Predicting stock market movement using twitter sentiment analysis.

Rebecca Baert

School of Mathematical and Physical Sciences

Background

- Sentiment analysis of tweets provides a useful insight to the public's attitude towards a specific discussion.
- This project considered sentiment analysis of tweets to attempt prediction of the S&P 500 Index using twitter mining and machine learning techniques.

Objectives

1. Use twitter mining to collect tweets relevant to the S&P 500.
2. Perform sentiment analysis on tweets to gain understanding of publics attitude towards the index at time.
3. Apply machine learning techniques to predict movement of S&P 500 based on data collected.

Twitter Scraping

Twitter scrapping is collecting tweets that pertain to a keyword, user, or location. Many organizations use twitter scraping as a tool to gain insight of the public's reaction to their operations. This project aimed to use twitter mining to understand the public's changing attitude towards the S&P 500 index.

1. This project used Tweepy, a Python package that allows access to twitter API.
2. Twitter was scrapped by date for key words such as 'SPX500', 'stocks', and several company symbols of the S&P 500.
3. These tweets were stored in a dataframe, analyzed, and converted into data useable for machine learning prediction.

Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) [3] was used for tweet sentiment analysis. VADER is open-source sentiment analysis tool specifically used for analyzing sentiment of social media text. Some cases it is equipped to consider include:

- Negations
- Emoticons
- Word emphasis (e.g., all caps)

Machine Learning

Machine learning is the process of training algorithms that can learn from data, with the goal of classifying, clustering, or labelling new data instances. This project used Scikit-Learn's Support Vector Machine Classifier [2]. A SVM classifier works by:

- Separating classes using decision boundaries.
- Maximizing space between support vectors.

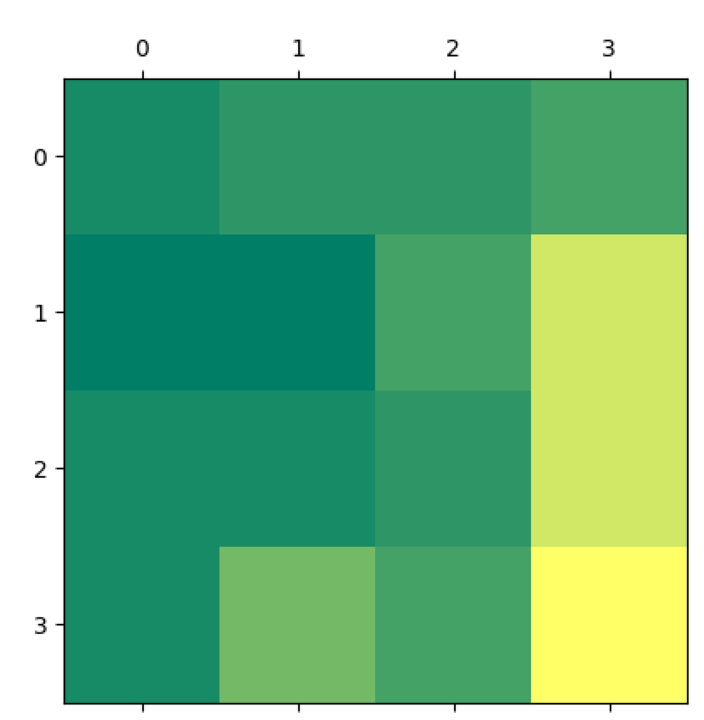
The model was trained on a data set of 5,000 tweets collected between April 9th and July 16th 2021. Several example data instances are show below. The algorithm trained on features:

- Number of positive tweets collected on the day.
- Number of negative tweets collected on the day.
- Ratio of positive to negative tweets.

The model was provided labels for the training set:

- S&P change of the given date.

Before fitting the model, the data was scaled using standard scalar. The model was fit to the data and evaluated using cross validation.



Confusion matrix. A confusion matrix is a way to compare predicted class to the actual class of instances. The predicted class is represented by the horizontal axis and actual class is the vertical axis. A strong confusion matrix would have a high-counts (here represented by light yellow) along the diagonal and low-counts (dark green) elsewhere. Confusion matrices are useful to determine instances are frequently misclassified.

Data

Date	Positive Count	Negative Count	Ratio of Positive to Negative Tweets	S&P Change
1586404800	27	15	1.800000	Large_increase
1586750400	46	17	1.529412	Small_decrease
1586836800	41	21	1.952381	Large_increase
1586923200	41	25	1.640000	Large_decrease
1587009600	40	5	8.000000	Small_increase
1587355200	54	23	2.347826	Large_increase
1587441600	29	15	1.933333	Large_decrease

Results

Using eight-fold cross validation, the model yielded the following array of accuracy scores:
[0.286, 0.429, 0.429, 0.143, 0.000, 0.167, 0.333, 0.333]
Average accuracy: 0.265
Standard deviation: 0.141
Confusion matrix shown to the left.
Moreover, the model was used to predict movement of a specific date from twitter scraping results.

Conclusion

The SVM model yielded an accuracy of 26.5%, roughly equivalent to random guessing. However, by viewing the confusion matrix, we can see high counts in the fourth column and low counts in the first column. The model is often predicting a large increase of the S&P 500 and rarely predicting a large decrease. This indicates the model is underfitting the training data: likely because the small data set doesn't provide enough variety in instances. The model was trained on roughly three months of data, providing only 54 training instances.

Additionally, the stock market is difficult to predict. It is very unlikely to accurately predict its movements based on public sentiment alone. However, it may be useful to consider public sentiment in conjunction with other features. Possible future features that could be implemented into the model include:

- Time series analysis
- Interest rates
- Price-to-earnings ratio of individual companies within S&P 500

References

[1] Beri, Aditya. (2020). *SENTIMENTAL ANALYSIS USING VADER.* Towards Data Science. Retrieved from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[2] Geron, Aurelien. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc., Sebastopol ,CA.

[3] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI..

[4] Taborda, A., Ana de Almeida, José Carlos Dias, Fernando Batista, & Ricardo Ribeiro. [2021]. *Stock Market Tweets Data*. IEEE Dataport. Retrieved from https://dx.doi.org/10.21227/g8vy-5w61.

[5] Roesslein, Joshua. (2022). *Tweepy Documentation*. Tweepy. Retrieved from https://docs.tweepy.org/en/stable/.