International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Subjectivity Sentence Level Sentiment Analysis and Classification using Correlation Based Embedded Feature Subset using Machine Learning

D. Geethanjali*[1], Dr. P. Suresh[2]

Submitted: 06/05/2023 **Revised**: 17/07/2023 **Accepted**: 08/08/2023

Abstract*:* Sentiment analysis with sentence level gains importance, as each sentence has a sentiment polarity word. However, in case of product reviews sometimes the review might be lengthy that describes the product fully. Therefore Correlation Analysis based Random Forest with Subjectivity Sentence level Sentiment Analysis and Classification is proposed here. Machine learning techniques have been included into sentiment classification to increase its accuracy and effectiveness. Here two types of analysis such as (i) Sentiment polarity based model is taken for Subjectivity Sentence level Sentiment Analysis and (ii) Classification with the evaluation measures and the proposed method CARF-SSSAC proves its efficiency for Fine Grained and Sentiment polarity model. From the analysis it is proved that the Sentiment Polarity model gives highest accuracy 88.72% for the data range 5000.

Keywords: *Correlation Analysis, Fine Grained, Random Forest Classifier, Sentiment Analysis, Sentiment Polarity,*

1. Introduction

Millions of online consumer ratings of goods are posted on various e-commerce and social media sites, allowing advanced data analytics to facilitate the detection of hidden patterns and potentially offer designers priceless insights into the design of products [1]. A sentiment is a behaviour, belief, or conclusion brought on by a sensation. Opinion mining and sentiment analysis both study the ways individuals feel about particular things. One of the key issues with sentiment analysis is the categorization of sentiments polarization [2].

The analysis of opinions is beneficial to individuals and particularly important for e-commerce companies. The number of individuals purchasing products from the web is growing, and there are an ever-increasing number of results kept there. As a consequence, the number of user evaluations or postings is growing daily [3]. On three levels, namely the document level, the sentence level, and the expression level, sentiment evaluation is carried out [4]. The goal of sentiment assessment at the document level is to categorize the entire document as a positive or negative one. As a result, in this type of categorization, the distinction between objectivity and subjectivity classification is crucial [5].

Objectives: Sentence level Sentiment analysis using TF-IDF with pruning, Text processing operators and Sentiment Classification using improved decision tree model are discussed here. This work examines two types of sentiment analysis namely Fine Grained and Sentiment Polarity model on one dataset namely Amazon Product Review and uses Rapidminer Studio for the implementation and assessment of performance.

2. Related Works

Some of the machines learning approaches are considered here for Sentiment Analysis (SA) and classification. These methods undergo classifying sentiment polarity in a big database of online Instant Video evaluations. This research makes use of a substantial data set of 500,000 internet reviews. Strongly Negative, Negative, Neutral, Positive, and Strongly Positive are the five categories. 3 aspects of polarity when classifying reviews at the review-level, verbs, adverbs, and adjectives are also taken into account [6]. For the SA monitoring reviews of products on the internet, the Local Search Improvised Bat Approach based Elman Neural Network (LSIBA-ENN) was presented as an improved machine learning approach [7]. After data pre-processing additional processing is done through term weighting and feature selection. The LSIBA-ENN, which separates the tone of customer evaluations as positive, negative, and a neutral value, is produced using hybrid mutations.

A cluster based classification model for online product reviews was proposed [8]. This method utilizes Support Vector Machine (SVM) algorithm for classification of product reviews. Confusion matrix is used to generate the possibilities of every consumer purchasing the product. K-

[1] *Research Scholar, Department of Computer Science, Periyar University, Salem-11.*
ORCID ID : 0000-3343-7165-777X
[2] *HOD, Department of Computer Science, Salem Sowdeswari College, Salem-10.*
ORCID ID : 0000-3343-7165-777X
** Corresponding Author Email: ageethaa7@gmail.com*

means clustering technique is applied for clustering the data in to groups and SA is applied for extracting the features. For the SA of internet-based product reviews, the Deep Learning Modified Neural Network (DLMNN) model and the Improved Adaptive Neuro Fuzzy Inferences System (IANFIS) were suggested. For forthcoming estimation, IANFIS performs a measure of weighting and categorization on the product [9].

A sentence-level ML model for online review sentiment classification was put out [10]. This technique takes the subjective phrases out of the reviews and uses a classifier called NB to categorize each sentence as either positive or negative based on its word level features. An annotated collection of sentences known as the Bag of Sentences (BoS) is made up of the labelled sentences. For the purpose of classifying the polarity of sentences, SVM is used to the BOS. Sentence Level Features (SLF) and Domain Sensitive Features (DSF) consider the semantics of terms in both sentence level and subject level evaluations of products, which are the two ways for extracting textual characteristics [11]. For extracting SLF, a phrase's connotation disambiguation-based technique was developed. The SentiCircle-based approach was improved to produce DSF for each similarities used for producing SLF.

Word-level attention was used in the sentence-level attentiveness technique, which results in unneeded sequential patterns and more complex sentence interpretation. In order to simplify sequential structures, a Sentence to Sentence Attention Network (S2SAN) was suggested [12] employing multi-head self-attention. To take advantage of the semantic relationships between phrases in document-level sentiment analysis, a model for Neural Networks (NN) with concealed layers for sentence representation was developed [13]. Sentence matrices are learned in the first layer to represent sentence semantics, and relations between sentences are represented in content description in the subsequent layer.

The relevance levels of sentences in documents are automatically established by gate operations in a document-level sentence categorization model based on deep NN. This technique successfully establishes the polarity of a document; each sentence deserves to be treated with varying degrees of priority [14]. A Sentence-level emotion assessment approach based on Word Embeddings (WEMB) was proposed [15]. The WEMB word vectors are used to calculate the WEMB statement vectors. Sentence vector and polarity are used to train the model for classification. The model can predict the reverse orientation of an unlabeled text after training. Integrating the two text techniques for sentiment analysis resulted in the creation of a weight distribution algorithm [16]. The sentence vectors produced can both emphasize words with sentimental connotations despite preserving the textual details of the original sentences.

3. Proposed Methodology – CARF-SSSAC

Correlation Analysis based Random Forest classifier with embedded feature selection for Subjectivity Sentence level Sentiment Analysis, Classification (CARF-SSSAC) is proposed here. The work flow of subjectivity sentence level sentiment analysis and classification is shown in figure 1. The proposed work starts with Amazon dataset with star rating (1-5). In case of classification, the correlated attributes leads to same information hence it becomes redundant. To avoid this redundancy, this work applies Correlation analysis to remove correlated attribute and this output applied for sentiment analysis and classification.

Sentiment analysis uses TF-IDF with pruning method to create word vectors. To split the documents into subjectivity sentences this work put forth text-processing operators namely tokenization with linguistic sentences, transform cases and POS Tagging with ratio. This word vector dataset implemented for sentiment classification with stratified cross validation. Random forest algorithm with gain ratio and majority voting is used for sentiment classification and as it is one of the embedded feature subset method that gives the subset of feature with tree importance.

3.1. Correlation Analysis (CA)

The statistical association between two variables referred as correlation and used in data analysis and modeling to better gets the knowledge of degree of relationships between variables. It ranges from -1 to +1.

The three types of correlation are:

- Positive-Correlation: Both attributes change in the same direction and conclude the same thing.
- Negative-Correlation: Attributes change in opposite directions.
- No Correlation: Attributes are unrelated.

The Pearson Correlation is calculated using equation (1) and equation (2) for the attribute X, Y with the Standard Deviation (S.D) as,

$$\mathrm{Pearson - Correlation = \ Covariance(X, Y)/(S.D\ (X) * S.D\ (Y)\ )} \qquad (1)$$

$$Covariance\ (X, Y) = ((Summation\ (X - mean(X\_i\ )) * (Y - mean\ (Y\_i\ ))))/(S.D\ (X) * S.D\ (Y)) \qquad (2)$$

The threshold value is set as greater than 0.5. The significance of removing the correlated attribute is avoiding the redundant analysis on the attributes that adds same information. In case of predicting, the value of one variable with the aid of another variable i.e. high correlated attribute is used.

3.2. Term Frequency – Inverse Document Frequency (TF-IDF) with pruning

In information retrieval system, TF-IDF is a numerical statistic method that finds the importance of a word/sentence in a whole document or in a corpus. A weighting factor, which increases in proportion to the number of times a term appears, and so it finds the frequency of it. The proposed method uses TF-IDF with pruning by percent. This will measure the importance of the terms and prune relatively unimportant terms.

Term Frequency – The proportion of a weight of a term occurs in a document.

Inverse Document frequency – Represents the logarithmically scaled inverse function of the corpus that contains the term.

Therefore TF-IDF is calculated as in equation 3.

$$\mathrm{TF-IDF}\ (\mathrm{t,c,C}) = \mathrm{tf}\ (\mathrm{t,c}).\,\mathrm{idf}(\mathrm{t,C}) \qquad (3)$$

The high term frequency gets the higher weight and lower gets lower weight. Hence, it is proved to get the frequency of the term. TF-IDF with pruning by percent – It ignore the words that appear less than the specified percentage of the entire corpus. The threshold value taken as lower limit – 30%, higher limit 100%

Subjectivity classification – Stochastic Part of speech (POS) Tagging by ratio is used to extract subjectivity classification that includes frequency or probability. This operator keep the token with specified number of verbs, nouns, adjectives otherwise it omit the sentence. Minimum ratio of nouns, verbs, adjectives are given as parameter (0.5, 0.2, 0.1). Stratified K-fold cross validation (K=3) is applied to have equal proportion of class in each granular based analysis.

3.3. Sentiment Classification

Generally, Decision tree based algorithm is used for this kind of approach. The proposed classifier incorporated with feature subset uses Tree importance model. The tree split takes place on a feature to find the correct variable. Random forest classifier is an ensemble learning method that operates by constructing a multitude of decision trees. It combines idea and random selection of features. Random forests reduce the variance of the model by trained on different parts of the same training set.

Granular based Subjectivity Sentiment Classification: The sentiment polarity with positive, neutral, negative gives precise details about user opinion. The work considers three polarities. The polarity 3 states positive opinion, 2 states neutral opinion and 1 states negative opinion. Sometimes, neutral polarity means that, ‘no opinion can be derived from the sentence’. Here, the analysis initially split the sentence based on subjectivity classification using POS tag and so each sentence has the user sentiment opinion. Hence, neutral polarity here states that ‘user opinion with average/ok description’.

Sentiment Polarity Model transforms the fine-grained model by using equation 4.

$$\mathrm{If\ (review.rating)} > 3\ then\ sentiment = positive$$

$$\mathrm{elseif\ (review.rating)} == 3\ \mathrm{then\ sentiment} = 2$$
$$\mathrm{else\ sentiment} = 1$$
$$(4)$$

Algorithm of CARF-SSSAC:

```
Input   Amazon product review dataset
Output  Performance on Granular based Classification
        with Sentiment Polarity model.

Step 1  Load the dataset in Rapidminer tool.

Step 2  Correlation analysis using equation 1 & 2

Step 3  For Fine Grained model do

Step 4  Process Documents with TF-IDF with pruning by
        percent using equation 3

        4.1  Apply Tokenization, Transform cases, POS
             Tag by ratio of adjective, noun, verb.

Step 5  Stratified sampling with Cross Validation (K=3)

        5.1  Apply random Forest algorithm

        5.2  Assessment with performance measures
             using confusion matrix.

Step 6  For Sentiment Polarity model using equation 4
        repeat Step 4, Step 5

Step 7  Create word list to data table.

Step 8  Filter the samples in prescribed range from the
        data table and present in word cloud.
```

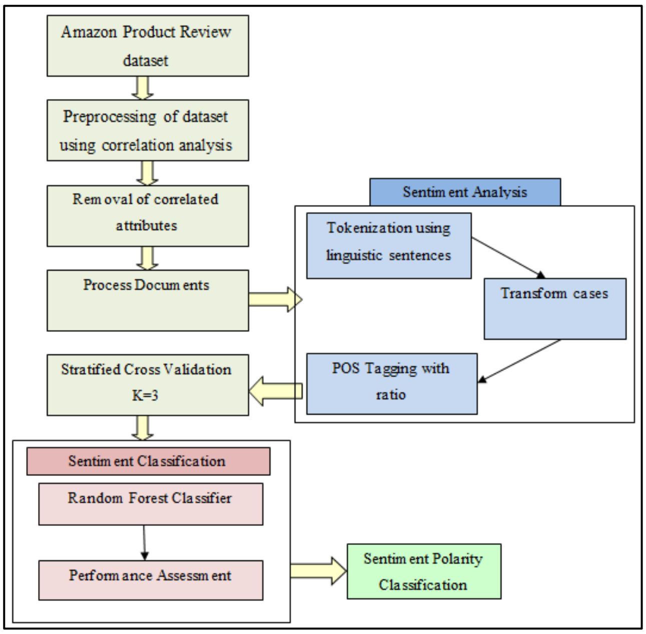
Amazon Product Review dataset
Preprocessing of dataset using correlation analysis
Removal of correlated attributes
Process Documents
Stratified Cross Validation K=3
Sentiment Analysis
Tokenization using linguistic sentences
Transform cases
POS Tagging with ratio
Sentiment Classification
Random Forest Classifier
Performance Assessment
Sentiment Polarity Classification

Fig. 1. Work flow for Subjectivity Sentiment Analysis, Classification

4. Results of Sentiment Analysis, Classification

The Amazon review dataset is taken from the Kaggle repository (https: //www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products) with twenty attributes and one sentiment polarity attribute

4.1. Subjectivity Sentences for Star Rating

Top frequent sentence displayed in word cloud form. Figure 2 shows the top most subjectivity sentence for rating 5 and the adjective with noun is filtered using POS tag to have the opinion of the user precisely. It comprise of the sentence such as 'awesome, better than I expected, great experience, awesome product, etc...' Mostly, each sentence has the word 'great'. It reveals the very positive opinion.

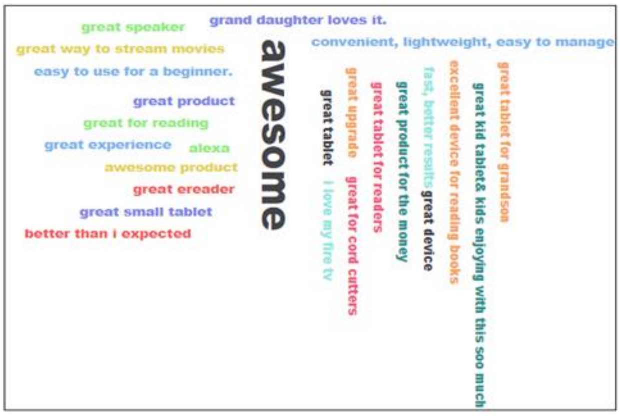

Fig. 2. Subjectivity sentence for star rating 5 – Very Positive

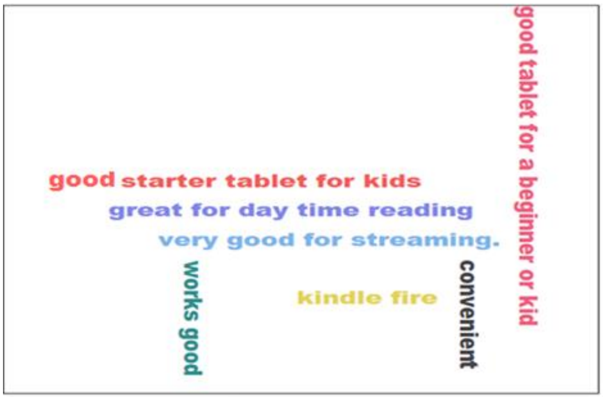
good starter tablet for kids
great for day time reading
very good for streaming.
works good
kindle fire
convenient
good tablet for a beginner or kid

Fig. 3. Subjectivity sentence for star rating 4 – Positive

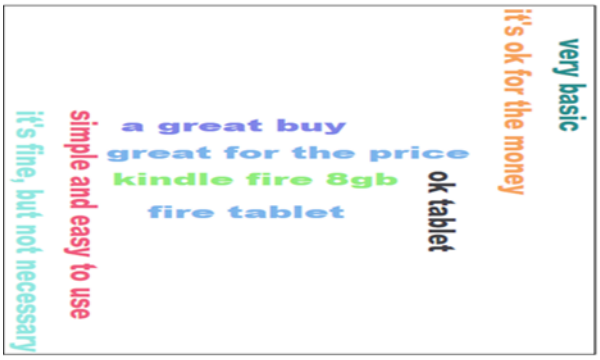

Fig. 4. Subjectivity sentence for star rating 3 – Neutral

Figure 3 shows the top most subjectivity sentence for rating 4 and it has sentence such as 'convenient, works good, good tablet for a beginner or kid, etc.'. It reveals the positive opinion. Mostly, each sentence has the word 'good'. Figure 4, shows the top most subjectivity sentence for rating 3 and it comprise of the sentence such as 'it's fine but not necessary, ok tablet, it's ok for the money, etc..'. It reveals the neutral opinion.

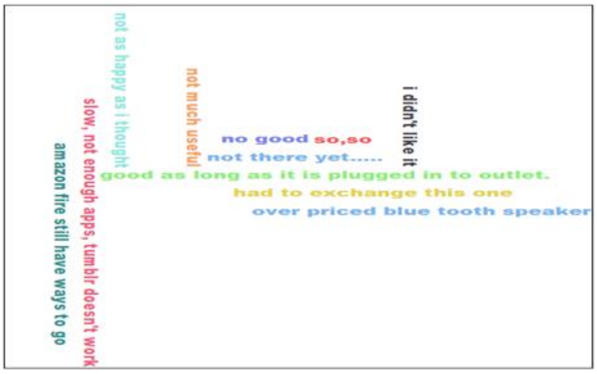
amazon fire still have ways to go
slow, not enough apps, tumblr doesn't work
not as happy as i thought
not much useful
no good
so,so
not there yet.....
i didn't like it
good as long as it is plugged in to outlet.
had to exchange this one
over priced blue tooth speaker

Fig. 5. Subjectivity sentence for star rating 2 – Negative

Figure 5, shows the top most subjectivity sentence for rating 2 and it has sentence such as 'slow, not enough apps, no good, etc..'. It reveals the negative opinion. Figure 6, shows the top most subjectivity sentence for rating 1 and it comprise of the sentence such as 'bad, hate it, horrible tablet, etc.'. It reveals the very negative opinion.

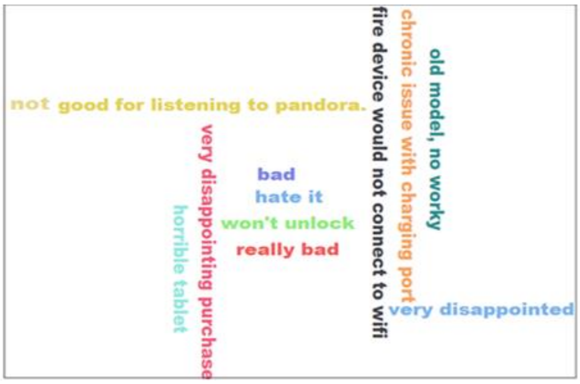
not good for listening to pandora.
very disappointing purchase
horrible tablet
bad
hate it
won't unlock
really bad
fire device would not connect to wifi
chronic issue with charging port
old model, no worky
very disappointed

Fig. 6. Subjectivity sentence for star rating 1 – Very Negative

Five types of measures were used in assessing the performance namely Accuracy estimated using equation 5, Root Mean Squared Error (RMSE) calculated using equation 6 and Processing Time is estimated by consider the total period for sentiment analysis and classification.

$$\sum_{i=1}^{n} \frac{Tp_i + Tn_i}{Tp_i + Tn_i + Fp_i + Fn_i} / n \quad (5)$$

$$\text{RMSE} = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(\text{Predicted}_{\mathrm{i}} - \text{Actual}_{\mathrm{i}})}{\mathrm{N}}}$$

(6)

4.2. Granular based Subjectivity Sentiment Classification with sentiment analysis - Sentiment polarity (positive, neutral, negative) model

The star rating (1-5) is given on user review of Amazon products is transformed to sentiment polarity model with positive (label 3), neutral (label 2) and negative (label 1). This analysis is to show how the model works for fine grained and sentiment polarity model. Table 1 has the performance values for Accuracy, in percentage. High accuracy obtains for the range 5000 with 88.72% for the proposed method CARF-SSSAC. The method CARF-SSSAC gives minimum 4.9 % higher accuracy than RF-SSAC. Figure 7 shows the analysis of accuracy and proves the proposed method CARF-SSSAC efficiency for the data range starts from 1000-9000. Among them, the values for the range 5000 gives better accuracy than RF-SSAC.

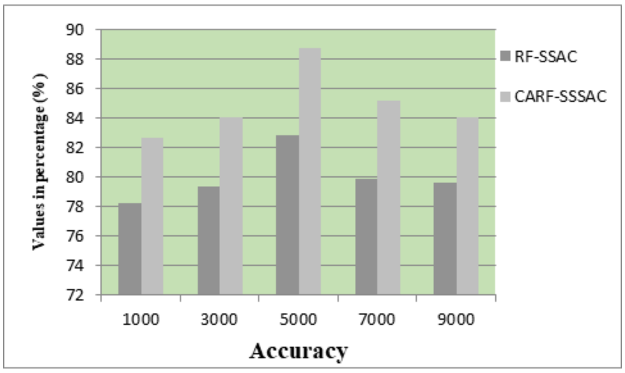

Fig. 7. Accuracy of Sentiment Polarity Model – Star Rating Model

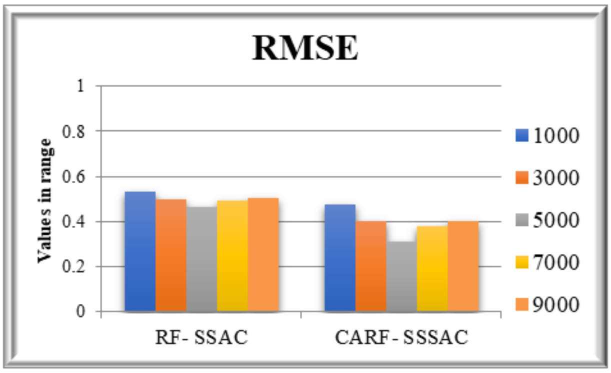

Fig. 8. RMSE for Sentiment Polarity Model

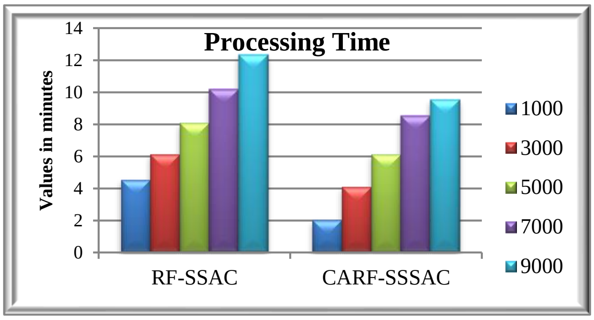

Fig. 9. Processing Time for Sentiment Polarity model

Table 1. Values obtained for Accuracy, RMSE and Processing Time for Sentiment Polarity Model

Data Range	Accuracy		RMSE		Processing Time	
	RF-SSAC	CARF-SSSAC	RF-SSAC	CARF-SSSAC	RF-SSAC	CARF-SSSAC
1000	78.22	82.68	0.532	0.478	4.55	2.05
3000	79.31	84.03	0.499	0.399	6.15	4.10
5000	82.78	**88.72**	0.462	**0.312**	8.10	**6.15**
7000	79.82	85.12	0.492	0.381	10.20	8.55
9000	79.56	84.05	0.502	0.402	12.35	9.55

From the values in Figure 8, it is observed that the RMSE values for CARF-SSSAC is less for the range 5000 that implies the model works better and above 5000 again the RMSE value begin to increase and hence it shows minimum 5000 range is needed for better analysis in case of fine grained model. Figure 9 shows the processing time for each range and the time increases for high data range. Hence, the granular based method is useful to identify the exact range for better accuracy with moderate processing time. Data range 5000 proves the analysis with better results. The total processing time for CARF-SSSAC and RF-SSAC is less for sentiment polarity model than the fine grained model.

5. Conclusion

The proposed CARF-SSSAC model is applied with sentiment polarity model that uses only three polarities (positive, neutral negative) for all types of sentences for providing error free analysis. Decision tree based classifier grow more trees with more number of attributes hence to avoid the complexity of Random Forest Classifier, this work initially removes the correlated attributes and then the classifier is substituted with the parameters gain ratio, number of trees, maximal depth, subset ratio for the embedded based subset, and majority voting strategy. Sentiment polarity analysis model is taken for Subjectivity Sentence level Sentiment Analysis, Classification with the evaluation measures and the proposed method CARF-SSSAC proves its efficiency for Fine Grained and Sentiment polarity model. In the overall analysis, Sentiment Polarity model gives better results with highest accuracy 88.72% for the data range 5000.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] R. Ireland and A. Liu, “Application of data analytics for product design: Sentiment analysis of online product reviews,” CIRP Journal of Manufacturing Science and Technology, vol. 23, pp. 128-144, 2018.

[2] X. Fang and J. Zhan, “Sentiment analysis using product review data,” Journal of Big Data, vol. 2, no. 1, pp. 1-14, 2015.

[3] K. S. Kumar, J. Desai and J. Majumdar, “Opinion mining and sentiment analysis on online customer review,” In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4, 2016.

[4] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, “Sentiment analysis on product reviews using machine learning techniques,” In Cognitive Informatics and Soft Computing: Proceeding of CISC 2017, pp. 639-647, Springer Singapore, 2019.

[5] M. D. Devika, C. Sunitha and A. Ganesh, “Sentiment analysis: a comparative study on different approaches,” Procedia Computer Science, vol. 87, pp. 44-49, 2016.

[6] S. Kausar, X. U. Huahu, W. Ahmad and M. Y. Shabir, “A sentiment polarity categorization technique for online product reviews,” IEEE Access, vol. 8, pp. 3594-3605, 2019.

[7] H. Zhao, Z. Liu, X. Yao and Q. Yang, “A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach,” Information Processing & Management, vol. 58, no. 5, pp. 102656, 2021.

[8] P. Vijayaragavan, R. Ponnusamy, R and M. Aramudhan, “An optimal support vector Machine based classification model for sentimental analysis of online product reviews,” Future Generation Computer Systems, vol. 111, pp. 234-240, 2020.

[9] P. Sasikala and L. Mary Immaculate Sheela, “Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS,” Journal of Big Data, vol. 7, pp. 1-20, 2020.

[10] A. Khan, B. Baharudin and K. Khan, “Sentence based sentiment classification from online customer reviews,” In Proceedings of the 8th International Conference on Frontiers of Information Technology, pp. 1-6, 2010.

[11] K. Gurumoorthy and P. Suresh, “Comparative Study of Recent Algorithms Used in Natural Language Processing “Parisodh Journal, ISSN NO: 2347-6648, vol. 9, no. 2, February 2020.

[12] B. S. Rintyarna, R. Sarno, R and C. Fatichah, “Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks,” Journal of Big Data, vol. 6, no. 1, pp. 1-19, 2019.

[13] P. Wang, J. Li and J. Hou, “S2SAN: A sentence-to-sentence attention network for sentiment analysis of online reviews,” Decision Support Systems, vol. 149, pp. 113603, 2021.

[14] S. Muthukumaran, P. Suresh, and J. Amudhavel, “A State of art approaches on Sentimental Analysis Techniques Journal of Advanced Research in Dynamic and Control Systems,(JARDCS) ISSN NO:1943023X, vol. 09, pp 1353-1370, August 2017.

[15] G. Rao, W. Huang, Z. Feng and Q. Cong, “LSTM with sentence representations for document-level

sentiment classification," Neurocomputing, vol. 308, pp. 49-57, 2018.

[16] G. Choi, S. Oh and H. Kim, "Improving document-level sentiment classification using importance of sentences," Entropy, vol. 22, no. 12, pp. 1336, 2020.

[17] T. Hayashi and H. Fujita, "Word embeddings-based sentence-level sentiment analysis considering word importance," Acta Polytechnica Hungarica, vol. 16, no. 7, pp. 7-24, 2019.

[18] H. Liu, X. Chen and X. Liu, "A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis," IEEE Access, vol. 10, pp. 32280-32289, 2022.

[19] Gandhi, L. ., Rishi, R. ., & Sharma, S. . (2023). An Efficient and Robust Tuple Timestamp Hybrid Historical Relational Data Model. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 01–10. https://doi.org/10.17762/ijritcc.v11i3.6193

[20] Pérez, C., Pérez, L., González, A., Gonzalez, L., & Ólafur, S. Personalized Learning Paths in Engineering Education: A Machine Learning Perspective. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/107