Prediction of Stock Price Movements using NLP and Financial Data (Dataset Generation)

Chinmay Singh, Sayantan Das, Varun Madhavan, Yash Kulkarni

chinmaysingh@iitkgp.ac.in, emailtosayantan@gmail.com, varun.m.madhavan221313@gmail.com , kulkarni.r.yash@gmail.com

The Idea

Our aim was to come up with a cross disciplinary approach of stock movement prediction that can combine both, the knowledge of computer science as well as Finance in order to generate an inclusive model for the aforesaid task.

Our model takes into account both technical, fundamental and other factors which are supposed to affect security prices and the mood (pessimism/optimism/outlook) of the investing community who ultimately determine security prices using NLP.

All commonly used technical indicators have been used, some with additional exclusive modifications based on manual trading experience. Some uncommon technical indicators are used which have shown to increase model efficiency.

Market inefficiency is due to the emotions/mood of market participants which is guided by publicly available financial news articles. Riding these inefficiencies to our advantage is possible through sentiment analysis, especially for shorter trades.

We have decided to group the movements into three categories like previous works in the field. An increase of more than one percent was marked +1 movement, a decrease of more than one percent the price was marked -1 and the other movements were marked as 0 or neutral.

Basic Terminology

Technical indicators - Mathematical quantities derived from security price movement history over a definite period of time

Correlation - Covariance of returns of two securities divided by the product of the standard deviation of their respective returns. Shows how closely two securities move together

Beta – Beta of a security is the slope of the regression line of the returns of the security against the retains of the market. Mathematically it is the covariance of returns of the security and the returns of the market divided by the square of standard deviation of the returns of the market.

Open Interest (OI) – The number of outstanding derivative contracts of a particular underlying. Max Pain – The controversial theory according to which option writers tend to profit in the long run at the expense of option buyers. Implied volatility – Not the actual volatility, but the value of volatility for which we get the market price of option contract using the Black Scholes equation. Data taken from NSE, so 10% interest rate assumed.

Momentum - A qualitative factor guiding stock movements in a particular direction on a short term basis guided by mood of the investing community and measured by certain momentum indicators

Sources of data Used

To estimate investor sentiment on a trading day we have used a collection of financial news articles collected from Reuters and Bloomberg by Huicheng Liu et al.

We have calculated the sentiment polarity score of each article and grouped article by date. The sentiment analyzer module has been trained on the Imdb movie review dataset.

For each article we have calculated three sentiment scores; one each for the title, the abstract and the main text.

The sentiments of all articles published on a particular day have been averaged out. Hence, we get three sentiment scores for each trading day.

In addition to this, we have obtained OHLCV data for S&P 500 from Yahoo Finance and have calculated a number of technical indicators from this data.

We have also obtained the closing price of market indices for markets that close before the S&P500, which reflect the overall trend of the global market.

Financial Word Embeddings

For analyzing the sentiment of articles, we needed to calculate word embeddings for each word. Instead of using models trained on generic text like Wikipedia articles, we have exclusively used a collection of financial articles for training.

This enables the model to learn more relevant contexts, such as the word 'apple' is mapped closer to words like 'iPhone' and 'AAPL' instead of fruits.

We have used the ConceptNet Numberbatch embeddings for initialization of the training of the same.

The Joint Model

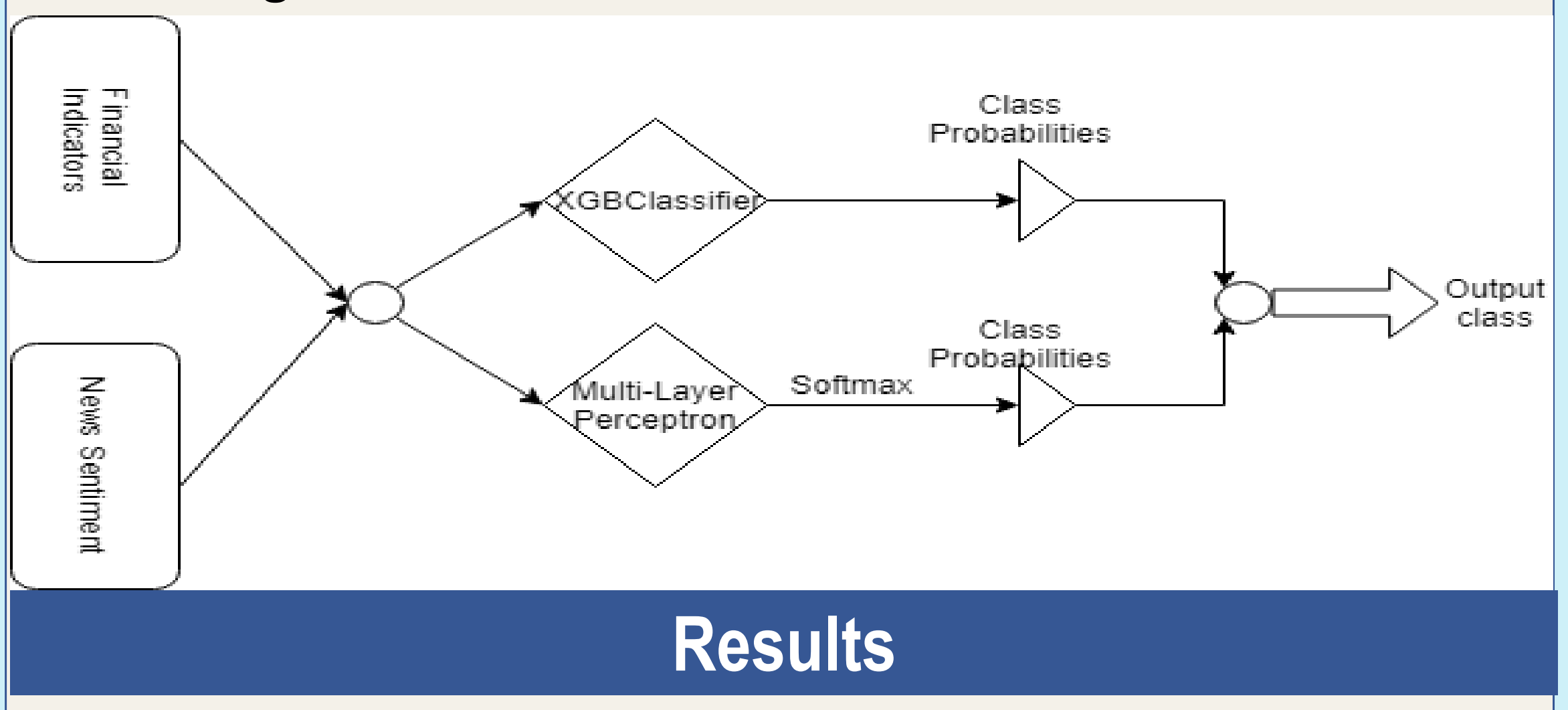
We have just finished the collection of data which would be made public through my GitHub account chinmay-singh The deep learning model to be used is an unfinished task though we have meddled with a few basic models for POC.

We have used a DNN with an LSTM with attention layer.

We tried a number of models types like
- Multi layer perceptron
- RNNs
- Decision Trees
- Boosted Trees

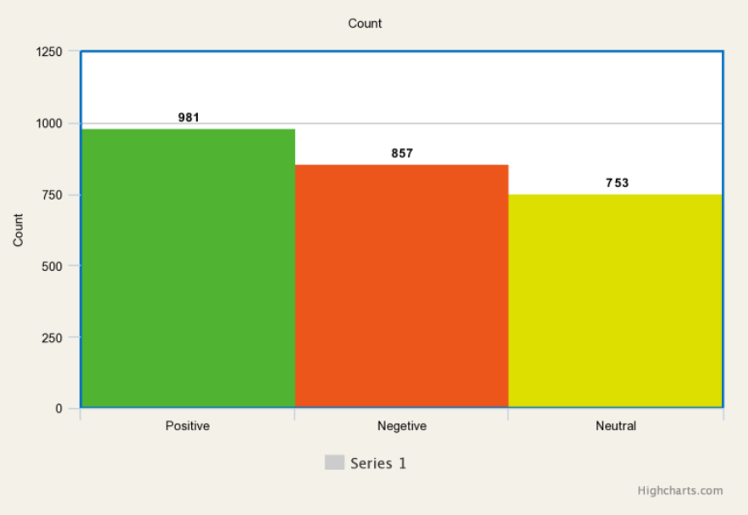
We used Bayesian hyperparameter optimization using hyperopt library to come up with an optimum set of hyperparameters

The model that gave the current best results has the following architecture:

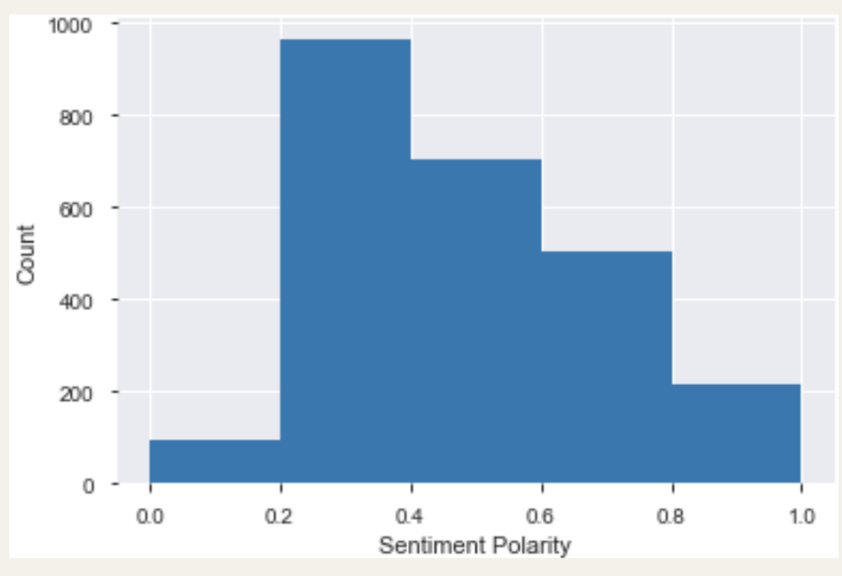


Results

We observed that the dataset was fairly evenly spread out with the following distribution of Positive, Negative and Neutral movements:



The average sentiments were found to be distributed as follows -



The following 5 indicators had the highest effect on prices of securities:
- ROI
- IMI
- Return
- Momentum
- UO

Ongoing work

- So far, we have used pretrained Glove word embeddings as well as the ConceptNet Numberbatch Embeddings however, as discussed above, pretrained word embeddings may not capture the relevant context for financial articles. Hence, we are obtaining custom made embeddings best suited for financial data.
- Parts of our efforts also include the development of a vast database that can be used in this domain of research. We would do this by scraping data wherever necessary from the various credible sources besides the data fetched from existing sources

Future Tasks Planned

Since the initial efforts of the project have mostly been focused on data identification, collection and processing. We would now be shifting to a more comprehensive prediction model for the movement prediction task.

Using more relevant data instead of the ImDb dataset for training the sentiment analysis model

Scraping up to date articles from news websites. Building a mechanism that scrapes all relevant news published in major news websites. Improving the filtering mechanism for choosing relevant articles

To use attention models for exploring the time of manifestation of the language based sentiment of the article into it's price

To use encoders that encode the news so as to avoid the loss of information that occurs because of taking into account only the sentiment

Incorporating a number of features derived from data in the option chain or future contract archives

Acknowledgements

- Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network - Huicheng Liu
- Thumbs up? Sentiment Classification using Machine Learning Techniques – Bo Pang, Lillian Lee, Shivakumar Viathyanathan
- Temporal Pattern Attention for Multivariate time series forecasting – Shun-Yao Shih, Fan-Keng Sun
- On the Importance of Text Analysis for Stock Price Prediction – Heeyoung Lee et al
- Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction – Ziniu Hu et al.
- Betting against Beta by Andrea Frazzini and Lasse Heje Pederson
- Post Earnings Announcement-Drift by Stephen J. Brown and Peter F. Pope
- Piotroski F-Score by Joseph D. Piotroski

Contact Information

Chinmay Singh
2nd Year Undergraduate Student
Chemical Engineering

Varun Madhavan
1st year Undergraduate Student
Department of Chemical Engineering

Sayantan Das
1st year Undergraduate Student
Department of Aerospace Engineering

Yash Kulkarni
1st year Undergraduate Student
Department of Aerospace Engineering