CS 410 Project Proposal

Michael Bencsik
bencsik2@illinois.edu

Team Members

Number	Name	ID
1	Michael Bencsik	bencsik2

Topic (Free Topic)

The goal is to perform stock market sentiment analysis by classifying text from news headlines to stock market prices to determine if there is a strong relationship between the direction of stock price changes and human emotion and if this outcome is predictable.

This topic is interesting due to the highly unpredictable nature of the stock market being influenced by many factors including the emotional nature of human beings. Overall, humans are emotional creatures with a biased perspective which influences their decisions. If there is a strong relationship between sentiment and stock prices, it would enable the ability to view human interactions based on emotions.

The input data will need to be cleaned of any special characters, split into a "bag of words" representation, and vectorized. The minimum success criteria is to train a model using classification to assign a binary sentiment (positive or negative) to a news headline based on the relationship between the text and market index, such as US indexes NYSE and NASDAQ. The model will then be used to predict if the market index will rise or fall based on news headlines and/or tweets. This approach will be performed multiple times using different approaches to the pre-process and model phase. For the pre-process phase, stop words can be removed and TF-IDF weighting can be applied to the vectorization of the text to determine if the model accuracy increases. For the model phase, different classifiers will be used and compared to determine if one approach is more accurate than the others.

The main frameworks that will be used are TensorFlow, PyTorch, and NLPK. Other sub libraries may be used in conjunction with the previous frameworks such as Keras, Word2Vec, numpy, pandas, matplotlib, and scikit. For classifiers, Naive Bayes, SVM, KNN, and neural networks will be used. The initial datasets to be used are "Daily News

for Stock Market Prediction" [1], "Daily Financial News for 6000+ Stocks" [2], "US Economic Performance" [3], and "Economic News Article Tone" [4]. More datasets will be used if found during the assignment.

I expect that the output of the project will show a relationship between news headlines and the stock market prices, although I doubt that the accuracy will be extremely high. To predict the stock market accurately would entail far more advanced NLP models combined with large amounts of additional data such as historical values, earnings reports, and current market analysis techniques. The models will be evaluated and compared using precision, accuracy, and F1 scores.

Programming Languages

Python is the intended main language, using the current stable version of 3.10.8. Other languages will also be used based on the necessity of toolkits and/or APIs.

Workload

Being the only member on the team, I expect the workload to be greater than 20 hours. The goal is to create the classifier initially on one framework and model. Then expand into different models for comparison. The amount of different models, frameworks, and overall variation will be dependent on the amount of time spent resolving any issues that arise from building and training models.

References

[1] Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [10/23/22] from https://www.kaggle.com/aaron7sun/stocknews.

[2] Daily Financial News for 6000+ Stocks, Retrieved [10/23/22] from https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests?select=analyst_ratings_processed.csv

[3] June 25, 2015. US Economic Performance Retrieved [10/23/22] from https://data.world/crowdflower/us-economic-performance

[4] December 8, 2015. Economic News Article Tone Retrieved [10/23/22] from https://data.world/crowdflower/economic-news-article-tone