STOCK MARKET PREDICTION USING NLP

Indiana University CSCI-B645 Project

Chirantana Krishnappa
Luddy School of Informatics,
Computing and Engineering
chikrish@iu.edu

Santhosh Manohar Gouda Patil
Luddy School of Informatics,
Computing and Engineering
sanpati@iu.edu

Abstract

This project is aimed to explore the use of Natural Language Processing (NLP) for predicting the market trends of the stock market using data scraped from the r/worldnews subreddit on Reddit to train a model. This subreddit is a prime source of news across the globe and provides an insight to all the events and trends that are going on. It offers a diverse and vast dataset which would cover the whole range of international politics and economic developments. Unlike focusing solely on financial data, this dataset would take into account other related events that could trigger change in stock market sentiment, thus acting as an ideal dataset for our market prediction model. We test out traditional statistical models along with the state of the art Financial-transformer based model (financial-BERT) to compare, analyze and forecast predictions, to find what works best for the Stock market prediction task.

1 Introduction

In this rapidly changing domain of financial markets, social media platforms, particularly Reddit, have emerged as very influential points of information that can play a significant part in influencing stock market trends. The WorldNews subreddit is well known for its global news coverage and discussion around diverse topics, which presents a unique opportunity to analyze and predict market changes. Our project tries to use this huge amount of textual data from WorldNews and find a correlation between news headlines and the movement of the Dow Jones Index (DJI) which is a representation of the overall state of the stock market.

The main focus of this project is to determine if the text content on the headlines of the WorldNews subreddit contains any predictive information about market movements. Our dataset comprises the top 25 headlines of news posts of each day from 2008 to 2016, capturing a variety of global events and their discussions within the subreddit. This period is particularly significant due to the numerous global events that have potentially affected market trends.

2 Related Work

There has been an increase in research focused on extracting information from sentences and documents, with importance on financial forecasting using NLP methods. Some Notable works include:

Harvard University Research: They have researched machine learning algorithms for financial market prediction, stressing on the use of NLP to analyze contextual information from unstructured data.[4]

Time Series Analysis and NLP for Market Trends project has integrated time series analysis with NLP, taking into account the complexity of predicting financial markets, due to a range of different factors that can influence it, which can include economic events and public sentiment expressed on social media platforms like reddit, twitter etc.,[5]

More systematic review has been conducted on the recent advancements in the field of AI, Machine Learning, and NLP, particularly focusing on how these technologies have been utilized in the financial industry.[3]

Focusing on the text-mining literature related to financial forecasting, banking, and the use of social media content, especially Twitter, for such forecasts. [1]

Natural Language-Based Financial Forecasting: This survey emphasizes the increased capability of NLP due to advancements in data availability and big data analytics techniques [2]

Our project, focusing on the WorldNews subreddit, would contribute to this expanding field by exploring the potential of NLP methods in analyzing social media content for financial forecasting. Reddit data is a relatively underexplored area compared to the frequently studied Twitter data.

3 Methodology

3.1 Data preparation

Effective data preparation is essential in the field of natural language processing in order to guarantee the accuracy and use of the input data for further analysis. The main methods used to preprocess and get the data ready for our stock market prediction model are described in this section:

- **Handling Missing Values:** The absence of data can have a substantial effect on the effectiveness of predictive models. To address this problem, we adopted a simple strategy of eliminating rows containing missing values. This will guarantee the completeness of our dataset and makes it possible for our model to be trained on reliable and robust data.
- **Tokenization:** Tokenization is an essential procedure in data processing for tasks related to natural language processing. This process involves breaking down a text into individual words or

tokens. This step is important for extracting meaningful information from textual data and is a prerequisite for implementing next steps in data preprocessing.

- **Removal of Special Characters, Numbers, Hyperlinks:** To ensure that our model focuses on relevant textual information, we removed special characters, numbers, and hyperlinks that may introduce unnecessary noise into the dataset.
- **Removal of Stopwords:** Stopwords are frequently used terms with little value, and can be eliminated from a dataset to cut down on noise. This stage helps in sharpening the model's emphasis to more informative terms for stock market forecasting.

Following the completion of these data preparation procedures, the dataset is now ready for further feature extraction, model training, and evaluation. The processed and tokenized textual data, cleared of irrelevant details, forms the basis for constructing a robust NLP-driven stock market prediction model.

3.2 Feature Extraction

The process of feature extraction plays an important role in converting raw textual data into a numerical format suitable for machine learning models. In this section, we discuss two distinct approaches for feature extraction utilized in our study.

- **Count Tokenizer for Statistical Machine Learning Model:** For traditional statistical machine learning models like Logistic Regression, Support Vector Machines (SVM) or Random Forests, we utilize the Count Tokenizer. This method entails transforming a set of text documents into a matrix of token counts, reflecting the frequency of each word in the document. By using the Count Tokenizer, we aim to capture the occurrence of words in the dataset, facilitating the model's ability to recognize patterns and associations between words and movements in the stock market.
- **BERT-Tokenizer:** For sophisticated models such as BERT (Bidirectional Encoder Representations from Transformers) [6], we utilize the BERT-Tokenizer. BERT represents a cutting-edge transformer-based model designed to grasp contextual information and inter-word relationships. The tokenizer transforms input text into tokens comprehensible by BERT, enhancing the model's capacity for a more nuanced comprehension of the textual content.

By incorporating these feature extraction methodologies, the stock market prediction model is aimed to efficiently analyze and interpret textual information, thereby increasing its overall predictive performance.

3.3 Model

In our approach to forecast stock market trends through Natural Language Processing (NLP), we utilized various conventional statistical machine learning algorithms and a sophisticated BERT model. This section explains our choices in selecting and implementing models, detailing the underlying idea behind these decisions.

- **Statistical ML Algorithms:** We chose a set of traditional machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbor (KNN). As these algorithms act as great baseline models. Their simplicity enables straightforward implementation and interpretation, making them widely used in various financial domains. These algorithms formed the basis of our analysis, serving as a benchmark against which the performance of more advanced models could be assessed.
- **FinancialBERT Model:** To Effectively use the capabilities of transformer-based models, we chose to utilize the pre-trained financialBert model. Originally designed for financial texts, we fine-tuned this model for our specific stock market prediction task. While its initial purpose was to predict three labels for stock price decisions (hold, buy, or sell), we customized it to predict two classes as per our needs, specifically forecasting whether the closing value of the market rises or falls.

4 Experiments

4.1 Data

The dataset used in this project is a collection of top daily news headlines from the Reddit WorldNews channel, which spans from August 8, 2008, to July 1, 2016. It contains 1989 entries in total, where each row corresponds to a single day's news. It is organized into 27 columns, where the first column denotes the date, and the second column, named as ‘'Label', denotes the movement of the Dow Jones Industrial Average (DJI) on that day. It has a value of 1 if the index rose at the end of the day and it has a value of 0 if it fell.

The subsequent columns: ‘Top1' to 'Top25', contain the top 25 headlines of the day. These headlines capture a large range of global events and issues, which provides a diverse and rich textual dataset for analyzing the impact of news on financial markets. This dataset allows us to explore the relationship between news sentiment and stock market performance over a period that includes significant global events.

4.2 Hyperparameters

In configuring the hyperparameters for our financialBERT model, we approached the task with careful consideration, given the complexity of the BERT model and the constraints posed by the limited dataset size. We chose a setting of three epochs to strike a balance between effective pattern learning and mitigating the risk of overfitting. Setting the learning rate at 1e-6 (0.000001) was important for delicately adjusting model weights, especially given the complexity of the BERT architecture, ensuring stability during the training process. To assess generalization and prevent overfitting, we employed a data split of 80% for training and 20% for testing. This allocation was designed to allow the model to be exposed to a diverse set of examples during training while providing an unbiased evaluation of its predictive capabilities on the testing set. These hyperparameter decisions were carefully made to optimize the financialBERT model's performance within the constraints of a relatively small dataset and its inherent complexity.

4.3 Results

We evaluated the above mentioned machine learning models for their effectiveness in predicting stock market movements based on news headlines. The models included Logistic Regression, Support Vector Classifier, Random Forest, Gradient Boosting, K-Nearest Neighbors, and Decision Trees. We assessed each model's performance using accuracy, precision, recall, and mainly F1 scores. The results put forth some significant variations in the effectiveness of these models, which showed the complexity that is inherently present in text-based financial forecasting.

The application of BERT for sequence classification demonstrated the potential of advanced NLP techniques in handling the financial news better. The preprocessing of data was pivotal in the models' performance, which stressed the importance of effective data cleaning and preparation for NLP tasks. These findings provided valuable insights into the limitations and abilities of the different algorithms and models with regards to financial textual data.

Model	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.5275	0.5450	0.5361
Support Vector Classifier	0.5230	0.9147	0.6655
Random Forest	0.5321	0.7867	0.6348
K-NN	0.5357	0.5687	0.5517

BERT	**0.5789**	**0.9649**	**0.7236**

Table 1: Results of Experiments for all models

In comparing the results, BERT stands out with the highest precision and recall, leading to the best F1 score. This indicates its effectiveness in accurately identifying relevant market movements and its consistency in capturing a broad range of market signals. Even though Support Vector Classifier has lower precision, it shows high recall, which is shown in its F1 score. Which suggests it's more inclined to extract most market-related events with less precision. Logistic Regression and K-Nearest Neighbours display a more balanced result, but moderate performance across all the metrics. Random Forest had a better recall than Logistic Regression and K-Nearest Neighbours, which indicates its higher capability in identifying more true positives.

5 Conclusion and Future work

In this project, we effectively illustrated the feasibility of predicting stock market movements to a useful extent by utilizing a transformer-based model with news headlines from reddit as its input. Our results show that the transformer-based model, which is specifically employing financialBERT, displayed effective capabilities in extracting complex relationships within textual data for stock market prediction. Moreover, our approach resulted in a major enhancement in the classification of stock movements, which easily overshadows

the performance of traditional statistical machine learning models. This achievement highlights the effectiveness of transformer-based models in handling the nuanced nature of language and the information embedded in news headlines.

Looking forward, we anticipate further advancements in our predictive capabilities through the fine-tuning of hyperparameters. This refinement process is designed to achieve even superior performance and optimize the model's predictive accuracy, creating a more robust tool for stock market forecasting. Moreover, the integration of both textual and numerical data into deep learning models emerged as a potential avenue for enhancing predictive accuracy. This multimodal approach allows for a thorough analysis, containing not only the semantic information of news headlines but also numerical indicators, thereby providing a more holistic understanding of market trends.To ensure the generalizability of our model, future efforts should concentrate on accommodating diverse news headlines from various sources. This expansion aims to augment the model's adaptability and effectiveness in capturing market dynamics influenced by a broad range of news outlets.

In conclusion, our study illustrates the promising potential of transformer-based models in stock market prediction using NLP. As we continue to refine our methodologies and explore innovative strategies, the convergence of natural language processing and financial forecasting opens up new frontiers for developing robust, adaptive models in the constantly evolving landscape of stock market prediction.

Code Link: https://colab.research.google.com/drive/1OfWi5eejygxS_GCQgU-HtOOtj5z3FMhS?usp=sharing

References

1. Xing, F.Z., Cambria, E. & Welsch, R.E. Natural language based financial forecasting: a survey. *Artif Intell Rev* 50, 49–73 (2018). https://doi.org/10.1007/s10462-017-9588-9

2. S. Yıldırım, D. Jothimani, C. Kavaklıoğlu and A. Başar, "Classification of "Hot News" for Financial Forecast Using NLP Techniques," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 4719-4722, doi: 10.1109/BigData.2018.8621903.

3. Javed Awan, Mazhar and Mohd Rahim, Mohd Shafry and Nobanee, Haitham and Nobanee, Haitham and Munawar, Ashna and Yasin, Awais and Zain, Azlan Mohd, Social Media and Stock Market Prediction: A Big Data Approach (2021). M. J. Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al., "Social media and stock market prediction: a big data approach," Computers, Materials & Continua, vol. 67, no.2, pp. 2569–2583, 2021, Available at SSRN: https://ssrn.com/abstract=3827106

4. Johnson, Jaya. 2023. Machine Learning for Financial Market Forecasting. Master's thesis, Harvard University Division of Continuing Education.

5. M. Kesavan, J. Karthiraman, R. T. Ebenezer and S. Adhithyan, "Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 477-482, doi: 10.1109/ICICCS48265.2020.9121121.

6. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.