Stock Analysis using Natural Language Processing

By: Bryan Archinega, Jacob Reyes, Harper Vans, Anas Alakkad, Tristan Scharfenstein-Montgomery, Andrew D.

Abstract

Stock price prediction is among one of the most uncertain predictions due to many volatile factors. One such important factor is the public sentiment about a certain stock at a certain time. One way to anticipate this is via how social media users react to certain stocks. In order to predict stock prices, many researchers have worked on training machine learning models using regression and classification to predict times series data. In this project, we will leverage those models, along with other available databases and tools, to improve prediction on stock prices based on analysis of social media and effectively execute stock trade.

Table of Contents

Introduction

Stock market has always been unpredictable. There are numerous volatile factors that account for how a stock will behave in a short amount of time. Many variables, both independent and dependent on the market, heavily influence a certain stock at a specific time that make it significantly challenging to anticipate how the stock will behave in both short and long period of time. One of the deciding factors of stock outlook is how the public view about the stock. It has always been known that stock market performance influences the public, but the reverse is also true. With the continuous growth of social media, people can easily express their opinions and emotions about any topic. Those opinions, especially those that have the majority support, will easily influence how the public views certain topics, including products and services. Understanding the nature of these stocks and their dependent factors will increase the chance of higher return for investors. Researchers have utilized public sentiment on these products to determine whether a product will gain or lose public support. Therefore, analyzing social media platforms will help predict how a stock will behave based on the trend of the current market.

The most popular way to analyze the ongoing trend among social media platform users is utilizing Machine Learning models and algorithms. Machine Learning has proved its potential in the financial world by providing significant insight to the stock market that helps traders and investors gain more understanding of how certain stocks and products behave. Machine Learning helps collect data from available databases, does analysis to find out the trend of the stocks, trains the model based on the analysis, and outputs the expected outlook for a certain stock based on the available history of the stocks. For this project, we utilized available algorithms, tools, and resources to predict how a certain stock outlook will be based on sentiment analysis on Twitter users. Stock trading will be executed if the stock is determined to have high return. For

implementation of the project, our model was trained using the machine learning algorithm in Python language using Jupyter Notebook and Google Colab's GPU. We then focus on sentiment analysis based on the Twitter platform by utilizing Tweepy API to train our model predicting the stock outlook. If the stock received high potential of selling, we would initiate stock trading for that specific stock using paper trading API Alpaca.

Project Implementation Procedure

To begin the training model, we need to collect databases from social media users. The social media platform that we chose to do our sentiment analysis on is Twitter. In order to collect data from Twitter, we needed access to the Twitter API. Application Programming Interface (API) is a magical tool to use as a software developer. They allow us to connect to endpoints from certain websites in this experiment. We created a Twitter account and requested developer access to their API. We then used their API to pull specific tweets related to stocks. Through the Twitter Tweepy API we will request a live stream of a fixed number of tweets that contain the ticker symbols of the stocks that we want to analyze. Using hashtags as the keywords to search and specific investing accounts as the main source, we can target more on certain stocks at a certain time. With this, we are able to get a decent consensus on the sentiment towards specific stocks outlook.

After we have a database of tweets from Twitter, we started to implement sentiment analysis in order to sort whether a certain stock is bullish or bearish. A bullish stock is one that investors believe is going to rise in value, while a bearish stock means that investors believe the share price will decrease.The sentiment analysis code was written in python using the helpful NLTK (Natural Language Toolkit) library in Python. The tweets were first cleaned up by

removing unnecessary symbols and numbers. We then labeled a set of tweets and used them to train the model. Each tweet that was pulled would then be run through the model and determine whether there was a positive or negative sentiment.

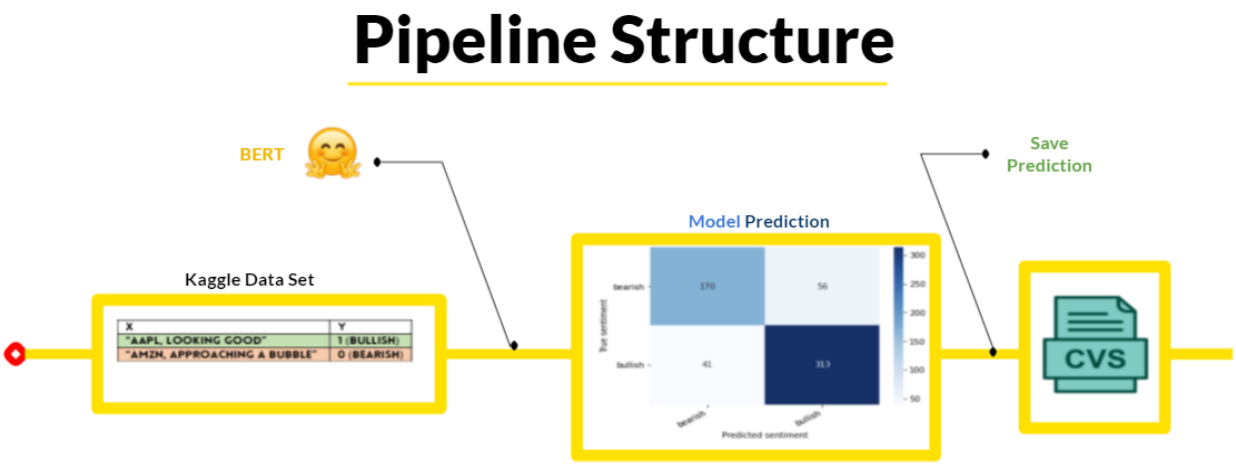

Figure 1. Pipeline Structure of our data collection process

In other to train the model, we utilized both Jupyter Notebook and Google Colab to work on our project. To first train our model we utilized Google Colab because this program allows us to train our model with the GPUs from Google. This saved us time as it only took about 25 min to train the model for 10 epochs on 5790 training tweets. After we trained the model we downloaded the model locally and ran it on Jupyter Notebook. This Jupyter Notebook was ran on a virtual linux machine to isolate it from my local python environment.

Once the sentiment is determined the sentiment and stock ticker is sent to an algorithm that sends an API call to a paper trading stock API that is able to either buy or sell the stock depending on the sentiment calculated. The paper trading stock API is called Alpaca which can track live portfolio performance. We will also use the Alpaca paper trading account to buy and sell certain stocks based on the bearish or bullish sentiment gathered from our training model.

Finally we ran the code from 6:30AM to 8:00AM and bought 11 stocks with a total of $197,151.71. As of November 23, we have lost 0.65% of our initial portfolio.

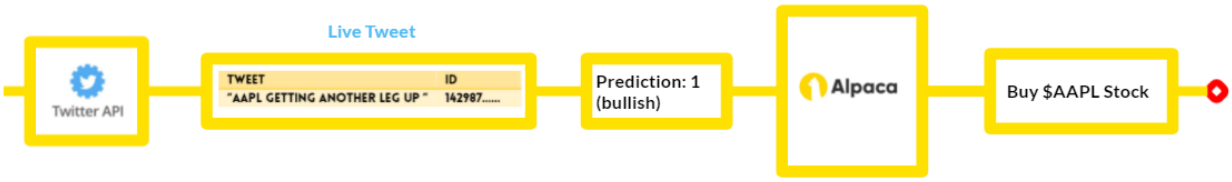

Figure 2. Pipeline Structure of Sentiment Analysis and Stock Trading Execution

Conclusion

In this project, we utilized the power of supply and demand of the stock market to improve a Machine Learning model to predict the stock price outlook and execute stock trading if needed. We used NLTK (Natural Language Toolkit) library in Python to implement sentiment analysis on Twitter users regarding specific stocks. We also utilized Jupyter Notebook and Google Colab to facilitate the model training. We tested the code by having it ran for about 2 hours in the early morning and traded some stocks. We planned to have the code ran for more days in order to gain more datas for our model and achieve a more accurate prediction.