APPLICATION OF SOFT COMPUTING

Sentiment analysis using treebank filtered preprocess with relevant vector boost classifier

P. Rajendiran1 · P. L. K. Priyadarsini1

Accepted: 6 November 2021 / Published online: 16 March 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
The procedure of identifying and classifying opinions in a piece of text to find out whether customer reviews toward a particular product or service are positive, negative, or neutral is termed as sentiment analysis. Stock market prediction is one of the most attractive topics in academic and real-life business. Many data mining techniques about sentiment analysis are suffering from the inaccuracy of prediction. The low classification accuracy has a direct effect on the reliability of stock market indicators. Treebank filtering Data Preprocessing based Ochiai-Barkman Relevance Vector Linear Programming Boost Classification technique is used for stock market prediction using sentimental analysis with higher prediction accuracy and lesser classification time for enhancing accuracy of stock market based on product review. Initially, the customer reviews and feedback on services or products are collected from the large database. After that, the collected customer reviews are preprocessed by performing the process such as tokenization, stemming, filtering. In order to achieve sentimental analysis through classifying customer reviews as positive and negative, Ochiai-Barkman Relevance Vector Linear Programming Boost Classification algorithm is used. The Linear Programming Boost Classification algorithm constructs with an empty set of weak classifiers as the Ochiai-Barkman Relevance Vector machine. The customer reviews are classified based on the Ochiai-Barkman similarity coefficient. The ensemble technique combines the weak classification results into strong by minimizing the error. In this way, the classification performance gets improved and the prediction of the stock market is carried out in a more accurate manner. Experimental evaluation is carried out on factors such as prediction accuracy, sensitivity, specificity, and prediction time versus amount of customer reviews.

Keywords Sentiment analysis · Stock market prediction · Linear programming boost classification · Relevance vector machine · Ochiai-Barkman similarity coefficient

1 Introduction
The stock market is a very important part of a country’s financial system. Forecasting the stock market movements is an important and demanding task through financial data environment. This type of stock market prediction is carried out through the sentimental analysis of consumer data. Fine-grained sentiment examination related to services with products plays a significant role in many applications.

NDMGA-XGB was developed in Al-Qudah et al. (2020) to efficiently predict and evaluate the customers’ reviews for online services. The designed model increases the accuracy but it failed to involve examining other online services with minimum time consumption. A new cluster-based classification model was introduced in Vijayaragavan et al. (2020) for analyzing the online product reviews based on SVM. The model was not efficient to perform the accurate classification with minimum error.

Machine learning classifiers were introduced in Khan et al. (2020) for stock market prediction. The designed
model increases the prediction accuracy, but an efficient technique was not applied for determining stock relevant keywords to minimize the time consumption.

Contextual analysis (CA) mechanism was introduced in Azwa Abdul Aziz (2019) for clustering sentiment terms. But it failed to consider improving prediction result. The text analysis system was designed in Sert et al. (2020) to predict the stock market movements based on news and social media data. But the system failed for enhancing prediction models.

A finer-grained textual and sentiment analysis was performed in Bouktif et al. (2020) for predicting the stock market movement trend. But an efficient machine learning technique is not utilized to improve sentiment analysis. Machine learning algorithms were designed in Yi and Liu (2020) for analyzing and categorizing product. But the algorithm was not combined for huge customer awareness.

A sentiment polarity categorization approach was introduced in Samina Kausar and Huahu (2019) using huge data set with the number of online reviews. But the approach failed to use the more advanced technology regarding the information by analyzing the online review. A multi-attribute decision-making (MADM) model was developed in Yang et al. (2021) to rank the dissimilar products using several online reviews. However, the model failed to increase its classification accuracy.

A model-independent approach was introduced in Wang et al. (2019) for forecasting the stock market based on logistic regression. But the accuracy was not high enough for stock market prediction.

1.1 Novel contributions

TFDP-ORVLPBC technique is introduced by novel contributions for solving existing problems.

- To improve the prediction accuracy, the TFDP-ORVLPBC technique is introduced using sentimental analysis with lesser time based on the two steps, namely preprocessing and classification.
- To minimize the prediction time, the TFDP-ORVLPBC technique executes the preprocessing step that includes tokenization, stemming, and filtering. On the contrary to the existing technique, the TFDP-ORVLPBC technique uses Treebank Word Tokenizer, Conditional light Stemming and Dynamic stop word filtering technique. This process of the TFDP-ORVLPBC technique extracts the main keywords for classification.
- Then, the Ochiai-Barkman Relevance Vector Linear Programming Boost Classification algorithm is employed in the TFDP-ORVLPBC technique to perform the sentimental analysis by analyzing the customer reviews as positive, negative, or neutral. The proposed ensemble algorithm constructs with an empty set of weak classifiers as the Ochiai-Barkman Relevance Vector machine. The customer reviews are classified based on the Ochiai-Barkman similarity coefficient. The ensemble technique increases the classification accuracy.
- At last, extensive testing is conducted to estimate the performance of our TFDP-ORVLPBC technique and other related works. The experimental result exhibits that the TFDP-ORVLPBC technique is analyzed with the various performance metrics with a number of customer reviews.

1.2 Paper organization

The rest of the paper is summarized as follows. Section 2 describes literature review of sentimental analysis of online products. Section 3 provides a brief description of the TFDP-ORVLPBC method with assist of architecture figure. Section 4 explains simulation settings. Section 5 illustrates results of simulations for five different methods. Section 6 explains conclusion of this paper.

2 Literature review

A hybrid approach was introduced in Pal et al. (2019) to find the time series of stock prices by using data discretization based on fuzzy rough set theory. But the approach was not efficient to minimize the error rate of the time series of stock price prediction. EMD2FNN was introduced in Zhou et al. (2019) to forecast the stock market movement. The designed network reduces the error, but the prediction time was not minimized.

A multi-source multiple instance method was developed in Zhang et al. (2018) for forecasting the stock market prediction. But the designed method failed to achieve higher prediction accuracy. Several machine learning approaches were designed in Rezapour (2021) for sentiment classification using examining the reviews. A novel pattern-based method was introduced in Rodrigues et al. (2020) for aspect extraction and sentiment analysis with higher accuracy. However, the method failed to improve the proposed method by considering the review sentences with dissimilar sentiment classification techniques.

A generic structure was developed in Zhou et al. (2018) using LSTM and CNN for involving high-frequency stock markets with higher accuracy and minimizes error. However, the framework failed to combine the predictive models under multistage conditions. A joint aspect-based sentiment topic (JABST) method was introduced in Tang et al. (2019) to classify the sentiment polarity. But the
method failed to analyze the more complex opinions with higher accuracy.

A novel deep learning-based solution was developed in Kiran (2020) for sentiment polarity categorization of reviews. But the time consumption of the sentiment polarity categorization of reviews was not minimized. The evolutionary strategy was introduced in He and Zhu (2020) with the influence of different factors using e-commerce platforms.

A polymerization topic sentiment model (PTSM) was developed in Huang et al. (2019) to perform the textual analysis based on online reviews. The designed model minimizes the error, but efficient preprocessing was not carried out to minimize the time.

### 3 Methodology

Sentiment analysis has become one of the most important procedures to predict the stock market behavior according to the customer reviews about a particular topic such as news, movie, event, and remarks related to the product. Due to the huge number of reviews generated from the customer, for analyzing information in an accurate manner. In order to detect general view of product, sentiment analysis technique is performed. Lately, the majority of research works is designed for sentiment analysis by application of organization and ranking techniques. But it suffers less accuracy of the accurate classification of the customer reviews. Based on the motivation, a novel technique called the TFDP-ORVLPBC technique is introduced for improving the classification accuracy.

Figure 1 indicates architecture diagram of TFDP-ORVLPBC method for improving accuracy of the customer reviews classification for stock market prediction. The proposed TFDP-ORVLPBC technique performs two major processes, namely data preprocessing and classification. The data preprocessing step of the proposed TFDP-ORVLPBC technique includes Tokenization, stemming, and filtering. After the preprocessing step, the classification process is performed using an ensemble technique called the Linear Programming Boost technique. The ensemble technique accurately classifies the reviews about the product with better accuracy and minimal time consumption. The different process of TFDP-ORVLPBC technique is discussed below.

#### 3.1 Treebank dynamic filtering-based data preprocessing

The proposed TFDP-ORVLPBC technique starts for achieving data preprocessing to reduce complexity. The data preprocessing steps include three different sub-processes, namely tokenization, stemming, and filtering.

Let us consider the sentiment dataset SD and the number of customer reviews is extracted from the dataset.

\[ R_i = r_1, r_2, r_3, \ldots, r_n \in SD \]

where \( R_i \) denotes the number of reviews \( r_1, r_2, r_3, \ldots, r_n \) collected from the dataset SD. After the review collection, the tokenization process is carried out using Treebank Word Tokenizer. Treebank Word Tokenizer is worked by partitioning the words into a number of words using punctuation and spaces.

\[ r_1 \rightarrow [\omega_1', \omega_2', \omega_3', \ldots, \omega_m' ] \]

where \( r_1 \) denotes a review, \( \omega_1, \omega_2, \omega_3, \ldots, \omega_m \) denotes words extracted from the review using Treebank Word Tokenizer.

##### 3.1.1 Conditional light stemming

The stemming is the process of removing the additional words from their root word. In other words, the word stemming process eliminates the suffixes and offers the root words. The conditional light stemming approach is used to perform the word stemming process.

As shown in Table 1, the word ends with ‘ing’, ‘ly’, ‘ed’ are called suffix that is removed and obtains the root word ends, sad, and finish.

##### 3.1.2 Dynamic stop word filtering technique

Then stop words are the words that are occurring continually in the documents and they did not provide any...
meaning. The filtering technique removes the stop words such as “are”, “the”, “a”, “an”, “in”, “and”, “our”, “this”, and so on. These words are removed from the given customer review.

### 3.2 Ochiai-Barkman relevance vector linear programming boost classification

Behind data preprocessing, an Ochiai-Barkman Relevance Vector Linear Programming Boost Classification algorithm is utilized in TFDP-ORVLPBC for achieving sentimental analysis. In the TFDP-ORVLPBC technique, Linear Programming Boosting is a supervised ensemble classification technique from the boosting family of classifiers. In machine learning, boosting is an ensemble algorithm which transforms weak classification into strong classification. Weak classification is base classifier that difficult to provide accurate results. The ensemble classifier combines weak learner into strong classification to provide the true classification.

Figure 2 displays the structural process of the Linear Programming Boost Classification technique to obtain final classification results. The ensemble technique considers the training sets as \((x_i, y)\) where \(x_i\) represents extracted sentiment words \(\omega_1, \omega_2, \omega_3, \ldots, \omega_m\) and ‘y’indicates ensemble classification results for the given inputs. Ensemble method constructs ‘b’ weak learners for classifying given input. The ensemble technique uses the kernel relevance vector machine as weak learner. Relevance Vector Machine is uses optimal hyperplane for classifying the reviews. The optimal hyperplane is decision boundary among two classes \(\{+1, -1\}\). Ochiai-Barkman relevance vector classifies the reviews on decision boundary.

\[
\phi \rightarrow q \cdot r_i + c = 0 \tag{3}
\]

where \(\phi\) symbolizes a decision boundary, \(q\) denotes the normal weight vector to training samples (i.e., reviews), ‘c’ represents a bias. The two marginal hyperplanes are selected either lower or upper side of the boundary.

\[
\begin{align*}
x_1 & \rightarrow q \cdot r_i + c > 0 \quad \text{i.e. } +1' \tag{4} \\
x_2 & \rightarrow q \cdot r_i + c < 0 \quad \text{i.e. } -1'
\end{align*}
\]

where \(M_1, M_2\) are the marginal hyperplanes for categorizing brain images into boundary. Hyperplanes use kernel function to measure the similarity between the sentiment

| Word extracted from review | Light stemming to remove the suffix | Root word |
|---------------------------|-----------------------------------|-----------|
| Ending                    | Ing                               | End       |
| Sadly                     | Ly                                | Sad       |
| Finished                  | Ed                                | Finish    |

The step-by-step process of the data preprocessing using the review dataset is explained in algorithm 1. Initially, a number of reviews are collected from the sentiment dataset. For each review, the Treebank Word Tokenizer is applied for partitioning the review into a number of words. After partitioning, conditional light stemming is applied to remove the stem words. The filtering technique is applied for removing the stop words. Finally, the significant words from the review are obtained for minimizing the classification time and improving the accuracy of review prediction about the products.
words using the Ochiai-Barkman coefficient. The similarity is estimated as given below,

\[ \vartheta(\omega_i, \omega_j) = \frac{|\omega_i \cap \omega_j|}{\sqrt{\sum \omega_i^2} \sqrt{\sum \omega_j^2}} \]  

where \( \vartheta \) denotes an Ochiai-Barkman similarity coefficient, \( \cap \) denotes a mutual dependence between the sentiment words \( \omega_i, \omega_j \), \( \sum \omega_i^2 \) symbolizes a squared score of \( \omega_i \), \( \sqrt{\sum \omega_j^2} \) denotes a signifies a squared score of \( \omega_j \). Based on the similarity, the hyperplanes classify the word above or below the decision boundary using the following expression.

\[ Z = \text{sign} \sum q_i R_i \vartheta(\omega_i, \omega_j) \]  

In (7), \( Z \) denotes predicted classification results, \( q_i \) denotes weights, \( R_i \) indicates dependent variable (i.e., output), \( \vartheta(\omega_i, \omega_j) \) indicates similarity between the words. ‘sign’ represent positive or negative or neutral.

Figure 3 illustrates the output results of the relevance vector machine. The hyperplane analyzes the sentiment words and returns the output ‘+ 1’ and ‘- 1’. Here, ‘+ 1’ indicates the positive review, ‘- 1’ indicates the negative review, the words fall onto the hyperplane indicates that the neutral review. The two sentiment words having a higher similarity, then the words are classified above the decision boundary. The two sentiment words having the lesser similarity, then the words are classified below the decision boundary. In this way, the sentiment words are correctly classified as positive, negative, and neutral reviews.

Ochiai-Barkman relevance vector classifier has some training errors and it hard to obtain accurate classification results. The ensemble technique combines all weak learners’ results to obtain the strong one.

\[ y = \sum_{i=1}^{b} Z_i \]  

where \( y \) symbolizes the output of ensemble classifier, \( Z_i \) represents an output of weak classifiers. The ensemble classifier initializes weights for each weak classifier.

\[ y = \sum_{i=1}^{b} Z_i \ast \mu \]  

where \( \mu \) indicates weights of weak classifier. In order to perform better accuracy, training error of every weak learner is measured after assigning weight. Error rate is measured as squared difference among actual classification and observed classification of weak classifier. It is expressed as follows,

\[ T_E = [Z_{\text{Actual}} - Z_{\text{Observed}}]^2 \]  

where \( T_E \) denotes an error, \( Z_{\text{Actual}} \) denotes actual classification results of weak classifier, \( Z_{\text{Observed}} \) denotes observed classification results of the weak classifier. Then, initial weights get updated for acquiring accurate classification. If the weak classifier wrongly classifies the reviews, initial weight is better. Otherwise, initial weight is lesser. Ensemble technique reduces error and enhances performance of classification. The linear program boosting technique enhances margin of two different classes of weak classifier. Therefore, classification results are exposed into margin,

\[ \sum_{i=1}^{b} Z_i \ast \nabla \mu + e \geq M \text{ Where } e \geq 0 \]  

where ‘\( e \)’ symbolizes non-negative vector of slack variable, \( M \) indicates margin between classes. If ensemble
classification was higher than margin, reviews are properly categorized into class resultant it reduces wrong classification. Therefore, iterated linear program boost ensemble technique obtains true classification. Algorithmic process of review classification using ensemble technique is given below.

Algorithm 2 portrays ensemble technique for improving review classification accuracy. Initially, ensemble technique constructs ‘b’ weak classifiers with number of extracted sentiment words. The relevance vector machine constructs the optimal hyperplane as a decision boundary for analyzing the extracted sentiment words using the Ochiai-Barkman similarity coefficient. Depend on similarity measure hyperplane categorize sentiment words as positive, negative, and neutral. The ensemble technique combines all the weak classification results. The weight is initialized for each weak classification result. Next, error is expressed on actual and predicted classification. Based on error value, the initial weight gets updated. The ensemble technique determines finest classification results with lesser error. Finally, ensemble classification obtains final classification results.
4 Experimental setup

Simulation of TFDP-ORVLPBC technique and NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) are implemented in Java using Consumer Reviews of Amazon Products dataset taken as Kaggle (https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products). This dataset consists of two CSV files. Among them, the Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products review file is taken for sentiment data analysis through the Consumer Reviews. The CSV files consist as 28,000 consumer reviews. In Brand and Manufacturer field, every product comprises name Amazon. From the 28,000 reviews, 1000–10,000 reviews are considered for conducting the experiment.

5 Performance results and discussion

The experimental results of the proposed TFDP-ORVLPBC technique and existing NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020), are discussed based on certain parameters such as accuracy, sensitivity, specificity, and prediction time with amount of reviews. Effectiveness and efficiency of the proposed and existing methods are discussed.

5.1 Impact of prediction accuracy

It is defined as number of reviews that are properly classified into different classes to the total number of reviews from the dataset. The prediction accuracy is calculated using the following expression,

$$PA = \left(\frac{P_T + N_T}{P_T + N_T + P_F + N_F}\right) \times 100$$  \hspace{1cm} (12)

where PA denotes prediction accuracy, $P_T$ symbolizes the true positive, i.e., number of reviews correctly classified, $N_T$ indicates true negative, $P_F$ symbolizes the false positive, $N_F$ denotes a false negative. The prediction accuracy is measured in percentage (%).

Table 2 reports experimental analysis of prediction accuracy versus the number of reviews in the ranges from 1000 to 10,000. The prediction accuracy is measured using three TFDP-ORVLPBC techniques and an existing NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification mode (Vijayaragavan et al. 2020). According to the observed results, the proposed TFDP-ORVLPBC technique obtains higher prediction accuracy than the other conventional methods. Let us consider 1000 reviews considered for calculating the prediction accuracy. By applying the TFDP-ORVLPBC technique, 93% of the prediction accuracy was observed. Whereas the prediction accuracy of the existing NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) are 89% and 85%, respectively. For each method, ten results are observed with respect to various counts of input reviews. TFDP-ORVLPBC technique is compared with existing methods. The average results designate accuracy of TFDP-ORVLPBC technique is increased as 6% and 8% compared with existing techniques.

In Fig. 4, prediction accuracy results under a varying number of reviews collected from the Amazon Products dataset. The graphical plot indicates that the numbers of reviews as input in horizontal axis and accuracy of three methods is observed in ‘y’ axis. The observed graphical results notice that the prediction accuracy of the TFDP-ORVLPBC technique is higher compared with two existing methods. The main reason was due to application of Ochiai-Barkman Relevance Vector Linear Programming Boost Classification algorithm. The ensemble technique uses the Relevance Vector classifier as a weak learner to analyze the extracted words from the reviews. The proposed Boost classification algorithm accurately analyzes the reviews and classifies the reviews as positive, negative, and neutral. Based on classification results, the prediction is said to be improved.

5.2 Impact of precision

Precision is calculated as ratio of amount of reviews that are properly classified to entire amount of reviews. The precision is formulated by,

$$P = \left(\frac{P_T}{P_T + P_F}\right) \times 100$$  \hspace{1cm} (13)

Table 2 Comparison of prediction accuracy

| Number of reviews | Prediction accuracy (%) |
|-------------------|-------------------------|
|                   | TFDP-ORVLPBC | NDMGA-XGB | New cluster-based classification model |
| 1000              | 93          | 89        | 85        |
| 2000              | 94          | 89.5      | 87        |
| 3000              | 92.66       | 90        | 87.66     |
| 4000              | 93.75       | 90.5      | 86.25     |
| 5000              | 94.6        | 89.6      | 87        |
| 6000              | 93.33       | 89.16     | 87.83     |
| 7000              | 93.57       | 88.28     | 87.14     |
| 8000              | 94          | 88        | 86.87     |
| 9000              | 94.44       | 87.77     | 86.11     |
| 10,000            | 94.5        | 87        | 85        |
where $P$ denotes precision, $P_T$ symbolizes the true positive, i.e., number of reviews correctly classified, $P_F$ symbolizes the false positive. The precision is measured in percentage (%).

Table 3 and Fig. 5 portray the comparative analysis of precision using three methods, namely TFDP-ORVLPBC technique and existing new cluster-based classification model (Al-Qudah et al. 2020), NDMGA-XGB (Vijayaragavan et al. 2020). The figure demonstrates that the TFDP-ORVLPBC technique with amount of reviews collected as Amazon Products. The figure summarizes the overall precision measures of three algorithms. Under the presence of 1000 reviews for conducting the experiments, the TFDP-ORVLPBC technique attains 95.60% of precision and precision of the existing NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) offers 93.02% and 90.36%. From the observed results, the presented TFDP-ORVLPBC technique offered better precision results. The comparison of ten results indicates that the precision of TFDP-ORVLPBC is considerably increased by 3% and 5% when compared to existing methods. The main reason for this significant improvement is to apply the ensemble classification technique. Ensemble technique integrates all weak learner results. Then, error identifies the best classification results. Finally, the ensemble classification is exposed into margin and obtains better classification results by increasing the true positives and minimizing the false positives.

### 5.3 Impact of recall

The recall is measured as proportion of reviews that are properly classified to entire number of reviews. Recall is measured as given below,

$$ R = \frac{P_T}{P_T + N_F} \times 100 $$

(14)

where $P$ denotes precision, $P_T$ symbolizes the true positive, i.e., number of reviews correctly classified, $N_F$ symbolizes the false negative. The recall is measured in percentage (%).

Table 4 and Fig. 6 depict the performance results of recall versus a number of reviews. The recall is measured using three methods TFDP-ORVLPBC technique, NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) with amount of consumer reviews is implemented by Java Language. Let us consider 1000 consumer for experiment, TFDP-ORVLPBC technique performs 96.66% of recall whereas the NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) achieves 94.11% and 91.46% correspondingly. Hence, recall of TFDP-ORVLPBC technique is comparatively high when compared with existing methods. The performance results of the proposed technique compared with existing methods. Recall of TFDP-ORVLPBC
technique is higher and it increased by 3% and 5% when compared to existing methods.

### 5.4 Impact of F-measure

The F-measure is the mean of precision as well as recall. It is formulated as given below,

$$FM = \left[ \frac{2 \times P \times R}{P + R} \right] \times 100$$  \hspace{1cm} (15)

where $FM$ denotes an F-measure, $P$ denotes precision, ‘$R$’ represents recall. It is calculated in percentage (%).

Table 5 given above illustrates the F-measure for varying numbers of reviews in the range of 1000 to 10,000 and the results are obtained using three TFDP-ORVLPCB techniques, (Al-Qudah et al. 2020) (Vijayaragavan et al. 2020). The obtained results designate that the F-measure of the proposed TFDP-ORVLPCB technique is increased compared with conventional methods. Let us consider the 1000 reviews. The F-measure is 96.13% using the TFDP-ORVLPCB technique, whereas the F-measure of exiting methods namely NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) is 93.56% and 90.90%, respectively. From the statistical analysis, the results indicate that the proposed TFDP-ORVLPCB technique maximizes the F-measure. The proposed technique is compared with conventional methods. Therefore, F-measure of proposed TFDP-ORVLPCB technique is considerably increased by 3% and 5% compared with existing methods. This confirms that proposed technique uses the ensemble boosting technique for minimizing incorrect classification as well as improves true positive (Fig. 7).

### 5.5 Impact of prediction time

It is calculated as number of times taken by algorithm for performing stock prediction through the review classification. Therefore, the overall prediction time is formulated as given below,

$$PT = n \times t[csr]$$  \hspace{1cm} (16)

From (16), PT denotes a prediction time, $n$ indicates number of reviews, $t$ represent time; csr indicate classification of a single review. The overall prediction consumption is calculated in milliseconds (ms).

Table 5 demonstrates the performance result of prediction time on dissimilar number of reviews ranges from 1000 to 10,000. There are three different methods are used for calculating the prediction time. Among three different methods, the proposed TFDP-ORVLPCB technique outperforms well than the existing methods. While considering 1000 customer reviews for sentiment classification, proposed TFDP-ORVLPCB technique as 14ms time consumption for prediction whereas the NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020) takes 16 ms and 18 ms, respectively. Similarly, ten different results are observed for each method. The obtained results of the TFDP-ORVLPCB are compared to the existing results. Compared to other existing methods, the TFDP-ORVLPCB provides better performance to achieve higher accuracy as 7% and 14% compared with NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020), respectively (Table 6).

Figure 8 provides the impact of prediction time for a different number of reviews taken by three methods. TFDP-ORVLPCB technique considerably outperforms the existing methods ’NDMGA-XGB (Al-Qudah et al. 2020), new cluster-based classification model (Vijayaragavan et al. 2020). Besides, while enhancing number of reviews, time taken for classification gets improved. But, TFDP-
ORVLPBC technique obtains reduced prediction time. Initially, a number of reviews are collected from the dataset. For each review, the Treebank Word Tokenizer is applied for partitioning the review into several words. After that, conditional light stemming is applied to remove the stem words. Followed by, the filtering technique is applied for removing the stop words. At last, the important words are extracted to perform the classification resulting it reduces classification time and improving the accuracy of products review prediction.

| Number of reviews | TFDP-ORVLPBC | NDMGA-XGB | New cluster-based classification model |
|-------------------|--------------|-----------|----------------------------------------|
| 1000              | 96.13        | 93.56     | 90.90                                  |
| 2000              | 96.77        | 94.15     | 92.52                                  |
| 3000              | 96           | 94.42     | 92.87                                  |
| 4000              | 96.64        | 94.73     | 91.90                                  |
| 5000              | 97.12        | 94.17     | 92.48                                  |
| 6000              | 96.39        | 93.89     | 93.02                                  |
| 7000              | 96.54        | 93.32     | 92.53                                  |
| 8000              | 96.78        | 93.18     | 92.36                                  |
| 9000              | 97.04        | 93.03     | 91.90                                  |
| 10,000            | 97.07        | 92.57     | 91.22                                  |

**Table 5** Comparison of F-measure

| Number of reviews | F-measure (%) |
|-------------------|---------------|
| 1000              | 96.13         |
| 2000              | 96.77         |
| 3000              | 96            |
| 4000              | 96.64         |
| 5000              | 97.12         |
| 6000              | 96.39         |
| 7000              | 96.54         |
| 8000              | 96.78         |
| 9000              | 97.04         |
| 10,000            | 97.07         |

**Table 6** Comparison of prediction time

| Number of reviews | Prediction time (ms) |
|-------------------|----------------------|
| 1000              | 14                   |
| 2000              | 20                   |
| 3000              | 24                   |
| 4000              | 28                   |
| 5000              | 30                   |
| 6000              | 33                   |
| 7000              | 35                   |
| 8000              | 38                   |
| 9000              | 41                   |
| 10,000            | 44                   |

TFDP-ORVLPBC technique is developed for predicting stock market movement and recognizes the importance of the customer reviews about the products simultaneously. Sentiment analysis is a well-known mining technique to demonstrate people’s reviews and sentiments about certain products or services. The major problem in sentiment analysis is the sentiment categorization that resolves whether a review is positive, negative, or neutral. An effective TFDP-ORVLPBC technique is introduced with the aim of enhancing the stock market prediction accuracy using sentimental data analysis. Initially, preprocessing step is performed by TFDP-ORVLPBC technique to remove unwanted words from customer reviews and improving the

---

**Fig. 7** Performance analysis of F-measure

**Fig. 8** Performance analysis of prediction time

**6 Conclusion**

TFDP-ORVLPBC technique is developed for predicting stock market movement and recognizes the importance of the customer reviews about the products simultaneously. Sentiment analysis is a well-known mining technique to demonstrate people’s reviews and sentiments about certain products or services. The major problem in sentiment analysis is the sentiment categorization that resolves whether a review is positive, negative, or neutral. An effective TFDP-ORVLPBC technique is introduced with the aim of enhancing the stock market prediction accuracy using sentimental data analysis. Initially, preprocessing step is performed by TFDP-ORVLPBC technique to remove unwanted words from customer reviews and improving the
prediction performance and minimum time consumption. After that, the classification step is carried out in the TFDP-ORVLPBC technique to analyze the extracted words. The accuracy of stock market prediction was increased. The experimental assessments are carried out with the Amazon product dataset. TFDP-ORVLPBC technique increases prediction accuracy, precision, recall, F-measure with minimum time compared with state-of-the-art works.

References

Al-Qudah DA, Al-Zoubil AM, Castillo-Valdivieso PA, Hossam F (2020) Sentiment analysis for e-payment service providers using evolutionary extreme gradient boosting. IEEE Access 8:189930–189944

Aziz AA, Starkey A (2019) Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access 8:17722–17733

Bouktif S, Fiaz A, Awad M (2020) Augmented textual features-based stock market prediction. IEEE Access 8:40269–40282

He H, Zhu L (2020) Online shopping green product quality supervision strategy with consumer feedback and collusion behavior. PLoS ONE 15(3):1–19

Huang L, Dou Z, Yongjun Hu, Huang R (2019) Textual analysis for online reviews: a polymerization topic sentiment model. IEEE Access 7:91940–91945

Kausar S, Huahu X, Ahmad W, Shabir MY, Ahmad W (2019) A sentiment polarity categorization technique for online product reviews. IEEE Access 8:3594–3605

Khan W, Ghazanfar MA, Azam MA, Karami A, Alyoubi KH, Alfaakeh AS (2020) Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Hum Comput 1–24

Kiran R, Kumar P, Bhasker B (2020) Oslcfit (organic simultaneous LSTM and CNN Fit): a novel deep learning based solution for sentiment polarity classification of reviews. Expert Syst Appl 157:1–12

Pal SS, Kar S (2019) Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory. Math Comput Simul 162:18–30

Rezapour M (2021) Sentiment classification of skewed shoppers’ reviews using machine learning techniques, examining the textual features. Eng Rep 3(1):1–13

Rodrigues AP, Chiplunkar NN, Fernandes R (2020) Aspect-based classification of product reviews using Hadoop framework. Cogent Eng 7(1):1–22

Sert OC, Sahin SD, Ozery T, Alhajj R (2020) Analysis and prediction in sparse and high dimensional text data: the case of dow jones stock market. Phys A Stat Mech Appl 545:1–45

Tang F, Luoyi Fu, Yao B, Wenchao Xu (2019) Aspect based fine-grained sentiment analysis for online reviews. Inf Sci 488:190–204

Vijayaragavan P, Ponnusamy R, Aramudhan M (2020) An optimal support vector machine-based classification model for sentimental analysis of online product reviews. Future Gener Comput Syst 111:234–240

Wang H, Shan Lu, Zhao J (2019) Aggregating multiple types of complex data in stock market prediction: a model-independent framework. Knowl-Based Syst 164:193–204

Yang L, Li Y, Wang J, Sherratt RS (2021) Sentiment analysis technique and neutrosophic set theory for mining and ranking big data from online reviews. IEEE Access 9:47338–47353

Yi S, Liu X (2020) Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers’ review. Complex Intell Syst 6:621–634

Zhang Xi, Siyu Qu, Huang J, Fung B, Philip Yu (2018) Stock market prediction via multi-source multiple instance learning. IEEE Access 6:50720–50728

Zhou X, Pan Z, Hu G, Tang S, Zhao C (2018) Stock market prediction on high-frequency data using generative adversarial nets. Math Probl Eng 2018:1–11

Zhou F, Zhou H-M, Yang Z, Yang L (2019) EMD2FNN: a strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. Expert Syst Appl 115:136–151

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.