Machine Learning Final Project Spring 2023

1st Alan Kassymkanov
Computer Science Dept.
University of Richmond
Richmond, VA, USA
alan.kassymkanov@richmond.edu
2nd Naron Chen
Computer Science Dept.
University of Richmond
Richmond, VA, USA
naron.chen@richmond.edu
3rd Tony Wang
Computer Science Dept.
University of Richmond
Richmond, VA, USA
luoli.wang@richmond.edu
4rd Wonyoung Kim
Computer Science Dept.
University of Richmond
Richmond, VA, USA
wonyoung.kim@richmond.edu

I. Minimum Requirements of Project

Please refer to the final project guidelines document available on Blackboard Learn. The document lists the minimum requirements, grading rubric, and a tentative project timeline. While working on the project plan, you can also review the lecture slides on machine learning system design and applications.

II. Project Title

Integrating Opinion Leader Sentiment for Stock Price Forecasting of Top Tech Companies

III. Project Aim

The aim of this project is to develop a more accurate stock price forecasting model for top tech companies by integrating the sentiments of opinion leaders. The dynamic nature of top tech companies led us to consider daily social media posts as an intuitive prediction feature. Different from other approaches that measure the overall mood from the social media as a feature, this approach recognizes the significant influence of opinion leaders in shaping public perception and expects that their opinions would have a more significant impact on the stock prices. We will leverage already-established sentiment analysis techniques to gather insights from opinion leaders' posts along with historical stock data to build our model.

IV. Overall Plan

A. Introduction

The stock market has always been a volatile and unpredictable domain. Investors and traders are constantly seeking ways to predict stock prices, and social media has emerged as a promising avenue for this purpose. The aim of this research is to develop a model that can forecast the stock prices of top tech companies using sentiment analysis of social media data. The study focuses on top tech companies, which are known for their fast-paced and dynamic nature. Thus intuitively, daily social media posts should be capable of reflecting some portion of the market.

B. Related Work

Paper Reference:

1) Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. "Sentiment analysis on social media for stock movement prediction." Expert Systems with Applications 42.24 (2015): 9603-9611.
2) Derakhshan, Ali, and Hamid Beigy. "Sentiment analysis on stock social media for stock price movement prediction." Engineering Applications of Artificial Intelligence 85 (2019): 569-578.

Code Reference:

1) https://www.kaggle.com/code/tommyupton/twitter-stock-market-sentiment-analysis/notebook
2) https://www.kaggle.com/code/hxtruong6/k417-how-twitter-affects-the-stock-market

Related work from student Project:

1) https://cs229.stanford.edu/proj2016/report/Tsui-PredictingStockPriceMovementUsingSocialMediaAnalysis-report.pdf

what we will do differently: we will use different datasets, different processing of datasets (eg. ranking the posts on popularity), different Machine learning Models which we will elaborate in later paragraph.

C. Dataset Used

1) Tweets about the Top Companies from 2015 to 2020
2) Values of Top NASDAQ Companies from 2010 to 2020
3) Financial Tweets
4) yahoo finance data

D. Approach

First, we'll process the data, linking each tweet's writer to the Twitter API to determine their number of followers and rank them. We'll then merge this with historical stock data, as outlined in the Data Analysis step.

One visualization option is a scatter plot, with sentiment analysis data represented by point size or color, and stock price movements represented by position. This will allow for a more detailed exploration of the relationship between sentiment and stock price.

Next, our model, detailed in the methodology section, will be trained for both regression and classification. The data will be split into 80% training and 20% testing, and we'll attempt to test the model on current stock market data through Alpaca.com, time permitting.

E. Data Analysis Steps

We will use the Twitter API to obtain the number of followers for each tweet's author and select the ones with the highest number of followers as "opinion leaders."

Then we perform sentiment analysis on the "opinion leaders" tweet to obtain their feeling about the market, then weight by the number of likes of that tweet as well as the number of followers. We will do this for all posts from "opinion leaders" on each day to get an overall "feeling" of "opinion leaders" for the market. Finally, we will merge this data with historical stock data by date to have the final data sets of all features.

Some ideas about visualization:

To visualize the final dataset, we can use a line chart or a candlestick chart to display the historical stock data by date. We can then overlay on this chart the sentiment analysis data for the "opinion leaders" by date, possibly as a line or bar chart.

We can also use a heat map or choropleth map to show the sentiment analysis results on a geographic basis, highlighting regions where the "opinion leaders" are particularly positive or negative about the market.

Another option is to use a scatter plot, with the sentiment analysis data represented by the size or color of the data points, and the stock price movements represented by the position on the chart. This would allow us to explore the relationship between the sentiment analysis data and the stock price movements in more detail.

F. Methodology

Since we will solve both regression and classification model, we will apply different models. To achieve a concrete price prediction for the next day:

1) Linear Regression,
2) Gamma Regression
3) Random Forest, XGBoost

To achieve a classification on price prediction trend for the next day:

1) Logistic Regression
2) Support Vector Machine
3) Gaussian Discriminant Analysis
4) KNN (with L1 and L2 metrics)

Evaluation Metrics for regression:

1) Mean Squared Error (MSE).
2) Root Mean Squared Error (RMSE).

Evaluation Metrics for classification:

1) F-1 score.
2) Accuracy.
3) Confusion Matrix.
4) Receiver Operating Characteristic (ROC) curve

G. Expected Results

Our hypothesis suggests a correlation exists between the swift fluctuations of top tech companies' stock prices and daily social media activity. Social media can provide a glimpse of the overall market sentiment, as it reflects a portion of it. By conducting sentiment analysis on social media, we can grasp the current market perception of a stock. Furthermore, the opinion leaders heavily dominate social media today.

Additionally, we believe that the most recent historical data, such as yesterday's stock price, has the greatest impact on today's stock price. Our perspective is that the market price trend resembles a Markov chain process, which is solely reliant on the most recent data.

V. Tentative Timeline and Responsibilities

will work on data collection and data processing

- Naron will do Data Collection and Data Processing
- Naron and Alan will build the linear regression and featuring engineering with sentiment analysis
- Alan will do Gamma Regression, Random Forest and Gaussian Discriminant Analysis
- Tony will work on Logistic Regression and SVM and KNN.
- Wonyoung will do testing and evaluation on all metrics

VI. Optional-Above and Beyond

1) Feature Engineering using Sentiment Analysis as mentioned above.
2) We will try to train the data on Neural Network Model