Machine Learning Approach based Sentiment Analysis, Classification: An Application of Natural Language Processing

1.D.Geethangili Research Scholar, Periyar University, Salem-11.

2.Dr.P.Suresh HOD Computer Science, Salem Sowdeswari College, Salem-10

Article Info
Page Number: 773-786
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

Many social websites and android applications is flooded with user reviews and data in this modernized world. Every enthusiastic social person tends to express his or her views in form of comments and this is referred as sentiments. These comments not only express the peoples' view but also able to know in depth knowledge mostly in case of any social media. Generally, these comments/reviews present in text format are unstructured in nature and hence text preprocessing technique is applied in prior to analysis. Natural language Processing is used to get the sentiments (positive, negative, neutral) with many feature extraction techniques. This paper analyses the Amazon product reviews with most popular feature extraction techniques and classification using machine learning approach for word and n-gram sentiment analysis levels. Term Frequency-Inverse Document Frequency (TF-IDF) with pruning is used for feature extraction and sentiment analysis then data mining decision tree based Random forest algorithm with feature weights is applied for sentiment classification. The work is implemented Rapidminer tool and suitable evaluation measure is used for assessing the performance of existing and proposed.

Keywords: Machine learning, Data mining, TF-IDF, Bag of Words (BOW), Random forest.

1. Introduction

Sentiment Analysis is the process to find the user centric opinion based on their experience. The main concept is to analyze the posted text by the user and to extract their feedbacks. Sentiment Analysis is used in any kind of social reviews such as healthcare, politics and movie industries, business, retail & wholesale, hotel industry etc. It is used to find out the quality of service (QOS). To perform the analysis, researchers are using various machine learning algorithms y to find out the result from people's sentiment.

1.1. Sentiment analysis approaches

There are two main approaches shown in Figure 1, one is [1] lexicon based and the other is machine learning based. The lexicon based is further divided into dictionary based and corpus based and the later is of two kinds statistical and semantic.

1.1.1 Lexicon based

This approach uses pre-prepared [2] sentiment lexicon that contains a word and corresponding sentiment score in order to score a document by aggregating the sentiment scores of all the words in the document. The dictionary of lexicons can be created manually or automatically generated. Initially, lexicons are captured from the whole document and then WorldNet or any online thesaurus can be used to discover the synonyms and antonyms to expand that dictionary. To calculate any text orientation, adjective and adverb combinations are extracted with their value.

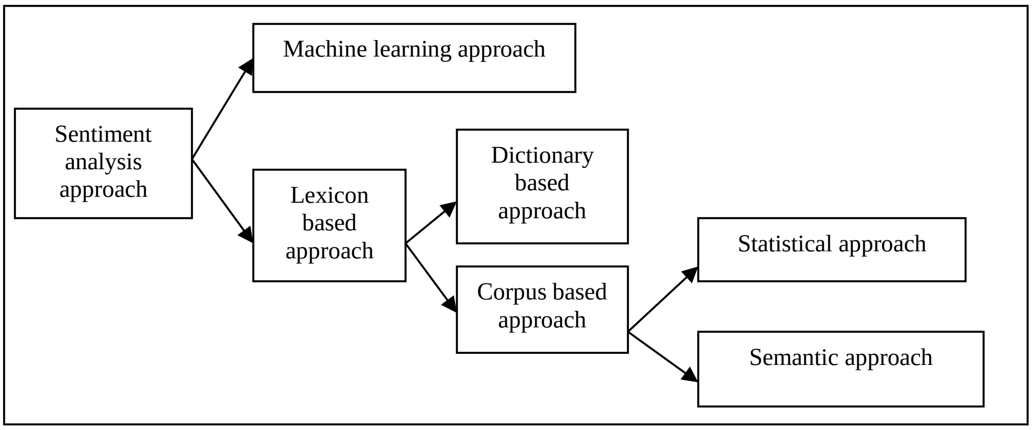

Figure 1. Sentiment analysis approaches

1) Dictionary based approach

A dictionary is created by taking a few words and can be refined by manual inspection periodically. An online dictionary, thesaurus or WordNet can be used to expand that dictionary till no new words can be added to that dictionary.

2) Corpus based approach

Corpus based finds sentiment orientation with context-specific words. Here, statistical approach considers positive, negative recurrence while the semantic approach assigns sentiment values to words by finding synonyms and antonyms to the appropriate word.

1.1.2 Machine learning based With the help of machine learning supervised algorithms, [3] the tools are made to detect sentiments without human input by analyzing the emotions in text. The algorithms such as naive bayes, support vector machine, deep learning, decision tree the model is created.

1.2 Sentiment analysis levels

Sentiment analysis can occur at [4] three different levels namely document level, sentence level and aspect/feature level. In the document level, the sentiment is extracted from the entire review, and a whole opinion is classified based on the overall sentiment as positive, negative and neutral. In sentence level, the objective sentence has factual information, while a subjective sentence has personal feelings, views, emotions, or beliefs. In aspect/feature level, the object features is extracted which is commented on by the opinion holder.

2. Literature Review

Sheikh Shah Mohammad Motiur Rahman *et al* [5] presents multiple ways of evaluations on Large Movie Review Dataset to achieve highest accuracy for classifying the sentiment of reviews. Two n-gram vectorization models - Unigram and Bigram individually were used with the features extraction method TF-IDF. Five ensemble machine learning algorithms namely Random Forest (RF), Extra Tree (ET), Bagging Classifier (BC), Ada Boost (ADA) and Gradient Boost (GB) were applied to determine which combination of vectorization models (Bigram, Unigram) along with feature extraction method (TF-IDF). From the observed results, it is noted that classifier ET gives the 100 % accuracy for all the combination and it is followed by RF with 99.5 % the least score was 80.5 % with ADA classifier. Also, all the classifiers gave improved results for the combination TF-IDF with Bigram for sentiment classification. But, Text preprocessing is not carried out also neutral review is ignored.

Ravinder Ahuja *et al* [6] analysed the impact of two features TF-IDF word level and N-Gram on SS-Tweet dataset of sentiment analysis. Six classification algorithms such as Decision tree, Support vector machine, K-Nearest Neighbor, Random forest, Logistic regression, Naive bayes were used for sentiment classification. Initially text preprocessing tokenization, normalization, stemming, lemmatization, stop word removal and noise removal was carried out. The discussed classifiers with Bigram and word level feature extraction was applied and noted the highest accuracy was obtained by random forest with 51% accuracy and the least accuracy for K-Nearest Neighbor and Support vector machine with 46 % on both word level and Bigram. The overall analysis reveals that Bigram gives improved result. However, the accuracy seems to be very less.

Bahrawi *et al* [7] conducted a sentimental analysis with Twitter data on six major US airlines with positive, negative and neutral using the Random Forest algorithm. In preprocessing, methods such as cleansing data, case folding, tokenization, stemming was carried out. To extract the feature or word vector TF-IDF method was used. Random forest algorithm was used to get the total number of positive, negative and neutral reviews on each airlines by the use of confusion matrix. The accuracy was obtained around 75.99%.

Karthika *et al* [8] implemented Random forest algorithm to assess rating from the online shopping flipkart.com based on the aspects of the product. Four feature are selected from the dataset namely product ID, product name, brand name and rating. The work was also assessed with SVM algorithm. Five classes (rating 1-5) was observed and the algorithm was applied. From the results, it is observed random forest gives highest accuracy 97%. But, the classification is done without sentiment analysis. The review is not analyzed with the rating.

Yassine AL Amrani *et al* [9] proposed a hybrid method based on Random forest (RF) and SVM for sentiment analysis on Amazon dataset. Initially, the dataset was preprocessed using text processing methods such stop word removal and tokenization that transform sentence to word. These words are formedto word vectors with Bag of Words using matrix. The work was assessed with hybrid method RFSVM and individual classifier RF, SVM and found the hybrid one outperforms the individual one with 83% accuracy. However, the accuracy should be improved.

ISSN: 2094-0343
2326-9865

Shoffan Saifullah *et al* [10] compared machine learning algorithms for Sentiment Analysis in Detecting Anxiety on COVD-19 data. The machine learning methods implemented include K-NN, Bernoulli, Decision Tree Classifier, Support Vector Classifier, Random Forest, and XG-boost for negative and positive data. TF-IDF method is used in prior to calculate the weight of each word. Among the classifiers Random Forest gave highest accuracy 84.99 %, but while it was combined with TF-IDF the accuracy was 82.63% hence it should be taken into consideration.

Suhasini *et al* [11] elucidated the Machine Learning classification approaches SVM with different feature extraction techniques to obtain a text analysis on detection of Fake News in Twitter Data using TF-IDF, n-grams. In preprocessing, letter casing, stop word removal, tokenizing, lemmatization was done. The proposed method applied SVM with the combination of TF-IDF and Bigram and proved with 90% accuracy. The accuracy was higher than SVM with TF-IDF and SVM with n-grams. However trigrams can be tried to further improve the accuracy.

3. Methodology

3.1 Existing methodology

The existing works [5] [6] applied text preprocessing techniques such as tokenization, normalization, stemming, lemmatization, stop word removal and noise removal. For feature extraction TF-IDF with bi- gram was used. Then Random forest algorithm with 100 trees was constructed for sentiment classification. But in this work the accuracy seems to be very low.

3.2 Proposed methodology

This work proposes TF-IDF [12] for word vector creation. Bi-Gram and Tri-Gram is used for sentiment analysis. Bag of Words (BOW) is created to get the frequency of word used and presented in word cloud form. Finally, random forest with feature subset based on weight is used for sentiment classification.

3.2.1 TF-IDF with pruning by rank

Frequency and words with a frequency less than 0.3 and higher than 1.0 will be pruned i order to get the most frequent word. This prior pruning ease the process while creating word vectors.

TF-IDF (Term Frequency-Inter Document Frequency)

It is a numerical statistic that gives how important a word is to a document in a corpus. The TF–IDF in Eq.(1) is a popular weighting scheme and its value increases proportionally to the number of times a word appears in the document.

TF-IDF is calculated as,

$$\mathrm{TF-IDF(t,d,D)} = \mathrm{tf(t,d)\,.\,idf\,(t,D)} \tag{1}$$

Term Frequency - tf (t, d) is the relative frequency of term t within document d.

The tf (t, d) in Eq.(1) is calculated as,

$$\mathrm{tf}\,(\mathrm{t},\mathrm{d}) = \frac{\mathrm{f}_{\mathrm{t,d}}}{\sum_{\mathrm{t'}\in\mathrm{d}} \mathrm{f}_{\mathrm{t',d}}} \tag{2}$$

Where, $f_{t,\,d}$ – count of a term in a document.

The Inverse document frequency – idf (t, D) in Eq.(1) is a measure of how much information the word provides.

idf (t, D) is calculated as,

$$\mathrm{idf}\,(\mathrm{t},\mathrm{D}) = \log\frac{\mathrm{N}}{|\{\mathrm{d}\epsilon\mathrm{D}:\mathrm{t}\epsilon\mathrm{d}\}|} \tag{3}$$

Where, N- total number of documents in a corpus (N=|D|) and $\{d\in D:t\in d\}|$ - number of documents in which the term *t* appears.

3.2.2 Text Preprocessing

Tokenization – The operator that breaks the whole document into sentence or to word by eliminating whitespace, punctuation, question mark, full stop etc..

Stopword removal - Generally articles and pronouns are classified as stop words.

Stemming – Truncate a word to its base or root.

n-Grams- A process to convert the whole sentence into sequence of words where n=1 means unigram i.e one word is extracted from the sentence, n=2 means bigram i.e a word with its frequent sequence of word in extracted. *n* is applicable to any number within a sentence. This work applies unigram, bigram, trigram for analysis.

Filter token by length- The length is fixed as minimum 4 to maximum 25 letters to avoid very small word as it doesn't provide enough knowledge.

3.2.3 Classifier – Random Forest with Weight based Feature Subset (RFWFS)

Initial process- Fine grained to general sentiment polarities

The Amazon product review dataset has fine grained polarities starts from the value 1-5 classes. To have a compact view, easy understanding, and for improved accuracy the five class dataset is transformed into three class (positive, negative, neutral) by using general if-then expression as,

$$\text{If rating} > 3 \text{ then sentiment } 3 \text{ else if rating} == 3 \text{ then sentiment } 2 \text{ else sentiment} = 1 \tag{4}$$

Where, 3, 2, 1 refers to positive, neutral and negative respectively. Here, 3- the positive reviews has good views about the products, 2 has neutral statement with average opinions and 1 has negative statements.

Weight based Feature subset (WFS)

Weight is assigned to each attribute based [13] on the label and its associated entries in each attribute. The attribute with highest weight is considered more relevant. This method enhances the accuracy of the classifier as the analysis is carried with most relevant attribute only.

Weight by Gini index- It is determined by deducting the sum of squared of probabilities of each class from one. The higher the weight of an attribute is considered as more relevant.

Gini index is calculated as,

$$\text{Gini index} = 1 - \sum_{i=1}^{n} (P_i^2) \quad (5)$$

Where P_i denotes the probability of an element being classified for a distinct class.

Random Forest Classifier (RF)

It is based on decision tree. Generally, decision [14] tree model has high variance and low bias which can give unstable output also it suffers from overfitting. Random forest combined with ensemble method 'bagging' that overcomes these high variance, low bias and overfitting. During training, it creates multiple decision tree and the output is the class that is selected by most of the trees. The advantages include, the algorithm is more accurate than decision tree with ensemble model, it avoids overfitting , high variance and low bias, also handle missing values, suits for both classification and regression, and offers feature/variable importance. But the limitations is Large number of trees make algorithm too slow hence fixing number of trees is crucial, and tuning of hyper parameters such as number of trees, maximum depth of tree, applying prepruning or pruning, criteria for splitting, voting strategy.

Working Principle

1. Choose *k* features from *m* features in the dataset.
2. With the highest Gini index select the root node among *m* features.
3. Split child node with class in the leaf for *n* times. Thus, forest is build with multiple decision trees.
4. Perform boot strapping (bagging) by combining all the decision trees to get the best results.

Algorithm of the proposed method RFWFS

Input	Amazon product review dataset
Output	Bag of words (BOW) in cloud form, Classified data as positive, negative and neutral.
Step 1	Creation of sentiments with three polarities positive, negative and neutral from fine grained polarities as in eq. (4)
Step 2	Weight based feature subset using Gini index as in eq.(5)
Step 3	Text preprocessing with stemming, stop word removal, n-grams, filter token by length.
Step 4	Feature extraction using TF-IDF with pruning as in eq. (1-3)
Step 5	Apply Random forest classifier with Gini index split criteria, 50 trees, maximum depth of the tree with 8, subset ratio for node split with 0.3, and voting strategy with majority vote.
Step 5	Performance assessment with classified dataset using the measures Accuracy, RMSE, Processing time.
Step 6	Visualization of top 50 frequent words in word cloud form created from Bag of Words.

Advantages of RFWFS – The advantages includes one more operator that prescribes the length (Filter token by length) of the word in Bag of words is used. Pruning is combined in TF-IDF that ease the filtering of words from the whole documents, weight based feature subset ease the classifier (Random forest) to have a good split based on Gini index. Minimized number of trees (50) in random forest classifier than existing that used 100 trees reduce the processing time.

4. Results and Discussion

4.1 Dataset description

The Amazon review dataset is taken from the Kaggle repository [15] with twenty one attributes. After feature selection method it is reduced to ten with most informative attributes and one special attribute named as sentiment to change the five classes to three classes.

The performance is assessed based on granularity model (the work is evaluated with range of data) and the range (500, 1000, 1500, 2000, 2500) is selected from the complete 9000 instances using stratified sampling.

4.2 Performance measure

ISSN: 2094-0343
2326-9865

For evaluation, three types of measures are used accuracy in percentage, RMSE in range and processing time in minutes[16].

1) Accuracy – It is defined as number of correctly classified instances to the total number of instances.

$$\text{Accuracy} = \frac{\text{Correctly classified instances}}{\text{N}} \tag{6}$$

2) Root Mean Squared Error (RMSE) – It calculates the error by taking the difference between predicted data and observed data to the whole number of data points. The range is between 0-1, in where, the '0' specifies the perfect model and the '1' specifies the erroneous model.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Predicted}_i - \text{Actual}_i)^2}{\text{N}}} \tag{7}$$

3) Processing time – The time taken to train and test the model.

4.3 Results

The application of Natural Language processing, sentiment analysis using machine learning approach with TF-IDF with random Forest Classifier is implemented in Rapidminer tool. The results is assessed for bigram and trigram for existing without feature subset and proposed with feature subset criteria in splitting the node. The results is assessed for variant range and the range is fixed with sampling technique.

4.3.1 Sentiment analysis using Bag of Words (BOW)

Bag of word is a creation of word vectors in where the words are listed with its frequency used in the documents in user reviews. The table may be presented in various formats such bar chart, word cloud etc. this work applies word cloud form and this analysis is useful to find the customer frequently used words for three type of polarities in sentiment analysis.

Figure 2. Word cloud for word level sentiment analysis

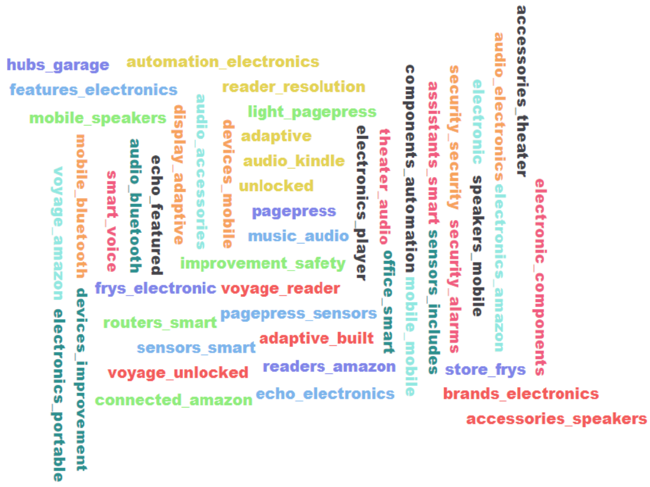

Figure 3. Word cloud for bigram level sentiment analysis

Figure 4. Word cloud for trigram level sentiment analysis

Figure 2, 3, 4 shows the most 50 frequent terms in word level, bigram level and trigram level respectively in word cloud form for the three polarities in combined. It shows the sentiment trend with user terms.

4.3.2 Performance analysis

The performance is assessed for RF, RFWFS with accuracy measure in Table I. The work is evaluated in granularity based (range of data starts from 500 to 2500) and this is because the enormous word or sequence of word generation for even a small range.

Table I. Granularity based Performance evaluation for Accuracy at Word, Bigram, Trigram levels

Range	RF - Accuracy in percentage	RFWFS - Accuracy in percentage

	Sentiment analysis Levels			Sentiment analysis Levels		
	Word	Bigram	Trigram	Word	Bigram	Trigram
500	80.35	84.21	85.33	84.55	87.12	89.22
1000	81.16	84.96	86.12	86.59	88.16	90.31
1500	83.02	86.25	88.52	87.47	90.01	92.12
2000	84.21	88.01	89.41	88.12	90.14	93.01
2500	85.95	90.12	92.03	90.12	92.36	94.80

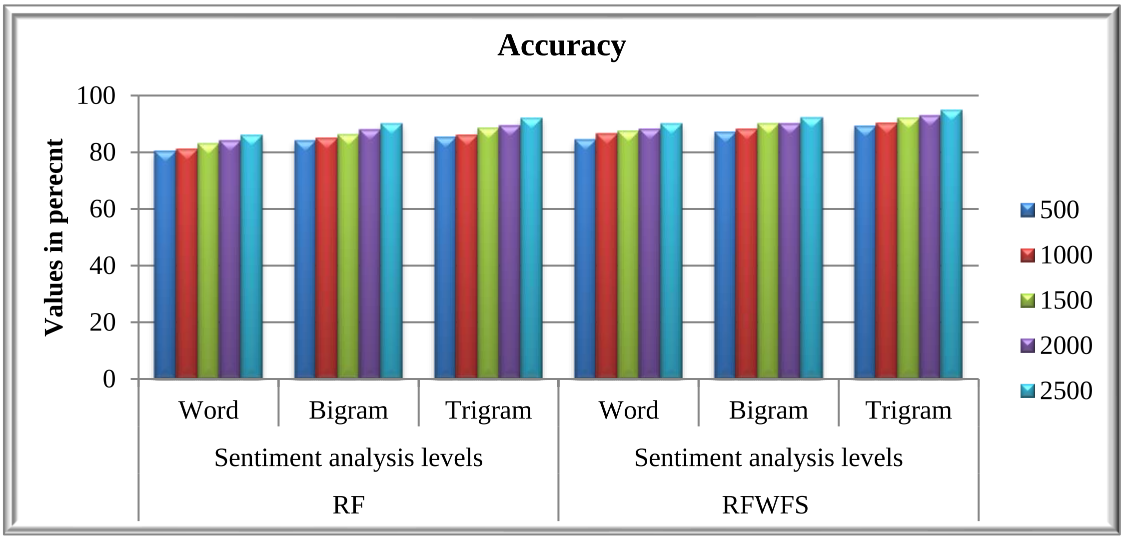

Figure 5. Granularity based Performance evaluation for Accuracy at Word, Bigram, Trigram levels

From the Table I, Figure 5, it is observed the overall existing work RF for all the levels (word, bigram, trigram) has less performance than the proposed method RFWFS. The maximum accuracy for RF is 92.03, which is for 2500 data with trigram level. The proposed method highest accuracy is 94.80, which is for 2500 data with trigram level. In addition, it is noted the highest accuracy obtained is in the order trigram level, bigram level and in last the word gram level for both methods RF, RFWFS. The higher the range of data, the higher the accuracy.

Table II. Granularity based Performance evaluation for RMSE at Word, Bigram, Trigram levels

Range	RF - RMSE in range between 0-1			RFWFS- RMSE in range between 0-1		
	Sentiment analysis Levels			Sentiment analysis Levels		
	Word	Bigram	Trigram	Word	Bigram	Trigram
500	0.149	0.125	0.115	0.121	0.112	0.098
1000	0.142	0.121	0.106	0.098	0.101	0.078

1500	0.131	0.103	0.097	0.105	0.089	0.063
2000	0.125	0.086	0.091	0.099	0.077	0.060
2500	0.104	0.071	0.061	0.086	0.063	0.042

Table II, shows the RMSE value with three levels for RF and RFWFS. The lowest range means the good model and the highest range resembles the model with less accuracy.

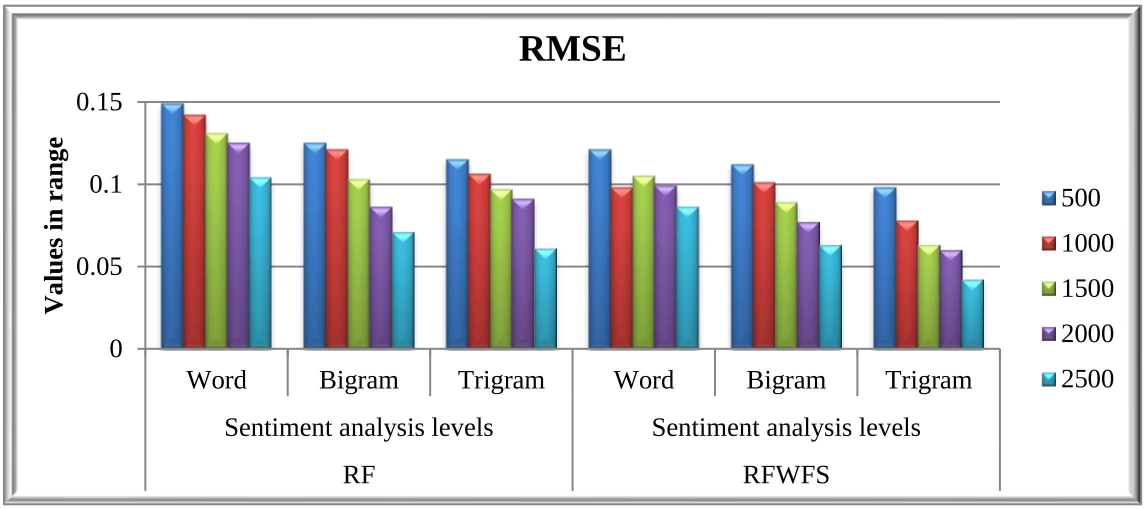

Figure 6. Granularity based Performance evaluation for RMSE at Word, Bigram, Trigram levels

From the Table II, Figure 6, it is observed the overall existing work RF for all the levels (word, bigram, trigram) has less performance than the proposed method RFWFS. The least RMSE (good model) for RF is 0.061 and the proposed method least RMSE is 0.042, which is for 2500 data with trigram level.

Table III. Granularity based Performance evaluation for Processing Time at Word, Bigram, Trigram levels

	RF-Processing time			RFWFS - Processing time		
	Sentiment analysis Levels			Sentiment analysis Levels		
	Word	Bigram	Trigram	Word	Bigram	Trigram
500	0.78	0.93	1.01	0.41	0.62	0.75
1000	0.98	1.12	1.39	0.55	0.78	0.91
1500	1.01	1. 33	1.50	0.87	1.04	1.21
2000	2.01	2.34	2.48	1.89	2.07	2.19
2500	3.12	3.46	3.51	2.97	3.12	3.25

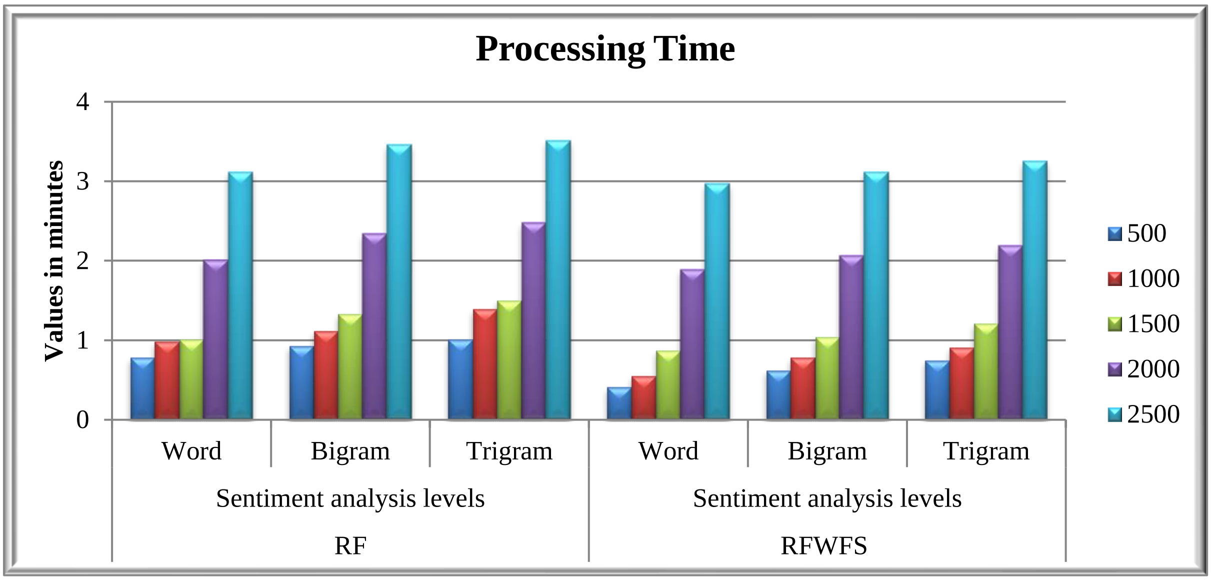

Figure 7. Granularity based Performance evaluation for Processing Time at Word, Bigram, Trigram levels

Table III and Figure 7, shows the Processing time with three levels namely word, bigram and trigram for RF and RFWFS. It is observed the overall existing work RF for all the levels (word, bigram, trigram) takes high processing time than the proposed method RFWFS. It is also observed that trigram (word with number of sequence 3) takes more time than other levels.

5 Conclusion

Sentiment analysis is one of the Natural Language processing (NLP) applications that responds to user text and assign polarities positive, negative, neutral or with fine grained ratings as 1-5 or 1-10. The sentiment polarities are assigned in prior using Natural Language Processing and it is analyzed or classified using variant approaches. This work uses machine learning approaches for sentiment analysis and classification. For sentiment analysis TF-IDF with pruning by ranking is used and for classification random forest with weight based feature subset RFWFS is used. Top 50 frequent words regarding three polarities namely positive, negative, neutral is generated in the form word cloud and classification performance with the existing RF is assessed with the evaluation measure accuracy, RMSE and processing time for three level of sentiment analysis such as word, bigram and trigram. From the overall observations, it is noted the proposed RFWFS gave high accuracy, less RMSE for trigram level sentiment analysis with high range of data. It is also noted in all the three level of analysis trigram needs more time to generate word vectors but it gives high level of accuracy and so the processing time is high.

In future, the work may be implemented for variant sectors. The work can be assessed with other evaluation measures. Prediction for a new review based on this proposed analysis may be tried.

References

[1] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142-150). Association for Computational Linguistics, 2011. https://ai.stanford.edu/~ang/papers/acl11-WordVectorsSentimentAnalysis.pdf

[2] https://www.sciencedirect.com/topics/computer-science/lexicon-based-approach

[3] https://www.globallogic.com/se/wp-content/uploads/2019/12/Introduction-to-Sentiment-Analysis.pdf

[4] Rao, Shivani, and Misha Kakkar. "A rating approach based on sentiment analysis." Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on. IEEE, 2017. https:// www.researchgate.net/publication/317418948_A_rating_approach_based_on_sentiment_analysis

[5] Sheikh Shah Mohammad Motiur Rahman, Md. Habibur Rahman, Kaushik Sarker, Md. Samadur RahmanNazmul Ahsan, M. Mesbahuddin Sarker, "Supervised Ensemble Machine Learning Aided Performance Evaluation of Sentiment Classification", IOP Conference Series,Journal of Physics, pp. 1-8, 2018. http://doi.org/10.1088/1742-6596/1060/1/012036.

[6] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja, "The Impact of Features Extraction on the Sentiment Analysis", International Conference on Pervasive Computing Advances and Applications – PerCAA 2019, Volume 152, pp. 341-348, Elsevier,2019. https://doi.org/10.1016/j.procs.2019.05.008.

[7] Bahrawi, "Sentiment Analysis using Random Forest Algorithm online Social Media Based", Journal of Information Technology and its Utilization, Volume 2, Issue 2, pp. 29-33, December-2019. https:// www.researchgate.net/publication /338548518_SENTIMENT_ANALYSIS_USING_RANDOM_FOREST_ALGORITHM_ONLINE_SOCIAL_MEDIA_BASED

[8] Karthika .P, Murugeswari .R, Manoranjithem .R, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm", 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), IEEE Publication, 2019. https://ieeexplore.ieee.org/document/8951367

[9] Yassine AL AMRANI, Mohamed LAZAAR , Kamal Eddine EL KADIRI, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis", The First International Conference on Intelligent Computing in Data Sciences, Elsevier, Volume 1227, pp. 511-520, 2018. https://www.sciencedirect.com/science/article/pii/S1877050918301625

[10] Shoffan Saifullah, Yuli Fauziah, Agus Sasmito Aribowo, "Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data", Jurnal Informatika, Volume 15, Issue 1, pp. 45-55, 2021. https://doi.org/10.48550/arXiv.2101.06353

[11] Suhasini V, Dr. N.Vimala, “A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data”, Turkish Journal of Computer and Mathematics Education Vol.12 , No.06, pp. 5710-5723, 2021. https:// turcomat. org/ index. php/ turkbilmat/article/view/10885

[12] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[13] Ahmad, F. K. Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels' Facebook Page. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9(1-2), pp. 13-17, 2017. https://core.ac.uk/download/pdf/288194723.pdf

[14] Hegde, Yashaswini, and S. K. Padma. "Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada." Advance Computing Conference (IACC), 2017 IEEE 7th International. IEEE, 2017. https://ieeexplore.ieee.org/document/7976894

[15] https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset

[16] S.Muthukumaran, Dr.P.Suresh “A Unified Framework of Sentimental Analysis for Online Product Reviews Using Enhanced Ant Colony Optimization Algorithm”, International Journal of Pure and Applied Mathematics [IJPAM], Vol. 119,pp 489-496, 2018,ISSN:1314-3395