Predicting S&P500 based on Twitter sentiments

Team: Yifeng Shen, Jonathan Shaw,Zihu Xu **Project Mentor TA:** Sihao Chen

1) Abstract

For this project, we will be using tweets as our input to predict S&P 500 movements.It's important since market sentiment might have a strong correlation with stock market which can help us make a better prediction. Our primary target contribution is to collect a new training dataset for this problem. We update the existing data with the most recent tweets, all three of us manually label 1000 tweets generated in recent months and balance the original dataset (POS,NEU,NEG). Perhaps most interestingly, we find that we can train a much better model with our new dataset since in the past few months the financial world had seen the biggest volatility since COVID due to rising rates and Russia's war on Ukraine. Tweets may reflect the correlation with the market way clearer, and these are all new data that perfectly fit our project.

2) Introduction

For this project, we will be using tweets as our input to predict S&P 500 movements. The input of the system will be the tweets. From these tweets, we will be labeling them with sentimental values, and looking at the overall sentiment correlation with same day market movements. Depending on our further research, maybe even discover any causations of future movements.

The training and testing data are simply split from the same dataset we have collected. Specifically, we collected tweets with twitter API that are related with label "SP500" or other relative labels, and we manually labeled 1000 of them to be one of the three states(pos,neg,neu) for further analysis. We also pull the opening and closing from yfinance with tick "spy" to find the correlation(or causality if exists) between the overall sentiment and the movement.

We will evaluate this system in two parts, which are the sentiment analysis accuracy and the movement prediction based on the sentiment. We will try to evaluate both parts with the F1 score. As we randomly choose the samples from the dataset and label them, it is not guaranteed that the distribution is fair. Also the market tends to be one sided during some period of time, it's fairer to use the F1 score instead of accuracy.

The motivation of this project is to help people invest smartly. To my knowledge, investing is one of the most feasible ways of FIRE, however, some people without proper knowledge may suffer from their operation, and even risk the fund they are living on. Some are also easily influenced by people around them, and operated irrationally, while the total environment is completely opposite. We want people to avoid these kinds of mistakes and be wise.

3) Background

A. Paper / blog post/ Kaggle submission etc.: Title. URL. What it does that is relevant to this project (2-3 sentences). Code link. Provide any links in text, like www.google.com, not like Google.

Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements".
Github implementation at:
github.com/you915/Sentiment-Analysis-of-Twitter-Data-for-predicting-Apple-stock-price.
This paper has the main framework as what we are expected to perform at the end. It introduced some ways on sentiment analysis.

B. More if needed. There should be at most 2 such closely related prior works in most cases.

1. Mayukh Sharma, Ilanthendral Kandasamy, B.Vasantha, "Comparison of neutrosophic approach to various deep learning models for sentiment analysis". This paper compared various models on sentiment analysis, and will be helpful when we conduct the first phase of our project.
2. Zhengjie Gao, Ao Feng, Xinyu Song, Xi Wu, "Target-Dependent Sentiment Classification With BERT". This paper introduces the BERT algorithm in context-free situations, which perfectly fits our need, since tweets are limited within 140 characters.

4) Summary of Our Contributions

Our main contribution is on updating the existing data with most recent tweets. In the past few months the financial world had seen the biggest volatility since COVID due to rising rates and Russia's war on Ukraine. These days, the public is pretty much in a high mood due to the war and FED's actions, so the tweets may reflect the correlation with the market clearer, and these are all new data that perfectly fit our project.

Furthermore, we are not using the original BERT model anymore. The original BERT model was trained using Wikipedia text, which is a different environment from twitter. To accommodate for this change, we are using BERTweet, a model with word embedding trained from twitter rather than Wikipedia.

5) Detailed Description of Contributions

5.1 Methods

Our data is retrieved from the API day by day. Every several days, we run the code to gather all tweets that satisfy the query day by day. The collected data is stored with the text itself, the time it is posted, and the retweet and like numbers into a table with the date as the table name in the database. We then filter out all retweets since its attitude may vary depending on the original post, and it’s very hard to relate all the retweets to the original ones. Then we also randomly selected 1000 to be manually labeled for later training.

During the process, we realized that retweet texts were cut off, meaning that from the twitter API, we were not able to grab the full retweet text, therefore we removed all retweets. Compared to previous papers, we also grabbed the like and retweet numbers as they might indicate the user's influence on the market. (The more people listen to you the more influential you are)

1	text	Label	like	rt	date
2	Hey! Looking for a new Trading server? I'm personally inviting you to Polar Trades a new server to the scene of trading. I will provide you with my notes, my daily day trades/scalps, my swings and Live Trading just ab		1	1	2022-03-20
3	#analysis #Bitcoin #DXY #SP500 #EGLD 1/12 Hello everyone, #Bitcoin and the whole crypto market passed yesterday another important pressure test. The FED Reserve raised the rates with 25bps. As expected and as we discussed previously, the effects were minimal on risk-C		329	96	2022-03-17
4	weak mkt me strong candidate mil gaya to bulls ko aur kay chiye #priceaction #StocksToTrade #StockMarketindia #stocks #StocksToBuy #StockMarket #Nifty #StockMarket #stocks #stockmarkets https://t.co/h1p6bRqtoo		1	1	2022-02-20
5	Today's Short Volume was 7,745,512, which is 60.59% of today's total reported volume. Over the past 30 days, the average Short Volume has been 61.67%. #AMC #AMCapes #AMCsqueeze #AMCNOTLEAVING #Investing #AMCTheatres #AMCStock #MOASS2022 #Stocks #StockMarket https://t.co/ZHFllVOiLP		52	8	2022-03-10
6	#BMN Just an indication of what can happen to the SP when #Vanadium prices rise dramatically. #FAR $LGO $AVL $VR8 $UUUU #JAN #TNG $NMT $TMT #Metals #Investment #Mining #InvestInAfrica #CriticalMetal #Stocks #Shares #Steel #VRFB #LDES #EnergyStorage https://t.co/Y4zGTD2fAI		29	6	2022-03-06
7	$DXYN - Dixie Group's (DXYN) CEO Dan Frierson on Q4 2021 Results - Earnings Call Transcript. https://t.co/81bdqG8hlb #business #economy #stocks		19	8	2022-03-10

Collected data

Labeling

For our tweet labeling, we are using the following rules for label:

1. **Rule of thumb**
   a. In general, we are labeling the tweets based on an investor's reaction after reading the tweet. If the investor is optimistic about the market, it is positive; if the investor is pessimistic, negative. If no effect, neutral.
2. **Regarding Shorts**
   a. If a person is happy because he/she’s short position (by betting the market goes down) made money, we will be marking it as negative, since someone making money from betting the market going down means a negative day for overall market
   b. Similarly, if a person is unhappy because he/she’s short position (by betting the market goes down) lost money, we will be marking it as positive, since someone losing money from betting the market going down means a positive day for overall market
   c. If a tweet simply claims he/she made money, we will just label it as positive. Since most people investing do not have short positions.
3. **Neutral Definition**
   a. If a tweet is stating an event occurring, for example, XXX is posting its earnings report tonight. We will label it as neutral.
   b. If the tweet only states the price level of a stock without mentioning it going up or down, we will label it neutral. However, if the tweet states the stock up/down by a percentage, our label will be corresponding to the stock price increase/decrease.
   c. If the tweet is regarding advertising for trading lectures. Even if the tweet is a phrases with clear sentimental values (for example “the market is in turmoil

times" or "market is climbing to new highs everyday"), if it is followed by advertisements such as "follow us to earn money" we will label it neutral

4. **Other markets/asset type**
    a. Many tweets referred to other markets or asset types such as oil, gold, or even foreign markets. In such cases, we follow the intuition of a general investor. For example, oil price is generally negatively correlated with the stock market, thus a rise in oil price negatively impacts investor sentiment, and is labeled negative.

NLP BERTweet Model

Preprocessing

We did draw inspiration from the previous paper "Predicting the S&P 500 Using Tweet Sentiment"[1], but because of Twitter's environment, we did make some changes to include more information that might have been wiped away in previous papers. These cleaning procedures are applied to both labeled and unlabeled data for consistency. Below are the procedures and justifications.

1. **All letters are lowercase**
    a. This is a standard procedure for NLP text cleaning, as capitalization "obfuscates the words more than it conveys sentiment"[1].
2. **"n't" was separated from previous word, "cannot" was split to "can not"**
    a. This procedure isolates the negative portion from the positive, giving the model more clarity in training.
3. **Emoji was tokenized as text strings**
    a. Many emojis have a strong feeling attached, in the previous paper, emojis were simply removed, which might cause a lot of confusion for the model. A classic text such as "the market is 🔥🔥🔥" is clearly positive, but once the emoji is removed, the model might put too much emphasis on other irrelevant words, therefore it should be included. Another example is rocket emoji, which is attached to the Wall Street Bet's to the moon, also positive.
4. **Website links were tokenized**
    a. Through our labeling process, we realized a majority of NEU labels were advertisements for stock trading lessons or lectures. Many of them have a website link attached. To give our model more information to identify this, the model tokenized all website links to HTTPURL.
5. **Tweet mentions of the form @name were tokenized**
    a. Similar to the above two reasons, mentions of names might contain extra information to determine tweet sentiment, therefore the model tokenized it.

5.2 Experiments and Results

There are two separate steps in our project, which are the sentiment analysis part and the part which finds the relationship between the average sentiment and the market performance. Basically we have different ways to evaluate these two parts.

The way we used to evaluate sentiment analysis is to use simple accuracy based on the given dataset. In the dataset, there are three types of sentiment, i.e. POS, NEU, NEG, and the base accuracy of guessing should be ⅓, but it would be ridiculous to set the target that low, so based on the paper we have previously viewed, the accuracy of sentiment analysis is 76%, and our target is then to beat the previous accuracy. Since the dataset is manually labeled and is not very large in size, all evaluation is done with the random split of the dataset. Finally we get an overall accuracy of 81%. This accuracy beats the target, and is a reasonable and acceptable rate to be used in the next step.

{'eval_loss': 0.7815760374069214,
 'eval_accuracy': 0.8175895765472313,
 'eval_runtime': 5.4898,
 'eval_samples_per_second': 335.531,
 'eval_steps_per_second': 42.078}

The second part of the project is to first check if there exist correlations between the average sentiment and the market movement. After having predicted all other unlabeled tweets with the fine tuned model, we first looked at the correlation between market movements and twitter activities.

change	predicted_label	count	avg_likes	avg_retweets
-1.0	-1	4204	18.176499	4.292103
	0	8622	11.583391	3.446996
	1	7406	18.733324	4.880232
1.0	-1	3503	16.904082	4.142449
	0	7578	12.013328	3.580232
	1	7095	16.720930	3.635095

As we can see, on a bad day, tweets regarding the market are significantly more than when the market is up, so is like and retweets. This might indicate that people will go to twitter more often when they lose money in the market, and focus both on the optimistic tweets and the negative ones, viewing and sharing them more. Neutral tweets contain a lot of advertisements and news, therefore even though the number of neutral tweets are the highest, both retweets and likes are significantly less than the optimistic ones.

For the next step, we have to combine the sentiment of a day into an average sentiment. Like numbers and retweet numbers are preserved during data collection, so these two features are used as the weight of tweets to reach a more accurate result. The correlation between sentiment and market movement is 0.53, but it drops to -0.14 between the market and the sentiment of the previous day. After also compared to the correlation value with more day offsets, the correlations become very random. This indicates that there probably does not exist causality between history sentiment and the market of the day, hence we decided to use only the sentiment of the day.

Since the market is fully stochastic, the baseline would be to beat the 50% random choice. Our accuracy is 61% after passing the average sentiment to a linear regression. We also did a simulation buy and sell on each day based on the sentiment, i.e. buy if it predicts a positive and sell short if negative. The simulation is the best way to evaluate the performance of the prediction, and after simulation from Feb.11 to Apr.24, we have a final 14% profit, while the average of doing a random choice everyday for 100 loops is merely less than 1%.

To experiment our results on more market data, we used the data from our previous paper[1], which included data from 2020. However, because their dataset didn't contain retweets and like data, we could only fit a model with the labeled average sentiment. Our results were a correlation of 2.02, and accuracy of 65%.

6) Compute/Other Resources Used

We used an original dataset that we collected ourselves, and manually labeled, hence the size of the dataset is relatively small, so it is fine tuned with a personal computer with GPU.

7) Conclusions

Not only did we gather a lot of new tweets from the past few volatile months, we successfully improved on the previous model of sentiment prediction, increasing accuracy from 74% to 82%. This model was trained for tweets regarding the financial market, and I believe it could be very useful in analyzing optimism of the general public regarding the market.

At the end of the day, we were not able to find any causation between tweet sentiment and the market movement of following days. However, we are successful in identifying the differences in overall tweet activities for different market conditions, as well as finding a correlation between market movement and same day tweet sentiment. The sentiment predictions were correct in most cases of major market swings; one potential reason for the low 61% accuracy was many of our observed days did not move that much, and twitter sentiment fluctuated during those days.

(Exempted from page limit) Other Prior Work / References (apart from Sec 3) that are cited in the text:

1. Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.