COMP4560 - Project Report

U7142680 - Srikanth Polisetty

Under the guidance of Dr Bernardo Periero Nunes,
Dr Artem Lenskiy

College of Engineering, Computing and Cybernetics

Australian National University

January 15, 2023

Acknowledgements

I would like to express my heartfelt gratitude to my professors for their invaluable guidance and support throughout the project. Their expertise and willingness to share their knowledge have been instrumental in helping me understand the intricacies of the research subjects and making this project a success.

I am particularly grateful for their guidance in providing me with important resources such as research papers and other literature that have been extremely helpful in understanding the subject matter and building the foundation for my project. Their constant encouragement and support have been a source of motivation and inspiration throughout the project.

I would also like to thank them for their patience and understanding while I navigated the research process. Their constructive feedback and suggestions have helped me to refine my ideas and improve the quality of my work.

Once again, I extend my sincere thanks to my professors for their invaluable contribution to the success of this project. Their support and guidance have been instrumental in helping me achieve my goals and I am truly grateful for their mentorship.

Contents

Abstract

The project aims to develop a web app titled Trade For Me (TFM), which utilizes sentiment analysis and time series analysis to predict stock market movements. The app will fetch tweets and train sentiment analysis models like LSTM, CNN+BiLSTM, and BiLSTM on a financial news dataset, to pick the best model based on accuracy and F1 score. The selected model will be used to get the sentiment of tweets of the required companies. Based on the sentiment score, the app will identify the most reliable Twitter user for each company and fetch his most recent tweets to inform the user whether to buy or sell shares of that company. If there are no recent tweets from the most reliable user about a company, the app will use time series analysis models like LSTM, RNN, CNN, BiLSTM, and Arima on the share prices to predict the future share prices of that company. The application is aimed to assist users in taking informed decisions based on the analysis of sentiments and trends of the market.

1 Introduction

News can provide important information about a company and its prospects for the future, which can influence investor opinion and demand for a stock. Positive news, such as strong financial performance, new partnerships, or supportive regulatory developments, may suggest that a firm is doing well and that its stock price is expected to rise. In contrast, bad news, such as inadequate financial performance, negative regulatory developments, or management changes, can suggest that a firm is encountering difficulties and that its stock price is likely to decline.

In the subject of behavioral economics, one of the hypotheses that is considered to be among the most well-known is the hypothesis that consumer sentiment has an effect on stock prices. According to this hypothesis, improved performance of the stock market can be accomplished by increasing the amount of investment made by those who are happy with their lives. Event-based stock movement is determined by the sentiment of the trader or investor as a result of the news or a collection of information they have read about particular stocks. If one engages in trading based on an event, the event-based trader's long-term return is significantly greater than the stock market index. (Makrehchi, Shah1, & Liao, 2013)

Numerous studies have demonstrated a significant impact on the stock market based on a particular event, and prior to the event, rumours and news tend to move the sentiment of stock traders in a particular direction. If a trader wishes to trade for a shorter period of time, it is easier to make a decision if that person is aware of the overall sentiment of traders. (Antonios Siganos, 2017). This study aims to leverage the sentiments of expert traders' tweets to generate trading signals for buying and selling stocks. This work contains the following:

- Generate effective trading signals for the end user in terms of BUY or

SELL, so that the user can make a decision faster based on their research and using our signals.

- Obtain valuable tweeter users based on backtesting of the tweets' sentiment signal and the actual price movement of the stock.
- Automation of tweets sentiment analysis of valuable users, to generate BUY or SELL signal
- Develop a full stack-based sentiment analysis-based system so that the user can easily use the developed system.

2 Background

Individuals have always understood the concept of emotion. In this sense, it signifies a specific concept or perspective. The term "sentiment analysis" refers to a variety of techniques from natural language processing, to text analysis, and text analytics that detects and extracts evaluative information from texts. (deHaaff, 2010).

2.1 Sentiment Analysis

In general, there are three categories of sentiment categorization methods: machine learning, lexicon-based, and hybrid. The Machine Learning (ML) approach is based on a combination of language characteristics and well-established ML techniques. The Lexicon-based methodology depends on a database of terms associated with specific emotions. The hybrid approach combines the two approaches and is widely applied. Sentiment lexicons are important to the majority of techniques. ML-based approaches to text categorization can be separated into supervised and unsupervised learning techniques. Multiple labeled training records are used by the supervised methods. In unsupervised learning, the dataset is unlabeled and researchers are unaware of which category a specific row data item belongs to. (Diana Maynard, 2011)

2.2 Sentiment Analysis for Stock Market Prediction

A lot of research has been made to give stock signal recommendations using natural language processing. Using the Twitter dataset (A. Pak, 2010), a classifier is developed for categorizing Tweets. Those categories are three types: positive, negative, and neutral. (W. Zhang, 2010), developed an easy market-neutral trading method utilizing trade signals derived from news and blog sentiment analysis. However, only the positive and negative sentiment dimensions are considered in the stock recommendation algorithm.

Twitter has professional traders' real-time tweets that contain a wealth of information about a particular company or market; the sentiment of these tweets may be useful for short-term trading. There are numerous accounts that provide incorrect information about a stock, which may result in a loss for the user who relies on the tweets' sentiment for trading. In this study, Twitter accounts that provide trading tips or stock-related data will be evaluated; the sentiment of such tweets will be extracted and based on the sentiment and the actual price movement of the stocks, the valuable Twitter users will be identified and used in generating buy or sell signals. This research also studies different type of neural networks and their performance for sentiment analysis.

3 Methodology

The purpose of this study is to evaluate the ability of natural language processing and sentiment analysis to produce buy or sell signals from the data of skilled traders.

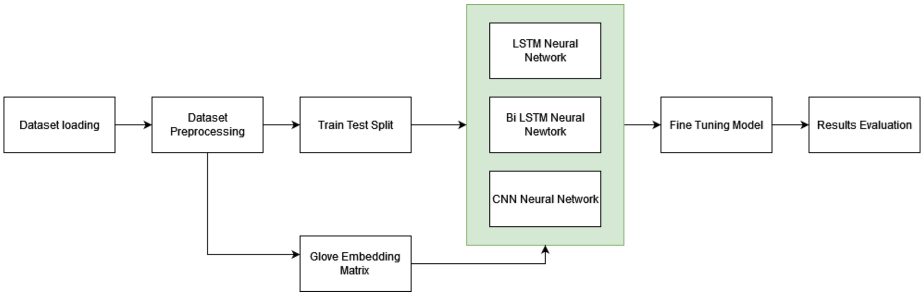


Figure 1: Sentiment analysis workflow

3.1 Workflow

The content of the tweets can be utilized by using natural language processing to identify various features from the text data. Text categorization is a process of allocating a category to text data based on features of the text. There are several disciplines that can benefit from the ability to classify texts based on particular attributes. Social networking, e-commerce, healthcare, the law, and marketing are among the examples. Text classification serves a variety of functions and has several applications in numerous fields, but the fundamental abstract problem is common.

3.2 Dataset

The data collection at Financial Phrase Bank includes the views of individual investors evaluating the content of financial news items. The dataset has two columns called "Sentiment" and "News Headline." The sentiment may be positive, negative, or neutral. This dataset presents the sentiments of retail investors with regard to various headlines in the financial news (Malo, 2004).

3.3 Dataset Loading and Preprocessing

The text data comprises English letters and words, which must be converted into a format using a vector representation so that deep learning algorithms may utilize the data and perform further computations. The preceding text data was inflated with unnecessary stop words, symbols, and hypertext links. Due to the fact that these data do not contribute to the classification of spam content, they have been eliminated to maximize the data's utility. This step involves cleaning the text data,

1. Removal of symbols
2. The textual data contains symbols such as @, ,

   For example:

   Input: This is a# sunny day

   Output: This is a sunny day.
3. Tokenizing the sentence Tokenization is the process of separating words and phrases from a text corpus. Sentences are tokens inside paragraphs, while words are tokens within sentences. The sentence is split into an array of words for further processing. For example, This is a sunny day. This sentence after tokenization will result in a list of words like [This, is, a, sunny, day]
4. Stop word removal Stop words are words that have been used so frequently that they have lost some of their original meaning. Stop words include the phrases ”of,” ”are,” ”the,” ”it,” and ”is.” In applications where keywords are more important than generic phrases, such as document search engines and document classification, it may be advantageous to exclude stop words.

   For example:

   Input: [This, is, a, sunny, day]

   Output: [sunny, day]
5. Lemmatization and stemming Stemming is the process of removing the final few letters of a word to get a shortened version. After stemming, the terms History and Historical will be shortened to histori. The downside of stemming is the loss of the original meaning of the term. Lemmatization addresses the drawback of stemming by reducing words to a shorter version that retains their meaning. After lemmatization, parrots will be reduced to parrot.

3.4 Glove: Global Vectors for Word Representation

Representation of words using global vectors (for ex.: Stanford's Word2Vec) is an unsupervised learning technique for generating word embedding from the global word-word co-occurrence matrix of a corpus. In vector space, the resulting embedding reveals the word's insightful linear substructures.

Quantitatively representing the precision required to distinguish between the genders, a model must give several numbers to the word pair "man" and "woman." The vector difference between the two word vectors is a logical and obvious choice for a wider array of discriminatory numbers. The objective of the design of GloVe was to ensure that these vector differences accurately represented the intended meaning of a pair of words. Example of Linear structures of Glove:

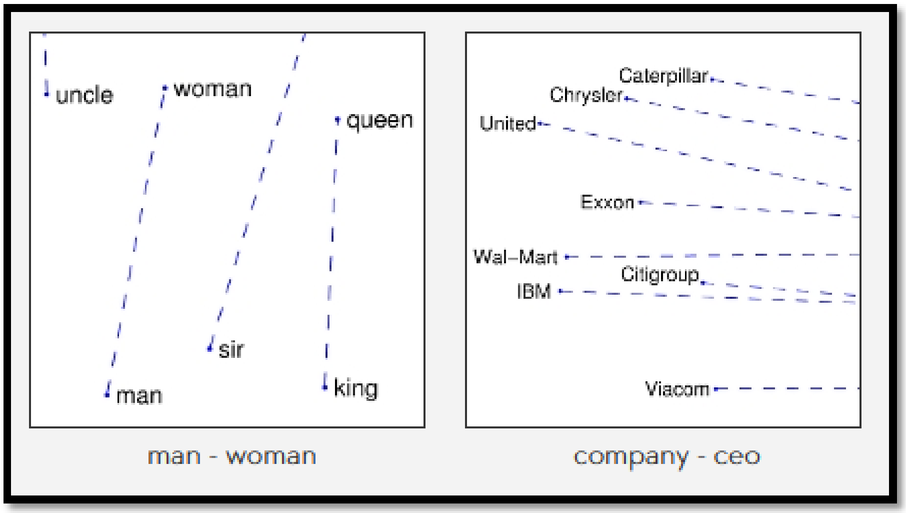


Figure 2: Vector differences and relation of Glove representation

Word embedding can assist in finding the relationship between two words with comparable meanings if the Glove approach vector format of the words has been produced with meaningful representation.

3.5 Train Test Split

Train, test, and split are conducted after preprocessing the data. The processed data and labels are then separated into training and testing sets. The train test brake method is used to evaluate the performance of machine learning algorithms required to generate predictions on data not utilized to train the model. The approach is simple and quick to execute, and its findings allow us to compare the predictive modeling output of machine learning algorithms. The process involves collecting and separating a dataset into two groupings. The first subset, also known as the training dataset, is examined in order to fit the model. Alternately, the input component of the dataset is delivered to the algorithm, and then projections are shown and compared to expected values. The second group does not participate in model training. This second collection of data is known as the test dataset.

3.6 Neural Network Model

TLong short-term memory (LSTM), Convolutional Neural Network (CNN)+LSTM, and BiLSTM (Bidirectional LSTM) are the preferred models for performing comparative analysis and determining the best neural network model. A CNN LSTM architecture is produced by combining LSTMs for sequence prediction with CNN layers for feature extraction on input data. The neural network is composed of many layers. The first layer is the embedding layer, which receives as input the embedding matrix. Next, the convolution 1D layer is combined with the Max Pooling layer to produce the convolution layer, which is then linked to the LSTM layer, therefore completing the CNN+LSTM neural network. The last layer is the output layer, and its activation function is softmax. Categorical cross entropy with the Adam optimizer is used as the loss.

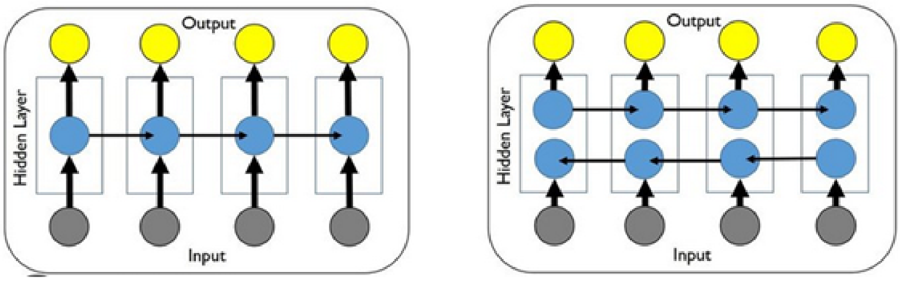


Figure 3: LSTM and Bi LSTM Architectures (Mohan and Gaitonde, 2018)

This diagram illustrates the fundamental construction of an LSTM Neural Network (NN) and a BiLSTM NN. In the hidden layer of the LSTM, the input to one cell is dictated by the computation performed by the cell processing data in the prior time instant. Due to its capacity to explicitly account for memory in a sequence, LSTM is a good sequential modeling technique. (Mohan & Gaitonde, 2018)

A BiLSTM is a sequencing processor model that uses two LSTMs, one to process data in the forward direction and another to process input in the reverse direction. Increased data processing in both directions enhances the neural network's context. When the input is flowing both ways, it is BiLSTM which is distinct from a standard LSTM. The Standard LSTM allows one to only choose one way for input to go, which is either backward or forwards. (Jang, et al., 2020)

3.7 Fine Tuning of the model

The neural network is optimized by performing hyperparameter tuning to find the best-performing model. The parameters fine-tuned are batch size, epochs, optimizers, learning rate for the optimizers, activation functions, and dropout rate. A different combination of the parameters was used to find out the best-performing fine-tuned model. The fine-tuned model is further trained on the training set and tested on the testing set to find the evaluation metrics of the model.

3.7.1 Evaluation metrics

1. Accuracy

   In machine learning classification models, the accuracy of a model is quantified as the proportion of accurate classifications to the total number of positive and negative cases. In other words, accuracy indicates the proportion of instances out of 100 in which our machine learning model properly anticipated the outcome.

2. F1 Score

   What proportion of positively predicted events was precise? In terms of precision and recall, the F1 score is a balanced harmonic mean, with a maximum value of 1.0 and a minimum rating of 0.0. Because F1 scores include memory and precision in their computation, they are superior to accuracy assessments.

3.8 GUI for users

3.8.1 Login Screen

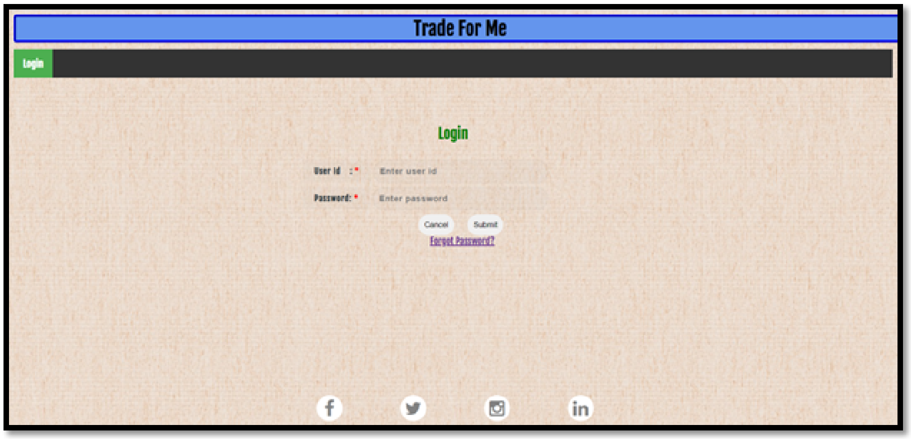
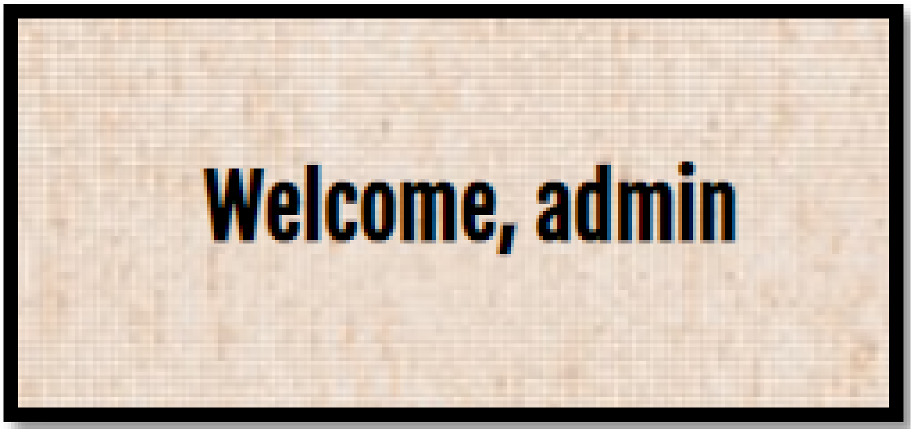


Figure 4: Login Page of “Trade for Me” website

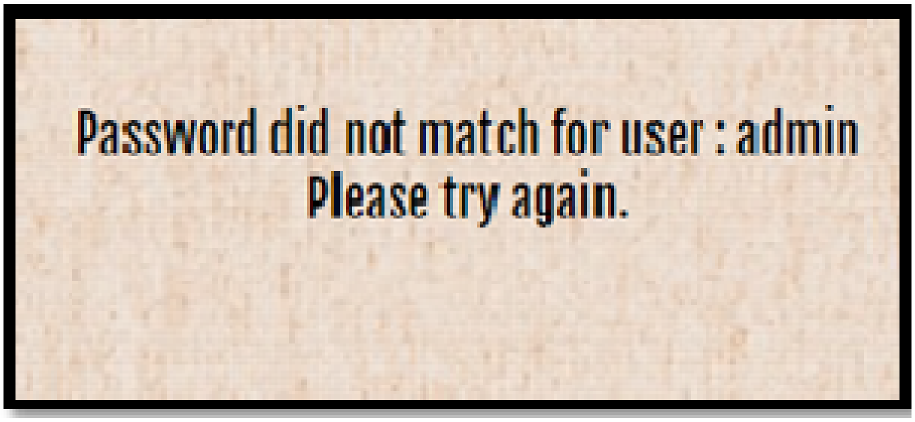
Depending on the login details, the user will be taken to the admin screen or to the customer screen. Upon entering the correct login details, a welcome screen will be displayed, if the details are correct.



Welcome, admin

Figure 5: Welcome message on successful login

Or an error page alerting about incorrect login details will be displayed.



Password did not match for user : admin
Please try again.

Figure 6: Error page on entering incorrect login details

If a user doesn't remember their password, a password reset email will be sent to their registered email address

3.8.2 Customer Login Management

If the logged-in user is Admin, then Customer Login Management, Get Reliable Tweeters, etc., options shall be shown. Customer Login Management shall allow the admin to add and delete customer IDs.

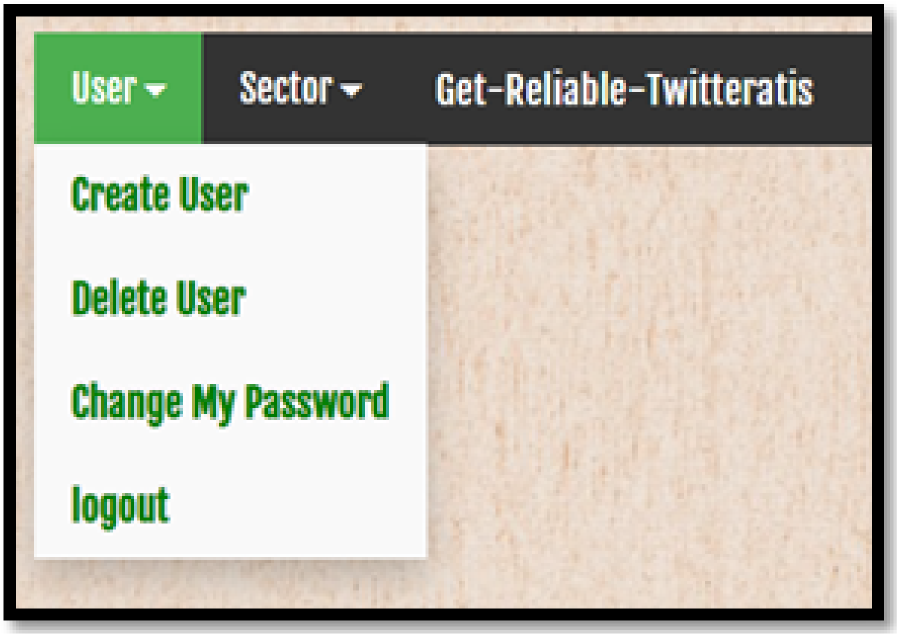


Figure 7: Options available for admin

For all the users created by the Admin user, user_type will be stored in the database (DB) as “Customer”.

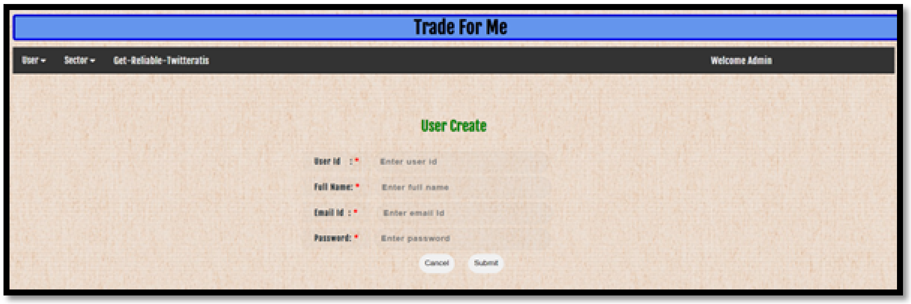


Figure 8: Create User Webpage

Admin user ID can’t be deleted. Customers will not be able to delete any user ID.

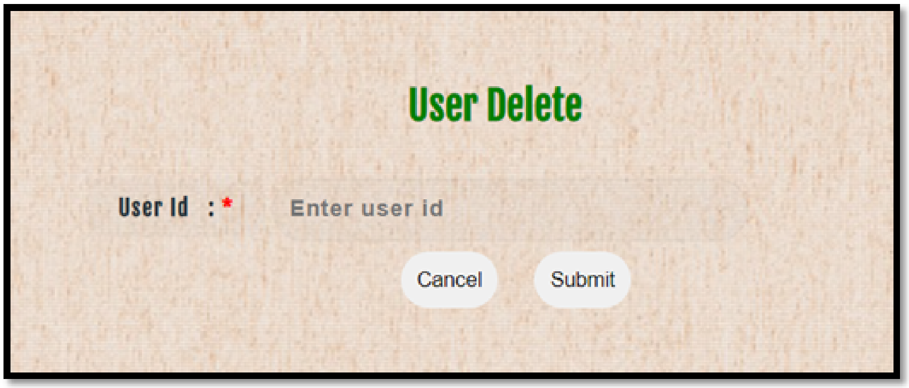
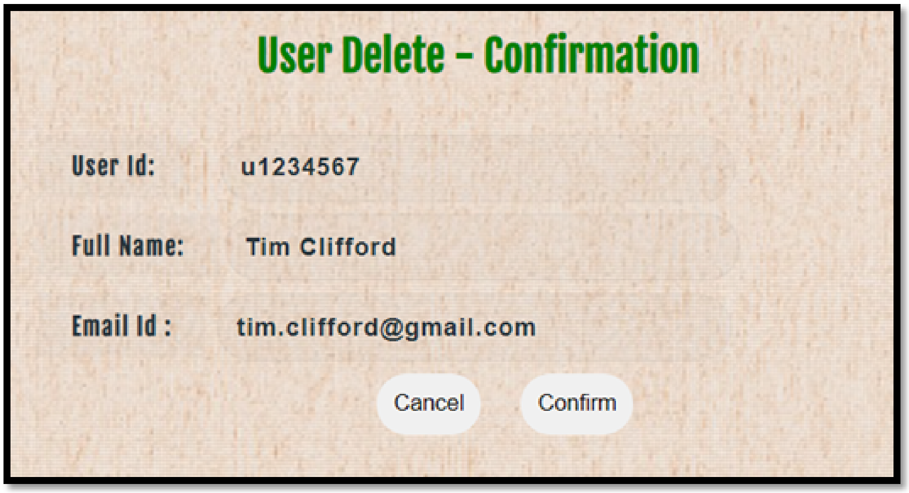


Figure 9: User Delete Webpage

Upon clicking the submit button in the User Delete screen, user-related details shall be shown to confirm the deletion of the selected User_ID.



User Id: u1234567

Full Name: Tim Clifford

Email Id : tim.clifford@gmail.com

Figure 10: User Delete Confirmation Webpage

3.8.3 Sector Management

If the logged-in user is Admin, then the Sector Management menu item shall be shown.

Sector Management shall allow admin users to add and delete Sectors.

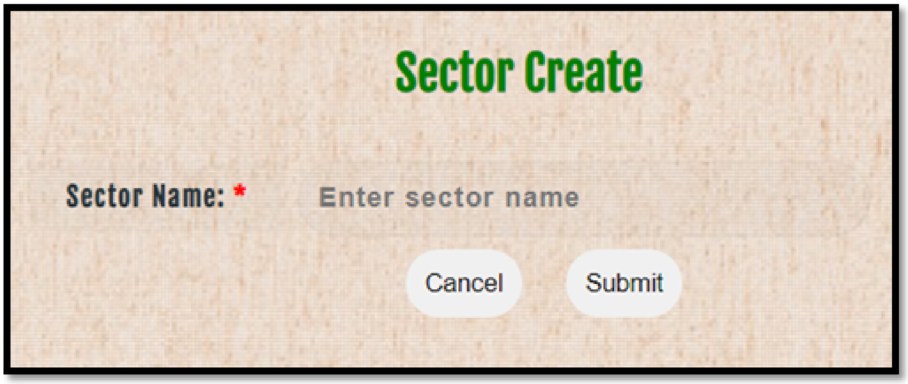
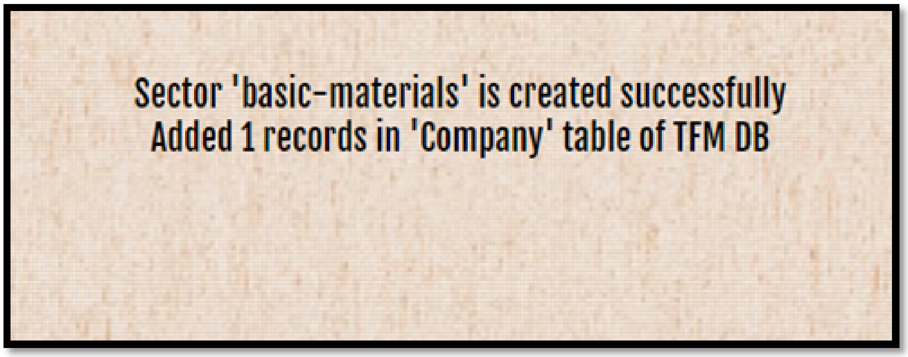


Figure 11: Sector Create Webpage

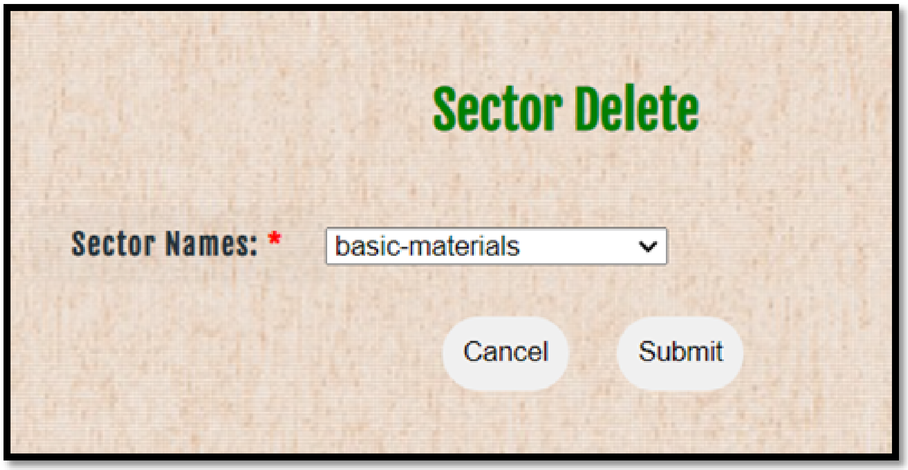
After adding a sector to the TFM database, corresponding company names shall be obtained from the stock market database and those company details shall be stored in the "company" Table in DB.



Sector 'basic-materials' is created successfully
Added 1 records in 'Company' table of TFM DB

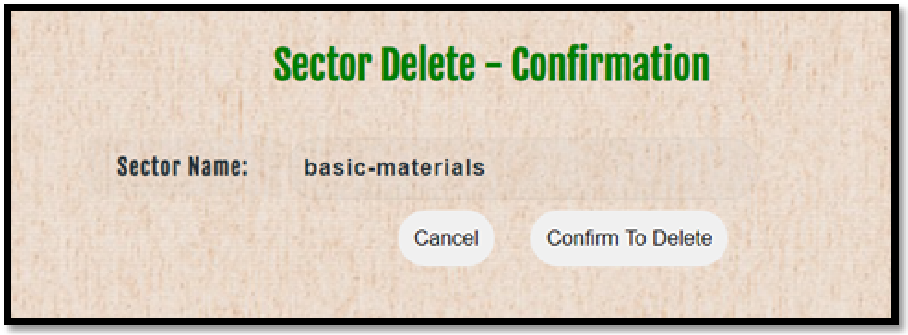
Figure 12: Create sector webpage which adds companies corresponding to sector into DB

After selecting a sector, upon clicking the submit button in the Sector Delete screen, a confirmation delete screen shall be shown.



Sector Names: * basic-materials

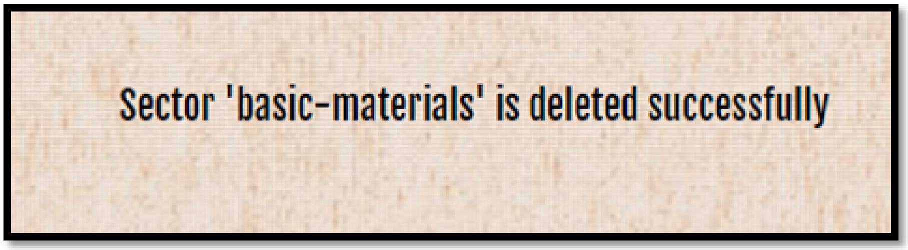
Figure 13: Sector Delete Webpage



Sector Name: basic-materials

Figure 14: Sector Delete Confirmation Webpage

Upon confirmation, the selected sector shall be deleted.

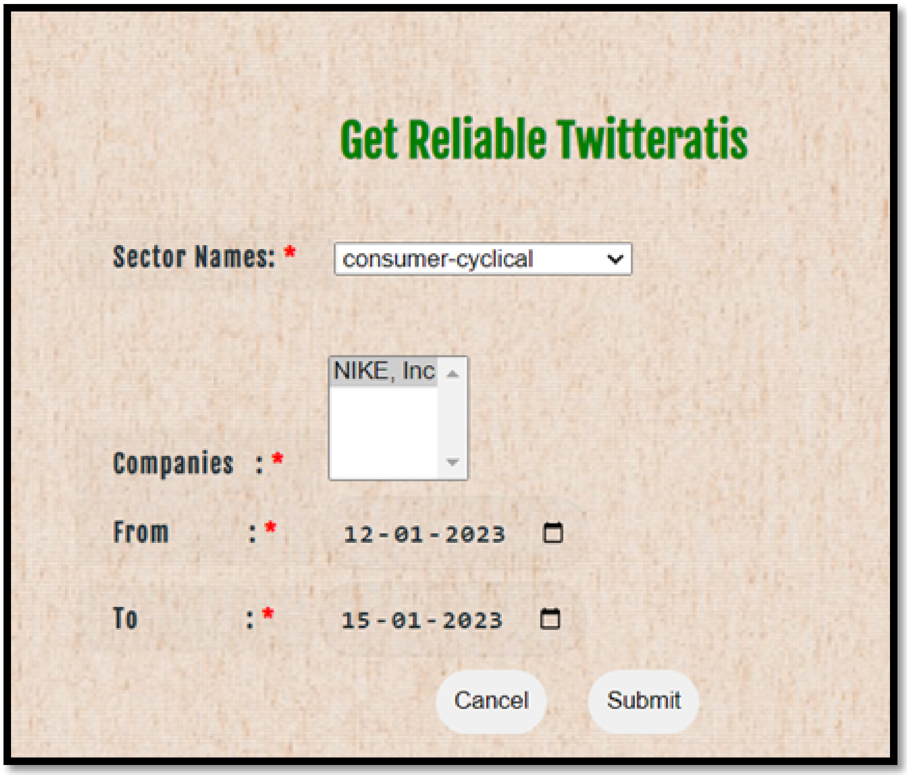


Sector 'basic-materials' is deleted successfully

Figure 15: Sector Delete Successful Webpage

3.8.4 Get Reliable Tweets

Get Reliable Tweeters screen is visible to the admin only.



Get Reliable Twitteratis

Sector Names: * consumer-cyclical

Companies : * NIKE, Inc

From : * 12-01-2023

To : * 15-01-2023

Figure 16: Get Reliable Twitteratis Webpage

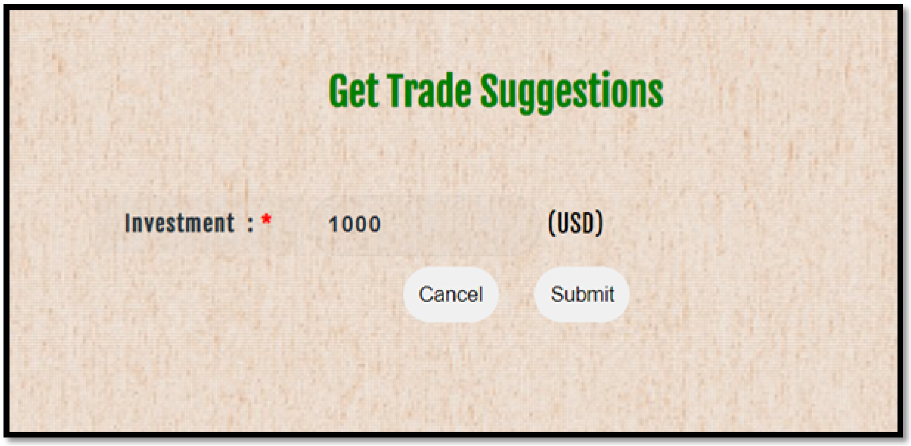
This screen allows the Administrator to initialize the database with reliable tweeters for a specified sector and selected companies. Admin can specify the duration for which the tweets need to be downloaded for processing and correlating against the market stock price variations.

When only one sector name is selected, the corresponding list of companies shall be fetched and shown in the next field. When submit is pressed, for all specified companies the reliable tweeter names shall be obtained and stored in the database.

Get Reliable Tweeters shall fetch the tweets done by various users on Twitter and correlate them with real variations in stock prices. Tweeter names with correlation factors shall be stored for each company as “Reliable Tweeters”

3.8.5 Get Trade Suggestions

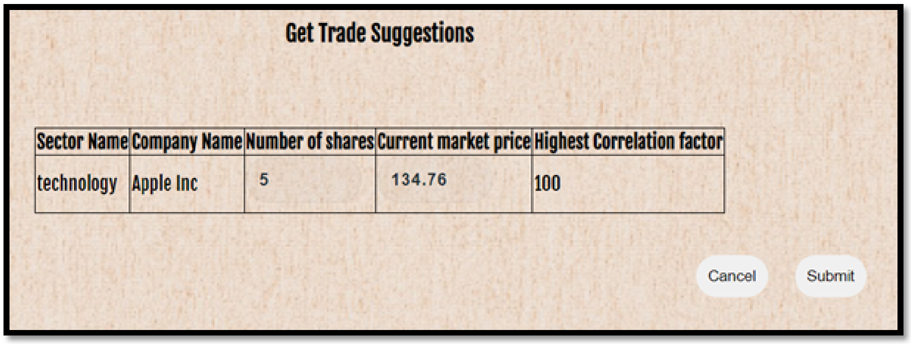
This screen is available for customers only. If the customer is new to the stock market, (s)he shall specify only the amount to be invested instead of specifying sector(s) and company name(s). Once the customer approves the suggested stocks, the BUY action will be executed.



Get Trade Suggestions

Investment : * 1000 (USD)

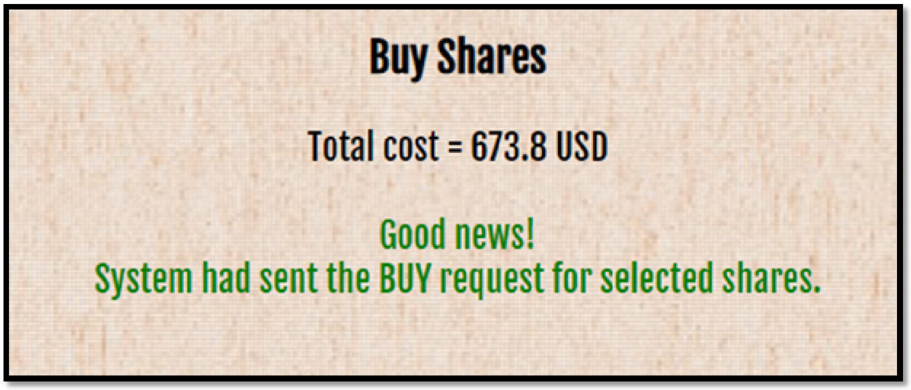
Figure 17.i: Enter Investment amount on Get Trade Suggestions Webpage



Get Trade Suggestions

Sector Name	Company Name	Number of shares	Current market price	Highest Correlation factor
technology	Apple Inc	5	134.76	100

Figure 17.ii: Enter numbers of shares to buy on Get Trade Suggestions Webpage



Buy Shares

Total cost = 673.8 USD

Good news!
System had sent the BUY request for selected shares.

Figure 18: Buy Shares Webpage

4 Results

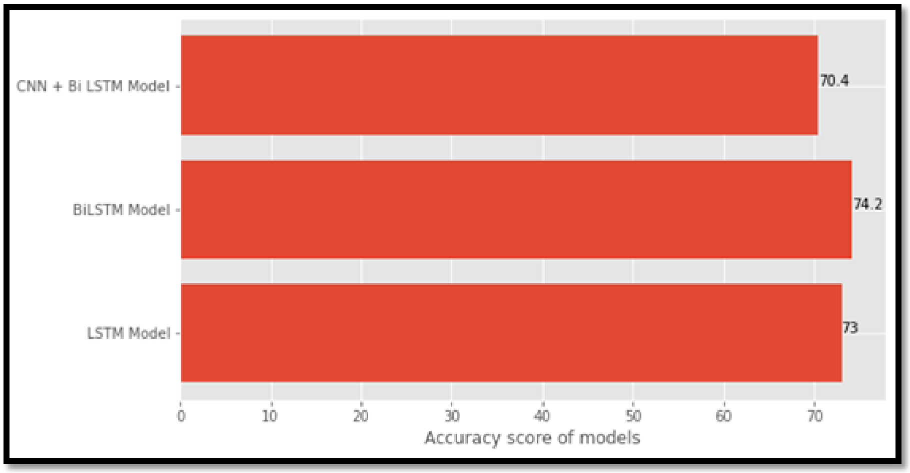


Figure 19: Accuracy score of various sentiment models

In comparison to the LSTM model and the CNN + LSTM model, the BiLSTM model that has been fine-tuned exhibits the greatest results, with an accuracy score of 74.2. The figure shows the accuracy score of the models that were utilized. The enhanced performance of the model was the direct outcome of the bidirectional flow of data that took place within the BiLSTM NN.

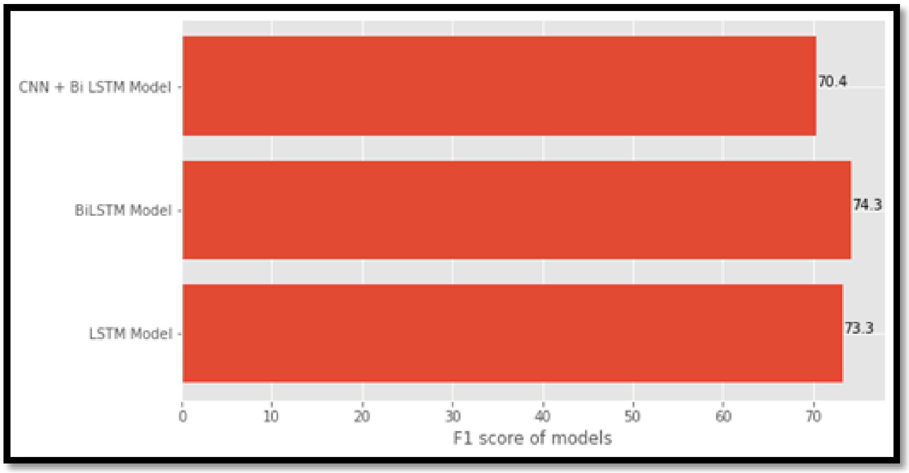
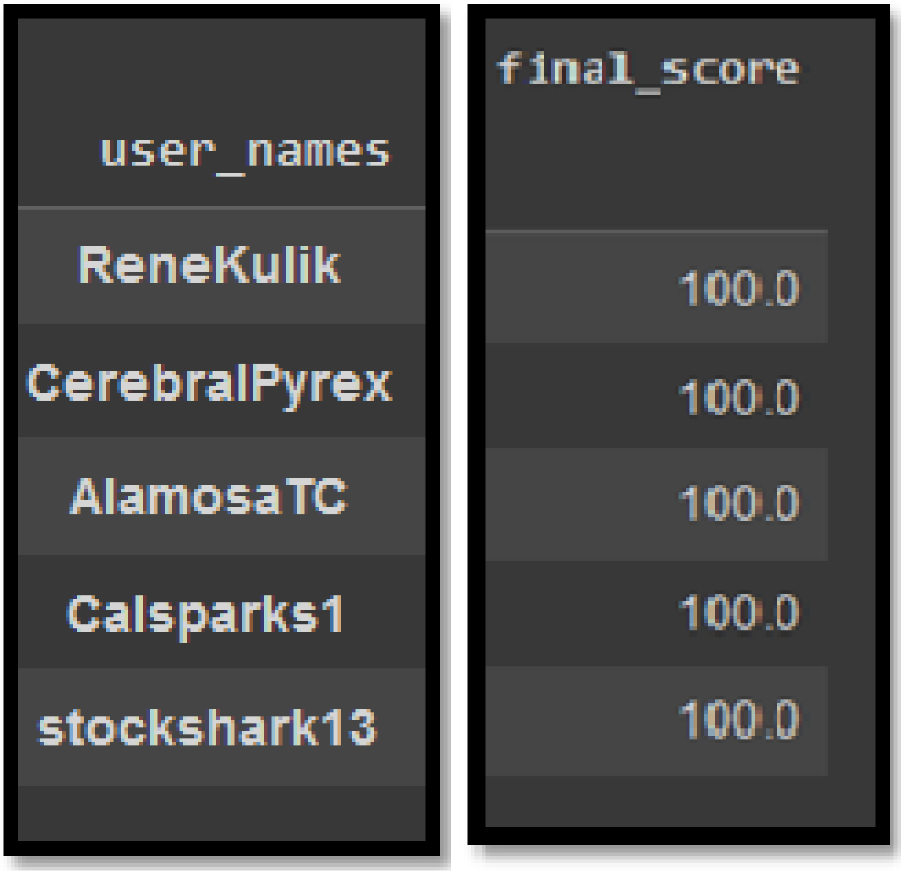


Figure 20: F1 Score of various sentiment analysis models

The f1 scores that the models have acquired are displayed in the graph that is situated above. The BiLSTM model is the one that has achieved the highest f1 score for the sentiment analysis application. The BiLSTM-based model is going to be utilized for the final application of the signals that are based on sentiment analysis.



user_names	final_score
ReneKulik	100.0
CerebralPyrex	100.0
AlamosaTC	100.0
Calsparks1	100.0
stockshark13	100.0

Figure 21: Filtered usernames using backtesting

These are the user accounts chosen based on the rate of accurate predictions made by the user's tweets. The score is computed using a specific formula:

$$FinalScore = \frac{Countofcorrectpredictions}{Totalpredictions}. \tag{1}$$

5 Conclusions

Machine learning and deep learning technology have been widely adopted to assist with tasks such as image recognition, speech recognition, item identification, and providing relevant search outcomes as well as widely used in financial markets. In this study, the sentiment analysis of the greatest experienced traders on Twitter is used to formulate a system that enables new users to quickly and simply buy or sell stocks. In this study, algorithms based on deep learning are combined with sentiment classification to produce a well-defined system for short-term trading based on the sentiment analysis of tweets. Any user utilizing the system's buy and sell signals will attain an accuracy of 74% based on the achieved outcome.

5.1 Limitations

This technique has been well tested, but if a Twitter user deemed as reliable makes an erroneous or misleading tweet, the end user would be provided with incorrect information resulting in the loss of the user. Although, it is important noting that it is highly improbable that a reliable user makes an erroneous tweet as the algorithm that is used to identify reliable users is foolproof.

5.2 Future Work

To improve the precision of the strategy, more complex algorithms such as Restnet and VGG16 can be employed. Text classification is a possible area for improvement. Here, the strategy may result in the temporary loss of a user. Hence, there is a need for improvement. When there are no tweets about a company from users with a high score in recent times, the previous prices of shares of that company are used to make predictions about future share prices. Various models like LSTM, RNN, CNN, BiLSTM, and Arima are used for time series forecasting. A bar plot comparing the root mean squared error of those models is as follows.

6 References

A. Pak, P. P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

Antonios Siganos, E. V.-N. (2017,). Divergence of sentiment and stock market trading. Journal of Banking & Finance, Pages 130-141.

deHaaff, M. (2010). Sentiment Analysis, Hard but Worth It! Customerthink.
Diana Maynard, A. F. (2011). Automatic detection of political opinions in tweets. Proceedings of the 8th international conference on the semantic web, 88–99.

Jang, B., Kim, M., Harerimana, G., Kang, S.-u., & Kim, J. W. (2020). Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. Applied Sciences, 10(17), 5841.

Makrehchi, M., Shah1, S., & Liao, W. (2013). Stock Prediction Using Event-based Sentiment Analysis. IEEE ACM Internation Conferences, 1.

Malo, P. S. (2004). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the American Society for Information Science and Technology.

Mohan, A., & Gaitonde, D. (2018). A Deep Learning based Approach to Reduced Order Modeling for Turbulent Flow Control using LSTM Neural Networks. . Research Gate.

W. Zhang, S. S. (2010). Trading strategies to exploit blog and news sentiment. Proceedings of the International Conference on Weblogs and Social Media.