MACHINE LEARNING MODEL TO PREDICT STOCK PRICES BASED ON TWITTER SENTIMENT ANALYSIS

Author: Vishaal Bakshi

OVERVIEW

1. Abstract

Stock price prediction is one of the oldest problems in the finance industry. Since the beginning of trading, there has been a need for predicting prices accurately in order to maximize/minimize profits. Yet the industry still lacks a standardized approach in solving this problem. I study one of the many parameters in this project to understand if there is a strong correlation between historical bank stock prices and short sellers who try to game the system. The issue is an important one since such speculative behavior led to the financial crisis of 2008/2009. The goal is to use a time-series model to predict the stock price in conjunction with short seller claims.

2. Project Scope

Introduction: *The purpose of the project is to build a model that can predict next day bank stock prices using prominent short seller twitter data. The approach shall be to use a time-series model: Long Short-Term Memory (LSTM). LSTM are extremely powerful since they can predict quiet far into the future. Dataset would be retrieved using standard Twitter API's and publicly available stock exchange trading data.*

Solution: *The LSTM approach has a strong case of success since it is explicitly built for time-series data. In fact, it is a time series model. Stock ticker data is done minute by minute which can be expanded to daily. The weighted daily average would be*

an easier way to train the model as it correlates with short seller twitter posts.

3. Challenge

Previous Work: *There have been numerous attempts at solving the stock price prediction problem. Some have used more established algorithms such as Linear Regression, K-NN or SVM and some have even used neural networks such as GANN's to solve this problem. However, there is no standard industry wide algorithm that has been able to address this. I am intrigued by the nature of the problem and to what degree short sellers can affect stock prices by making public statements.*

The problem if solved with a reasonable degree of accuracy, would benefit not only traders but also regulators. Regulators would gain a better insight into the effect short sellers have at swinging next day stock prices. If the effect is severe then policies can be used to curtail their behavior.

4. Approach and Methodology

Steps and Expected Problems*: First step would be to collect the data. Stock exchanges have large datasets available to the public free of charge. However, the data is not structured. The datasets will likely require wrangling to make it structured. It is my hope that the datasets are not severely unstructured as this is the most time-consuming portion of a data scientist. If the data requires a lot of" cleaning", I plan on limiting my hyperparameters for LSTM to finish the project in time.*

Another challenging task will be to scrape short seller data using their twitter handles. Finding influential short sellers who influence stock prices is a time-consuming process. The step requires

delving into bank terminology and stock market literature for the banking industry.

5. Metrics

Measure of Success: *Training the model will be another time-consuming task as I do not have access to powerful GPU's or online frameworks such as AWS or Google TensorFlow. Therefore, running large dataset calculations will have to be done on my desktop. Acceptable accuracy of the model will depend on how long it takes to train the model. I would like to aim for 90%+ accuracy, however project deadline submissions will determine how deep I am able to get into the model and re-train it.*

6. Summary

It is my aim to achieve a better understanding on how to implement a time-series algorithm. The problem outlined above has remained largely unsolved since the inception of stock trading. I hope to show some degree of correlation between stock prices and human input that industry experts can quantify and predict.

7. Specific Exclusions from Scope

There are several things that shall be excluded from the dataset intentionally. While this is not an exhaustive list, the following are some that should be stated:

1) *Central Bank interest rate announcements*
2) *SEC investigations*
3) *Responses to short seller statements*

These have been excluded as short sellers will often use the same data to make public statements. However, since these tend to come out before short seller statements, they might effect the pricing prediction. The goal of this project is to analyze the direct relationship between short sellers and bank stock prices.

8. High-Level Timeline/Schedule

1) *The aim is to have the entire dataset collected and merged by end of January to early February.*
2) *Once the data is merged, it will require data wrangling in order to structure it. This is the most time consuming portion and therefore will probably require the month of February to complete this.*
3) *Training the model and code implementation should take up to 2 weeks (hopefully). Once this is achieved I will have the remainder of the time to train and re-train the model. As mentioned earlier, the level of acceptable accuracy shall depend on how long it takes the model to run.*