terms, and emojis used widely within the subreddit community. Examples of updated and added words are shown in Table 3.

Word	Polarity	Word	Polarity
'moon'	+4.0	'uppies'	+3.0
'pump'	+3.0	'downies'	-3.0
'dump'	-3.0	🚀	+4.0
'bear'	-4.0	🌈🐻	-4.0

Table 3: Updated VADER Lexicon

Pre-trained BERT and RoBERTa transformers were obtained using Hugging Face's open-sourced transformers library. Comments were fed directly into the model for evaluation of the testing set and for calculating inference speed.

Transfer learning was conducted on BERT and DistilBERT transformers by fine-tuning the final layers of the model using the training data and evaluation data on only 3 training epochs. The final performance metrics were computed using the testing set.

Tables 4 to 6 show the F1-scores for each of the models and their respective performances on each of the sentiment labels: positive, negative, and neutral on the testing set. Table 7 shows the macro average F1-scores on the testing set.

Model	Positive – F1
Logistic Regression	0.62
Naïve Bayes	0.63
Tuned VADER	0.61
Untuned VADER	0.49
BERTweet	0.37
RoBERTa-Twitter-Base	0.54
WSBERTs	**0.69**
WSBERTs-DistilBERT	**0.69**

Table 4: F1-Score for Positive-Labeled Comments

Model	Negative - F1
Logistic Regression	0.51
Naïve Bayes	0.53
Tuned VADER	0.60
Untuned VADER	0.53
BERTweet	0.63
RoBERTa-Twitter-Base	0.62
WSBERTs	**0.64**
WSBERTs-DistilBERT	**0.64**

Table 5: F1-Score for Negative-Labeled Comments

Model	Neutral - F1
Logistic Regression	0.49
Naïve Bayes	0.42
Tuned VADER	0.47
Untuned VADER	0.45
BERTweet	0.44
RoBERTa-Twitter-Base	0.49
WSBERTs	**0.61**
WSBERTs-DistilBERT	**0.57**

Table 6: F1-Score for Neutral-Labeled Comments

Model	Macro Avg F1
Logistic Regression	0.54
Naïve Bayes	0.53
Tuned VADER	0.56
Untuned VADER	0.49
BERTweet	0.49
RoBERTa-Twitter-Base	0.55
WSBERTs	**0.64**
WSBERTs-DistilBERT	**0.63**

Table 7: Macro F1-Score for All Comments

Inference time was measured in seconds for each model using a holdout set consisting of 1-month of subreddit data comprising of 18,500 user comments. Trials were re-run multiple times to generate confidence intervals around an average.

Inference times do not reflect the time it took to scrape the data or ingestion into the database. Data pre-processing steps for each respective model were considered for each model's inference time. Any additional steps needed for a model to perform inference directly from the comments data was also considered for their inference time.

Model	Inference Time (s)
Logistic Regression	0.63 ± 0.10 secs
Naïve Bayes	0.54 ± 0.12 secs
Tuned VADER	2.38 ± 0.44 secs
Untuned VADER	2.05 ± 0.48 secs
BERTweet	710.66 ± 20.25 secs
RoBERTa-Twitter-Base	630.24 ± 38.63 secs
WSBERTs	357.20 ± 24.60 secs
WSBERTs-DistilBERT	345.85 ± 23.14 secs

Table 8: Inference Time for 1-Month of Comments

6 Analysis

The outcomes of the experiments yielded both expected and unexpected results for both the macro average F1-scores and inference time metrics measured for all the models.

Looking at the F1-scores across all class labels, we see in Table 6 that Neutral-labeled comments were the most difficult to classify. Every model except for our fine-tuned transformers had an F1-score under 0.50 suggesting that it would most likely underperform even a random model. Our fine-tuned models WSBERTs and WSBERTs-DistilBERT clearly scored much higher on these neutral labels suggesting the effectiveness of fine-tuning.

The best performing model for macro average F1-score was WSBERTs with a score of 0.64, which was the BERT transformer fine-tuned on our dataset comprising of 4,300 labeled tweets and reddit comments. It completely outperformed the traditional models used as baseline and both rule-based VADER models in all of the label categories as shown in Tables 4 to 6.

The smaller, fine-tuned transformer WSBERTs-DistilBERT performed slightly worse than our primary WSBERTs transformer but still managed to outperform all the other models as well. This is to be expected because the goal of DistilBERT is to offer a smaller-sized language model with the tradeoff of being less performant which we saw in the results of the experiments.

Our VADER models showed surprising results from our initial hypothesis. The tuned-VADER model was top 3 in macro average F1-scores which surprisingly outperformed the pre-trained RoBERTa-Twitter-Base model. The untuned-VADER model was also the worse performing model which is surprising because its lexicon was tuned specifically to social media data.

The pre-trained BERTweet and RoBERTa-Twitter-Base models showed mixed results as well. The RoBERTa-Twitter-Base that was pre-trained on labeled tweets underperformed against our tuned, rule-based VADER model, however, not by a significant amount. Expectingly, it outperformed the logistic regression and naïve bayes classifiers. However, the pre-trained BERTweet model that was also trained on labeled Twitter data was one of the worse performing models and did not outperform any of the baselines.

The traditional machine learning models logistic regression and naïve bayes classifiers used as baselines performed as expected, some might suggest even better than expected.

In terms of model inference speed, the results aligned accordingly with our initial hypotheses. We expected our rule-based VADER models to exceptionally fast which our results clearly show with inference speeds of 2.38 ± 0.44 and 2.05 ± 0.48 seconds, respectively.

Surprisingly, our traditional machine learning models performed the fastest with the logistic regression classifier having an inference speed of 0.63 ± 0.10 seconds and the naïve bayes classifier with 0.54 ± 0.12 seconds. Even with extra pre-processing steps, these models inferred faster than our rule-based VADER models.

The pre-trained transformer models inferred the slowest which was expected from our initial hypothesis. Surprisingly, inference speeds were about two times slower than our fine-tuned transformers which could undergo further investigation to identify the cause for these discrepancies.

The fine-tuned transformers WSBERTs and WSBERTs-DistilBERT had inference speeds of 357.20 ± 24.60 and 345.85 ± 23.14 seconds, respectively, performing worse than our traditional models and rule-based models. These inference times were expected in our initial hypothesis.

7 Conclusion

Our strategy of fine-tuning a BERT transformer to data specific to our target domain proved to be a successful tool for gauging the sentiment of an online, stock trading community. Both our fine-tuned WSBERTs and WSBERTs-DistilBERT models outperformed traditional machine learning techniques, popular rule-based techniques for sentiment analysis, and other pre-trained transformers.

Furthermore, our results showed the usefulness of fine-tuning models to your domain-specific task. Our rule-based VADER model, whose lexical dictionary was fine-tuned using their specific community terms, slangs, and emojis, was able to classify sentiment much better than even traditional machine learning models and some pre-trained transformers. Therefore, it is extremely important to understand the language of the target population when attempting to construct models for such language tasks.

Another important metric when selecting a final model to deploy into production is inference speed. Unfortunately, there seems to be a clear tradeoff between model performance vs inference speed. Our fine-tuned transformer model was ranked first in our primary macro average F1-score but ranked second to last in inference speed. On the other hand, the fine-tuned rule-based VADER model ranked third in macro average F1-score but had much faster inference speed (357.20 ± 24.60 seconds vs 2.38 ± 0.44 seconds, respectively). If applying sentiment analysis to a high-frequency trading strategy, it might be worthwhile to use the latter. If applying to a longer-term trading strategy, a fine-tuned transformer may be the optimal choice.

References

Araci, Dogu. 2019. "FinBERT: Financial sentiment analysis with pre-trained language models." arXiv https://arxiv.org/pdf/1908.10063.pdf

Devlin, Jacob, et. al. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”arXiv https://arxiv.org/pdf/1810.04805.pdf

Hutto, C.J. and Gilbert, Eric. 2014. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.” eegilbert https://eegilbert.org/papers/icwsm14.vader.hutto.pdf

Kabbani, Taylan and Enes Usta, Fatih. 2022. “Predicting The Stock Trend Using News Sentiment Analysis and Technical Indicators in Spark.” arXiv https://arxiv.org/pdf/2201.12283.pdf

Liu, Yinhan, et. al. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv https://arxiv.org/pdf/1907.11692.pdf

Sanh, Victor, Debut, Lysandre, Chaumond, Julien, and Wolf, Thomas. 2019. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.”arXiv https://arxiv.org/pdf/1910.01108.pdf

Taborda, Bruno, et. al. 2021. "Stock Market Tweets Data", DataPort, https://dx.doi.org/10.21227/g8vy-5w61

Vaswani, Ashish, et. al. 2017 “Attention Is All You Need.” arXiv Araci, Dogu. 2017. arXiv https://arxiv.org/pdf/1706.03762.pdf