Project Proposal
MPCS 53113: Natural Language Processing

Shobhit Verma, Ashish Verma
shobhitv@uchicago.edu, ashishv@uchicago.edu

June 27, 2022

1 Title and Team Members

Our project is 'Prediction of stock price of a company via sentiment analysis of tweets'. The team members are - Shobhit Verma (shobhitv@uchicago.edu) and Ashish Verma (ashishv@uchicago.edu).

2 Abstract

Our project aims at tackling the age old problem of predicting stock prices. Our assumption (which is widely accepted) is that the share price is hugely correlated to 'public sentiment'. In this project, we first aim at quantifying public sentiment by sentiment analysis of related tweets, and then use that to predict the change in stock price. The project is composed of two parts:

1. Building a classification model for sentiment analysis of tweets - given a tweet, the model classifies the tweet into one of the following 5 classes: extremely negative, negative, neutral, positive, extremely positive.

2. Building a regression model that maps the sentiment (or a group of sentiments of all the tweets made in, say, a day) to the change (delta) in stock price.

The classification task allows us to choose from many models - RNNs/LSTMs, transformers, etc. and similarly, in regression, we will experiment with linear/logistic regression, neural networks, SVMs, etc. We build our own dataset from scratch - tweets are fetched from the Twitter API and the stock prices are fetched from yahoo finance API. We use pre-existing datasets to train our sentiment classifier, and then use that to classify our tweets into sentiments. Once we have the aggregated 'sentiment' and the corresponding fluctuation of stock price in a day, we use this to train our regression model. Tweets and stock prices will be fetched for only a single company at first to reduce the complexity of our task.

3 Related work and literature survey

According to [1], a stock market, equity market, or share market is the aggregation of buyers and sellers of stocks (also called shares), which represent ownership claims on businesses. From [2], the share price is determined by 'demand and supply' of the shares in the market. Typically, shares with higher demand have higher prices and shares with higher supply have lower prices. In this section, we first give a brief overview of existing works in sentiment analysis and then we discuss previous attempts in utilizing sentiment analysis for prediction of stock prices.

[3] gives a perfect survey of existing sentiment analysis algorithms and applications. It organizes these algorithms into two broad categories - the machine learning approach and the lexicon based approach. We adopt the former category in this project. Although not used in this project, it is worth describing lexicon based approaches - these rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. Further division is into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. [3] surveys and compares the following ML approaches - Probabilistic classifiers (Naive Bayes classifier, Bayesian networks, Maximum Entropy classifier), Linear classifiers (Support Vector Machines, Neural Networks), Decision Tree classifiers and rule based classifiers. They, however, do not mention models used commonly in NLP tasks - RNNs/LSTMs, GRUs and transformers.

As stated by [4], the main problems of the supervised methods mentioned above are that they need a large amount of training data and are usually slow.

4 Plan of action

- WHy are we not using existing datasets
- Curate dataset specific to comapanies - using twitter API fetch tweets, and using yahoo finance api fetch stock price.
- Use different models in sentiment analysis annotator - RNN/LSTM, Transformer
- Develop a regression model for sentiment -¿ stock price fluctuation

5 Evaluation criteria

- Train till day T - k, predict from T - k to T, compare with actual/ground truth
- Train till day k, predict stock price at end of k+1, compare with ground truth.

6 Division of work

- Dataset curation - company specific division,
- Different models for SA - study them by dividing
- Regression model - same