International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I12P240

Volume 70 Issue 12, 421-430, December 2022
© 2022 Seventh Sense Research Group®

Original Article

Aspect Based Polarity Extraction in Tamil Tweets using Tree-Based Recursive Partitioning Techniques

S. Rajeswari[1], S. Gokila[2], K. Thinakaran[3], R.Surendiran[4]

[1]PG Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Chennai, India,
[2]Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, India,
[3]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences Chennai, India,
[4]School of Information Science, Annai College of Arts and Science, Kumbakonam, India.

[1]Corresponding Author : vrajee2008@gmail.com

Received: 20 October 2022 Revised: 30 November 2022 Accepted: 20 December 2022 Published: 24 December 2022

Abstract* - The overall outcome of the emotional statement about one particular discussion falls into two positive or negative that can be identified by the word/words and their synonymous that are closely connected with the theme of the topic. This work aims to identify the impacting word of the motion and analyse the performance of the Tree-based Machine Learning (ML) classifiers to classify the Tamil Tweets into two polarities (positive or negative). All the models are separately trained and tested with both Non-Weighted Vector and Weighted Vectors and analysed to freeze the accuracy. The prelabelled 1015 Tamil tweets are pre-processed to remove the noises to form a word dictionary. The words in the dictionary are tagged with weight to indicate the impact. The structured corpus with various lengths of statements is experimented with using a Decision tree, XGBoost and Random Forest classifiers with varying parameters. The comparative study report shows that Random Forest performs well by showing 78.81% of accuracy with Weighted Vector, which is better compared with Decision Tree and XGBoost classifiers.

Keywords* - Decision tree, XGBoost, Random forest, Natural Language Processing, Classification.

1. Introduction

Learning technology in our mother tongue creates more creativity. The Mother's tongue is only in the region. So almost all are working to include regional languages in technology [1]. Nowadays, a large volume of information is available in online documents, social media, and various resources. The development of the internet led to the exponential growth in the number of electronic documents in various regional languages. One of them is Tamil [2], a Dravidian language with no standard corpus for sentiment analysis and the work for Tamil in Natural Language Processing (NLP) is very limited. Therefore, an automatic text classification of the Tamil language with the help of NLP and ML is to be generated [3].

Machine Learning is an application of AI that provides systems to automatically learn and improve from experience without being programmed explicitly. Supervised and Unsupervised learning belongs to the Machine Learning algorithms. Classification or Predictive analysis belongs to supervised learning. Natural Language processing is also a part of machine learning, with the ability of a system to understand, analyse, manipulate and potentially generate human language. Text sentiment analysis is a very compelling topic in the field of NLP. It is focused on public ideas, feelings, and attitudes on several products, services, organisations, individuals, events, and themes such as entity emotions tend to make effective analysis [4, 5]. The automatic recommendation systems in many domains based on past reviews and feedback necessarily analyse the exact cause for positive and negative feedback. The contextual emotion expressed in the review statement has to be identified for the same [6,7]. Opinions on any topic from the common public shall be expressed in a single word or multiple sentences [8]. The proposed system identifies the exact word expressing the emotion and directly correlates with the polarity of the statement. The word that matches with polarity proposes a high accuracy of prediction. Handling the various sizes of statements during the vectorisation process is another challenging task, which depends on word embedding. Some of the work is fixing the limitation in the size of the statement. The proposed method handles the statement with various sizes without any minimum constraints on the number of words in the statement. Even short-size statements are strongly supportive of polarity identification.

Tree-based classification models are considered to be best in supervised machine learning. It empowers predictive models with high accuracy, stability and ease of interpretation. These types of algorithms are built by recursively splitting training data using different attributes from the dataset at each node that splits them effectively.

This splitting refers to learning simple decision rules taken from the training samples. Tree-based classifiers with supportive logic create relevancy among the words in statements in terms of semantic-based features [1]. Prediction is also a kind of decision; consequently, it has to be possible to use tree structures to represent prediction models. The technique of the creation of a tree entails recursive partitioning of data. This is where predictions reside in leaf nodes [27]. The proposed model focuses on this to identify the word that exactly expresses the emotion and the proposition between this word and the prelabeled polarity of the statement. Some tree-based classifiers, Decision Trees, Random Forest, and XGBoost, are applied to reviews to classify the final emotion class.

Emotion AI or Sentiment Analysis in Tamil tweets or reviews merges Natural language processing with Tree based Machine learning classifiers to predict positive and negative comments. The proposed work deals with the dictionary of 7933 tokens, and the model is trained and tested with 1015 tamil tweets with the grounded emotion expressed.

2. Argumentative Review

Sentiment analysis of Twitter with the case of the Anti-LGBT campaign in Indonesia using Naïve Bayes, Decision Tree and Random Forest machine learning classifiers. They concluded that the Naives Bayes produces an accuracy of 86.43% more than the other two algorithms [10].

The rating from the online shopping of the product will be posted as positive, negative and neutral reviews which helps the customer to purchase. Therefore, a feature-based opinion extraction concept is used to extract the comments, and it is classified with Random Forest and Support Vector Machine (SVM) classifiers. They observed that the Random Forest classifier gives better accuracy of 97% than the SVM [11].

In analysing the product reviews offered by Amazon, the RFSVM, a hybrid approach of Random Forest and Support Vector Machine for generating rules in classification technique, is applied. They focused on finding the positive and negative comments from the original text collected from the search and proved that the RFSVM gives better results than the individual classifiers [12]. Case studies of public comments on Nokia's product were stated as positive, negative and neutral opinions. The author conducts the sentiment analysis with the help of the classifier, such as a Decision tree and Random forest, to find whether the product is Good or Bad. It is noted that the Decision tree classifier gives higher accuracy of 89.4% than the Random forest algorithm[13].

Random Forest ensemble prediction for mobile product review written in Kanada was analysed, and a Random Forest classifier was used to classify the multiclass prediction [14]. The methodology produced 72% of accuracy. The accuracy of the proposed method was compared with Naives Bayes, which produced only 65% accuracy.

The Decision Random Forest method was used to select the feature based on Inverse Document Frequency. The projection matrix of the feature vector is attained using Principal Component analysis [15]. The sentiment analysis was performed to identify the polarity of the review to classify the sentiment. The ordinal classification was used to give more clarity to the review. The ML-based Support Vector Machine and Random Forest algorithms are used to suggest a recommendation system [16].

Rule-based sentiment analysis with Word2Vec and FastText embeddings for creating sentiment lexicon expansion method. They use three models: SentiWordNet-based Sentiment Analysis method, UJ_Lex_Pos and UJ_Lex_Neg lexicons-based Sentiment analysis method and Rule-based sentiment Analysis method with negation and conjunction with Tamil text of 10537 positive & 12664 negative words [17]. There are many dictionaries for sentiment analysis, but they are for a single domain. Multiple cross-domain dictionaries have been formed from the review content. The Enhanced Sentiment Dictionary has been formed for both positive and negative words. The bigram tree has been used to identify context relations. The SVM trained using this dictionary to analyse Multi-domain sentiment [18].

The classification algorithm finds the polarity and decides the class of the sentence. The recommendation system required additional output to match the request to suggest the exact match. The core of the content that decides the emotion on the review has to be analysed. The proposed method identifies the core concept of emotion, which is applied with a weighting factor.

3. Research Methodology

The emotion in a review statement concludes and bags the conclusion either as 'Good' or 'Bad' The proposed work determines the best Tree based on the ML algorithm to classify the Tamil reviews about the movies. The informal Tamil statement in a review is a challenging input in the pre-processing step. The sentence analysis stage of methodology focuses on different emotions handled in a movie and identifies the emotional category that decides the review's final class. The complete configuration of the proposed methodology is shown in Fig 1.

3.1. Data Pre-processing

The main purpose of processing is to reduce the dimensionality of the corresponding raw Tamil reviews and make them feasible for further feature representation and manipulation of the sentiments [28]. The proposed work is only for Tamil content, so the pre-processing phase is activated to remove the unwanted symbols and English characters using expression matching. The Tamil-only review is tokenised into individual elements with index values to form the Bag of words (BoW). Here the review comments are processed with emoticons and labelled with proper polarities, which is then continued with different Tree based recursive partitioning classification techniques.

3.2. Feature Extraction

The feature selection criteria directly impact the accuracy of the classification. Though the boosting technique of the Tree-based classifier scrutinises the feature, the emotional classification needs extra input on the feature to identify the exact cause of the emission on the content. The word directly connected with the overall expression of the emotion of all the reviewers is considered. The additional weight to emboss the exact feature improves the model's accuracy or reduces time complexity [20]. All possible emotions are identified, and focus is given to the one handled more in the entire review.

Table 1. Argumentative review of Tamil Sentiment Analysis

Ref. No	Dataset	Methodologies Used	Key Findings
[2]	Online tamil movie reviews	**Feature extraction** - TamilSentiWordNet **Classification** - SVM, Maxent classifier, Decision Tree, Naive bayes	SVM gives better accuracy of about 75.9% for classifying tamil reviews
[4]	Chinese language public comments	**Pre-processing -** Denoising, Word segmentation, filtering, stopping word **Feature Analysis** - TF-IDF **Classification -** EL with kernels and SVM	ELM with kernels method was more effective in classifying text emotion in Chinese language comments with 88.74% than SVM (88.54%)
[10]	Anti-LGBT 3744 twitter comments	**Pre-processing** - Tokenisation, Stop word removal and Stemming **Classification** - Naive Bayes, Decision Tree and Random Forest	Naive Bayes gives 83.43% accuracy when compared to Decision Tree and Random Forest
[11]	Flipkart dataset - 20,000 reviews	**Classification** - Random Forest and Support Vector Machine (SVM)	Random Forest produces an accuracy of 97% than SVM (92%)
[12]	Amazon dataset - 1000 instances	**Classification** - Random Forest (RF), Support vector Machine(SVM) and RFSVM	Hybrid RFSVM accuracy - 83.4% gives better results than the other two with 834 reviews of correct classification.
[13]	Youtube Nokia Mobile Channel - 2000 comments	**Pre-processing** - Translation, cleansing and removing URL, Auto labelling with VADER tool. TF-IDF **feature extraction classification** - Decision Tree & Random forest	Random Forest accuracy - 89.4%, F1measure - 82.2%, recall - 79.5%, precision - 86.5%
[14]	Weekly Mobile product reviews - GadgetLoka	**Classification** - Random Forest Ensemble classifier	RF Ensemble Multi-class Kannada classification for comparative and conditional statements with an accuracy of 72%
[15]	Twitter dataset	**Feature selection** - IDF, PCA, Decision tree based feature extraction, Decision Forest based feature extraction **Classification -** CART, Naive Bayes and LVQ	Decision Forest with LVQ classification produces 81% accuracy than the PCA and decision tree.
[16]		**Pre-processing** - Tokenisation, stop word removal, punctuation removal and streaming **classification of polarity** - Random forest & SVM Creating recommendation system using consumer reviews and profiles.	The author concluded that SVM is suitable for ordinal classification of public opinion reviews and Random forest for robustness.
[17]	noolaham.org, ta.wikipedia.org, Twitter, Facebook - 1377412 sentences, film reviews-629 reviews	Pre-processing - Word2vec and fastText word embeddings, assigning values using cosine similarity for the various polarities	Rule-based sentiment analysis method with UJ_Lex_pos and UJ_Lex_Neg algorithm with an accuracy of about 88%.
[18]	Amazon Mutli-domain dataset (Tamil reviews). Movie dataset	**Pre-processing -** English comments translated into Tamil using Google Translator toolkit, POS tags by RDRPOS	
[24]	-	**Classification -** Random Forest, Boost Trees, Max Entropy, Naive bayes	Comparative study of the following classifiers

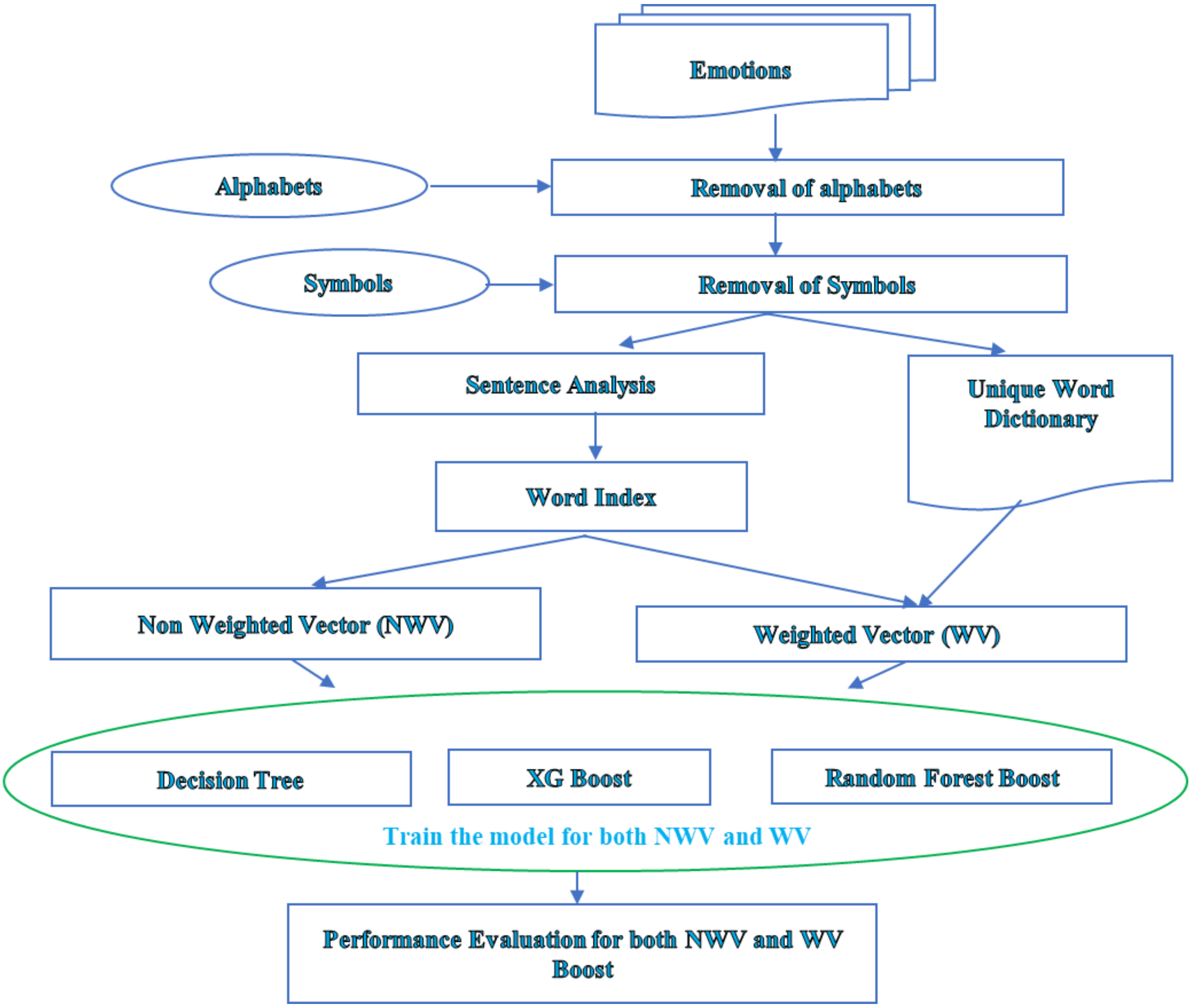

Fig. 1 Architecture of Emotion AI

The above has been identified using the following sequence of processes.

Step 1: list of words in the review statements DL.

Step 2:

$$f(W,C) = W_i^n \in DL \qquad (1)$$

Which finds the word vector with the number of occurrences of the word in the entire review.

Step 3:

f(W,C,R) = f(W,C) <=> Max(Re(Po))

Mapps the word vector with the prelabeled polarity of the sentence in which it occurs and assigns the vector with mapping with maximum polarity. f(W, C, R) is maintained as a Non-Weighted Vector (NWV).

Step 4:

$$WV = [hin\ \ldots.\ddot{h}_{i_1}] * f(W,C,R) \qquad (2)$$

The weighting factor has been applied to the word vector to the words directly correlated with polarity, which embossed the possible polarity. Eqn 1 and Eqn 2 are populated with all three classifiers experimented with in this work to fix the model with high performance. Each classifier is analysed for its performance for both NWV and WV.

3.3. Decision Tree

Among the three components of the Tree, the features are denoted in internal components (internal nodes), rules are in branches and output in leaves which do not have any further branches. It can handle both categorical and numerical data. To predict the class label for the record attribute, it is started from the root node, comparing the root attributes with the record's attribute by following the values of the corresponding branches and directing them to the next node [21]. Each node in a tree act as a test case, and each edge descendent belongs to the possible solution to the test case. The recursive process gets repeated for every subtree, starting with a new root node.

The accuracy of the decision tree is based on the attribute assigned in the root node, so choosing the root attribute is a challenging process in implementing the DT model. There are various attribute selection methods available; in this work Entropy method is used for the above-said challenge. It is a measure of randomness in the information being processed.

Entropy for multiple attributes is represented as:

$$E(N, S) = \sum_{i \in N} \quad \sum_{i \in N} \quad P(i)E(i) \tag{3}$$

where N - Current State and S- Selected attribute

3.4. XGBoost

Extreme Gradient Boosting is also said to be Regularised Boosting technique and is applied for supervised machine learning problems, where training data is used to predict the target variable. It is based on a gradient-boosting framework. It contains several regularisations that decrease overfitting and improve the performance of classifiers with higher predictive power than the Gradient Tree boosting technique [28]. This approach is where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. For this machine learning classification model, max_depth and n_estimators were the two hyperparameters applied.

3.5. Random Forest

Random Forest is an extension of recursive partitioning that grows multiple trees instead of one and is used to train the classification algorithm [23]. It is a supervised machine-learning algorithm that builds the forest with an ensemble of decision trees. The input phase produces the multi-altitude decision trees among two-phase processes. In the second phase, multi-decision trees are generated as an output [24]. Correlation adds randomness to the model to find the subset of the best attributes while splitting the node and increasing the prediction rate and efficiency. The parameters of the random forest included in this model are the total number of trees, minimum split and splitting criteria. All these three Tree based classifiers are trained and tested with the non-weighted and weighted feature vectors.

4. Results and Discussions

4.1. Data Set

The emotional statement about the Tamil movie contains 1015 reviews that were taken for this analysis. The review statements had some special characters and non-tamil letters. All that was removed in preliminary pre-processing activities. The statements were pre-labelled under two classes: 'Good' and 'Bad'. The number of statements falling under each label and the number of statements taken for the training and testing phase are shown in Table 2.

Table 2. Emotion Dataset

No. of Emotions labelled 'Good'	No. of Emotions labelled 'Bad	Total
506	509	1015

The emotional statements are in varying lengths. There are 5 lengthiest emotions with 41 words and 3 shortest emotions with 2 words. All the varying sizes of statements are taken and handled with balance by applying vectorisation.

The word 'பாசம்' is used a maximum of 379 times in the entire corpus, which states that maximum reviews are based on the concept ‘'பாசம்' handled in the movie. Those 147 emotions are reviewed as 'Bad' and 217 as 'Good'. The analysis identifies that the concept relevant to the maximum used word was reviewed among the audience, and their emotions were expressed. The weight matrix based on the frequency of the word has been generated, and the sentence vector has created a copy with applied weight.

Among 1015 statements, 812 were taken for training the model, and 203 were taken to test the model. The emotions for training and testing were picked in 19 random selections, and the models were trained and tested for all the possibilities. The number of 'Good' and 'Bad' reviews in the testing and training phases of all the random combinations applied in the proposed work is shown in Table 3.

Table 3. Emotions in the Training and Testing phase

Training Data		Testing Data	
No. of Emotions labelled 'Good'	No. of Emotions labelled 'Bad'	No. of Emotions labelled 'Good'	No. of Emotions labelled 'Bad'
393	419	113	90
395	417	111	92
396	416	110	93
397	415	109	94
398	414	108	95
399	413	107	96
402	410	104	99
403	409	103	100
404	408	102	101
405	407	101	102
406	406	100	103
407	405	99	104
408	404	98	105
411	401	95	108
412	400	94	109
414	398	92	111
415	397	91	112
416	396	90	113
417	395	89	114
422	390	84	119

The pre-processed Tamil emotional statements were split into 80% of training and 20% of testing data. It is fed into the tree-based classifiers such as Decision Tree, XGBoost and Random Forest algorithms. The individual high-performance models were trained separately with a randomly selected training data set. The accuracy analysis of all three concludes that RF produces better classification.

4.2. Performance of DT

The 1015 Tamil tweets are processed into 7933 tokens or features fed into the Decision tree classifier to classify the reviews. Instead of a different attribute selection method, Entropy is used as a parameter in the DT classifier to improve the accuracy.

4.3. Performance of XGBoost

In the XGBoost classifier, by including the hyperparameters as max_depth and n_estimators, the accuracy of the algorithms varies, as depicted in the graph (Fig. 2). The model produces high accuracy when the n_estimator is 14. The model was fixed with this parameter for further analysis.

The XGBoost is frozen at the epoch of 13 with high accuracy when a weighted word vector is taken as input. The early mapping of the feature vector makes the model perfect Fig 2(b).

4.4. Performance of Random Forest

The overfitting problem is prevented, and the accuracy is directly proportional to the number of trees in a forest. The random forest with a non-Weighted vector model produces higher accuracy, more than 77.2%, only when the n_Estimator is 900. The accuracy was less in other estimator values (Fig 4). The model was finalised with n_Estimator 900 with NWV to analyse and compare the performance with other Tree-based models.

Though the bonus RF classifier performs best among other models with non-weighted vectors, it performs even better with weighted vectors Fig 4(b). The model produces high accuracy of 78.81% at n_estimator 800. RF has been fixed to test the testing phase with WV.

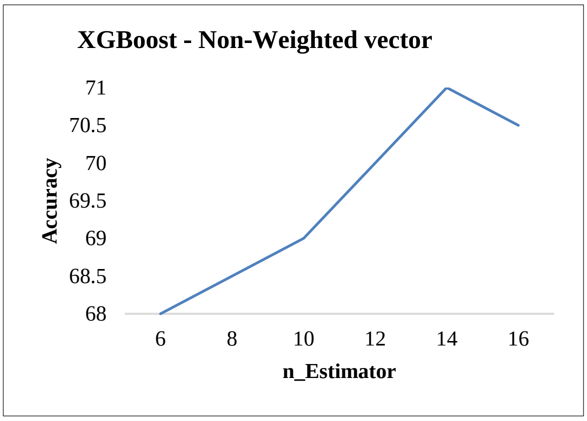

Fig. 2(a) XGBoost with Non-Weighted vector

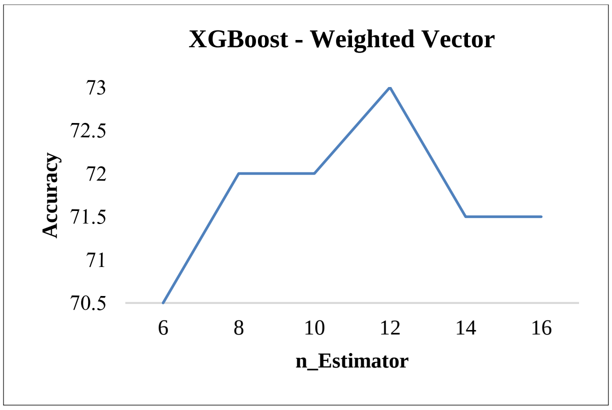

Fig. 2(b) XGBoost with Weighted vector

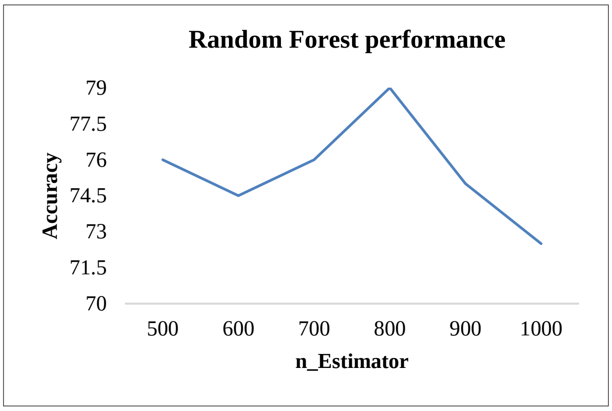

Fig. 3(a) RF with Non-Weighted vector

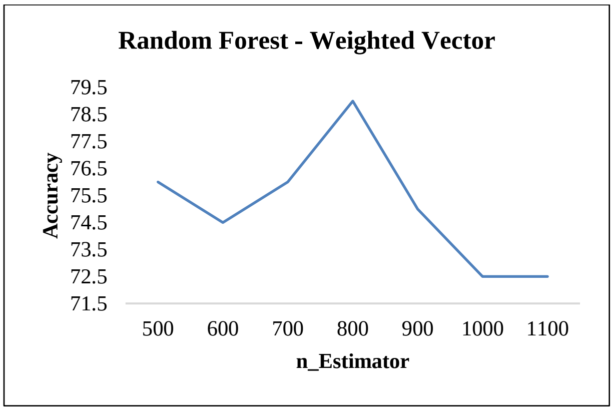

Fig. 3(b) Random Forests model with WV

4.5. Comparative Analysis

Upon all three models' high accuracy is attained in the earlier epoch when the weighted vector is given as an input for all possible random selection of training and testing sets as per Table 3. Though the difference in the accuracy of the model with both Weighted Vector (WV) and Non-Weighted Vector (NWV) is very minimal, the early epoch may improvise the overall optimality of the algorithm by reducing the time consumption for the increasing number of input records and the exponential increase in unique words dictionary. The least accuracy of each of the DT, XGBoot and RF with weightage is 64.54, 72 and 72, respectively. This worst case is also one percent more than the same model s without weightage values on the feature. The average weighted accuracy is almost 2% more than the non-weighted of all three models.

Table 4. Performance of DT, XGBoost and Random Forest with Non-weighted values (P-Precision, R-Recall, F1- F1 Score, A- Accuracy)

Decision Tree				XGBoost				Random Forest			
P	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
69.11	57.01	62.48	63	82.31	60.60	70.03	68.71	81.11	63.52	71.32	71.16
76.08	59.40	65.90	65.45	77.12	59.13	67.10	65.75	80.4	62.75	70.34	69.72
74.31	60.51	67.9	67.25	86.07	61.07	71.32	68.71	86.17	64.33	73.32	72.16
74.27	63.15	68.32	68.22	83.09	63.19	73.10	72.6	83.09	71.16	76.71	76.47
71.25	62.01	66.62	66.75	87.25	62.16	72.44	69.71	85.22	66.02	74.32	72.44
76.02	64.96	69.10	69.03	85.2	61.05	71.36	67.72	81.11	65.52	72.52	71.29
77.27	69.9	73.4	72.6	80.6	68	73.7	72.08	84.4	69.02	76.09	74.23
80	65.32	72.66	69.31	88	63.43	73.71	69.55	89	68	77.07	74.05
59.15	67.44	63.26	65.33	84.10	64.20	73	69.10	83.07	69.33	75.62	73.43
71.32	59.19	65.05	61.32	84.16	62.23	71.42	66.35	85.15	65.73	65.19	70.7
74.32	66.21	69.07	66.23	74.37	64.72	69.37	66.31	76.31	70.72	73.5	72.17
68.10	67.31	67.43	67.46	74.71	66.14	70.16	67.45	79.71	68.42	73.36	71.17
59.7	68.43	63.71	63.77	79.10	70.05	74.66	72.18	80.32	73	76.18	74.73
76.7	73.55	75.11	73.21	82.11	70.6	76.03	72.75	78.22	72.15	75.11	72.34
60.07	68.65	64.13	64	72.21	68.13	70.13	67.24	76.38	71.2	74.09	71.21
68.11	69.33	68.5	66.11	72.55	72.0	72.22	69.72	77.01	73.42	75.21	72.42
56.60	74.52	64.33	65.23	77.23	71.22	74.16	70.7	77.18	78.5	78.09	76.16
60.62	73.32	66.38	65.76	74.77	72.11	73.43	70.21	74.09	74.32	74.22	71.43
57.33	71.71	64.01	63.51	67.13	71.62	69.26	66.77	64.32	76.12	70.04	68.72
71.09	74.77	73	69.21	81.07	71.33	76.13	70.3	76.15	73.3	74.35	70.21

Table 5. Performance of DT, XGBoost and Random Forest with weighted vector (P-Precision, R-Recall, F1- F1 Score, A- Accuracy)

Decision Tree				XGBoost				Random Forest			
P	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
71.11	59.81	64.97	66	84.44	62.8	72.03	70.93	83.33	65.78	73.52	73.39
76.08	61.4	67.96	67.48	79.34	61.34	69.19	67.98	82.6	64.95	72.72	71.92
77.41	63.71	69.9	69.45	88.17	63.07	73.54	70.93	88.17	66.66	75.92	74.38
76.59	65.45	70.58	70.44	85.1	67.22	75.11	73.8	85.1	73.39	78.81	78.81
73.68	64.81	68.96	68.96	89.47	64.39	74.88	71.92	87.63	68.03	76.49	74.87
78.12	66.96	72.11	71.42	87.5	63.15	73.36	69.95	83.33	67.79	74.76	73.39
79.79	71.8	75.6	74.8	82.8	70	75.9	74.38	86.8	71.07	78.18	76.36
82	67.21	73.87	71.42	90	65.69	75.94	71.92	91	70	79.13	76.35
61.38	69.66	65.26	67.48	86.13	66.41	75	71.42	85.14	71.66	77.82	75.86
74.5	61.29	67.25	63.54	86.27	64.23	73.64	68.96	87.25	67.93	67.39	72.9
76.69	68.47	71.17	68.47	76.69	66.94	71.49	68.96	78.64	72.97	75.7	74.38
70.19	69.52	69.85	69.85	76.92	68.37	72.39	69.95	81.93	70.83	75.89	73.39
61.9	70.65	65.98	66.99	81.9	72.26	76.78	74.38	82.85	75	78.33	76.84
78.7	75.89	77.27	75.36	84.25	72.8	78.11	74.87	80.55	74.35	77.33	74.87
62.38	70.83	66.34	66	74.31	70.43	72.32	69.45	78.89	73.5	76.1	73.39
70.27	71.55	70.7	68.47	74.77	74.1	74.43	71.92	79.27	75.86	77.53	74.87
58.92	76.74	66.66	67.48	79.46	73.55	76.39	72.9	79.46	80.9	80.18	78.32
62.83	75.53	68.59	67.98	76.99	74.35	75.65	72.41	76.1	76.78	76.44	73.89
59.64	73.91	66.01	65.51	69.29	73.83	71.49	68.96	66.66	78.35	72.03	70.93
73.1	76.99	75	71.42	83.19	73.88	78.26	72.9	78.15	75.6	76.85	72.41

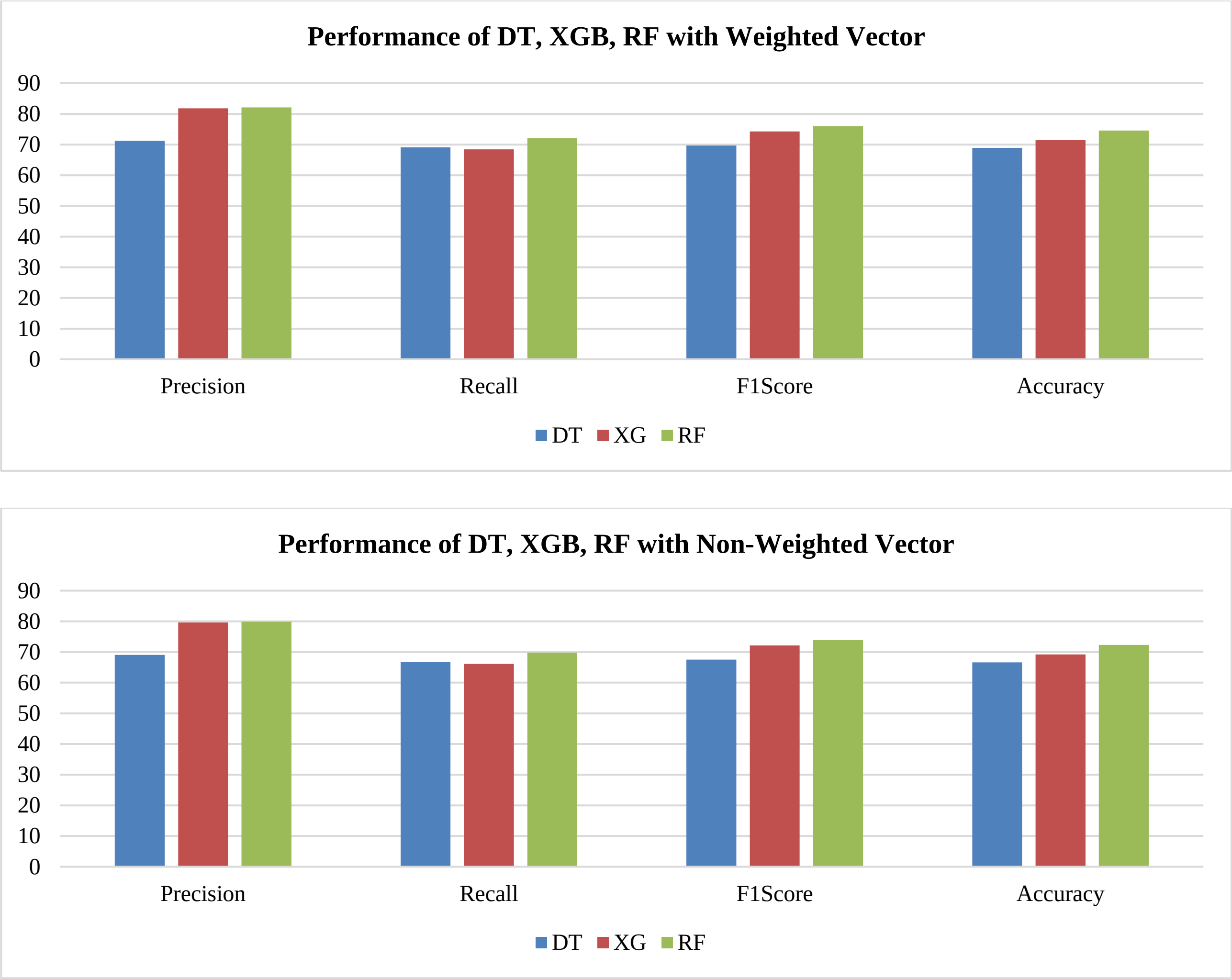

Fig. 4 Performance analysis of DT, XGB and RF (a) With Weighted Vector (b) Non-Weighted vector

The finalised DT, XGB and RF models with apt parameters which produced the best and stabilised accuracy are trained and tested for both Weighted Vector (WV) and Non-Weighted Vector (NWV). The Decision Tree in sentiment analysis performs less when compared to SVM classifier [2, 25], but the same with WV is high and the epoch also low. The performance of Tree-based classifiers produces considerable improvement while trained with WV when compared to other ML classifiers [20]. Testing and training data are executed for 20 possible randomly selected statements from the data sets mentioned in Table 4. The Precision, Recall, F1 score and accuracy of each of the 20 epochs are shown in Tables 4 and 5 for NWV and WV. The average of all these parameters with accuracy is shown in Fig 4 (a & b).

The performance of Random Forest with WV is remarkably high in all the possible training data. The boost-up value to the perfect feature increases the model's performance [20]. The testing accuracy of RF ranges from 70.93 to 78.81. The accuracy of all three models is equal only in one case. The number of emotional statements of each Binary Class is not influencing the result. But the polarity of the words relevant to the maximum appearance played a vital role in deciding the prediction.

5. Conclusion

The unique feature of humans is expressing their emotions, which could be well expressed when regional languages are used as mediums. The proposed method finds the best classifiers to penetrate the emotional text and classifies the conclusion category. The category of emotion reviewed maximum in a statement is identified, and the influence of the same in deciding the final class of review is also identified. The chained factor of words related to emotion and the deciding factor of the final class is weighted. The weighted Word vector of size equal to a Word dictionary has been given as input in a proposed method to analyse the performance of DT, XGB and RF, a tree-based machine learning classifiers model. All three model parameters had frozen at saturated accuracy level. The Random Forest classifier produces high accuracy in all possible training and testing data sets. The accuracy, along with the emotion factor, shall be the criteria to map the recommendation result with the query requesting the content. The same shall be applied to a large corpus to match the content searched for a particular emotion.

Authors' Contribution

The authors confirm their contribution to the paper as follows: Study conception and design- SR, SG and KT; Collecting existing work and argumentative review for prediction SR, SG, KT and RS; Argumentative review for Feature extraction: SR, SG and RS; Methodology for feature weightage SG, SR and KT; Data collection and analysis RS and KT; Analysis and Interpretation of results SR, SG, KT; Draft manuscript preparation SR, SG, KT and RS All authors reviewed the results and approved the final version of the manuscript.

References

[1] N. Rajkumar et al., "An Efficient Feature Extraction with Subset Selection Model Using Machine Learning Techniques for Tamil Documents Classification," *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 11, pp. 66-81, 2020.

[2] Shriya Se et al., "Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms," *Indian Journal of Science and Technology*, vol. 9, no. 45, pp. 1-5, 2016. *Crossref*, http://doi.org/10.17485/ijst/2016/v9i45/106482

[3] Diksha Khurana et al., "Natural Language Processing: State of the Art, Current Trends and Challenges," *Multimedia Tools and Applications*, 2022. *Crossref*, https://doi.org/10.1007/s11042-022-13428-4

[4] Xueying Zhang, and Xianghan Zheng, "Comparison of Text Sentiment Analysis based on Machine Learning," *15th International Symposium on Parallel and Distributed Computing*, pp. 230-233, 2016. *Crossref*, https://doi.org/10.1109/ISPDC.2016.39

[5] Anuj Gupta et al., "*Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*," United States, O'Reilly Media, 2020.

[6] Samheeta Gourammolla, and S Gokila, "HCB Machine Learning Approach for Movie Recommendation System," *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE*, pp. 1186-1190, 2022. *Crossref*, https://doi.org/10.1109/ICICCS53718.2022.9788163

[7] Wei Li et al., "Bieru: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis," *Neurocomputing*, vol. 467, pp. 73-82, 2022. *Crossref*, https://doi.org/10.1016/j.neucom.2021.09.057

[8] Joni Salminen et al., "Creating and Detecting Fake Reviews of Online Products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, 2022. *Crossref*, https://doi.org/10.1016/j.jretconser.2021.102771

[9] K. Kavitha, and Suneetha Chittineni, "Efficient Sentimental Analysis using Hybrid Deep Transfer Learning Neural Network," *International Journal of Engineering Trends and Technology*, vol. 70, no. 10, pp. 155-165, 2022. *Crossref*, https://doi.org/10.14445/22315381/IJETT-V70I10P216

[10] Veny Amilia Fitri, Rachmadita Andreswari, and Muhammad Azani Hasibuan, "Sentiment Analysis of Social media Twitter with Case Anti LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree and Random Forest Algorithm," *Procedia Computer Science*, vol. 161, pp. 765-772, 2019. *Crossref*, https://doi.org/10.1016/j.procs.2019.11.181

[11] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1-5, 2019. *Crossref*, https://doi.org/10.1109/INCOS45849.2019.8951367

[12] Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, vol. 127, pp. 511-520, 2018. *Crossref*, https://doi.org/10.1016/j.procs.2018.01.150

[13] Mohammad Aufar, Rachmadita Andreswari, and Dita Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 1-7, 2020. *Crossref*, https://doi.org/10.1109/ICoDSA50139.2020.9213078

[14] Yashaswin Hegde, and S.K.Padma, "Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada," *IEEE 7th International Advance Computing Conference*, pp. 777-782, 2017. *Crossref*, https://doi.org/10.1109/IACC.2017.0160

[15] Jeevanandam Jotheeswaran, and S. Koteeswaran, "Feature Selection using Random Fores Method for Sentiment Analysis," *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1- 6, 2016. *Crossref*, https://doi.org/10.17485/ijst/2016/v9i3/75971

[16] Gayatri Khanvilkar, and Deepali Vora, "Sentiment Analysis for Product Recommendation Using Random Forest," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 3.3, pp. 87-89, 2018. *Crossref*, https://doi.org/10.14419/ijet.v7i3.3.14492

[17] Sajeetha Thavareesan, and Sinnathamby Mahesan, "Sentiment Lexicon Expansion using Word2Vec and Fast Text for Sentiment Prediction in Tamil Texts," *2020 Moratuwa Engineering Research Conference (MERCon), IEEE*, pp. 272-276. *Crossref*, https://doi.org/10.1109/MERCon50084.2020.9185369

[18] E. Sivasankar, K. Krishnakumari, and P. Balasubramanian, "An Enhanced Sentiment Dictionary for Domain Adaptation with Multi-Domain Dataset in Tamil Language (ESD-DA)," *Soft Computing*, vol. 25, no. 2, pp. 3697-3711, 2021. *Crossref*, https://doi.org/10.1007/s00500-020-05400-x

[19] Jerry Wood, "COVID-19: The Pandemic's Impact on the Dissemination of Data in Virtual Teams using Computer-Mediated Communication Technology," *International Journal of Computer Trends and Technology*, vol. 68, no. 12, pp. 26-30, 2020. *Crossref*, https://doi.org/10.14445/22312803/IJCTT-V68I12P106

[20] Thevatheepan Priyadharshan, and Sagara Sumathipala, “Text Summarization for Tamil Online Sports News Using NLP,” *2018 3rd International Conference on Information Technology Research (ICITR), IEEE,* pp. 1-5, 2018. *Crossref,* https://doi.org/10.1109/ICITR.2018.8736154

[21] Bharathi Raja Chakravarthi et al., "Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text," *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL),* pp. 202-210, 2020.

[22] Balaji Karumanchi, "An Unsupervised Clustering Approach for Twitter Sentimental Analysis: A Case Study for George Floyd Incident," *International Journal of Computer Trends and Technology,* vol. 68, no. 6, pp. 46-50, 2020. *Crossref,* https://doi.org/10.14445/22312803/IJCTT-V68I6P10

[23] Thorvardur Jon Love, Tianxi Cai, and Elizabeth W.Karlson, "Validation of Psoriatic Arthritis Diagnoses in Electronic Medical Records Using Natural Language Processing," *Seminars in Arthritis and Rheumatism,* vol. 40, no. 5, pp. 413-420, 2011. *Crossref,* https://doi.org/10.1016/j.semarthrit.2010.05.002

[24] Amit Gupte et al., "Comparative Study of Classification Algorithms used in Sentiment Analysis," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 5, pp. 6261-6264, 2014.

[25] Xiaohui Liang et al., "Evaluating Voice-Assistant Commands for Dementia Detection," *Computer Speech & Language*, vol. 72, p. 101297, 2022. *Crossref*, https://doi.org/10.1016/j.csl.2021.101297

[26] Md. Sirajul Huque, and V. Kiran Kumar, "A Study on Sentiment Analysis of Movie Reviews using ML Algorithms," *International Journal of Computer Trends and Technology*, vol. 70, no. 9, pp. 33-37, 2022. *Crossref,* https://doi.org/10.14445/22312803/IJCTT-V70I9P104

[27] Abhijeet Mankar, and Sudhakar Bhoite, "Review of Literature on Recursive Partitioning and its Applications in Various Area," *Proceedings of the International Conference on Emerging Trends in Artificial Intelligence and Smart Systems, THEETAS 2022,* 2022. *Crossref,* https://doi.org/10.4108/eai.16-4-2022.2318071

[28] Roza Hikmat Hama Aziz, and Nazife Dimililer, "SentiXGboost: Enhanced Sentiment Analysis in Social Media Posts with Ensemble Xgboost Classifier," *Journal of the Chinese Institute of Engineers*, vol. 44, no. 6, pp. 562–572, 2021. *Crossref,* https://doi.org/10.1080/02533839.2021.1933598