Sentiment Analysis of Unstructured Data Using Spark for Predicting Stock Market Price Movement

Miss Dhara N. Darji, Satyen M. Parikh, and Hiral R. Patel

Abstract In this digital era, social media generates a large quantity of online financial data, which includes a substantial amount of investor sentiment. On the other hand, only technical and fundamental indicators are no longer adequate to forecast the stock price movement. The investors' sentiments on social media likes, tweets on twitter, comments and post on Facebook as well as other online financial information like online news, google trend, and forum discussion are also affecting the stock price movement. In particular, researchers have gained a lot of interest for analyzing the financial tweets on Twitter and online financial news to study public sentiments. This would be extremely helpful to develop an efficient solution for automating the sentiment analysis of such vast quantities of online financial texts. Henceforth, the proposed sentiment analysis model aims to predict the stock price movement based on the unstructured data like financial tweets on Twitter and news data used, and this research work also introduces Spark NLP-based text preprocessing pipeline to remove noise data and extract the features using the TFDIF by organizing the text in structured format. For sentiment analysis, two library Textblob and Vader are used. Further, the performance comparison has been carried out. The main aim of the proposed sentiment analysis model is to understand the perspective of the writer from a piece of text whether it is positive, negative, or neutral. In an extensive manner, news and tweets about a security will certainly inspire individuals to invest in that company's stocks, and as a result, the company's stock price will increase.

Keywords Spark NLP pipeline · Sentiment analysis · Data preprocessing · Stock price movement · TFIDF · Textblob · Vader · Logistic regression · Naïve Bayes · Random forest

M. D. N. Darji (✉) · H. R. Patel
DCS, Ganpat University, Mehsana, India
e-mail: dnd01@ganpatuniversity.ac.in

H. R. Patel
e-mail: hrp02@ganpatuniversity.ac.in

S. M. Parikh
FCA Ganpat University, Mehsana, India
e-mail: satyen.parikh@ganpatuniversity.ac.in

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_39

1 Introduction

Text mining is the process of autonomously extracting unique, non-trivial information from unstructured text sources by combining data mining, machine learning (NLP), information retrieval, and knowledge management approaches. Popular text mining tasks include classification of documents, summarization, clustering of similar documents, extraction of concepts, and sentiment analysis. Recently, text mining has leveraged a large variety of applications. To carry out the proposed experimental study, Spark NLP has been utilized.

As social media grow more popular and reach a wider range of users, the data available on these sites gradually represents the real life and the market [1] Since this data is available in an unstructured manner, it is highly required to organize it and utilize the data to infer about future relationships between markets and opinions shared in the network [2]. As a data source, the Twitter microblogging site and financial news are used for sentiment analysis and big data platform Apache Spark is used for text preprocessing and identify the correlation between stock and social media data [3].

The natural language processing (NLP) is considered as a key component in several data science systems that require an understanding about a text. The popular use cases are question answering, language modeling, paraphrasing, and sentiment analysis. In the broader field of NLP, there are several more libraries, but here, we emphasized on general-purpose libraries and not on the ones that cater to particular use cases. The only Spark NLP is a single unified solution to include all the NLP and all-in-one solution to ease the burden of preprocessing text and link the dots between multiple phases to solve the NLP-related challenges in data science. Spark NLP is developed on top of Apache Spark, and Spark ML is an open-source natural language processing library, which covers several popular NLP tasks, including tokenization, speech tagging, stop-word removal, lemmatization and stemming, sentiment analysis, text classification, spell checking, named entity recognition, and more. The core components of the Spark NLP are annotators, pipelines, transformers, and pre-trained models.

- Term frequency–inverse document frequency (TF-IDF);
- Spark's machine learning (ML) library (Spark MLlib);
- Spark's natural language processing (Spark NLP);
- Logistic regression (LR);
- Random forests (RF);
- Naïve Bayes (NB).

2 Related Work

Derakhshan et al. [4] discuss about the growth of social media sites, which have provided space for many individuals to share their views. This research work discusses about the most sensitive field in the world is financial market, where people can share their opinion, and it changes the trend of the overall market. In fact, there are several variables that influence the movement of the stock market, and one of them is the sentiment of investors, who drive the market.

Haddi et al. [5] analyze the importance of the text preprocessing in the sentiment analysis, and the resulted outcome shows that the appropriate feature extraction, and interpretation has improved the accuracy of the sentiment analysis using SVM. They also point out that sentiment analysis is a daunting field to obtain valuable insights from the opinion has expressed on social media requires natural language processing.

The work proposed in [6] by Ashish Pathak et al. discusses about the advantages of implementing machine learning on historical data and sentiment analysis on news headlines that builds the fuzzy logic module, which improves the accuracy of the stock market predication and also describes the limitation of the conventional stock market analysis methodology. This research uses text preprocessing techniques on textual data that is news headlines and finds out the most effective feature, which is categorized as positive and negative.

Elagamy et al. observed in [7] that data mining approaches using historical data to anticipate stock price movement are limited to making judgments within the context of current knowledge and are unable to detect random stock market activity or give triggers behind events. Thus, this research added that the huge financial data available on textual format focuses on the random behaviors of the stock market events, but this data is unstructured so that the text mining is applied on it to provide combined approach of random forest (RF) algorithm with text mining (TM) to study the critical indicator for predicting the abnormal movement of the stock market.

Wang et al. [8] have used social media mining technology to quantitatively determine the market segment and forecast the short-term stock price movement in conjunction with the other indicators.

Ho et al. [9] stated that the sentiments of financial news play an important role in investors' decision-making processes.

Das et al. [10] claim that the sentiment analysis of public's opinion obtained from social media feeds can be used to predict individual stock price fluctuations in the future.

Pagolu et al. [11] reported that incorporating Twitter sentiment analysis adds useful data to the prediction model and enhance accuracy. There is a strong correlation between the rise and fall of the company stock price and public opinions expressed on twitter via tweets about that company.

Pradha et al. [12] conducted research in which they proposed a suitable preprocessing approach for textual information, which plays a significant role in the classifier's prediction accuracy and efficiency while utilizing unstructured data. To remove the noise from the data and rendering such unstructured data into organized and meaningful, text data preprocessing is considered as one of the successful methods. This research work compares various preprocessing techniques for the textual data and their effect on the generated sentiment.

3 Proposed Model

The proposed approach is built on the Apache Spark big data framework, in which Spark NLP is used for data preparation and Spark MLlib is used to categorize news and Twitter data using a machine learning algorithm. The news data collected from the Moneycontrol Web site for BSE top 100 stock companies is based on the market capitalization for 2010–20, and the Twitter data is collected by using the Python Tweepy API. The specific company data or security-wise tweets data is collected by passing the relevant hashtags.

The Spark NLP pipeline created for the data preprocessing includes different phases. The clean text data is converted into the vector by using the TF-IDF for feature selection and extraction, and finally, the Textblob and Vader library are used to check the sentiment impact of the news and tweets. Finally, machine learning algorithms, logistic regression, naïve Bayes, and random forest are applied to classify the data as positive, negative, and neutral for performing sentiment analysis. Furthermore, the accuracy score generated from different libraries as mentioned is compared.

4 Phases of Sentiment Analysis

Sentiment analysis: This is called as logical mining of text, which distinguishes and extricates the abstract data in source material and helping a context to comprehend the social estimation of their phenomena or administration while observing on the web discussion. Henceforth, in stock market, the insiders and outsiders give some extent of information to understand the market movement, which really gives improvement in the prediction of price movement. In the proposed model, the expert and user reviews are observed in terms of news feed and tweets. To classify the reviews in terms of positively added statement, complement statement and no effect statement, the sentiment analysis method is used. The most widely recognized content characterization device examines an approaching message and tells whether the basic supposition is positive, negative, or neutral. It is also helpful to investigate the intention behind the reviews given by users. For stock market, the proposed model is utilized for handling both structured and unstructured data. Structured data preprocessing is followed by statistical data analytics and preprocessing. Unstructured data is processed by using sentimental analysis (Fig. 1).

The sentiment analysis is incorporated by using the following phases.

Tokenization: Tokenization is the common task involved in natural language processing (NLP). Tokenization is generally considered as a way for separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or sub-words. The sentences or tweets are converted into words.

Stemmer: It is fundamentally eliminating the postfix from a word and decrease it to its root word. For instance: “Moving” is a word and its addition is “ing”, in the event

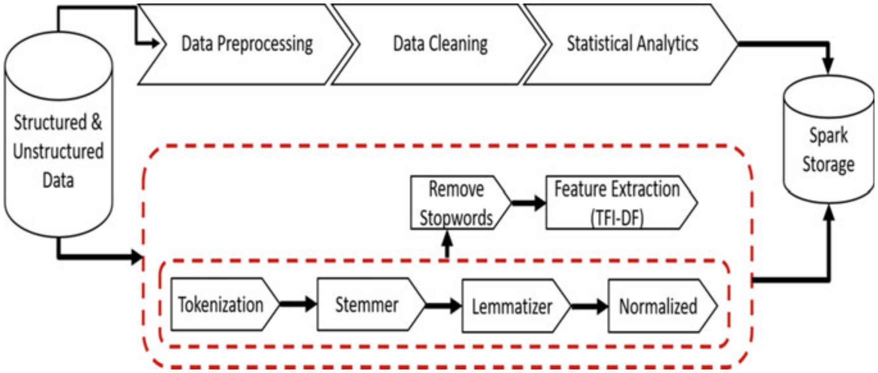

Fig. 1 Proposed model using Spark NLP

that we eliminate "ing" from "Moving" at that point we will get base word or root word which is "Move." Model utilizes these additions to make another word from unique stem word.

Lemmatization: For linguistic reasons, the records will utilize various types of a word, for example, sort out, puts together, and arranging. Also, there are groups of derivationally related words with comparative implications, for example vote-based system, majority rule, and democratization. As a general rule, it appears as a quest for one of these terms and it might be beneficial in restoring the reports that contain another word in the set. The goal of stemming and lemmatization is to reduce inflectional structures and occasionally the derivationally related word types to a standard base structure. The result of this text mapping will be something like: The stock's drives are different, and the stock drive may differ.

Normalized: Text standardization is the way toward changing a content into an accepted (standard) structure. For instance, "gooood" and "gud" can be changed to "acceptable," its standard structure. Another model involves planning the related indistinguishable words, for example "stopwords," "stop-words," and "stop words" to simply "stopwords."

Stop Word Removal: Stop words are a bunch of normally utilized words in a language. Stop words are ordinarily utilized in text mining and natural language processing (NLP) to dispense the words that are so generally utilized to convey next to no helpful data.

Feature Extraction (TF-IDF): In data recovery, TF-IDF or term recurrence converse record recurrence is a mathematical measurement that is planned to reflect how significant a word is to an archive in an assortment or corpus. The TF-IDF weight is a weight regularly utilized in data recovery and text mining.

5 Model Implementation

```
model_pipeline = Pipeline(
   stages=[document_assembler,
         tokenizer,
         normalizer,
         stopwords_cleaner,
         stemmer,
         finisher,
         hashingTF,
         idf,
         label_stringIdx,
         ml_classifier,
         label_to_stringIdx])
sentimentmodel = model_pipeline.fit(traindata).transform(testdata)
```

The above pipeline contains three parts.

The first part contains document_assembler, tokenizer, normalizer, stopwords cleaner, stemmer, and finisher, which is the process of implementing Spark NLP for data cleaning and data preprocessing.

Second part contains hashingTF, idf, and ml classifier, which is process of Spark MLlib for the implementation of machine learning, feature extraction as well as the implementation of machine learning algorithms for sentiment classification.

label_stringidx and label_to_stringIdx just for the string labeling.

The last part is the implementation of single execution plan for the testing and training data.

6 Results and Discussion

This section shows the results of a sequence of stages of the text preprocessing and steps for building the Spark NLP pipeline. Spark NLP comes with a number of annotators and transformers, and it also seamlessly integrates with Spark MLLib to build a data preprocessing pipeline.

Step 1: Initialize Spark.

Step 2: Load the Twitter and news data.

Step 3: Build NLP pipeline using Spark NLP [This pipeline can include feature extraction modules like HashingTF and IDF and machine learning classifier model].

Step 4: Implement and evaluate the model.

Original Text: In Spark NLP, the first stage transforms raw data into document type for further process. A special transformer DocumentAssembler() with desired parameters is used for that.

Tokens: The next stage identifies tokens with tokenization standards. Tokenizer annotator splits the documents into token according to the parameters like min max width of the token, case sensitivity of the text, and character list that is used to separate from the inside of tokens based on the patterns it will be separated from inside tokens and many more.

Normalized Text: This stage removes all dirty characters from text using the normalizer annotator, which has followed the regex pattern and transform the words based on the provided directory.

Clean Tokens: This stage obtains the clean tokens by removing the stop words using the StopWordsCleaner because in NLP process these are useless words.

Stemmer Text: This stage performs stemming process for removing a part of a word or reducing a word to its stem or root and for that stemmer and annotator is used.

Token Features: This is an important stage, where NLP pipeline is ready to go, we might as well put our annotation results to use somewhere else. Annotation values are the output obtained by the finisher as a string.

TF Features: Finally, the feature token form the documents. The TF-IDF feature extraction technique converts the token into vectors. TF and IDF are implemented in HashingTF and IDF. This stage utilizes the HashingTF to convert the documents to fixed size vectors.

Features: The last and final stage generates the inverse document frequency. The minDocFreq variable of the IDF supports filtering out terms, which do not appear in a minimum number of documents. For terms that are not in at least minDocFreq documents, the IDF is found as 0, resulting in TF-IDFs of 0.

7 Sentiment Analysis Comparison

This research work utilizes company wise NEWS data of BSE 100 stock company based on the market capitalization. It is difficult to represent scratch view of all stock modeling, where the paper show top 5 companies [Reliance Industries Ltd, Tata Consultancy Services Ltd, HDFC Bank Ltd, Infosys Ltd, and Hindustan Unilever Ltd.] data (Figs. 2, 3 and 4).

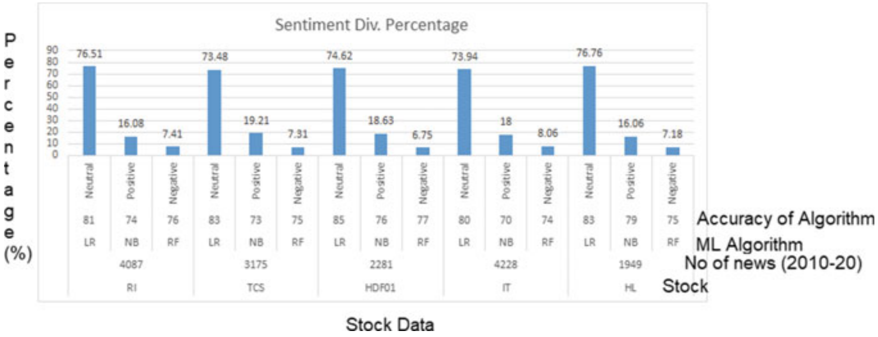

Fig. 2 Top 5 stock company sentiment analysis using Textblob

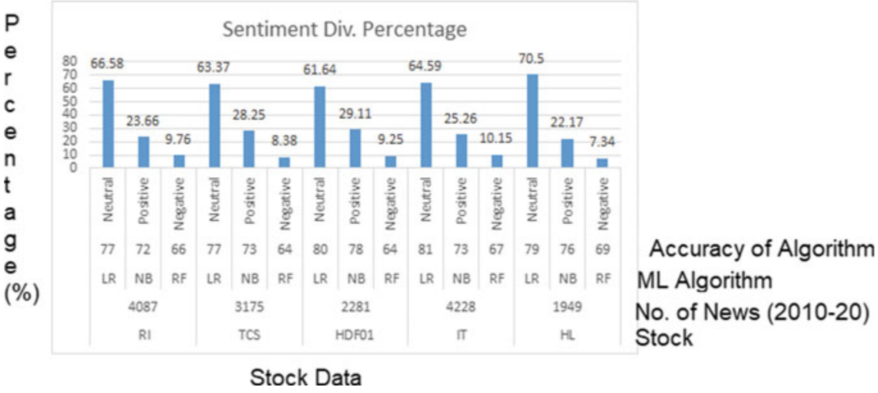

Fig. 3 Top 5 stock company sentiment analysis using Vader

8 Conclusion and Future Scope

The proposed research study has successfully implemented the sentimental analytics for performing stock market predictive analysis. This paper clearly shows the methods and model implementation of sentiment on Twitter's tweets and news for stock. The proposed model includes a script to fetch online news from online verified sources and tweets from twitter, and then, the text normalization is performed, and finally, Textblob and Vader library are used to obtain the sentiment impact score of the financial text. After that, the sentimental analysis steps are incorporated and each step results are discussed. According to experiments, the Spark NLP gives best performance for calculating the sentiments from text. As mentioned, in this experimental study, different algorithms and libraries were applied. As per the result study, the Textblob library gives better result in the context of sentiment analysis. After generating sentimental results, the machine learning algorithms were applied and among that logistic regression gives accuracy between 80 and 85% for all companies

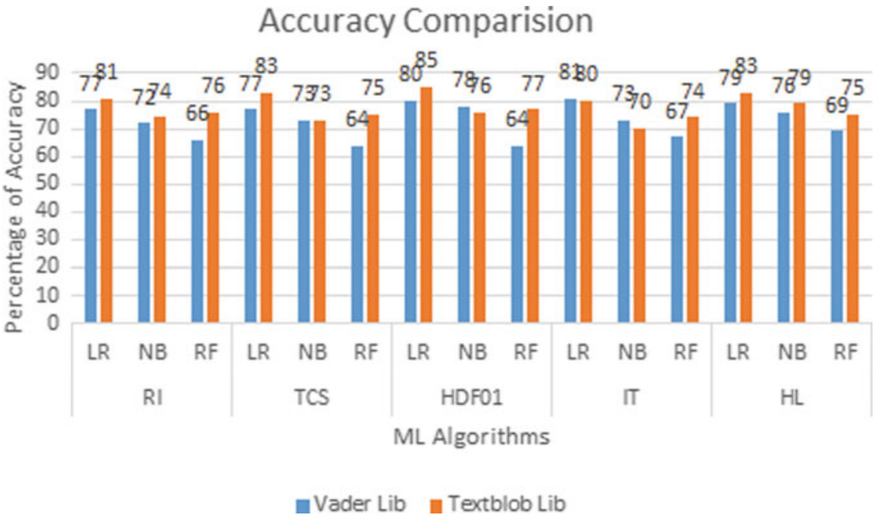

Fig. 4 Top 5 stock company sentiment analysis accuracy comparison for Textblob and Vader

stock. The experiment is carried out with sample data if more data is applied with advance Spark NLP methods than increasing the accuracy.

References

1. T.H. Nguyen, K. Shirai, J. Velcin, Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. **42**(24), 9603–9611 (2015). https://doi.org/10.1016/j.eswa.2015.07.052
2. A. Romanowski, M. Skuza, Towards predicting stock price moves with aid of sentiment analysis of Twitter social network data and big data processing environment. Adv. Bus. ICT: New Ideas Ongoing Res. **658**, 105–123 (2016). https://doi.org/10.1007/978-3-319-47208-9_7
3. C. Lee, I. Paik, Stock market analysis from Twitter and news based on streaming big data infrastructure, in *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)* (2017). https://doi.org/10.1109/icawst.2017.8256469
4. A. Derakhshan, H. Beigy, Sentiment analysis on stock social media for stock price movement prediction. Eng. Appl. Artif. Intell. **85**, 569–578 (2019). https://doi.org/10.1016/j.engappai.2019.07.002
5. E. Haddi, X. Liu, Y. Shi, The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. **17**, 26–32 (2013)
6. A. Pathak, N.P. Shetty, Indian Stock Market prediction using machine learning and sentiment analysis. Adv. Intel. Syst. Comput. Comput. Intel. Data Mining 595–603 (2018). https://doi.org/10.1007/978-981-10-8055-5_53
7. M.N. Elagamy, C. Stanier, B. Sharp, Text mining approach to analyse stock market movement, in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018) Advances in Intelligent Systems and Computing*, pp. 661–670 (2018). https://doi.org/10.1007/978-3-319-74690-6_65

8. Y. Wang, Y. Wang, Using social media mining technology to assist in price prediction of stock market, in *2016 IEEE International Conference on Big Data Analysis (ICBDA)* (2016). https://doi.org/10.1109/icbda.2016.7509794
9. K. Ho, W. Wang, Predicting stock price movements with news sentiment: an artificial neural network approach. Artif. Neur. Netw. Model. Stud. Comput. Intel. **628**, 395–403 (2016). https://doi.org/10.1007/978-3-319-28495-8_18
10. S. Das, R.K. Behera, M. Kumar, S.K. Rath, Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia Comput. Sci. **132**, 956–964 (2018). https://doi.org/10.1016/j.procs.2018.05.111
11. V.S. Pagolu, K.N. Reddy, G. Panda, B. Majhi, Sentiment analysis of Twitter data for predicting stock market movements, in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)* (2016). https://doi.org/10.1109/scopes.2016.7955659
12. S. Pradha, M.N. Halgamuge, N.T. Vinh, Effective text data preprocessing technique for sentiment analysis in social media data, in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)* (2019). https://doi.org/10.1109/kse.2019.8919368