Project Proposal

LSTM Neural Network for Sentiment Analysis of Twitter Stock Messages

1 Introduction

The impact of Social Media on influencing the prices of traded assets appears to be growing rapidly in recent years. While Social Media platforms such as Reddit, Facebook, and Twitter have been around for more than a decade, their effects on market prices has become very apparent just recently as there has been a sudden influx of young retail traders to the markets. User friendly trading applications, such as Robinhood, have a median age of 31 years old [1]. This demographic are heavy users of social media and often will exchange stock recommendations that can rapidly gain momentum. The recent SEC investigation into the Gamestop price surge of January 2021 found that most price gain was due to purchases by individual retail traders, who were almost exclusively communicating through public social media, primarily on the Reddit "Wall Street Bets" forum.

There are also individual users with large influence through platforms such as Twitter. Elon Musk's positive tweets about "Dogecoin" caused sudden buying frenzies that has helped drive prices to 11,000% gain over 12 months. Conversely, his negative tweets about Bitcoin on June 3rd, 2021 caused an intraday drop of 4.3%.

As shown in the previous cryptocurrency example, social media popularity plays a large role in affecting market prices, but also the sentiment with the content needs to be examined to help improve odds of accurately predicting future price movements.

While the inherent volatility and unpredictability of Stock price movements will always create risk for any trader, sentiment analysis of social media media has been demonstrated to improve overall accuracy of investing strategies. The 2016 published research paper "Stock Price Forecasting via Sentiment Analysis on Twitter" found a substantial correlation between sentiment and future price movements [3]. This correlation may likely grow in strength over time as Social Media continues to increase in influence and the average investor demographic will further overlap with "Millenial" and "Zoomer" age brackets, who are heaviest users of social media.

2 Dataset

The "Stock Market Tweets Lexicon Data" dataset publicly shared on Kaggle and IEEE DataPort will be used to train-test sentiment analysis models. Tweets were collected using the "Tweepy" Twitter Rest API from April 9 to July 16, 2020 [4]. All tweets were

tagged with #SPX500 and #stocks. Of the 900,000 tweets collected, 1300 of them were manually reviewed and classified with a positive, neutral or negative sentiment rating. Multiple users reviewed tweets in the data set, reducing impact of individual bias when classifying as positive or negative.

If more samples are needed in order to train the model, larger data sets can quickly be generated by using positive emojis to annotate tweets for supervised learning.

A typical 70/30 train-test split will be used for developing the classification model.

3 Evaluation Method

Fortunately, research on the field of sentiment analysis is very active due to its many real world applications. The consensus in recent years has been favoring use of Neural Networks rather than statistical methods. There are instances where simpler statistics methods such as Naive Bayes classifier perform well, and do not run the risk of overfitting. However, improvements in computer hardware have allowed greater development of Deep Learning neural networks that continue to outperform Naive Bayes. Recurrent Neural Networks, specifically LSTM models (Long-Short-Term Memory) are commonly used, and will be the first method applied for sentiment analysis of the Twitter dataset.

The LSTM will be trained under supervised learning on the 70% data split. Gradient descent most likely will be used to back propagate and optimize results. Testing on the final 30% of data will demonstrate the overall performance of the model.

While it is expected from research that LSTM should outperform all other methods, classification via methods such as Decision Tree or Naive-Bayes Bayesian Network will be used for performance comparison. Currently, many hybrid models are being tested to some degree of success for sentiment analysis. Thus, it is beneficial to explore multiple ways to assess the meaning of a text, especially given that different media platforms have distinct characteristics that may be classified better with different models. The inherently short nature of tweets may give the simpler Naive Bayes approach better results in this instance.

References

[1] Keshner, Andrew.”A burning question to ask before buying Robinhood IPO stock, Will Users ‘Age Out’ of the App?”. 29 July, 2021. https://www.marketwatch.com/story/a-burning-question-to-ask-before-buying-robinhood-ipo-stock-will-users-age-out-of-the-app-11627567521

[2] Browne, Ryan. "Bitcoin falls after Elon Musk tweets breakup meme“. 4 June, 2021. https://www.cnbc.com/2021/06/04/bitcoin-falls-after-elon-musk-tweets-breakup-meme.html

[3] J. Kordonis & S. Symeonidis & A. Arampatzis. “Stock Price Forecasting via Sentiment Analysis on Twitter.” November 2016. https://www.researchgate.net/publication/311843931_Stock_Price_Forecasting_via_Sentiment_Analysis_on_Twitter

[4] Bruno Taborda, Ana de Almeida, José Carlos Dias, Fernando Batista, Ricardo Ribeiro, April 15, 2021, "Stock Market Tweets Data", IEEE Dataport, doi: https://dx.doi.org/10.21227/g8vy-5w61.