A Sentence Level Classification of Telugu News Document using Sentiment Analysis

Jalaja kumari Bygani [1*], *Dr.Yella* Venkateshwaralu[2], and *Dr.K.V.*Ramana[3]

[1]Assistant Professor, Department of Computer Science Engineering, International School of Technology and Sciences for Women, Rajahmundry.

[2]Computer Science Engineering, Kakinada Institute of Technological Sciences, Ramachandra Puram, Andhra Pradesh.

[3]Department of Computer Science Engineering, University College of Engineering, JNT University, Kakinada, Andhra Pradesh

Abstract. In recent years, sentiment analysis-based categorization in low-resource languages and regional languages has become a hot topic in natural language processing. Researchers are more interested in categorizing sentiment in Indian languages such as Hindi, Telugu, Tamil, Bengali, Malayalam, and others. To the best of our knowledge, no microscopic study on Indian languages has been published to yet due to a lack of annotated data. Using Telugu sentiment analysis, we presented a two-phase classification technique for Telugu news phrases in this work. It first recognizes subjectivity categorization, which categorizes statements as Positive, Negative, or Neutral. Sentiment Classification is the next step, which divides subjective statements into positive and negative categories. We get an accuracy of 81 percent for sentiment analysis categorization using this method.

1. Introduction

With the introduction of the World Wide Web, the amount of data available on the internet skyrocketed. Although such a large amount of data is beneficial, and the majority of it is in the form of texts, people have a difficulty or a challenge in identifying the most relevant data or knowledge. As a result, text classification aids in the resolution of this problem. Text categorization is the process of classifying a collection of input documents into two or more classes, with each document belonging to one or more of them [1]. Text classification is a text mining approach for categorizing text content into predetermined groups. Manual or automatic classification is possible.

* Corresponding author: byganijalaja@gmail.com

Automated Text Classification, in contrast to human classification, which takes time and demands high accuracy, makes the classification process faster and more efficient by automatically categorizing documents.

Sentiment analysis may be used on text at three levels: phrase, document, and aspect. The goal of sentence-level analysis is to find the polarity value of each sentence in a text. The polarity value is determined using document level analysis, which considers the entire content. Aspect level analysis determines the polarity of each aspect (word-by-word) in a text.

After Hindi, Telugu is India's second most popular language. Telugu is ranked fifteenth on the Ethnologic list of most-spoken languages in the world, with 85 million native speakers [2]. Several e-Newspapers, such as Eenadu, Sakshi, Andhra Jyothy, Vaartha, and Andhra Bhoomi, are accessible in Telugu and publish news on a regular basis.

Because of the relevance of document classification and sentiment analysis, we integrated the two in this study and developed a unique two-phase process termed classification of documents using sentiment analysis.

2. Literature

There are various emotion analysers for the English language [4-8], but little work has been done in the context of Indian languages [9-25]. The fundamental reason for this is a scarcity of materials in Indian languages.

Patra et *al.* [1] objective of this task was to classify the tweets into positive, negative, and neutral polarity. For training and testing purpose, the tweets from each language were provided. Each of the participating teams was asked to submit two systems, constrained and unconstrained systems for each of the languages. We ranked the systems based on the accuracy of the systems. Total of six teams submitted the results and the maximum accuracy achieved for Bengali, Hindi, and Tamil.

Dipankar et *al.* [2] proposed an alternate way to build the resources for multilingual affect analysis. They have prepared WordNet affects for the three Indian languages such as Hindi, Bengali, and Telugu, and used English as a source language. For translation into target languages, they used WordNet of every language which is publicly available over the internet.

Kumar SS et al., [3] proposed method used binary features, statistical features generated from SentiWordNet, and word presence (binary feature). Due to the sparse nature of the generated features, the input features were mapped to a random Fourier feature space to get a separation and performed a linear classification using regularized least square method. The proposed method identified more negative tweets in the test data provided Hindi and Bengali language. In test tweet for Tamil language, positive tweets were identified more than other two polarity categories. Due to the lack of language specific features and sentiment oriented features, the tweets under neutral were less identified and also caused misclassifications in all the three polarity categories.

Prasad et *al.[4]* describes the system they used for Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets, at MIKE-2015. They take the help of a twitter training dataset in Indian Language (Hindi) and apply data mining approaches for analyzing the sentiments. They used a state-of-the-art Data Mining tool Weka to automatically classify the sentiment of Hindi tweets into positive, negative or neutral.

Sarkar et *al.* [5] developed a sentiment analysis system for Hindi and Bengali tweets using multinomial naive Bayes classifier that use unigrams, bigrams and trigrams for the selection of features. The system has been trained and tested on the dataset released for SAIL TWEET CONTEST 2015. This system obtains accuracy of 50.75 %, 48.82 %, 41.20 %, and 40.20 % for Hindi constrained, Hindi unconstrained, Bengali constrained and Bengali unconstrained run respectively.

Venugopalan M, Gupta D [6] proposed work explores Sentiment Analysis on Hindi tweets in a constrained environment and hence proposes a model for dealing with the challenges in extracting sentiment from Hindi tweets. They have used raw corpus provided by Indian Languages Corpora Initiative (ILCI) to train the Doc2Vec model and for pre-processing, Doc2Vec tool that gives the semantic representation of a sentence in the dataset. The model has exhibited an average performance with cross validation accuracy for training data around 56 % and a test accuracy of 43 %.

Mukku SS [7] produces a more focused and accurate sentiment summary of a given Telugu sentence which is useful for the users. They explore various Machine Learning techniques for the classification of Telugu sentences into positive or negative polarities.

Sarmah, Saharia and Sarma [8] presented an approach for classification of Assamese documents using Assamese WordNet. This approach has accuracy of 90.27 % on Assamese documents. Frequent terms of Assamese document are searched in Assamese WordNet. For each frequent term sysnet is found in WordNet, extended form of it is found and is associated with it. It searches each pre-defined class for extended terms present in testing document. It assigns the test document a class with which it has highest number of matching terms.

Dhar A et al., [9] presenting a methodology for developing an automatic system for solving the problem of classifying Bangla text document into their respective text categories. It introduces a hybrid approach (i.e., PART) for classification of Bangla text documents based on 'term association' and 'term aggregation' as baseline feature extraction methods.

Sharma R et al. [10] proposed.a Hindi language opinion mining system. The system classifies the reviews as positive, negative and neutral for Hindi language. Negation is also handled in the proposed system. Experimental results using reviews of movies show the effectiveness of the system

Gupta V et al. [11] represents the advanced NLP learning resources in context of Indian languages: Hindi and Urdu. The research is based on domain-specific platforms which covers health, tourism, and agriculture corpora with 60 k sentences. With these corpora, some NLP-based learning resources such as stemmer, lemmatizer, POS tagger, and MWE identifier have been developed.

3. Methodology

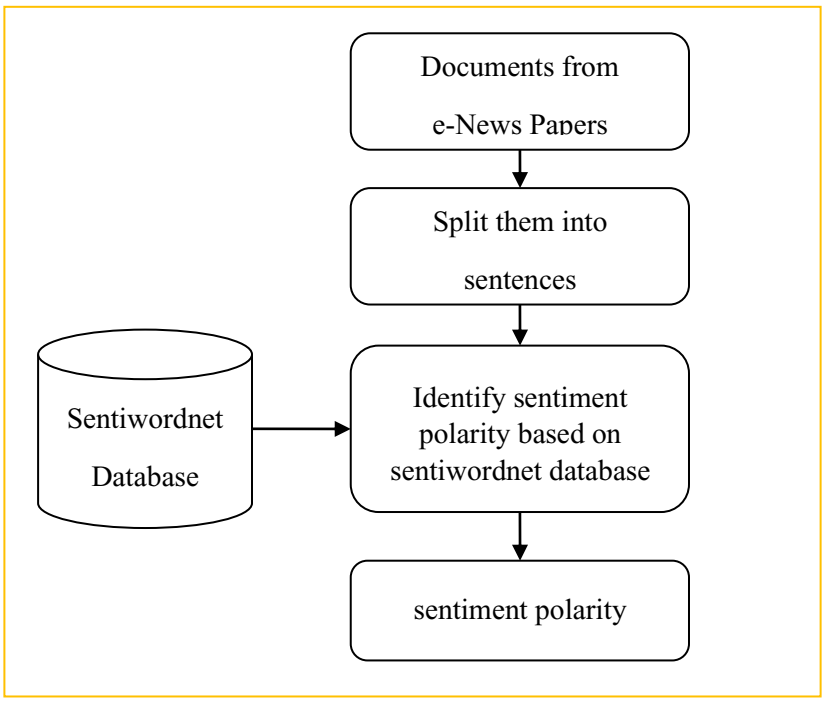

Fig 1 sentence labeling process

In this research, we present a sentence-level categorization for Telugu news using a sentiment analyser. Subjectivity and sentiment analysis are two steps in the sentiment analysis process. We categories the subjective and objective texts in a corpus using subjectivity analysis. We also look at the tone of subjective statements, both favourable and negative. Because objective sentences do not

have any feeling value, they are viewed as neutral statements. As a result, the system identifies the statements as subjective (good, negative) or objective in the first phase (neutral). The algorithm then categorizes the subjective statements as favourable or negative in the second step.

As indicated in Figure 1, the first phase is further divided into subphases, each of which has its own value and requirement. During this step, data is collected from e-newspapers, then divided into sentences and a corpus is constructed. Sentiwordnet assigns a positive, negative, or neutral classification to each sentence.

As illustrated in figure 2, the second phase is the classification process, which consists of several subphases, each of which has its own relevance and requirement. Data collection, data cleaning, the pre-processing phase, and the machine learning process, which includes model testing, model training, training datasets, testing datasets, and performance evaluation.

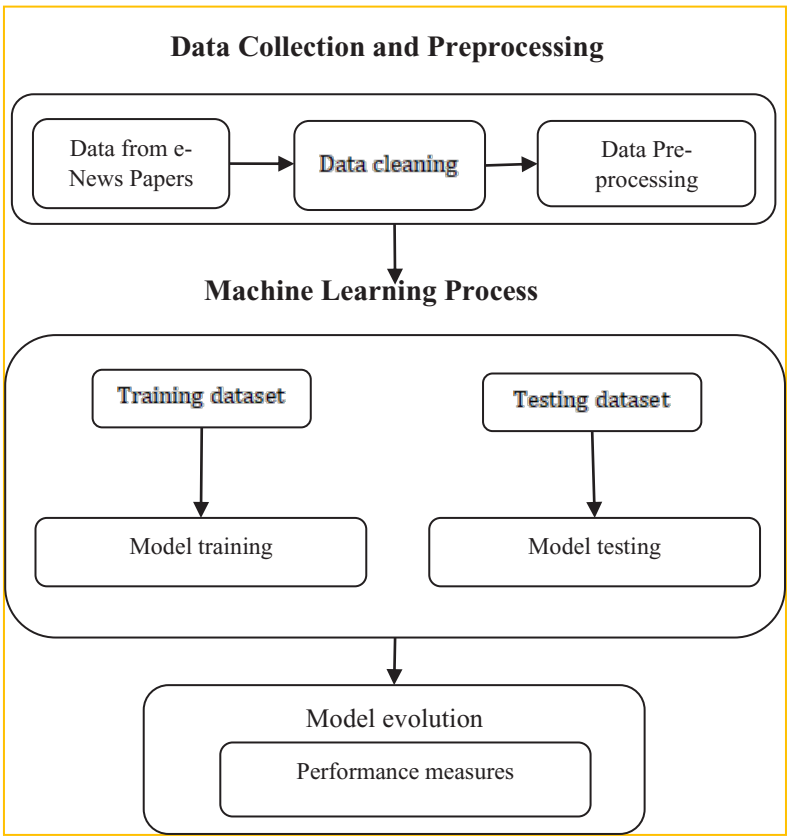

Fig 2 classification process

Data Collection: The data was collected from different resources in different formats and later it will used. Data cleaning refers to discovering incomplete, erroneous, imprecise, or unnecessary pieces of data and then replacing coarse data; after cleansing, a phrase should be consistent.

Pre-Processing

The initial stage in pre-processing is to convert text documents into a readable word format. There are a lot of characteristics in the papers that are ready for the next phase in text categorization. Typically, the following actions are taken:

Tokenization: A document is parsed into a list of tokens after being regarded as a string.

Removing stop words: Because stop words are commonly used, unimportant terms must be eliminated.

Stemming word: Using a stemming method to transform diverse word forms into canonical forms that are comparable. Conflating tokens to their root form, e.g. connection to connect, computing to compute, is the procedure of this phase.

Machine learning process: Model training, training dataset, testing dataset, and model testing are all included in this step.
Training dataset: Training data is labelled data that is used to improve the accuracy of machine learning algorithms by training them. A model is usually given a data collection of known data, referred to as the training data set.
Model Training: It is the process of train the model with training data.

Testing dataset: It is a set of unknown data against which the model is evaluated. It is a subset of data that is used to offer an unbiased assessment of a final model's fit on the training dataset.
Model testing: For testing purposes, the testing dataset is provided to the simulation model. We employed TWO distinct categorization mechanisms in the testing, which are known as Decision Tree and Naive Bayesian.
Model evolution: In this phase, benchmark model and some other models are evaluated with performance measures like precision, recall, accuracy, F1 scores.

Decision Tree: Actual process of this approach is depicted in Algorithm 1.

Algorithm 1: Decision Tree	
Input: Telugu sentences with polarity dataset Output: Sentiment prediction	
1	I/P data set was read
2	Information gain is calculated for each feature for root attribute selection $Infogain(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$ *Here pi denotes nonzero probability that an arbitrary tuple in D belongs to class Ci and* $\|C_{i,D}\|/\|D\|$. $Info_A(D) = \sum_{j=1}^{v} \frac{\|D_j\|}{\|D\|} * Info(D_j)$. *Where the term* $\frac{\|D_j\|}{\|D\|}$ *acts as the weight of the jth partition.* $Info_A(D)$*is the expected information required to classify a tuple from D based on the pa* . $Gain(A) = Info(D) - InfoA(D)$ *Where* $Gain(A)$ *tells us how much would be gained by branching on A.*
3	Above step is repeated for tree building.

Naïve Bayesian: Actual process of this approach is depicted in Algorithm 2.

Algorithm 5: Naïve Bayesian	
Input: Telugu sentences with polarity dataset Output: Sentiment prediction	
1.	I/P data set was read
2.	Create a frequency table from the data set.
3.	Find the probabilities and create a Likelihood table.
4.	Calculate the posterior probability for each class using the NB equation. The equation is $P(H\|X) = \frac{P(X\|H)P(H)}{P(X)}$
5.	The outcome of prediction is the class with the highest posterior probability.
6.	Final results are returned

4. Results

Datasets on politics, sports, technology, entertainment, and business news were gathered from the websites of several Telugu newspapers like Andhra Bhumi (www.andhrabhoomi.net), Andhra Jyothi (www.andhrajyothy.com), Eeenadu (www.eenadu.net), Namasthe Telangana (www.ntnews.com), Prajasakthi (www.prajasakti.com), Sakshi (www.sakshi.com), Surya (www. suryaa.com), Vartha (www.vaartha.com).

The execution of the sentence Four metrics may be used to evaluate a classification system: Accuracy, precision, recall, and the F1 measure.

5. Data Analysis

Number of documents in each category is depicted in Figure 3. From Fig 3, it is observed that, politics having 417 documents, sports of having 511 documents, technology of having 401 documents, entertainment of having 386 documents, business of having 510 documents.

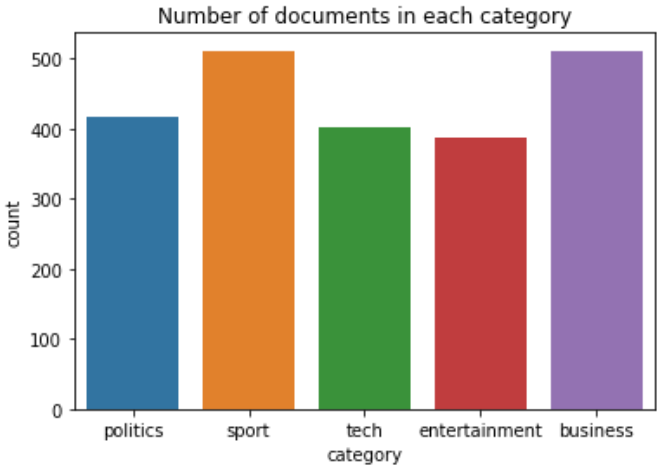

Fig 3 Number of documents in each category

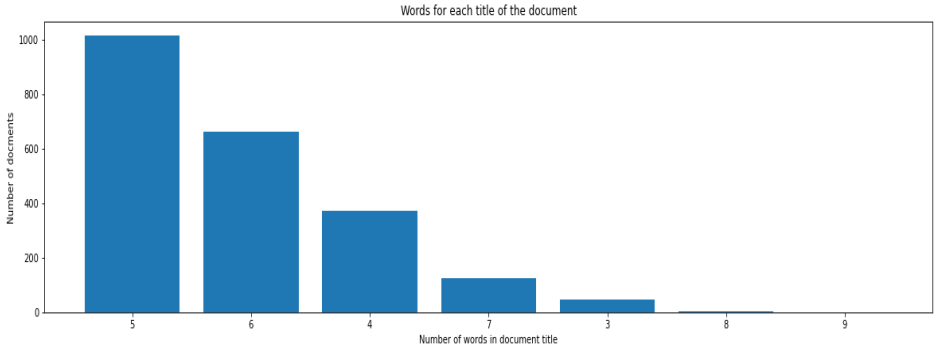

Fig 4 The no. of words in document title

Figure 4 shows the no. of words in document titles. It can be seen that many project names have a word count of 4-6, while a small initiatives have titles with a word count of 3 or 7 to 9.

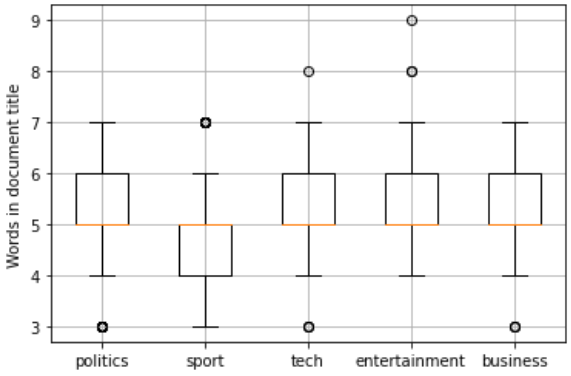

Fig 5 words in document title

Figure 5 shows the words in document titles. It can be seen that the average title word count is 5 for all categories. Sports category titles appear to be smaller; all other categories, with the exception of sports, have a 25% to 75% le between 4 and 7.

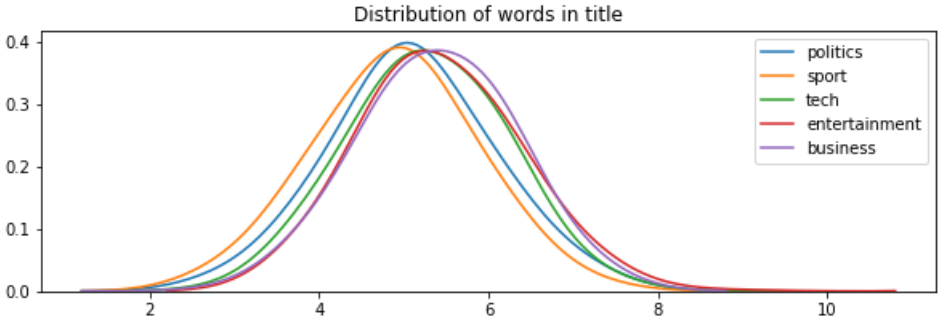

Fig 6 words distribution in title

Figure 6 shows the distribution of terms in the title. It can be seen that the words are scattered in groups of four to six.

Figure 7 shows the distribution of terms in tales across categories. Figure 7 shows that the words are dispersed in a range of 0 to 1000.

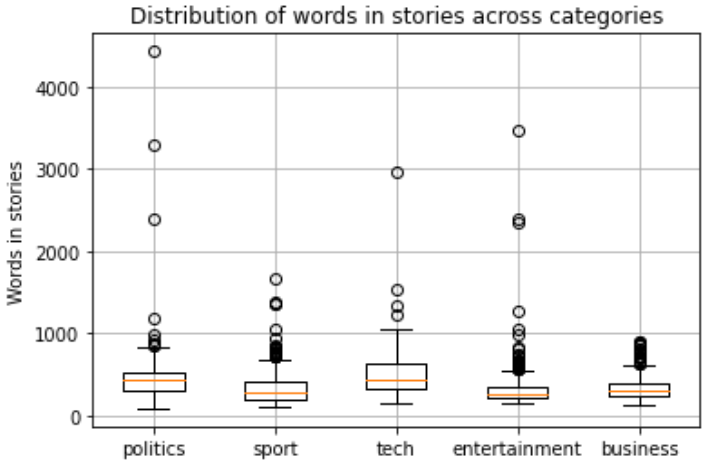

Fig 7 Distribution of words in stories across categories

Table 1 shows the NB and DTC classifiers' precision, recall, and F1 score values. Figure 8(a), (b), (c), (d), (e), (f) and (g) show graphs for these values (f). Figures 8(a) and (b) show that the greatest precision of NB in the technology category is 0.896, while the highest precision of DTC in the sports and entertainment category is 0.89.

Figures 8(c) and (d) show that the greatest value for NB recall in the politics category is 0.898, while the highest value for DTC recall in the technology category is 0.895.

Finally From the Figure 8(e) & (f), it is observed that the greatest F1-score of NB for entertainment is0.94, while the highest F1-score of DTC for politics is0.95, as shown in Figures 8(e) & (f).

Table 1 Precision, Recall and F1 Score for Naïve Bayes and Decision Tree classifier

	Naïve Bayes			**Decision tree**		
Dataset Name	Precision	Recall	F1 score	Precision	Recall	F1 score
Politics	0.83	0.898	0.93	0.881	0.83	0.95
Sports	0.84	0.85	0.89	0.89	0.795	0.899
Technology	0.896	0.84	0.87	0.825	0.895	0.93
Entertainment	0.81	0.82	0.94	0.89	0.884	0.92
Business	0.82	0.87	0.858	0.825	0.89	0.88

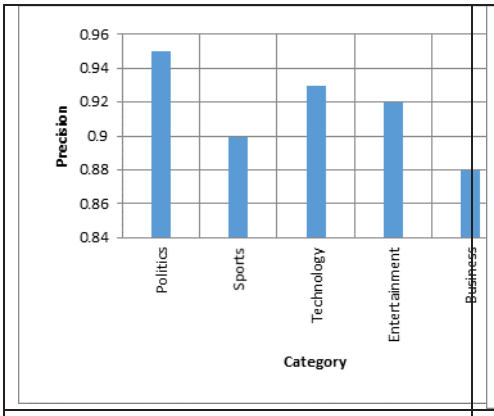

Fig 8(a) precision graph of NB

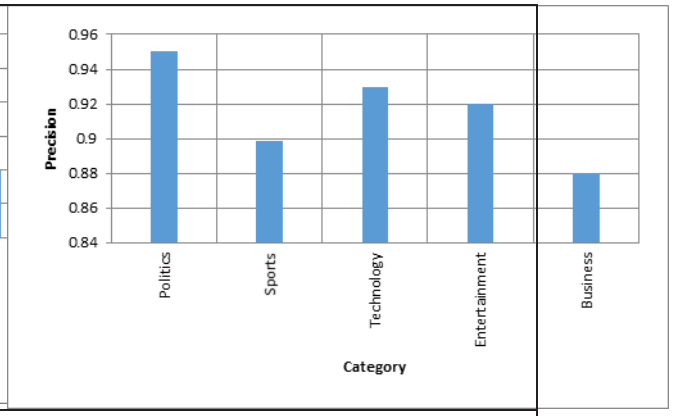

Fig 8(b) precision graph of DTC

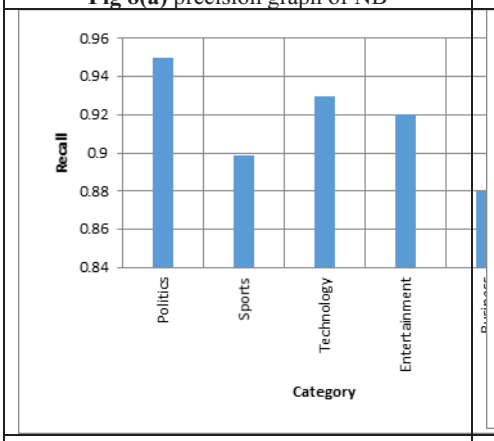

Fig 8(c) recall graph of NB

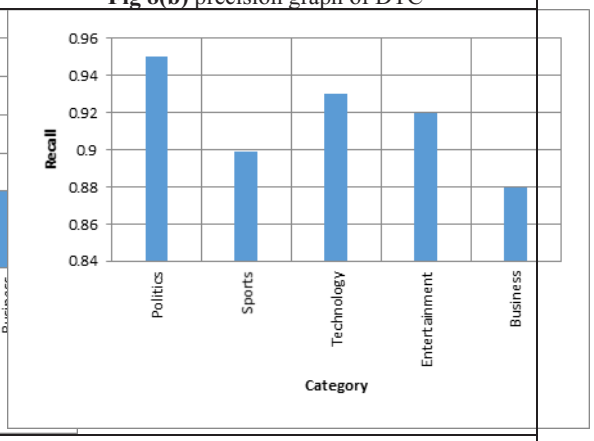

Fig 8(d) recall graph of DTC

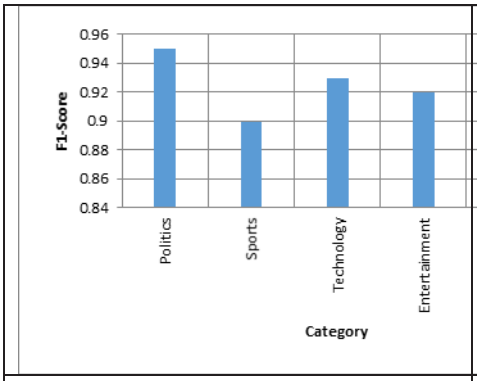

Fig 8(e) F1-score graph of NB

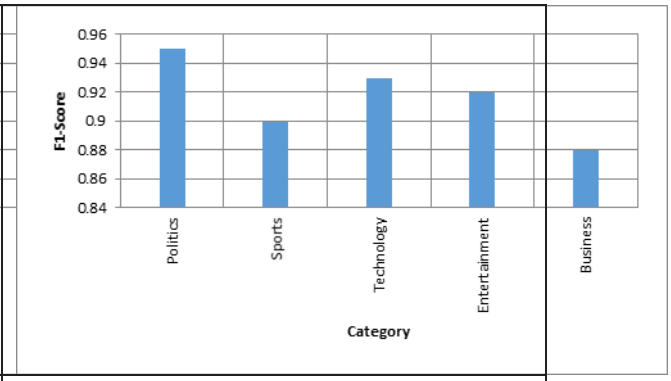

Fig 8(f) F1-score graph of DTC

6. Conclusions

Sentence classification is crucial in the field of data mining, particularly in the field of text mining. Because there is a great amount of material available on the internet, it must be arranged by content for simple retrieval. Due to its relevance, we presented two level methods in this paper: phrase level classification and sentiment analysis, both of which are important nowadays. To test our model, we gathered data from several e-newspapers, divided it into intelligible phrases, and then categorized it as positive, negative, or neutral. Finally, we partition these phrases into training and testing scenarios and put our suggested system to the test. The results of our tests revealed an accuracy of 81 percent. We must increase this accuracy and minimize the number of errors in the future by applying the deep learning process.

References

1. Patra BG, Das D, Das A, Prasath R. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In International Conference on Mining Intelligence and Knowledge Exploration 2015 Dec 9 (pp. 650-655). Springer, Cham.
2. Das D, Poria S, Dasari CM, Bandyopadhyay S. Building resources for multilingual affect analysis–a case study on hindi, bengali and telugu. In Workshop Programme 2012 (p. 54).
3. Kumar SS, Premjith B, Kumar MA, Soman KP. AMRITA_CEN-NLP@ SAIL2015: sentiment analysis in Indian Language using regularized least square approach with randomized feature learning. In International Conference on Mining Intelligence and Knowledge Exploration 2015 Dec 9 (pp. 671-683). Springer, Cham.
4. Prasad SS, Kumar J, Prabhakar DK, Pal S. Sentiment classification: an approach for Indian language tweets using decision tree. In International Conference on Mining Intelligence and Knowledge Exploration 2015 Dec 9 (pp. 656-663). Springer, Cham.
5. Sarkar K, Chakraborty S. A sentiment analysis system for Indian language tweets. In International Conference on Mining Intelligence and Knowledge Exploration 2015 Dec 9 (pp. 694-702). Springer, Cham.
6. Venugopalan M, Gupta D. Sentiment classification for Hindi tweets in a constrained environment augmented using tweet specific features. In International

Conference on Mining Intelligence and Knowledge Exploration 2015 Dec 9 (pp. 664-670). Springer, Cham.
7. Mukku SS, Choudhary N, Mamidi R. Enhanced Sentiment Classification of Telugu Text using ML Techniques. In SAAIP@ IJCAI 2016 Jul 10 (pp. 29-34).
8. Sarmah J, Saharia N, Sarma SK. A novel approach for document classification using assamese wordnet. In 6th International Global Wordnet Conference 2012 (pp. 324-329).
9. Dhar A, Dash NS, Roy K. An Innovative Method of Feature Extraction for Text Classification Using PART Classifier. In International Conference on Information, Communication and Computing Technology 2018 May 12 (pp. 131-138). Springer, Singapore.
10. Sharma R, Nigam S, Jain R. Polarity detection movie reviews in hindi language. arXiv preprint arXiv:1409.3942. 2014 Sep 13.
11. Gupta V, Joshi N, Mathur I. Advanced Machine Learning Techniques in Natural Language Processing for Indian Languages. In Smart Techniques for a Smarter Planet 2019 (pp. 117-144). Springer, Cham.