INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 11, Issue - 08, August - 2025

DOIs:10.2015/IJIRMF/202508020 --:-- Research Paper / Article / Review

Feature-level Rating System for Telugu Amazon Customer Reviews using Hybrid XGB-RF Classifier

Dr. Palli Suryachandra
Lecturer in Computer Science, Government Degree College, Pattikonda, Andhra Pradesh
Email: chandravj111@gmail.com

Abstract: *Sentiment Analysis (SA) is the opinion extraction that studies the attitude, sentiments, opinions and emotion of people. The huge number of active users will provide information about their opinions in E-Commerce websites which gives the effective review about products. Various sentiment analysis approaches were presented to classify the sentiments as positive, negative, and neutral. However, the existing methods are not effective and ignore the subtle sentiment classification among various text. But, the supervised learning methods was achieved some satisfactory performance on dimensional sentiment analysis, although they needed multiple labels to train the system, that are cost effective and consumes time for annotation of data. In order to overcome such an issue, proposed an Hybrid XGB-RF (XG Boost-Random Forest) Classifier method for the sentimental analysis of the Amazon telugu reviews. The proposed, Hybrid XGB-RF classifier is used to classify the product reviews and ratings are generated. When the ratings are generated the reviews are categorized as Terrible (1star), Poor (2stars), Average (3stars), Very Good (4stars) and Excellent (5stars).The proposed Feature Level Rating with Hybrid XGB-RF obtained accuracy of 97.22 % better when compared to the existing SentiPhraseNet model that obtained 78% of accuracy.*

Key Words: *Sentimental Analysis, E-Commerce, Enhanced Hybrid XGB-RF , Structured data, Amazon Telugu Reviews.*

1. INTRODUCTION:

SA is also known as opinion extraction, it is the field of study which examines the people's attitude, sentiments, opinion, and emotion [1]. SA is often researched domain in Natural Language Processing (NLP), it has multiple applications in the field of data mining such as extraction of information, summarization of data, answering the questions, and recommendation system [2]. The fastest development of this field is because of multiple usage of social media on the internet such as reviewing the products, blogging, discussion of forums, etc. [3] The huge number of active users will provide information's about their opinions in social media which gives the effective review about product or movies [4]. SA is applied to classify the reviews in three categories such as positive, neutral and negative [5]. SA helps to study the data which are opinioned and to extract some important features, that helps other users to make decision [6].The social media users write reviews in English or in their native language. So, the sentiment analysis was carried out in English and non-English textual data. Due to the complexity of textual content in Indian languages, a good system to perform sentimental analysis is required [7]. Various sentiment analysis approaches were presented to classify the sentiments like positive, negative and neutral also categorizes the emotions like joy, anger and sad from texts[8].The standard methods like SVM, random forest were utilized for sentiment analysis it provided considerable performance [9]. However, the existing methods are not effective and ignores the subtle sentiment classification among various text. But, the supervised learning methods was achieved some satisfactory performance on dimensional sentiment analysis, although they needed multiple labels to train the system, that are cost effective and consumes time for annotation of data [10].In order to overcome such an issue, proposed an Hybrid XGB-RF Classifier method for the sentimental analysis of the amazon telugu reviews. Initially, the amazon telugu reviews are collected and required features are selected. Then, preprocessing is carried to convert the unstructured data into structured data. Further, relevant sentences are extracted from the structured data. The contribution of the proposed Hybrid XGB-RF classifier is used to classify the product reviews and ratings are generated. The proposed Feature level rating with hybrid

XGB-RF showed well decision making it kept both the customer well informed and improved the product buying by other customers. As the customer reviews will be based on different customers and their different interest in features. The proposed feature-level ratings made buying decisions personalized. The proposed Feature Level Rating with Hybrid XGB-RF obtained accuracy of 97.22 % better when compared to the existing SentiPhraseNet model that obtained 78% of accuracy. When the ratings are generated the reviews are categorized as Terrible (1star), Poor (2stars), Average (3stars), Very Good (4stars) and Excellent (5stars).

The paper is organized as follow, survey of the existing methods is given in section 2, the proposed Hybrid XGB-RF classifier method for the sentimental analysis of amazon telugu reviews is explained in section 3, experimental results are discussed in section 4. The conclusion of this research is given in section 5.

2. Literature Review:

Pravarsha Jonnalagaddaet.al. [11] developed rule based method for sentiment analysis of telugu language by using telugu sentiwordnet. Initially, annotated corpus for telugu sentiment analysis data sets were collected. Then, Parts of Speech (PoS) tagger is used to divide the sentences into various PoS and also identified and removed the stop words. At last, sentiments were classified by using sentiwordnet, for unknown words sentiments were classified by using textblob. The rule based method has reduced the difficulty of natural language processing in telugu language. However, the rule based method has not provided the correct sentiments for the entire sentences.

Regatte Yashwanth Reddy et.al. [12] developedaspect based sentiment analysis for telugu movie reviews. In aspect based method, the data were collected from the websites of different movie reviews, preprocessed and annotated corpus was created. The corpus includes annotated data for the identification of terms, classification of polarity and categorization. Then, utilized deep learning approaches to demonstrate the reliability and usefulness. The aspect based method has successfully solved the sequence label tasks such as POS tagging and entity recognition. However, the performance of aspect based model was low because the size of vocabulary was less.

Kumar RG et.al. [13]developed sentiment analysis approach by using Bi-directional Recurrent Neural Network (BRNN) for telugu movies. Initially, tweets related to telugu movies were collected. Then, character trigram is used to solve the grammatical errors present in the sentences and morphological analyzer is employed for the vector representations. At last, the inputs are given to BRNN for the sentimental analysis of the reviews. The developed method represents the high and low resources in the same space and classified them on the basis of similarities among the annotated tags. However, the representation of texts and optimization method need to be effective to increase the performance of the sentiment analysis.

Santosh Kumar Bhartiet.al.[14] developed hyperbolic features based sarcastic sentiment identification method for telugu conversation sentences. Initially, the telugu sentences were collected from the different sources were collected and annotated, PoS tag were identified for every sentences. The tagged data was used for classification of sarcastic sentences by using hyperbolic features such as interjection, intensifier, question marks and exclamation marks. The hyperbolic feature based method showed a higher accuracy in the detection of sarcastic sentiments. However, the hyperbolic feature method considered lesser datasets.

Abhinav Garapatiet.al. [15] developed sentiphrasenet method for the sentimental analysis of telugu language. Initially, annotated corpus for telugu sentiment analysis datasets were collected. Then, Parts of Speech (PoS) tagger is used to divide the sentences into various PoS and also identified and removed the stop words. Further, the phrases were identified by using sentiphrasenet on the basis of bigram and trigram rules. Finally, sentiments were classified by using textblob. However, the sentiphrasenet method has not provided the sentiments for the entire sentences.

3. Proposed Method

In this research, proposed an Hybrid XGB-RF Classifier method for sentimental analysis performance by using amazon telugu reviews. The data is collected fromamazon.com which includes product reviews based on telugu language clothes, mobiles phones, laptops etc. The data extracted from the amazon website includes product level. After collecting the datasets, feature selection process is carried out, the features on which the people is interested are identified. The words having frequency lesser than 0.02% are neglected from the entire reviews in the datasets, because those reviews rarely includes features related words. Then, preprocessing is undergone to convert the unstructured data into structured data. Then, preprocessing is carried to convert the unstructured data into structured data. Further, relevant sentences are extracted from the structured data. Then, Hybrid XGB-RF classifier is used to classify the product reviews and ratings are generated. When the ratings are generated the reviews are categorized as

Terrible (1star), Poor (2stars), Average (3stars), Very Good (4stars) and Excellent (5stars). The block diagram of proposed method is shown in Figure1.

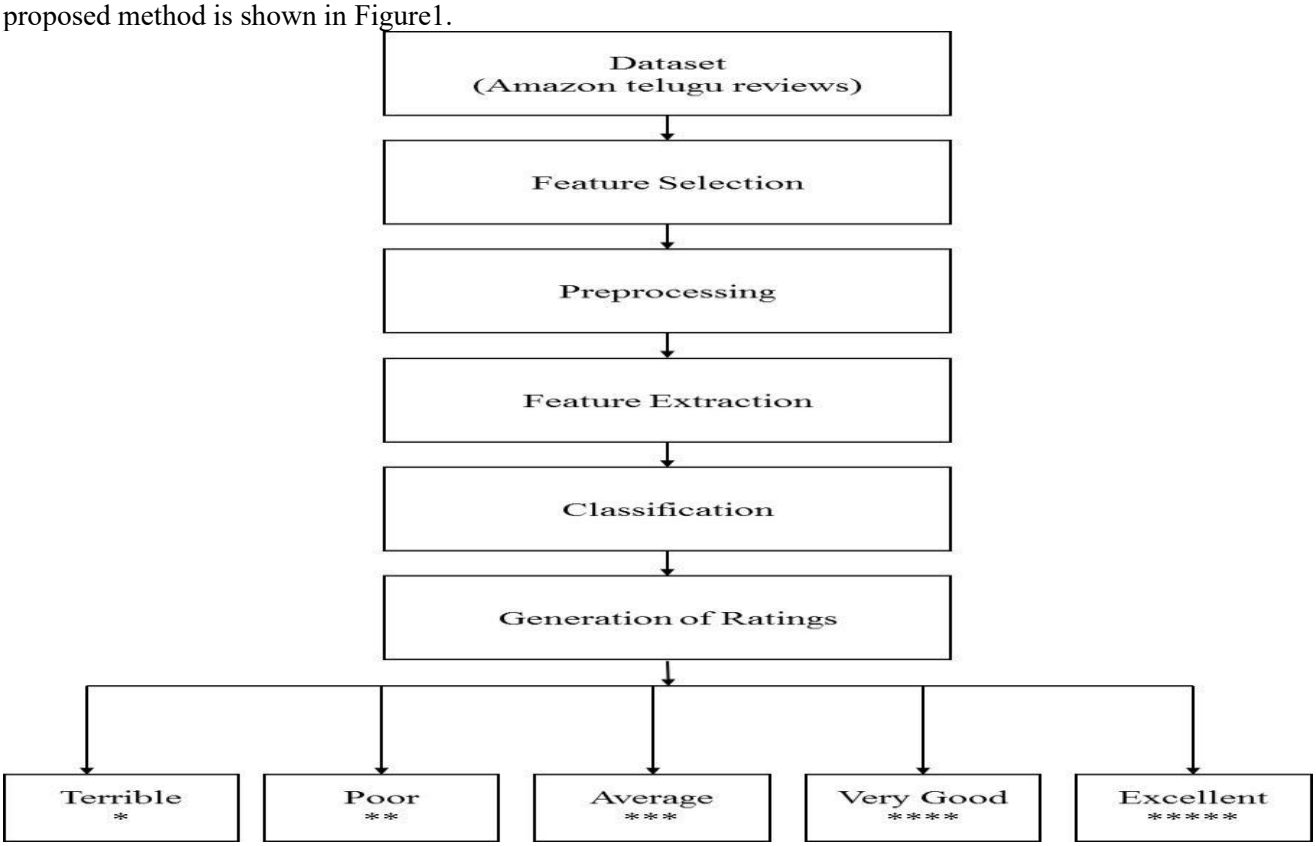

Figure 1. Block diagram of proposed sentiment analysis method.

3.1 Dataset

Dataset The data is collected from amazon.com which includes product reviews based on telugu language clothes, mobiles phones, laptops etc. The data extracted from the amazon website includes product level. The Amazon.com has various different types of products such as camera, phone, jewelry, ratings, reviews, and voting for reviews. The reviews include reviewer id, product id, review texts, ratings and review timings. Every reviews in the amazon.com consists of text comments that will be posted by product users with the accurate timestamp. The reviews includes the scales having rating up to 5 stars which is associated with textual comments. The 5stars review will be as a scalerates for each product along with the label categories and descriptions. The telugu review from the amazon datasets were utilized in the proposed method and collected data sets will be undergone for feature selection.

3.2 **Feature Selection**

After collecting the datasets, feature selection process is carried out. Let us consider, collecteda datasets of N and features related words that are denoted as W is shown in equation (1). By manually considering the frequency of words for the whole customer reviews on the products. From this the features on which the people is interested are identified. The words having frequency lesser than 0.02% are neglected from the entire reviews in the datasets, because those reviews rarely includes features related words.

$$W = \{w_1, w_2, \dots\dots\dots, w_N\} \quad (1)$$

The corresponding frequency of the features related words form is denoted as Z, and is shown in equation(2). The particular features related to features related words are grouped into multiple feature.

$$Z = \{z_1, z_2, \dots\dots\dots, z_N\} \quad (2)$$

The Relationizer function is defined as R, that returns a group of every related words of wi in W and is shown in equation (3). The Relationizer function notation is manually taken and further defined an feature datasets that is denoted by F as a set of distinct groups of similar feature words which is shown in equation (4).

$$R(W, w_i) \quad (3)$$

$$F = \{R(W, w_i)\}_{i\epsilon\{1,\dots,N\}} \quad (4)$$

Every words in W is iterated and formed a distinct group of similar words by using Relationizer function. The different related words will form the similar groups, others will be removed to make the groups that are lefted distinct.

Let F_k be the kth feature words groups in the F feature datasets, any F_k features word is taken to find the entire feature words group which is called as feature keywords. The often feature words in the group is selected as feature keyword to assign the name to the groups as shown in equation (5)

$$F_k \leftarrow w_i | i = \max(\{Z(i) | w_i \epsilon F_k\}) \quad (5)$$

Where, the feature group is given a name with word which has the highest frequency in set. The abusing notation bit $F_k$means both kth feature group and keyword based on the context.

1.2 Pre-processing

After selecting the features from the datasets, the pre-processing is undergone. The review data are unstructured, in pre-processing step the unstructured data is converted to structured data. The important data are extracted, corrected and disintegrated. The data that are present after removing the useless characters are important data. By retaining the characters such as punctuations, emoticons and word formation, the other characters including numbers and entries which are empty is removed.

3.3 Feature extraction

After converting the unstructured data into structured data, relevant sentences are extracted.Let product review data be D which includes m important comments it is shown in equation (6). The corrected data means the obtained data after correcting the similar feature words problem and correcting the spelling in the important that are extracted.

The python NLTK package is used to separate the review into words and to use them separately, package helped separating the reviews from words as token by neglecting the space. The period (.) is also considered as tokens which is useful for splitting the comments to sentence.

$$D = \{C_1, C_2, \dots\dots, C_m\} \quad (6)$$

Every comments is represented as tokens as shown in equation (7). Any useful token $t_j^i$is considered as shown in equation (8).

$$C_j = \{t_j^1, t_j^2, \dots\dots\dots t_j^{|C_j|}\} \quad (7)$$

Where, $|C_j|$ is the total tokens that are obtained in the comments C_j, t_j^i is the ith token in the jth comment.

$$t_j^i = \begin{cases} F_k, \ if\ S(t_j^i \epsilon F_k\ or\ t_i \epsilon F_k, \forall\ F_k \epsilon F \\ S(t_j^i), \quad Otherwise \end{cases} \quad (8)$$

Where, $S(.)$ Is the spelling correction function. The tokens after or before correcting the spellings will match with the members of any feature word groups, that is replaced with the feature keywords of the set or with corrected tokens. So, the data that are extracted, corrected and disintegrated the reviews of a product has become structured. Thus, t_j^i is the ith token of jth comments of D.

The continuous tokens that are present in the reviews are grouped as sentences as shown in equation (9) of a group of sentences X_j that are obtained from C_j.

$$Xj = \{(t_j^u, \dots, t_j^v) | (t_j^v, t_j^{u-1}) = '.', (t_j^u, \dots, t_j^{v-1}) \neq '.', (u, v) \in \{1, \dots, |C_j|\}\ and\ v > u\} \quad (9)$$

Where, a set of continuous tokens as a sentence is called if the last and previous to the first token are of the period (.), and if other tokens in the set are not periods of the form (.). But, not every sentence are related to rating based on features. The sentence is defined of the form X_j^l, lth sentence in the X_j is related or not as shown in the equation (10).

$$P(X_j^l) = \begin{cases} 1, \ if F_k \in X_j^l\ for\ any\ F_k \in F \\ 0, Otherwise \end{cases} \quad (10)$$

Where, $P(.)$is the relevant function for sentence which gives output as 1, if any keywords are present in the sentence.

3.4 Classification using Hybrid XGB-RF classifier

XGBoost is used for training gradient-boosted decision trees and other gradient boosted models. Random Forests use the same model representation and inference, as gradient-boosted decision trees but is used with different

training algorithm. After extracting the relevant features product review classification is carried out by using the Hybrid XGB-RF classifier. The XGBoost algorithm is effective for a wide range of regression and classification predictive modeling problems. It is an efficient implementation of the stochastic gradient boosting algorithm and offers a range of hyperparameters that give fine-grained control over the model training procedure. Although the algorithm performs well in general, even on imbalanced classification datasets, it offers a way to tune the training algorithm to pay more attention to misclassification of the minority class for datasets with a skewed class distribution.XGBoost is short for **Extreme Gradient Boosting** and is an efficient implementation of the stochastic gradient boosting machine learning algorithm.

Although the XGBoost algorithm performs well for a wide range of challenging problems, it offers a large number of hyperparameters, many of which require tuning in order to get the most out of the algorithm on a given dataset.The implementation provides a hyperparameter designed to tune the behavior of the algorithm for imbalanced classification problems; this is the **scale_pos_weight** hyperparameter.XGBoost is trained to minimize a loss function and the "*gradient*" in gradient boosting refers to the steepness of this loss function, e.g. the amount of error. A small gradient means a small error and, in turn, a small change to the model to correct the error. A large error gradient during training in turn results in a large correction. The small gradient is the small error or correction to the model. The larger gradient is the large error or correction to the model.The gradients are used as the basis for fitting subsequent trees T added to boost or correct errors made by the existing state of the ensemble of decision trees T. The advantage of the Hybrid XGB-RF classifier is it provides a highly efficient implementation and access to a suite of model hyperparameters designed to provide control over the model training process. Then, the most important factor behind the success of Hybrid XGB-RF is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a multiple machine and scales to billions of examples in distributed or memory-limited settings. The sample weights ratio class and average of classes is calculated by using equation (11) and (12)

$$Sample\ weight\ ratio\ class[i] = \frac{Number\ of\ sample\ class[i]}{Total\ number\ of\ samples - Number\ of\ sample\ class[i]} \quad (11)$$

$$Average\ of\ classes = Mean(Sample\ weight\ ratio\ class[i]) \quad (12)$$

Where, i=1,2…5.

The random forest is a non-parametric classification technique which effectively reduces the issue of probability density complexity. In random forest, each tree is assumed as a distinct classifier which are used for attaining better decision making. Further, the growth rules of each tree is examined to develop a robust random forest classifier.

An advantage of using hybrid XGBoost-Random forest model as a final model and make predictions for classification. The XGBoost random forest ensemble is fit on all available data, then the predict() function can be called to make predictions on new data.An ensemble classification technique (random forest) work based on the principle of bagging which utilizes decision tree as a basic classification technique. Initially, the extracted feature vectors are randomly sampled for training sets N.Then, select the sub feature vectors from the extracted feature vectors if$m(m < M)$, where M is represented as the extracted feature vectors. In addition, select m feature vectors from the M features and then split the nodes by using best spilt on the m dimensional feature vectors. Therefore, the random forest classifier's error rate depends on two factors such as the correlation among trees need to be low to diminish the error rate, and the tree strength need to be high to diminish the error rate.

Using the Random forest, the generalization error is bound to be dependent mainly on the tree strength that achieves correlation among them. Based on the maximum voting approach, the elements such as i and j are voted in the RF model thereby classifies the reviews using the following Eq. (13).

$$prox(i,j) = \frac{\sum_{t=1}^{n\ tree} I(h_t(i) = h_t(j))}{n\ tree} \quad (13)$$

where I(·)represents the indicator function,
h_t represents the tree of the forest
h_t(i) iis the value which is predicted for all the values of i
If $prox(i,j)$= 1 then the classes i and j of the same classes are classified

Therefore, RF provides the important rank which will be variable and that is used to select the important features. Pseudo code of random forest classifier is given below.

Pseudo code for the Hybrid XGB-RF algorithm

Input: D: Telugu Amazon Review dataset
N : The Number of Trees in XGBOOST
Output: Discrete Data

For w_i=0; $i < D$; i++
Preprocess the data for the text participated
Extract the Similar text key words
Using Hybrid XGB-RF discretize the feature
Calculate the mean value Sample weight ratio class
End for

3.5 Generation of ratings

The relevant extracted features will go through every sentence for particular feature F_k, emotion of the sentence is extracted to score the sentence. So, sentiment analysis for each sentences are extracted, it acquires emoticons while performing the analysis. The compound score is used to analyze the sentiment scores, it ranges within 1 and -1. The range is divided into 5 equal parts and assigned the ratings. let the function that calculates the scores of sentiment and assigns the rating be$\psi(.)$. Then cumulative rating $Q(.)$ is calculated for every features by the data of product review D as shown in equation (11).

$$Q(F_k) = \sum_{C_j \epsilon D} \sum_{X_j^l \epsilon X\_j}^{p(X_j^l)=1} \psi(X_j^l) \times \delta(F_k \epsilon X_j^l) \times \left(\phi(C_j) + 1\right) \quad (13)$$

Where, $\delta(.)$ is the logical function to find the concerned features present in the sentence or not. The total votes obtained to know the sentence which it belongs to, these votes informs the strength of opinions that are associated with the reviews. Where, $\phi(C_j)$ is the total votes obtained for C_j and any sentences are equally contributed to the strength of opinions. The votes are adjusted by adding 1 for the own votes of the customers who wrote the reviews. Calculated the final rating $A(.)$ for the features F_k by using equation (12)

$$A(F_k) = \frac{Q(F_k)}{\sum_{C_j \epsilon D} \sum_{X_j^l \epsilon X_j}^{p(X_j^l)=1} \psi(X_j^l) \times \delta(F_k \epsilon X_j^l) \times (\phi(C_j)+1)} \quad (14)$$

Where the $Q(F_k)$ is cumulative number of stars is divided by the overall number of votes obtained during the accumulation. By this, the average weights are computed that are determined by the votes received. Thus, the rating of feature level for the features F_k of a product by utilizing the customer review votes and reviews is obtained. When the ratings are generated the reviews are categorized as Terrible (1star), Poor (2stars), Average (3stars), Very Good (4stars) and Excellent (5stars).

4. Results and discussion

Sentiment Analysis is also known as opinion extraction, it is the field of study which examines the people's attitude, sentiments, opinion, and emotion. various sentiment analysis approaches were presented to classify the sentiments. However, the existing methods are cost effective and consumes time for annotation of data .The present research performs descriptive analysis on Telugu Amazon reviews for classification approach.In this research, Hybrid XGB-RF Classifier method is proposed for the product review classification and rating form the amazon telugu reviews datasets. The proposed method is evaluated by the python3 in windows 10, i7 core processor, 16 GB RAM, and 6 GB 2080Ti NVDIA GTX edition GPU environment.

4.1 Performance metrics

The performance of the proposed is evaluated using the following metrics:

❖ **Accuracy**

Accuracy is the measure used to predict the exactness of machine learning model. The equation for accuracy is shown in Equation (15)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

❖ **Precision**

Precision is the ratio of correctly predicted positive observation to the total predicted positive observation. The equation for Precision is shown in Equation (16)

$$Precision = \frac{TP}{TP+FP} (16)$$

❖ **Recall**

The ratio of correctly predicted as fault-modules is defined as recall. The proportion of actual positives is correctly predicted by using recall, which is shown in Equation. (17)

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

❖ **F-Measure**

F-Measure is the measure of test accuracy and defined as weighted harmonic mean of the precision and recalls the test. The equation for F-Measure is shown in Equation (18).

$$F - measure = \frac{2PR}{P+R} (18)$$

Where,
P = Precision
R = Recall
TP =True Positive
TN =True Negative
FP = False Positive
FN = False Negative

4.2 **Quantitative Analysis**

The values obtained for the proposed Hybrid XGB-RF Classifier with other classifiers such as XGB and RF method for the evaluation of results is shown in the table1. The table 1 consists of results obtained for sentimental analysis in terms of accuracy, precision, recall and f-measure. The plotted graph for the obtained values are as shown in the figure 2

Table1: Performance evaluation of the proposed Hybrid XGB-RF Classifier interms of accuracy, precision, recall and f-measure.

Feature Selection	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
TF-IDF	XGB	71.02	73.45	86	87
	RF	92.25	89.93	94.5	96.21
	Hybrid XGB-RF	93.02	97.32	96.2	97.32
LSA	XGB	81.32	74.56	88	87
	RF	90.12	90.13	95.56	96.21
	Hybrid XGB-RF	95.12	96.23	97.23	97.32
Feature level rating	**XGB**	**83.02**	**78**	**89**	**88**
	RF	**92.12**	**89.13**	**97.56**	**97.21**
	Hybrid XGB-RF	**97.22**	**96.23**	**98.30**	**98.32**

The table 1 shows that the XGB takes longer to train the model and thus the trees were built subsequently. Similarly, the RF results shows that XGB are better learners when compared to RF. Thus, both XGB and RF are hybridized in the proposed method. The accuracy value is lowered by using LR as it generates results only for Independent data and whereas the sentiment analysis requires dependent data as well. As DT was unstable, when small change in the can lead to a large change in the structure of the optimal decision tree showed often inaccurate results. But XGB is more prone to overfit the data when compared to RF but XGB is robust enough for conserving higher accuracy. Thus, the accuracy values obtained for the proposed Hybrid XGB-RF as 83.02 %, 92.12%, and 97.12 %.

Table 2: The results obtained by using Feature level ratingwith Hybrid XGB-RF

Feature Selection	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
TF-IDF	LR	71.2	74.45	87	88
	DT	93.25	90.93	95.5	96.21
	Hybrid XGB-RF	94.02	98.32	97.2	97.32
LSA	LR	82.32	75.56	89	88
	DT	91.12	91.13	96.56	97.21
	Hybrid XGB-RF	96.12	97.23	98.23	97.23
Feature level rating	LR	83.02	78	89	88
	DT	92.12	89.13	97.56	97.21
	Hybrid XGB-RF	97.22	96.23	98.30	98.32

The present research work compares various feature selection algorithms such as TF-IDF, LSA with the Feature level rating algorithm. The Term Frequency Inverse Document Frequency (TF-IDF) computes document similarity directly in the word-count space, which may be slow for large vocabularies. Similarly, the Latent Semantic Analysis (LSA) showed relatively difficult when large amount of data was used showed computation more and required space for data saving. Whereas, the proposed Feature level rating with hybrid XGB-RF showed well decision making it kept both the customer well informed and improved the product buying by other customers. As the customer reviews will be based on different customers and their different interest in features. The proposed feature-level ratings made buying decisions personalized.

4.3 Comparative analysis

The comparative analysis for the proposed Feature Level rating method is undergone and the values are tabulated as shown in the table 2.

Table 3: Comparative Analysis of the existing methods with the proposed enhanced Feature level rating in terms of Accuracy

Methods	Accuracy (%)
SentiWordNet [11]	74.74
SentiPhraseNet [15]	78.002
Feature Level Rating with Hybrid XGB-RF	97.22

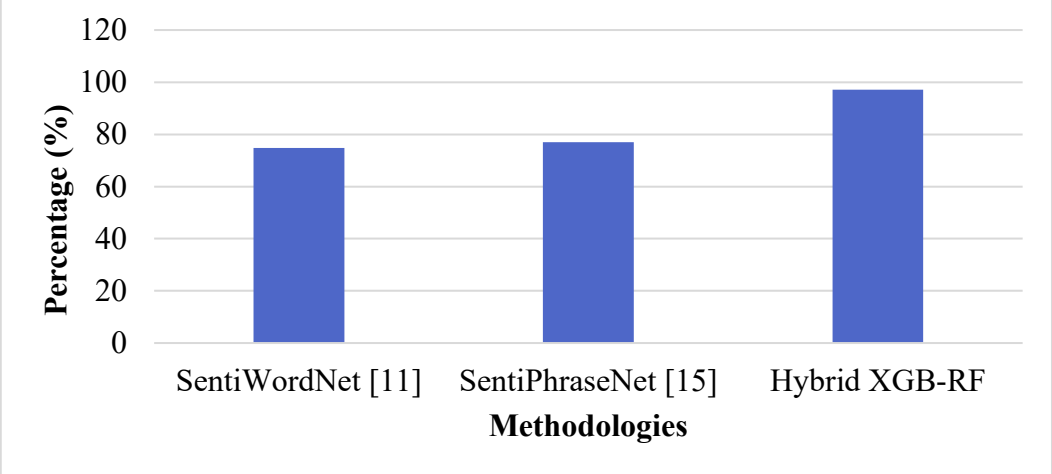

Figure 2. comparative analysis result for the proposed method interms of accuracy

The existing methods such as SentiWordNet[11] approach and SentiPhraseNet[15] are compared with the proposed Feature level rating method. The overall accuracy obtained for the proposed enhanced method is obtained as97.12 % that shows 20 % improvement when compared with other methods. The graphical representation of accuracy values obtained for the proposed and existing methods are shown in the figure 3.

5. Conclusion

Sentiment Analysis (SA) is also known as opinion extraction, it is the field of study which examines the people's attitude, sentiments, opinion, and emotion. various sentiment analysis approaches were presented to classify the sentiments. However, the existing methods are cost effective and consumes time for annotation of data. In this research, proposed an Hybrid XGB-RF classifier for the sentimental analysis of the amazon telugu reviews. Initially, the amazon telugu reviews are collected and required features are selected. Then, preprocessing is carried to convert the unstructured data into structured data. Further, relevant sentences are extracted from the structured data. The proposed hybrid XGB-RF was used as XGB are better learners when compared to RF XGB is more prone to overfit the data when compared to RF but XGB is robust enough for conserving higher accuracy. Thus, both XGB and RF are hybridized in the proposed method to improve the accuracy.Then, Hybrid XGB-RF classifier is used to classify the product reviews and ratings are generated. When the ratings are generated the reviews are categorized as Terrible (1star), Poor (2stars), Average (3stars), Very Good (4stars) and Excellent (5stars). The overall accuracy obtained for feature level rating method is 97.12%, precision is of 96.23 %, recall is of 98.30%, and f-measure is of 98.32 %.

REFERENCES

1. Reddy, G.R.R., 2020. *Enhancing Sentiment Prediction and Bias Detection for Telugu Language across Multiple Domains using ML and Deep Learning* (Doctoral dissertation, International Institute of Information Technology Hyderabad).
2. Yang, C., Zhang, H., Jiang, B. and Li, K., "Aspect-based sentiment analysis with alternating coattention networks". *Information Processing & Management*, *56*(3), pp.463-478., 2019
3. Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H. and Kwak, K.S., Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, *174*, pp.27-42.,2019.
4. Diamantini, C., Mircoli, A., Potena, D. and Storti, E., 2019. Social information discovery enhanced by sentiment analysis techniques. *Future Generation Computer Systems*, *95*, pp.816-828.
5. Rehman, A.U., Malik, A.K., Raza, B. and Ali, W., 2019. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, *78*(18), pp.26597-26613.
6. Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N., 2019. Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing* (pp. 639-647). Springer, Singapore.
7. Kumar, S.S., Kumar, M.A., Soman, K.P. and Poornachandran, P., 2020. Dynamic mode-based feature with random mapping for sentiment analysis. In *Intelligent systems, technologies and applications* (pp. 1-15). Springer, Singapore.
8. Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018. Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), p.11.
9. Liang, Z., Du, J. and Li, C., 2020. Abstractive Social Media Text Summarization using Selective Reinforced Seq2Seq Attention Model. Neurocomputing.
10. Wu, C., Wu, F., Wu, S., Yuan, Z., Liu, J. and Huang, Y., 2019. Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems*, *165*, pp.30-39.
11. Jonnalagadda, P., Hari, K.P., Batha, S. and Boyina, H., 2019. A rule based sentiment analysis in Telugu. *International Journal of Advance Research, Ideas and Innovations in Technology*.
12. Regatte, Y.R., Gangula, R.R.R. and Mamidi, R., 2020, May. Dataset Creation and Evaluation of Aspect Based Sentiment Analysis in Telugu, a Low Resource Language. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5017-5024).
13. Kumar, R.G. and Shriram, R., Sentiment Analysis using Bi-directional Recurrent Neural Network for Telugu Movies.
14. Bharti, S.K., Naidu, R. and Babu, K.S., 2020. Hyperbolic Feature-based Sarcasm Detection in Telugu Conversation Sentences. *Journal of Intelligent Systems*, *30*(1), pp.73-89.
15. Garapati, A., Bora, N., Balla, H. and Sai, M., 2019. SentiPhraseNet: An extended SentiWordNet approach for Telugu sentiment analysis.