Is Wisdom of the Crowd a Viable Methodology to Predict Financial Markets?

DSC 180B WI 2023 Capstone Project - Justin Cun, Dakshh Saraf

This research paper aims to answer whether the "wisdom of the crowd" methodology is a viable approach to predicting financial markets. The wisdom of the crowd approach is based on the idea that the aggregation of opinions from many individuals can lead to more accurate and unbiased decisions and, in the context of financial markets, more accurate predictions of future market movements. We analyze batches of tweets from Twitter and leverage machine learning models to conduct sentiment analysis. Through our analysis, we aim to evaluate the effectiveness of this approach. Our findings may have implications for investors, financial analysts, and policymakers, as well as for future research in this area. The results of our study suggest that the wisdom of the crowd approach can provide valuable insights into market trends, but also has limitations that need to be carefully considered. Ultimately, our research aims to contribute to the ongoing discussion surrounding the use of social media data and machine learning models for financial market predictions.

---

1. Introduction

Predicting financial markets is an important task that can have a significant impact on investment decisions and economic policies. In recent years, the "wisdom of the crowd" approach has gained attention as a methodology that uses collective intelligence to predict future market movements. This approach suggests that by aggregating the opinions of many people, more accurate and unbiased predictions can be made.

This research paper will delve into whether the wisdom of the crowd methodology is a viable approach for predicting financial markets. To do this, we will analyze batches of Twitter data and build predictive models to extract insights and conduct sentiment analysis. By comparing this approach to traditional methods, we can provide insight into the effectiveness of social media data and machine learning models for financial market predictions, along with shedding light on the benefits and limitations of the wisdom of the crowd.

2. Literature Review

Our study builds on Shawndra Hill's findings in "Expert Stock Picker: The Wisdom of (Expert in) Crowds" (2011). Hill explores the potential of crowdsourcing as a method for identifying expert stock pickers. The study compares the performance of financial analysts with that of non-expert participants in a stock-picking competition. The results indicate that, on average, the non-expert participants outperformed the financial

analysts in their stock recommendations, suggesting that the wisdom of the crowd can be an effective approach to identify expert stock pickers.

Overall, the research suggests that the wisdom of the crowd can be a viable approach to investing, particularly when there is diversity of opinion and independence of decision-making. The findings of Hill's paper suggest that diversity of opinion and independence of decision-making are important factors in the success of the wisdom of the crowd methodology. Therefore, it is important to carefully consider the context and characteristics of the crowd when applying this approach to investing.

3. Methodology

3.1 Financial Tweets dataset

To collect data for our analysis, we used a Kaggle dataset of financial tweets collected between 2015 and 2020 for the top 5 tickers in S&P 500. There are 3.6 million tweets in this dataset. We also get their respective timestamps and tweet ids.

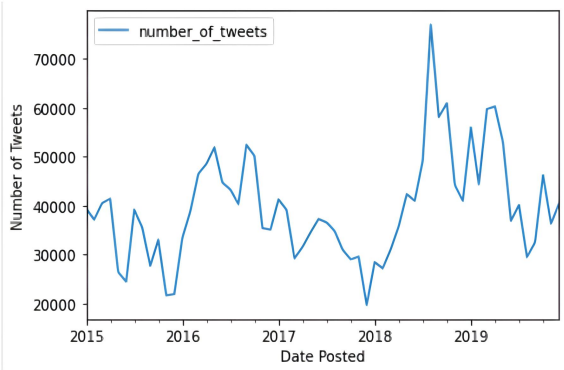


Figure: Line chart of number of tweets from top 5 tickers 2015-2019

We did some extensive data cleaning to make sure we only analyzed clean tweets about the stocks and not ads or spam. Financial Twitter is buzzing with Telegram channel ads about trade calls. We removed all URLs from the tweets and removed all duplicates. We calculated some in-house heuristics such as the number of tickers present in a tweet. If a tweet had more than 3 tickers present, we removed it from our analysis on the assumption of it being too ambiguous or either spam.

3.2 Stock price dataset

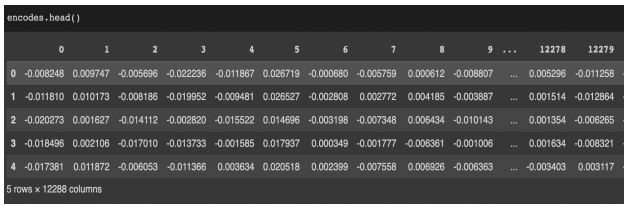
We also used Yahoo Financial historical API to get historical prices for the respective stock tickers. We especially needed to observe if the stock was positive, negative, or neutral the next day of the prediction. A “neutral” day can be defined as fairly arbitrary by looking at the price movements in the table. It’s rare for a day to be exactly flat, i.e, have a percent change of 0. Because of this, we made a threshold for neutral labels. If percentage change in stock prices were [-0.5,0.5], that day would be labeled as neutral.

Date	MSFT	TSLA	AAPL	AMZN	GOOG
2014-12-31	-0.599184	-0.304813	-2.162739	-0.385174	-0.912943
2015-01-02	0.214313	-1.597337	-1.849356	-1.298868	-0.793936
2015-01-05	-0.086256	-2.078772	-1.883831	-1.569983	-1.794517
2015-01-06	-1.573953	0.580785	-0.262811	-2.299502	-2.532040
2015-01-07	0.543715	-1.124918	0.513063	0.309241	-1.163704
...	...	...	...	...	...
2016-03-18	-2.603781	1.588829	-0.394958	-1.579494	-0.574225
2016-03-21	1.145541	1.266253	-0.018877	0.923649	0.758993
2016-03-22	0.858047	-1.252055	1.396676	2.819612	0.446119
2016-03-23	-0.258731	-4.213105	-0.328706	1.538329	-0.579234
2016-03-24	0.687219	5.547319	0.189624	2.793108	0.449451

Figure: % change in stock price everyday

3.3 Classification with GPT-3

We used GPT-3 field embeddings API to embed these tweets, capturing the context and meaning of the words in the domain of social media. Using OpenAI's embedding endpoint, we were able to transform each tweet into 1,536 high-precision floating point numbers. These vectors or embeddings will serve as features in our machine learning model.



encodes.head()

	0	1	2	3	4	5	6	7	8	9	...	12278	12279
0	-0.008248	0.009747	-0.005696	-0.022236	-0.011867	0.026719	-0.000680	-0.005759	0.000612	-0.008807	...	0.005296	-0.011258
1	-0.011810	0.010173	-0.008186	-0.019952	-0.009481	0.026527	-0.002808	0.002772	0.004185	-0.003887	...	0.001514	-0.012864
2	-0.020273	0.001627	-0.014112	-0.002820	-0.015522	0.014696	-0.003198	-0.007348	0.006434	-0.010143	...	0.001354	-0.006265
3	-0.018496	0.002106	-0.017010	-0.013733	-0.001585	0.017937	0.000349	-0.001777	-0.006361	-0.001006	...	0.001634	-0.008321
4	-0.017381	0.011872	-0.006053	-0.011366	0.003634	0.020518	0.002399	-0.007558	0.006926	-0.006363	...	-0.003403	0.003117

5 rows × 12288 columns

Figure: 12,288 dimensions field embedding for each tweet, obtained from GPT-3

Next, we asked GPT-3 to label a random subset of 10,000 tweets as Positive, Negative, or Neutral, providing us with a set of labeled data for training our model. We used OpenAI's latest text-davinci-003 model for the labeling part.

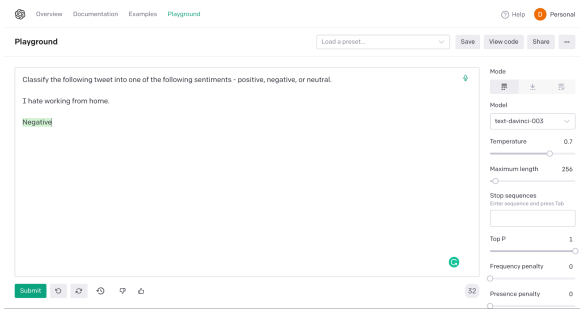


Figure: Ask GPT-3 to classify tweets into Negative, Positive or Neutral

These labels were used as the training input for our classifier.

The field embeddings were a watershed breakthrough in NLP; they provide an easier way of finding semantic similarity between two text strings. For example, we have 3 tweets -

- "$TSLA calls printing right now! Up up and Away!" (Tweet A)
- "All the way green on $MSFT" (Tweet B)
- "The bears are tearing $AMZN up" (Tweet C)

Each of these tweets would be transformed to a 1,536 dimensional vector, namely Vector {A,B,C}. If we were to find euclidean distances within those 3 vectors, we noticed that the Euclidean distance between Vectors A and B is much smaller than the distance between Vectors A and C or Vectors B and C in a hyper-dimensional space.

3.4 SVM Model

Using these GPT-3 embeddings as our input value and GPT-3 labels as output labels, we trained a Support Vector Machine (SVM) to classify the remaining tweets. We chose to train our own classifier because having GPT-3 label every single tweet of the >2 million tweets would cost over $1000. Hence, we thought of obtaining a small labeled set at a fraction of the cost and then train our classifier on top of it. We ultimately passed all the 2 million embedding vectors into the classifier, and were able to label every tweet into "bullish", "bearish" or "neutral".

3.4 Model Evaluation

Ultimately, we did our data aggregations where we aggregated the data on the date and ticker columns.

	date	ticker_symbol	labels	number_of_tweets
0	2015-01-01	AAPL	negative	26
1	2015-01-01	AAPL	neutral	115
2	2015-01-01	AAPL	positive	42
3	2015-01-01	AMZN	negative	35
4	2015-01-01	AMZN	neutral	30
5	2015-01-01	AMZN	positive	9
6	2015-01-01	GOOG	negative	4
7	2015-01-01	GOOG	neutral	27
8	2015-01-01	GOOG	positive	6
9	2015-01-01	GOOGL	negative	4
10	2015-01-01	GOOGL	neutral	11
11	2015-01-01	GOOGL	positive	5
12	2015-01-01	MSFT	negative	6
13	2015-01-01	MSFT	neutral	15
14	2015-01-01	MSFT	positive	11
15	2015-01-01	TSLA	negative	14
16	2015-01-01	TSLA	neutral	39

Figure: Tickers with number of bullish, bearish, and neutral labels for each day

The ultimate goal of our analysis was to train a classifier to accurately predict the stock market based on the tweets we collected. We ran the classifier on each tweet on a day-over-day basis, and based on the classifications we assigned a rating to the stock tickers. To evaluate the accuracy of our predictions, we tracked the performance of the stocks on the following day after the rating was assigned. For example, if $TSLA saw a +7% price movement, it would be labeled as a positive or “bullish” day. We would then compare the accuracy by tracking our model’s prediction for that day.

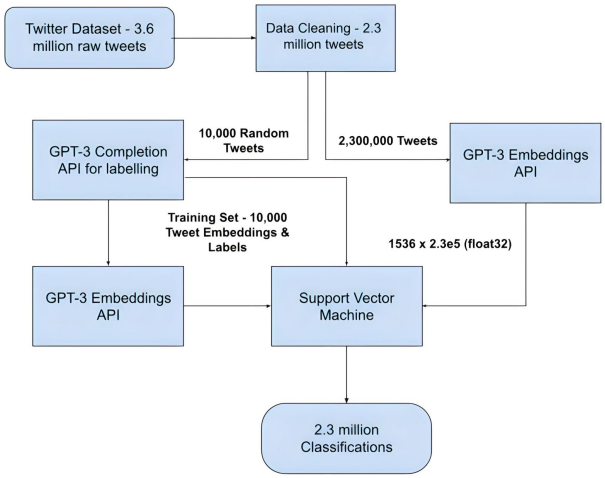


Figure: Model Architecture

4. Results

4.1 Visualizing Labels

After classifying all tweets, we used Uniform Manifold Approximation and Projection (UMAP) for dimensionality reductions to reduce our 1,536 dimension vector embeddings to 2 dimensions for ease of visualization. We used UMAP over other dimensionality reduction techniques like t-SNE and PCA because it has been proven to be useful at reducing very high dimensional (> 1000) data. The data points are also color coded by their labels.

Bullish Tweets are in yellow, Bearish Tweets are in yellow and Neutral tweets are in Blue. As observed, the most dense and frequent label is the Neutral label.

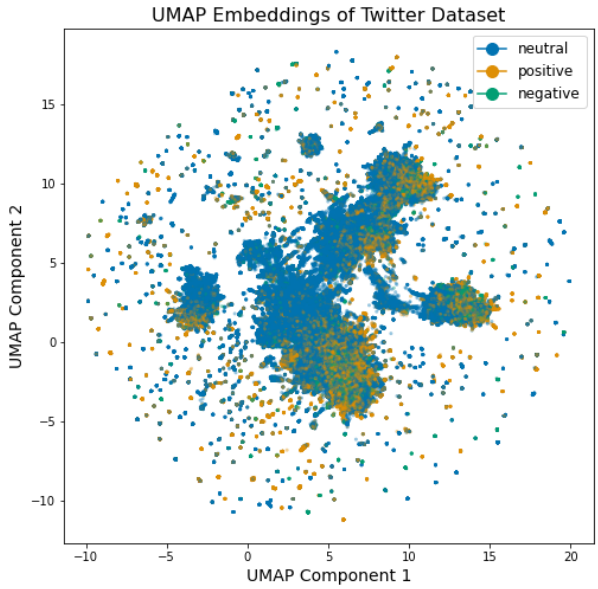


Figure: Scatterplot figure of bullish, bearish, and neutral tweets

4.2 Dense Neutral Labels

Upon aggregating the results, we found that the Neutral tweets overpowered the results. Of the initial 10,000 tweets classified using GPT-3, 70% were labeled as neutral. This bias likely permeated when we trained our SVM classifier on the rest of the 2 million tweets. In other words, on a certain day for $AAPL, most predictions were likely neutral because of the high density of "Neutral" tweets that the model was trained and tested on.

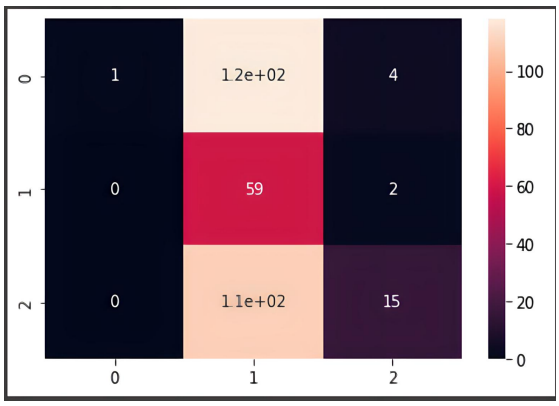


Figure: Heatmap of the Correlation matrix on $TSLA predictions and 3 labels

In the correlation matrix above for $TSLA, the column in the middle represents the Neutral predictions. Upon this, we thought of removing the neutral tweets and received the following confusion matrix for $TSLA:

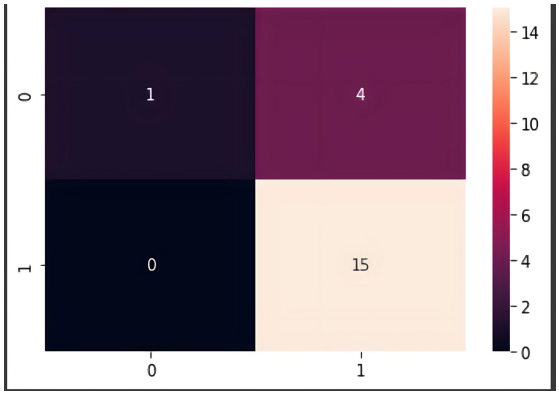


Figure: Heatmap of $TSLA predictions after removal of neutral labels

As one can observe, the "True Positive" box on the lower right corner has the highest frequency, implying that our model predicted 75% the right label when labeling "Positive" or "Negative".

Our results improved significantly from this change. However, the removal of neutral labels left us with notably fewer data points.

5. Future Improvements

In future iterations of this methodology, there is an opportunity to utilize real-time streaming clients such as Astra DB and Twitter API to produce more relevant and useful predictions. Since February 9th, 2023, Twitter has ended free access to its API. As a result, with current limitations in resources and funding, projects involving the collection of data from Twitter

(via wrappers such as Tweepy or Twython) would not be possible at this time.

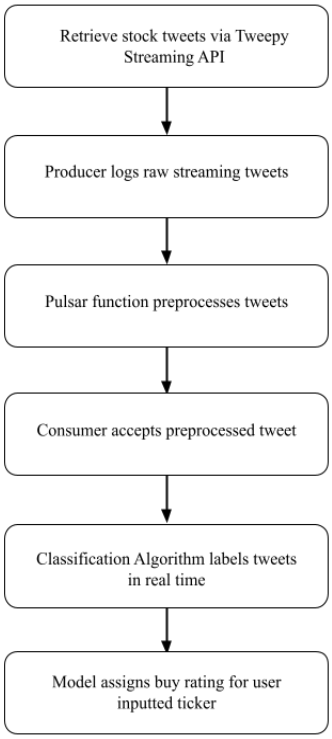


Figure: Block diagram of the proposed methodology utilizing streaming clients

However, using streaming clients in the future, we would be able to stream and classify tweets in real-time. Approximately hundreds of tweets relating to financial markets are posted every second. We would be able to filter specific tickers with Tweepy to view tweets of tickers such as $TSLA, $GOOGL, $APPL, $MSFT, and $AMZN. After a set length of time of streaming and classifying tweets, we could perform a similar buy rating for each ticker. Assuming that roughly 1,000 relevant tweets about $TSLA are posted within a minute, we could classify each tweet and give a buy rating for $TSLA within a minute timeframe. This would provide more relevant and useful information for the user since everything is up-to-date and can influence trading strategies such as day or swing trading.

6. Conclusion

Our research has shown that advanced natural language processing techniques can provide valuable insights into the wisdom of the crowd phenomenon. By utilizing an extremely large dataset and GPT-3 field embeddings, we were able to classify tweets into Bullish, Bearish, and Neutral sentiments with the help of GPT-3 and SVMs. Furthermore, by utilizing overall sentiments to make a decision on buying, selling, or holding a particular stock ticker, we were able to show the potential of the wisdom of the crowd.

Nonetheless, further research is needed to improve the accuracy and usability of our analysis, particularly with the use of real-time streaming clients. Despite these limitations, our study highlights the potential of a popular phenomenon and advanced natural language processing techniques for sentiment analysis, and the potential applications of such analysis in fields such as economics and financial markets.

7. Appendix

The main source of data for this project can be found in the following **Kaggle dataset.** This dataset contains 3.6 million unique tweets with their information such as tweet id, tweet author, post date, the text of the tweet, and the number of comments, likes, and retweets of tweets matched with the related company. These tweets come from the top companies from 2015 to 2020 which include Tesla, Google, Apple, Microsoft, and Amazon. Each tweet isn't indexed by a specific company since the tweets themselves often mention several companies or tickers.