International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

AI ML BASED NEWS CLASSIFIER FOR SHARE MARKET

Chetan Patil
G. H. R. I. E. T, Pune,
B. Tech in Computer Engineering.
chetanpatil112000@gmail.com

Sankalp Yedave
G. H. R. I. E. T, Pune,
B. Tech in Computer Engineering
sankalpyedave17@gmail.com

Vineet Patil
G. H. R. I. E. T, Pune,
B. Tech in Computer Engineering.
vineetnitinpatil@gmail.com

Bhimashankar Pujari
G. H. R. I. E. T, Pune,
B. Tech in Computer Engineering
bhimashankarp99@gmail.com

ABSTRACT

The stock exchange is a public market where you can buy and sell shares of public companies. There are many techniques us ed to predict a company's stock price, one of which is to use th e news to make the prediction. According to the polarity of the text, news can be classified as good, moderate or bad.

1 INTRODUCTION

In this project, we need to extract news data from websites like money check, FT etc. Gather information about a company, classify the same news as good news or neutral news or bad news, based on which the system predicts whether the stock of this company will rise, fall, or remain stable.

The process of determining the future value of a company's stock is called stock market forecasting. We'll use an AI and ML technique called sentiment analysis to categorize news into three categories, good, bad, or neutral. After the classification, the rise and fall of the stock price is directly related to the classification.

Research shows that stock price movements correlate with public perception of companies. When user sentiment was taken into account to make predictions, the accuracy of the predictive model increased by 20%. For example, a company's perceptions in the media, industry reports, social media commentary, or investor opinions can provide good insight into how stock prices are moving.

Companies can use AI and ML techniques, such as sentiment analysis, to understand how market prices will move over time and take action to buy, sell or hold their stocks accordingly.

For the model to work properly and provide accurate results, pre-processing such as data labeling is required.

Text, images, or speech from the algorithm has no meaning unless they are tagged with meaningful labels and sorted into different groups. Therefore, labeled data is one of the building blocks of a machine learning model, which learns these labels so that further classification can be performed.

2 SYSTEM DESIGN

2.1 SYSTEM ARCHITECTURE

System Architecture: The design of movie review sentiment analysis system adopts Unified Modeling Language (UML) as the modeling language and the concept of object-oriented programming.

A system architecture describes the structure of the system and its behavioral components. The proposed stock prediction system consists of a classifier, a machine learning predictor, a preprocessor component, and historical stock price data. The

only users of the system will be the stockbrokers of the respective companies. Stock price prediction starts with a user or stockbroker entering a keyword to retrieve tweets related to a particular company's stock price. After the tweets are retrieved, the Twitter Search API matches the keywords and sends them to a preprocessor for cleaning.

The processed tweets are transformed into document term matrices suitable for machine learning algorithms and tweet classifiers. Based on the machine learning algorithm used in the study; tweets are classified as positive (1), negative (-1) or neutral (0).

Combines tweets with historical stock prices in the database, based on classification, to obtain stock price predictions for the current time period. The results are then submitted to stockbrokers for analysis or decision-making.

2.1.1 Algorithm: SVM performance

	Precision	Recall	F-score	Support
Positive (1)	0.84	0.96	0.90	1349
Negative (-1)	0.78	0.46	0.58	488
Total average	0.78	0.46	0.58	488

2.1.2 Naïve bayes performance

Feature	Accuracy	Precision	Recall	F-score
Unigram	0.768	0.81	0.77	0.69
Bigram	0.762	0.82	0.76	0.67
Trigram	0.759	0.82	0.76	0.66

2.1.3 Random forest performance

Feature	Accuracy	Precision	Recall	F-score
Unigram	0.802	0.80	0.79	0.79
Bigram	0.798	0.79	0.79	0.78
Trigram	0.783	0.78	0.78	0.77

2.2 Data Flow Diagram

2.2.1 DFD level 0 diagram

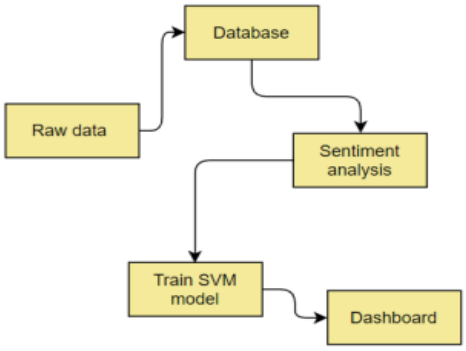

Figure 1 DFD level 0

2.2.2 DFD level 1 diagram

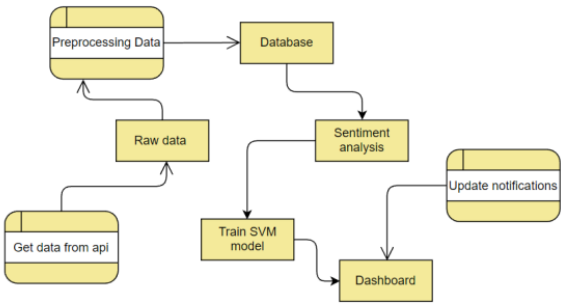

Figure 2; DFD level 1

2.2.3 DFD level 2 diagram

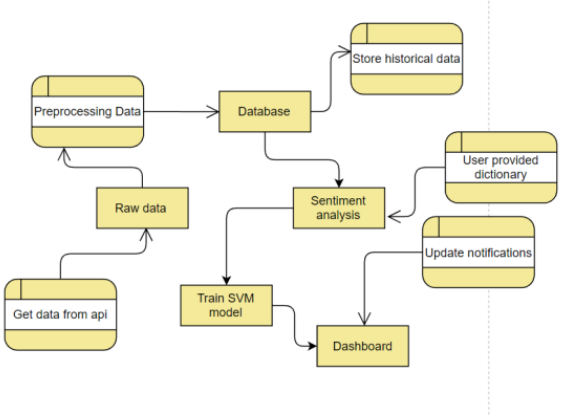

Figure 3; DFD level 2

3 OTHER SPECIFICATIONS

3.1 ADVANTAGES

Linear SVM classifiers dominate accuracy results compared to Naive, Perceptron, and Random Forest arrays.

Different strategies can be combined with the SVM model to further improve the accuracy rate.

Strategies that can be used to further improve the accuracy rate are: Hold and wait, i.e. buy back shares according to forecast Momentum, buy more stocks is a forecast that indicates the upward trend of stocks. They

can make real-time predictions, using user-created word dictionaries for more accurate sentiment analysis. Accuracy scores sometimes reach 70% to 80%, and some algorithms predict quickly.

3.2 Applications

The program is limited to shares of companies listed on the NSE. Additionally, companies must have been in business for at least five years to ensure data consistency. The five languages to use in the sentiment analysis process are all English. The use of slang in this context will disregard English and the vernacular.

Currently in the Indian context brokers using trend pattern-based approaches may not be effective

The methods are currently not predictive, they are based on supply and demand.

Helps provide accurate results in stock price predictions by including news and people's sentiment (in this case considered key external factors not presented in a quantifiable format)

3.3 System implementation plan

The implementation starts with the preprocessing of the corpus, followed by the training of the model.

The sentiment value is then correlated with the stock price for that day to build a forecast model for the next 30 days.

Tested several experiments with different algorithms with different characteristics to choose the best model.

Pre-processing: The resulting text data is in an unstructured data format and cannot be used to build machine learning models. Data contains tweet ids, create timestamp and text fields on tweet ids. Textual data contains more noisy words and symbols which do not help classification. Text data contains numbers, spaces, tabs, punctuation marks, stop words,

Labelling: after preprocessing, we need tags as training and test data sets. Labels are considered the most important in model development.

Each tweet in this dataset is labeled 1 (if positive) or -1 (if negative) or 0 (if neutral). Categorize these tweets. This process is done manually.

4 SOFTWARE REQUIREMENTS

ASSUMPTIONS & DEPENDENCIES

We fetch data from Yahoo Finance using pandas. Since our target values are closing values, the created target data frame contains only the closing column.

This project uses the stock values of Apple, Google and Microsoft. Their market value is public.

training and prediction is performed using a long-short-term memory (LSTM) architecture.

The LSTM architecture is a subset of recurrent neural networks, commonly used in the field of deep learning. LSTMs have feedback connections, making them useful for processing the full range of data. The data must be processed and normalized before being transmitted to the LSTM.

4.2 System Requirements

Database requirements: Data is provided in CSV, MySql database can also be used

Software requirements: Python, Linux, Anaconda, Python libraries numpy, pandas, matplotlib

Hardware requirements: 2.4 GHZ, 1.5 GB, HDD for installation of python 512 MB memory Bluetooth

5 ACKNOWLEDGMENTS

I would like to express my profound gratitude to Prof. Rachna Sable, HOD Computer science department, and Prof. Madhavi Netke (Guide) of GH Raisoni institute of engineering and technology for their contributions to the completion of our project titled AI ML Based news classifier for share market.

I would like to express my special thanks to our mentor Prof. Madhavi Netke for her time and efforts she provided throughout the year. Your useful advice and suggestions were helpful to us during the project's completion. In this aspect, we are eternally grateful to you.

I would like to acknowledge that this project was completed entirely by us and not by someone else.

REFERENCES

[1] Chun-Ming lai, Mei-Hua chen, Endah kristiani, Vinod kumar verma, Chao-tung yang "Fake news classification based on content level features" (2022).

[2] Faten alzazah, Xiaochun Cheng, Xiaohong gao "Predict market movements based on the sentiment of financial video news sites" (2022)

[3] Gianluca Anese, Marco corazza, Michele costola, Loriana pelizzon "Impact of public news sentiment on stock market index return and volatility" (2021)

[4] Chetan Gondaliya. "Sentiment analysis and prediction of Indian stock market amid covid-19 pandemic" (2020).

[5] Deepak kumar, Prakash kumar sarangi "A systematic review of stock market prediction using machine learning and statistical techniques" (2020)

[6] Smita Deshmukh, Suyash Saxena, Pooja devrukhkar, Sagar sanil "AI and ML based news classifier for share market" (2020).