International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2020) pp. 276-285


Received: 17 February 2020; Accepted: 15 July, 2020; Published: 1 August, 2020

# An Adaptable Scheme to Enhance the Sentiment Classification of Telugu Language

**Midde. Venkateswarlu Naik[1], D. Vasumathi[2] and A.P. Siva Kumar[3]**

[1]Research Scholar, Department of Computer Science and Engineering,
JNTUA, Ananthapuramu, Andhra Pradesh - 515002, India
Hyderabad - 501401, Telangana, India
*middevenkateswarlunaik@gmail.com*

[2]Professor, Department of Computer Science and Engineering,
JNTUH College of Engineering, Kukatpally,
Hyderabad - 500085, Telangana, India
*vasu_devara@jntuh.ac.in*

[3]Assistant Professor, Department of Computer Science and Engineering,
JNTUA, Ananthapuramu, Andhra Pradesh - 515002, India,
*sivakumar.cse@jntua.ac.in*

***Abstract*: Nowadays, the big data is ruling the entire digital world with its applications and facilities. Thus to run the online services in better way some of the machine learning model is utilized, also the machine learning strategy is became a trending field in big data; hence the success of online services or business is based upon the customer reviews. Almost the review contains neutral, positive, and negative sentiment value; this specification is done using natural Language Processing (NLP). Manual classification of sentiment value is a difficult task so that the Natural Language Processing (NLP) scheme is used which is processed using a machine learning strategy. Moreover, the part of Speech Specification for different language is difficult. To overcome this issue, the current research developed a CatBoost machine learning model with Less Error Pruning (LEP)-Shortest Description Length (SDL) and Ant Lion Optimization (MOALO) approach to classify the sentiment values in Telugu reviews. The purpose of using LEP-SDL is to remove unwanted characters and make the classification process easier. Several error removing models are available for machine learning process but those models are ineffective when it comes under to remove the error in Telugu Language, so that LEP-SDL model is developed here. Moreover, the fitness function of ALO is used in the catboost classification module improves the accuracy of sentiment classification. In addition, the proposed approach is implemented using python; the efficiency of the proposed model is compared with recent existing works and achieved better results by attaining high accuracy and precision rate of sentiment classification. The obtained results were justified that the proposed model is applicable for online services or businesses to classify the sentiment rates of each customer.**

***Keywords*:** Big data, Natural Language Processing, Sentiment analysis, Ant Lion Optimization, Telugu language, CatBoost classifier**.**

## I. Introduction

One of the greatest technologies was introduced to deal with structured and unstructured data called big data [1]; it carries the social media data like Facebook, Google, twitter...etc., [2]. Moreover, the big data is the massive amount of shared storage model for query processing in all applications. Thus the conventional data separation method is not enough to deal the big data in social media, to resolve that machine learning strategy is introduced [3]. In that, sentiment or opinion analysis is used to structure the review comments in social media. Hence, it is the approach to define the emotion or feeling of a text. Moreover, Sentiment analysis accuracy comes under two forms, specifically degree and polarity [4]. Large data in NLP is another trending opportunity because it is in the form of artificial intelligence which supports machines to read the text by replicating the human facility to understand the verbal communication. The NLP techniques are incorporate with various methods including statistics, semantics, linguistics and machine learning, relationship, and understand context [5] which enables an understanding of what's being written in an inclusive way. Moreover, the document can retrieve through the standard NLP techniques; hence it is the scientific discipline worth concerned.

Many of the applications have been developed to annotate the part-of-speech [6]. However, the challenging task in machine learning is sentiment analysis which covers different techniques and approaches that directly enable the opinion oriented data seeking system [7]. The NLP is a superset of opinion mining however; there are several types of problems in different classes. Normally, sentiment analysis is used to classify the text [8] which is positive or negative. Consequently,

deep learning in sentiment analysis helps to find detailed information about the particular information. The important task of senti-worth relies upon to detect the hidden subjective expression [9] in the text or data. Thus the identification of sentiment value is evaluated using the verb framework model. The main focus of Sentiment evaluation is to categorize the sentiment in the file or document, for that it specifies the given text into three levels that are document level, sentence level, feature level [10]. The computer has no natural language to communicate or understand human behavior. For this reason, machine learning scheme is introduced. Moreover, for the sentiment or opinion categorization the system must know the emotion of human behavior like delight, annoyance, fury, etc., [11]. Opinion evaluation in NLP is used to classify human emotions with the use of the machine. Furthermore, sentiment examination is a method which is used to examine the opinions of users or customers towards a cinema, product, reports, events or organization, etc. This model attempts the machine that has proficiently understood the context, sentence formation, also more attentive on the series of code words [12]. NLP approaches are worn to label parts of speech, entity specification, target words, etc.

Sentiment classification is also termed as Opinion mining. It is a type of text mining that classifies opinion identification, sentiment categorization; feature extraction and result summarization have been evaluated in three cases which are neutral, positive, and negative [13]. In recent years, many researchers have been presented opinion mining in various levels, like naïve Bayes [14], map reduce [15], benchmarking on datasets [16], RBEM deep learning [17], Emotion tracker [18], etc., still selection of suitable word in the large data set is impossible. So the current research work planned to develop a hybrid machine learning model to classify the sentiment values in Telugu news review data. The rest of this article is categorized as follows; Section 2 elaborates related works of NLP in Big data. Section 3 explains the problem statement. Section 4 elaborates proposed methodology, section 5 demonstrates the outcome of the proposed work and comparison and section 6 concludes the paper.

## II. Related Work

Some of the recent literature related to sentiment specification in big data is summarized below;

In the last few years, researchers have shown their interest in opinion mining, in the context of Indian languages such as Telugu, Hindi, Marathi, Punjab, Bengali, etc. Moreover, the microscopic work is done in sentiment analysis but its performance became poor. For that Reddy Naidu *et al* [19] deployed a computational technique "two-phase sentence-level sentiment analysis" to develop Indian languages in SentiWordNet. For the subsequent work, they have elaborated their work in Telugu SentiWordNet via an interactive gaming approach. Thus the proposed scheme attained 80% accuracy in sentiment classification.

Sentiment analysis or opinion mining is an important area of NLP within large and growing literature in big data, such as Twitter, Facebook, and social media hosting websites such as YouTube and Flickr have become familiar in recent years. This emerging field has been attracted a large research interest, and the work mainly focuses on customer reviews, but sentiment analysis in Telugu reviews are some more complicated. So Benarji Tharini and Dr. Vishnu Vardhan Bulusu [20] proposed CLUTO (Software for Clustering High-Dimensional Datasets) a standard partition clustering algorithm. So they mined the product description from the raw reviews also enhance sentiment classification accuracy.

Big data is filled with a volume of structured or unstructured data. Moreover, the success of the online service is depending upon the feedback of the customers, thus the feedback procedure is more important in online services. Mandhula Trupthi *et al* [21] proposed structured query language based feedback collection. Moreover, to classify the false and positive reviews the decision trees are used, hence the query based model is utilized for feedback collection.

In a big data approach, opinion mining becomes as an important task for classification or decision purposes. To enhance and improve the efficiency of the sentiment model various machine learning algorithm is used. Thus Badr Ait Hammou *et al* [22] proposed a neural network scheme in sentiment analysis; from this they classified the customer emotions with high accuracy.

This literature deals with medical datasets, NLP and machine learning scheme is used to evaluate and investigate drug delivery. Finally, it has provided the investigation report to the user, in some cases the classification may difficult because of different drug categories. Kajal Negi *et al* [23] proposed the suspect drug classification scheme based on the machine learning model. Finally, the developed scheme achieved a high classification accuracy rate.

NLP used a large amount of data with the help of computer programs, this literature focus machine learning for a grammatical purpose. Haiqin Yang *et al* [24] proposed word embedding schemes with recurrent and convolution neural networks to performing the automatic grammar correction. The automatic grammar correction theory is validating over the benchmark dataset. Finally, the proposed work can helpful for many computer vision applications.

Opinion mining or Sentiment analysis of Telugu social media texts have to face several challenges because Telugu is a morphologically complicated language. Most of the sentiments analyzing strategies are used like SVM or expert-defined lexicons since it doesn't well in polarity classification. To address this problem Gangula Rama Rohit Reddy and Radhika Mamidi [25] created a corpus for multi-domain and perform naïve Bayes for data mining also Scikit learns a framework for polarity classification. Thus they can enhance the multi-domain model in sentiment analysis also perform the classification of the polarity for suitable emotions in a hybrid manner.

In recent years, the opinion evaluation in traditional data is too difficult because of its complexity. So, Johannes von Bloh *et al* [26] made sentiment analysis in a newly developed data set which is based on German press organization. Thus the opinion classification accuracy is achieved in different categories like the relationship between business partners, product review, movie categories, etc. The key contributions of the research are summarized as follows:

- Telugu news dataset is taken for this research.
- Initially, the machine is trained with the Telugu dataset with the help of the CatBoost machine learning algorithm.
- The data is scrubbed by the LEP-SDL which is present in the CatBoost mechanism.
- Moreover, the presence of the LEP-SDL error pruning mechanism is to prune error (similar sentence, unwanted characters) which is present in the dataset.
- Subsequently, the sentiment classification is processed by CatBoost with MOALO the fitness function of Ant Lion model enhanced the classification accuracy.
- Finally, the output is evaluated in Python and its efficiency is compared and justified with recent existing approaches.

## III. System Model and Problem Statement

Nowadays, the big data rule the computerized worlds with its huge application. A large number of structured and unstructured data is defined as big data. So organizing the big data is a challengeable task for that many of the NLP techniques are processed over the big data to classify the reviewers or customer opinions for each and every application.

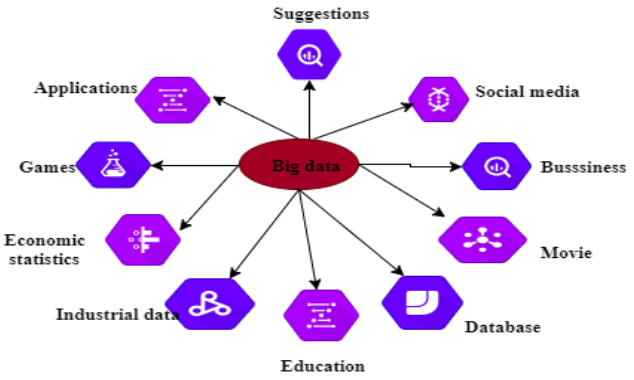


**Figure 1.** Big data system model

The system model of structure and unstructured data is defined in fig. 1; this huge number of data causes the sentiment classification as a difficult task. So that several efficient machine learning algorithms are discovered to end this problem, but still suitable solution is not found. Thus to end this opinion classification issues, the machine learning boosting approach with heuristic mechanism and error pruning model is planned in this paper.

## IV. Proposed Methodology (CatBoost with LEP-SDL and MOALO)

Initially, the datasets are trained to the system with the use of CatBoost machine learning model. Subsequently, unwanted characters, other symbols, etc., are reduced by LEP-SDL. After that cleaned data are preprocessed to the CatBoost with MOALO approach initially training and preprocessing errors are removed after that it classified the sentiment for each sentence in fig. 2. Furthermore, the classification process of the CatBoost mechanism is carried out with the fitness value of the ALO approach.

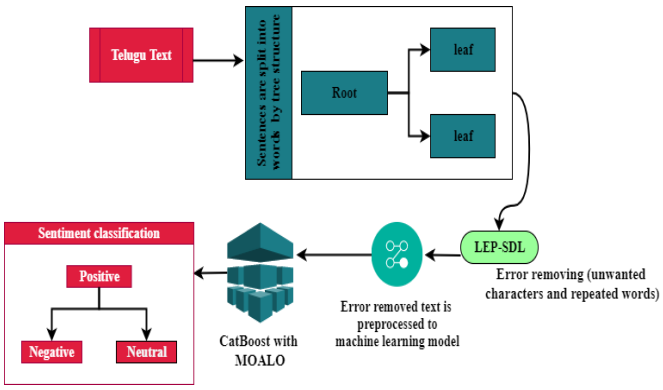


**Figure 2.** Proposed Methodology (CatBoost with LEP-SDL and MOALO)

### A. *LEP-SDL*

Initially, the trained data is forward to the preprocessing step which is the error pruning model. The function of this LEP-SDL model is to remove the error such as non-Telugu words, repeated words, extra spaces, and dots, tec. The initial process of error removing is elaborated in eqn. (1) [30], here $l$ is the total sentence and $j$ is the specified sentence and $k$ represent each word.

Algorithm 1: LEP-SDL

int $(j, l, K, \Pr)$

If $K = [l_{j,k} = l_{j,k}]J$

//analyzing parameter K searching for the repeated words and unwanted characters`

{

$\Pr = (m-1)$

// $\Pr$ is the pruning parameter and $m$ is the repeated word, $\Pr$ prunes the repeated word

$\Pr = o$ // neglect

// $o$ is the unwanted characters, when $\Pr$ finds $o$ while searching it neglects the unwanted character like extra dots, comma, etc.

End if

}

$$\frac{\sum_{j=1}^{n}[l_{j,k} = l_{j,k}]J}{\sum_{j=1}^{n}[l_{j,k} = l_{i,k}]} \quad (1)$$

The error removing process is diagrammatically shown in fig. 3.

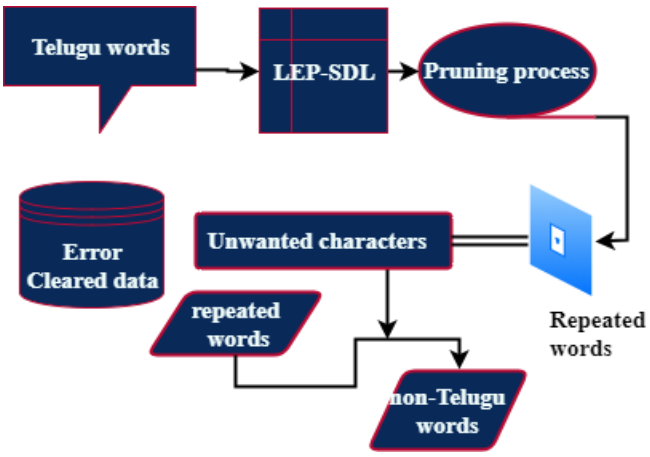


**Figure 3.** Error Pruning LEP-SDL

The novel LEP-SDL Model is utilized to remove unwanted characters such as non-Telugu words, extra dots, additional spaces, repeated characters, etc. The purpose of using this LEP-SDL is to minimize the length of the decision trees by removing the unnecessary characters which contain less parameter values. Moreover, the sentiment classification process is easier while errors are removed; the pruning function is shown in eqn. (2) [30],

$$\Pr = \frac{l_j}{l_k} + (m-1) \tag{2}$$

From this process, it removes the unwanted characters and reduced the word set length to improve the classification process. The process of LEP-SDL is elaborated in algorithm 1. Moreover, here the LEP-SDL is the novel approach; the purpose of developing this model is the conventional error removing model is not enough to remove the error in Telugu data like repeated and unwanted characters. So that, the LEP-SDL model is used to prune the error thus it enhances the classification model.

### *B. CatBoost-MOALO*

Boosting is the classification model, which is utilized to make the classification process easier. In this current research work, the classification model of the CatBoost model is enhanced with Ant lion approach [29]. The Ant lion is the optimization approach also it is the hunting mechanism, which is often utilized to solve the optimization problem. After the error pruning the text is arranged in order by eqn. (3). The process of Catboost with MOALO is elaborated in algorithm 2.

$$Y_g = [0; b^*(1); b^*(1) + b^*(2) : \ldots\ldots : \sum_{h=1}^{b-1} b^*(h^*); h = 1 \sum_{h=1}^{M} b^*(h^*)] \tag{3}$$

Here, $b^*$ is the sentiment word and $h*$ is the classification element, at the beginning of the classification process the sentiment word[1] is trained to the ant lion model as input function. To process the machine learning model, the training samples of sentiment word is arranged in the form of 0's and 1's and -1 represent the neutral value which is detailed in eqn. (4).

$$b^* = \begin{cases} 0 & if random \rightarrow negative \\ 1 & if random \rightarrow positive \\ -1 otherwise & \rightarrow neutral \end{cases} \tag{4}$$

While the opinion classification it randomly takes the sentences. If the sentence contains a negative value then its result became '0'. If the sentence contains a positive value then its result became '1'. The sentences which could not fall under both condition then it is termed as n neutral sentences and its value became '-1'.

Algorithm 2: Pseudo code for Catboost with MOALO

Initialize $L, M, K, t_i$

//Here $L$ is the collection positive key words and $M$ is the collection of negative key words also $t_i$ is the next iteration element.

$\{(L_k, M_k)\}k^{n}_{=1}$

*Label* $\{0,1,-1\}$

Model=CatBoost MOALO classifier $(iteration = no \quad of \quad sentences)$

Initialize K as the Telugu words and $t_i$ is the iteration of next sentence

establish the best while $(K = 1) \rightarrow L$

//Then that specified sentence is classified as positive

Initiate iteration number $t_i = next$

// $t_i$ is the iteration of the next sentence

While $(t_i = 0 \rightarrow M)$

{

for every sentence $(M = Negative)$

}

If $(t_i \neq 0,1) \rightarrow -1$ //Neutral

end while

Return the finest classification

The process of the proposed strategy is shown in fig. 4, initially, a set of Telugu sentence is trained to the system. In the first step, the training and dataset errors are removed by LEP-SDL model. Consequently, the error free data is trained to the CatBoost MOALO Classification module, the presence of ant lion fitness function in classification model enhances the sentiment classification accuracy.

[1] Aspect terms

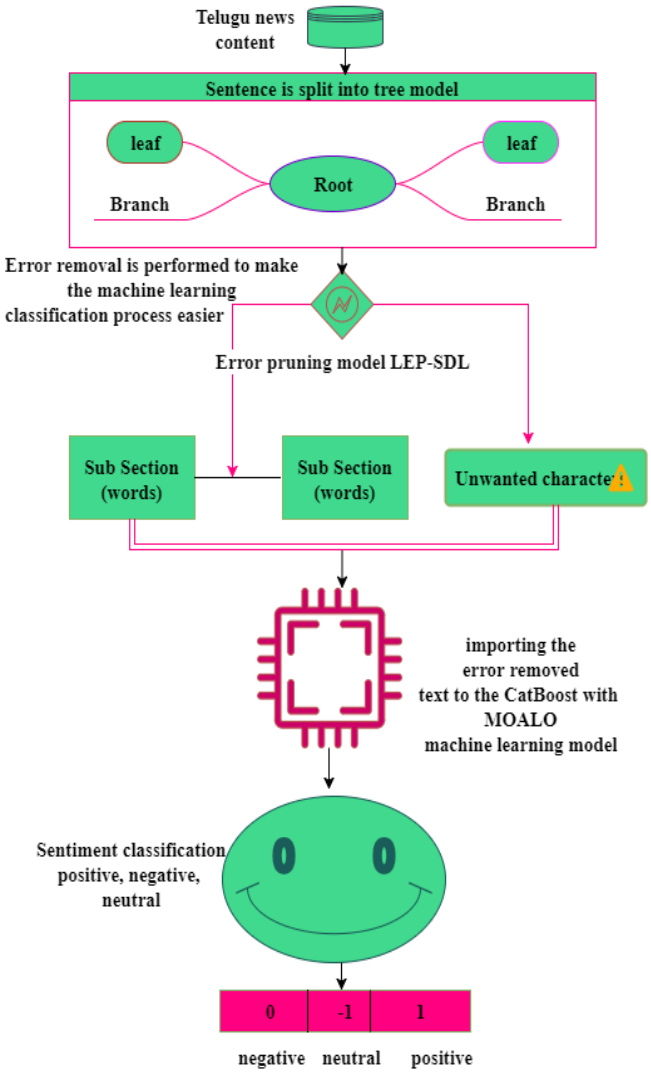


**Figure 4.** Work flow of CatBoost with LEP-SDL and MOALO

## V. Result and Discussion

The proposed CatBoost with (LEP-SDL and MOALO are elaborated in Python 3.7.0, running in windows 10 platforms. The proposed novel algorithm is utilized to improve the sentiment classification such as negative, positive, and neutral.

Dataset Evaluation-Dataset generation: One of the Dravidian languages is Telugu which is native to India; it ranks third by the number of inhabitants speakers in India. Telugu dataset requires indispensable processing to discover the word embedding model, and for sentiment information removal. Here Telugu data set is collected from different sites in social media as shown in Table 1 so it is hard to find labeled information of sentences for Telugu sentiment analysis.

### *A. Case study*

To estimate the performance of the proposed methodology, Telugu news is taken. Thousands of Telugu news sentences are taken for an evaluation process. Initially, it is split up into tree structures with the use of CatBoost classifier. Then by the error pruning LEP-SDL model, the training errors are pruned. Some of the Telugu Text is elaborated in Table 1.

*Table 1.* Telugu News Contents.

| S. No | Telugu Text | Meaning in English | Positive | neutral | Negative |
|---|---|---|---|---|---|
| 1. | బుర్హాన్వనికాశ్మీర్లో ఒకస్కూలుహెడ్మాస్టర్కొడుకు | Burhan is a school headmaster son of Kashmir | - | -1 | - |
| 2. | సామాజికన్యాయంకోసంఆయనఅవిశ్రాంతంగాకృషిచేశారు. | He has worked hard for social justice. | +1 | - | - |
| 3. | అనేకకుట్రలకు పాల్పడింది. | Many of the conspiracies. | - | - | 0 |
| 4. | ఒక్కముక్కలోఅర్థంచేసుకోవాలంటేఆయనఫిలిప్పీన్స్ట్రంప్. | To understand one piece, he is a Philippine trump | - | -1 | - |
| 5. | దీంతోతాన్యానాయక్తాండాలోవిషాదంనెలకొంది. | The tragedy has occurred in Tanyaayak antanda. | - | - | 0 |
| 6. | 2012 నుంచిఅల్ఖైదాఉగ్రవాదులుతమకార్యకలాపాలనుఫ్రాన్స్లోఎక్కువచేశారు. | Since 2012, Al Qaeda militants have done their best in France. | +1 | - | - |

Let us consider 10 sentences are adopted from the Telugu newspaper and its word specification is 200. $l=10$, $k=200$ and $j=1$. While substituting these values in eqn. (1), eqn. (5) is obtained.

$$\frac{\sum_{1=1}^{n}[10_{1,200}=10_{1,200}]1}{\sum_{1=1}^{n}[10_{1,200}=10_{1,200}]} \quad (5)$$

Here m is the repeated words, $pr$ is the error pruning parameter, if $l_j=1$, $m=2$, and $l_k=200$. These imagined values are substitutes in eqn. (2), eqn. (6) is obtained pr=1.05.

$$\mathbf{Pr}=\frac{1}{200}+(2-1)=1.05 \quad (6)$$

Thus one word repeated by two times that is m=2 by the

pruning mechanism one word is removed $\Pr = 1.05$. From this process, it reduced the word set length and it helps to improve the classification process. The error removed data is again trained to the CatBoost MOALO mechanism for the sentiment classification process. Initially, the classifier categorizes the aspect terms after that sentiment classification is done in an effective manner by the use of fitness function mentioned in algorithm 2.

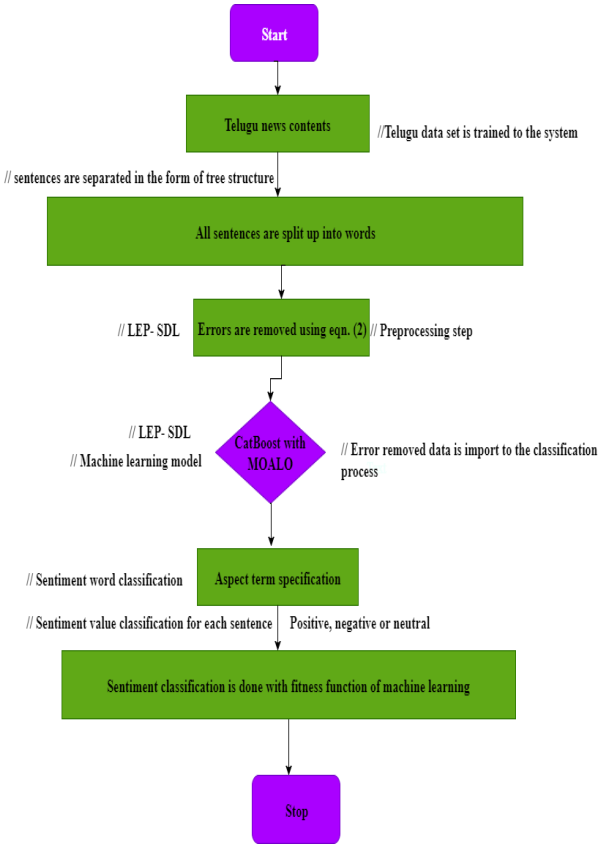


**Figure 5.** Working Flow of CatBoost with LEP-SDL and MOALO

The working flow of the proposed model is represented by the flow model in fig. 5. Here the positive word is represented as 1 and the negative word is represented as 0 and neutral sentiment word is represented as -1. Thus, the sentiment classification is processed successfully.

### *B. Performance Metrics*

The efficiency of the proposed method is analyzed with existing work such as SentiWordNet (SWN) [19], Support Vector Machine (SVM) [28], benchmark [20] and Sentiment Analysis of Code-Mixed Text (SACMT) [27]. Thus to evaluate the efficiency of the proposed approach some of important metrics should be validated such as accuracy, precision, error rate, recall, F-measure, aspect terms, and opinion specification.

In this current research work the trained total compound sentence as 0.4215%, the positive sentence as 0.114%, the negative sentence as 0.065%, and neutral sentence as 0.822%, the dataset specification is shown in fig. 6.

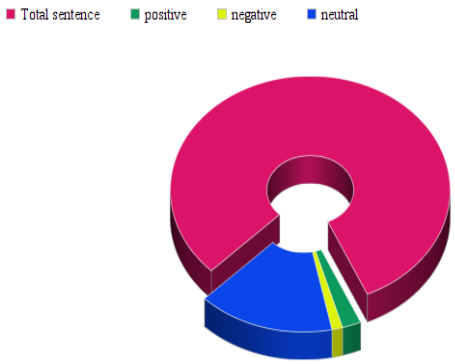


**Figure 6.** Dataset specification

#### *1) Accuracy*

The accuracy of the sentiment value can be calculated using eqn. (7). Here $TN$ is True Negative, $TP$ is True Positive, $FN$ is False Negative and $FP$ is False Positive. The comparison of accuracy measures with existing techniques is shown in fig. 7 and Table 2.

$$accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \quad (7)$$

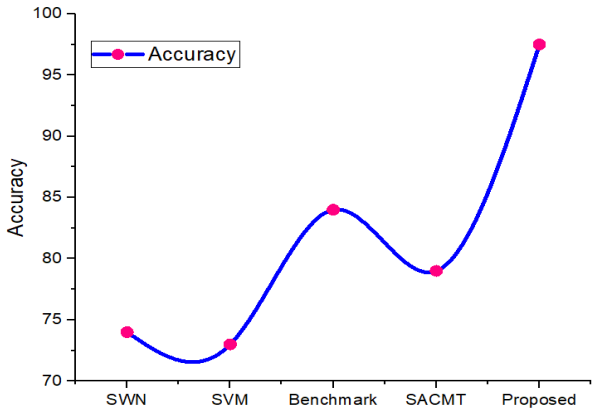


**Figure 7.** Accuracy

SentiWordNet (SWN) attained accuracy measure as 74%, SACMT earned 79% as the accuracy of classification, SVM pertained accuracy rate as 73% and benchmark strategy gained accuracy rate as 84%. Subsequently, the developed strategy gained accuracy and precision rate as 97.5%.

*Table 2.* Accuracy Comparison.

| | **Accuracy** |
|---|---|
| SWN | 74 |
| SVM | 73 |
| Benchmark | 84 |
| SACMT | 79 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 97.5 |

*2) Precision*

The precision of processed data is evaluated as the number of correct positive predictions alienated by the total number of specified opinion sentences which is validated by eqn. (8). The comparison precision with recent existing models is shown in fig. 8 and Table 3.

$$Precision = \frac{no.of\ correct\ opinion\ \ classification}{Total\ \ number\ \ of\ \ particular\ sentiment\ \ sentences} \quad (8)$$

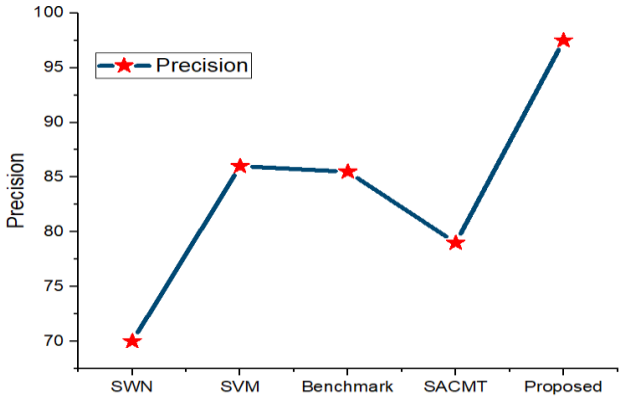


**Figure 8.** Precision comparison

SentiWordNet attained precision rate as 70%, SACMT earned 79% of precision rate, SVM pertained precision measure as 86% and benchmark strategy gained precision rate as 85.5% subsequently, the developed strategy gained the precision rate as 97.5%.

*Table 3.* Precision comparison.

| | **Precision** |
|---|---|
| SWN | 70 |
| SVM | 86 |
| Benchmark | 85.5 |
| SACMT | 79 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 97.5 |

*3) Opinion Classification*

The other word of sentiment classification is opinion classification; the sentences are classified under three kinds of sentiment values such as positive negative and neutral. The validation of opinion specification is elaborated in fig. 9 and Table 4.

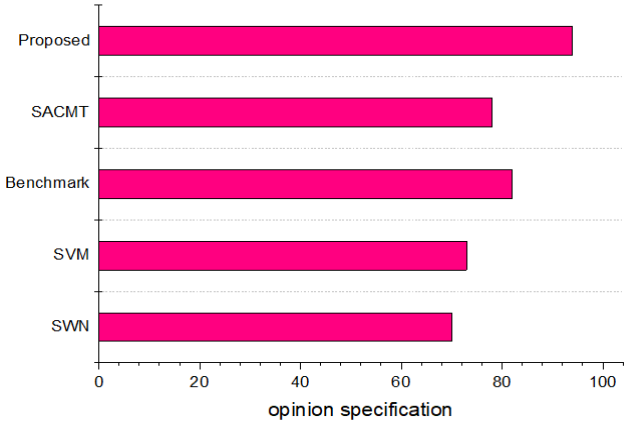


**Figure 9.** Opinion specification

By the validation the SentiWordNet attained aspect word opinion classification rate as 70%, SACMT earned 78% of opinion classification, SVM attained opinion identification measure as 73% and benchmark strategy gained opinion classification rate as 82%. Moreover, the proposed strategy achieved word opinion specification as 94%.

*Table 4.* Opinion specification.

| | **Opinion specification** |
|---|---|
| SWN | 70 |
| SVM | 73 |
| Benchmark | 82 |
| SACMT | 78 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 94 |

*4) Aspect terms classification*

The aspect terms specification is the key metric for the sentiment specification process, the separation of sentiment word is defined as aspect term specification. The comparison measure of aspect term specification is defined in Table 5 and fig. 10.

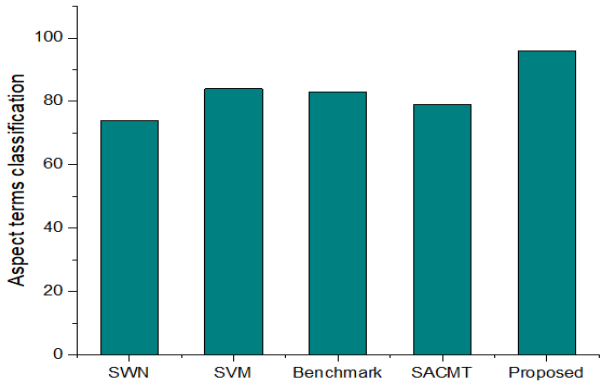


**Figure 10.** Aspect term classification

The aspect term specification rate of different approaches are shown in fig. 10 and in Table 5 In that the SentiWordNet attained aspect term specification as 74%, SACMT earned 79% of aspect term detection, SVM attained aspect term recognition rate as 84% and benchmark strategy gained aspect term ratio as 83%. Moreover, the proposed strategy achieved the aspect term classification as 96%.

*Table 5.* Aspect term classification.

| | **Aspect term classification** |
|---|---|
| SWN | 74 |
| SVM | 84 |
| Benchmark | 83 |
| SACMT | 79 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 96 |

*5) Recall*

Recall is validated as the number of correct sentiment

specifications by the total number of sentences that are shown in fig. 11.

$$\text{Recall} = \frac{\text{no.of correct sentiment specification}}{\text{number of total sentences}} \quad (9)$$

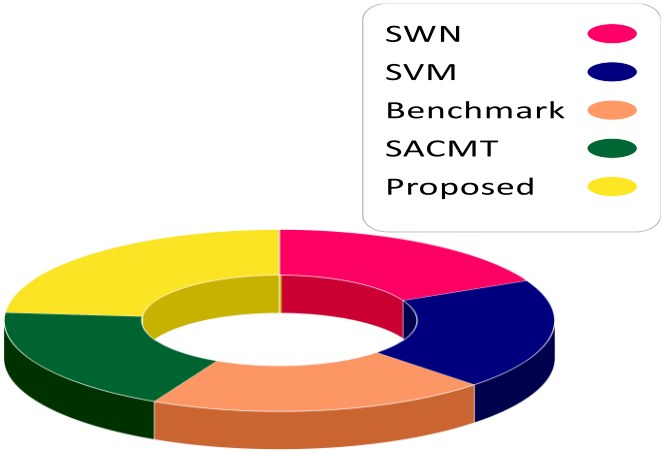


**Figure 11.** Recall measure

The recall measure of the proposed strategy 96.5 consequently, the existing model SVM attained recall as 84.3 %. Benchmark achieved 82.5% of recall SWN attained recall rate at 70% and SACMT gained recall measure as 77%. These all validation is shown in Table 6 and fig. 11.

*Table 6.* Recall measure.

| | **Recall** |
|---|---|
| SWN | 70 |
| SVM | 84.3 |
| Benchmark | 82.5 |
| SACMT | 77 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 96.5 |

*6) F-measure*

The F measure is calculated to find the mean average between precision and recall, the F measure is calculated by eqn. (10) and elaborated in fig. 12 and Table 7.

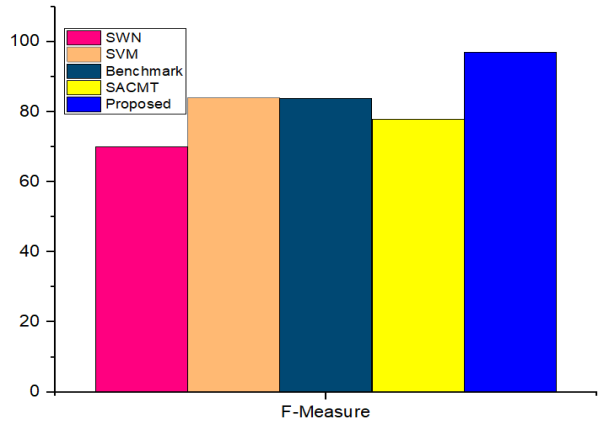


**Figure 12.** F-measure

$$F-measure = 2 \times \frac{\text{Precision} \times recall}{precision + recall} \quad (10)$$

The F-measure of the proposed strategy 96.9 consequently, the existing model SVM attained F-measure as 84%. Benchmark achieved 83.9% of F-measure SWN attained F-measure at 70% and SACMT gained F-measure as 77.9%. These all validation is shown in Table 7 and fig. 12.

*Table 7.* Comparison of F-measure.

| | **F-measure** |
|---|---|
| SWN | 70 |
| SVM | 84 |
| Benchmark | 83.9 |
| SACMT | 77.9 |
| Proposed (CatBoost with (LEP-SDL and MOALO) | 96.9 |

*C. Discussion*

The main aim behind in sentiment analysis is to improve the online business and social media services, so the suggested reviewers are more important for that this sentiment strategy is applied in big data. Because the identification of reviewer suggestions by manual is impossible in big data so that machine learning strategy is applied. The current research developed a novel pruning and machine learning model for sentiment specification of the Telugu news content. The obtained results proved that the proposed methodology attained high accuracy rate while comparing other techniques and it is applicable for big data industry to classify the opinion value.

## VI. Conclusion

This research elaborated the opinion mining in big data towards the sentiment specification of social media datasets. For that the Telugu news dataset was taken to analyze the sentiment value; the sentence might include neutral, positive, and negative. This research aimed at a brief study about opinion mining and categorization methodology to estimate the sentiment value for online customer needs in the product review, movie reviews, etc. Thus, a novel CatBoost with LEP-SDL and MOALO model enhanced the opinion classification. Moreover, LEP-SDL is a pruning mechanism which is frequently used in decision trees. So the involvement of this pruning strategy in machine learning improved the sentiment classification process. Thus, the proposed system is helpful to investigate online customer reviews in a proficient way also it helps to improve the online services. Finally, the proposed approach achieved an accuracy of 97.5% and 97.5% as the precision rate for subjectivity and sentiment categorization respectively. Furthermore, in some rare cases the specification of sentiment classification is difficult because a sentence with positive sentiment word may give negative meaning. So, in future, the development of hybrid deep learning model with a hybrid heuristic model will improve the sentiment classification of Telugu language by analyzing the meanings of the sentences.

## Acknowledgment

None

## References

[1] Y. Kumar, K. Sood, S. Kaul, R. Vasuja. "Big Data Analytics and Its Benefits in Healthcare", *Big Data Analytics in Healthcare*, Springer, Cham, pp. 3-21, 2020.

[2] K. Umapavankumar, S.V.N. Srinivasu, S.S.N. Rao. "Machine Learning Usage in Facebook, Twitter and Google Along with the Other Tools", *Emerging Research in Data Engineering Systems and Computer Communications*, Springer, Singapore, pp. 465-471, 2020.

[3] M. Chen, S. Mao, Y. Liu. "Big data: A survey", *Mobile networks and applications*, 19(2), pp. 171-209, 2014.

[4] T. Chumwatana. "COMMENT ANALYSIS FOR PRODUCT AND SERVICE SATISFACTION FROM THAI CUSTOMERS'REVIEW IN SOCIAL NETWORK", *Journal of Information and Communication Technology*, 17(2), pp. 271-289, 2020.

[5] Z. Khan, N. Iltaf, H. Afzal, H. Abbas. "DST-HRS: A topic driven hybrid recommender system based on deep semantics", *Computer Communications*, 2020.

[6] G.C. Banks, H.M. Woznyj, R.S. Wesslen. "A review of best practice recommendations for text analysis in R (and a user-friendly app)", *Journal of Business and Psychology*, 33(4), pp. 445-459, 2018.

[7] G. Alwakid, T. Osman, T. Hughes-Roberts. "Challenges in sentiment analysis for Arabic social networks", *Procedia Computer Science*, 117, pp. 89-100, 2017.

[8] A. Shrivastava, S. Pawar. "A Theoretical Concept on Opinion Mining Process and Sentiment Analysis in Data Mining", 2019.

[9] D. Kandé, F. Camara, S. Ndiaye. "FWLSA-score: French and Wolof Lexicon-based for Sentiment Analysis", In *2019 5th International Conference on Information Management (ICIM)*, IEEE, 2019.

[10] C.A. Dhawale, V.V. Chaudhari. "Sentiment Analysis Techniques, Tools, Applications, and Challenge", *Exploring the Power of Electronic Word-of-Mouth in the Services Industry*, IGI Global, pp. 35-48, 2020.

[11] M.D. Munezero, C.S. Montero, E. Sutinen. "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text", *IEEE transactions on affective computing*, 5(2), pp. 101-111, 2014.

[12] Z. Quan, Z.J. Wang, Y. Le, B. Yao, K. Li. "An efficient framework for sentence similarity modeling", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4), pp. 853-865, 2019.

[13] A. Abdi, S.M. Shamsuddin, S. Hasan, J. Piran. "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion", *Information Processing & Management*, 56(4), pp. 1245-1259, 2019.

[14] H. Gao, X. Zeng, C. Yao. "Application of improved distributed naive Bayesian algorithms in text classification", *The Journal of Supercomputing*, 75(9), pp. 5831-5847, 2019.

[15] K. Sivakumar, N.S. Nithya, O. Revathy. "Phenotype Algorithm based Big Data Analytics for Cancer Diagnose", *Journal of medical systems*, 43(8), pp. 264, 2019.

[16] K. Weiying, D.N. Pham, Y. Eftekharypour. "Benchmarking NLP Toolkits for Enterprise Application", *Pacific Rim International Conference on Artificial Intelligence*, Springer, Cham, 2019.

[17] E. Tromp, M. Pechenizkiy, M.M. Gaber. "Expressive modeling for trusted big data analytics: techniques and applications in sentiment analysis", *Big Data Analytics*, 2(1), pp. 5, 2017.

[18] D. Panda, D.D. Chakladar, T. Dasgupta. "Multimodal System for Emotion Recognition Using EEG and Customer Review", In *Proceedings of the Global AI Congress 2019*, Springer, Singapore, 2020.

[19] R. Naidu, S.K. Bharti, K.S. Babu, R.K. Mohapatra. "Text summarization with automatic keyword extraction in telugu e-newspapers", *Smart Computing and Informatics*, Springer, Singapore, pp. 555-564, 2018.

[20] D. Benarji Tharini, V.V. Bulusu. "Development of a Micro Telugu Opinion WordNet and Aligning with TELOWN Ontology for Automatic Recognition of Opinion Words from Telugu Documents".

[21] M. Trupthi, S. Pabboju, N. Gugulotu. "Deep Sentiments Extraction for Consumer Products Using NLP-Based Technique", *Soft Computing and Signal Processing*, Springer, Singapore, pp. 191-201, 2019.

[22] B.A. Hammou, A.A. Lahcen, S. Mouline. "A Distributed Ensemble of Deep Convolutional Neural Networks with Random Forest for Big Data Sentiment Analysis", *International Conference on Mobile, Secure, and Programmable Networking*, Springer, Cham, 2019.

[23] K. Negi, A. Pavuri, L. Patel, C. Jain. "A novel method for drug-adverse event extraction using machine learning", *Informatics in Medicine Unlocked*, pp. 100190, 2019.

[24] H. Yang, L. Luo, L.P. Chueng, D. Ling, F. Chin. "Deep Learning and Its Applications to Natural Language Processing", *Deep Learning: Fundamentals, Theory and Applications*, Springer, Cham, pp. 89-109, 2019.

[25] R.R.R. Gangula, R. Mamidi. "Resproposedce Creation Towards Automated Sentiment Analysis in Telugu (a low resproposedce language) and Integrating Multiple Domain Sproposedces to Enhance Sentiment Prediction", *LREC*, 2018.

[26] J. von Bloh, T. Broekel, B. Özgun, R. Sternberg. "New (s) data for entrepreneurship research? An innovative approach to use big data on media coverage", *Small Business Economics*, pp. 1-22, 2019.

[27] N. Choudhary, R. Singh, I. Bindlish. "Sentiment Analysis of Code-Mixed Languages leveraging Resproposedce Rich Languages", arXiv preprint arXiv:1804.00806, 2018.

[28] S.S. Mukku, S.R. Oota, R. Mamidi. "Tag me a label with multi-arm: Active learning for telugu sentiment analysis", *International Conference on Big Data Analytics and Knowledge Discovery*, Springer, Cham, 2017.

[29] M. Mishra, S.K. Barman, D. Maity, D.K. Maiti. "Ant lion optimisation algorithm for structural damage detection using vibration data", *Journal of Civil Structural Health Monitoring*, 9(1), pp. 117-136, 2019.

[30] B.T. Pham, I. Prakash, S.K. Singh, A. Shirzadi, H. Shahabi. "Landslide susceptibility modeling using reduced error pruning trees and different ensemble techniques: Hybrid machine learning approaches", *Catena*, 175, pp. 203-218, 2019.

## Author Biographies

**Midde. Venkateswarlu Naik** received B.Tech and M.Tech Computer Science and Engineering from JNT University, Hyderabad, India. Currently, he is pursuing Ph.D. in Computer science and Engineering in the area of Big data analytics from JNT University, Ananthapuramu, India. He has around 14 years of experience in academia from various engineering colleges. Currently he is working in CMR College of Engineering & Technology, Hyderabad. To his credit, he published 17 papers in international journals and conferences. He is also a member of various professional of technical bodies like MIETE, MCSI, and MIEEE. His areas of interest are Big data analytics, Informational retrieval, Natural language processing, Machine learning.

**D. Vasumathi** received B.Tech, M.Tech Computer Science & Engineering from JNT University, Hyderabad, India and she has been awarded Ph.D. in Computer Science from JNT University, Hyderabad, India. At present, she is working as a Professor in the Department of Computer Science and Engineering, JNTUH College of Engineering, Hyderabad. Her area of Research includes Data mining and Data warehousing, Information Retrieval, Big data analytics, Cloud computing, Network security, and DBMS. To her credit, she is having more than 100 publications in reputed International Journals and Conferences. She is also a Member of Various Technical Associations including ISTE, CSI, and IEEE.

**A.P. Siva Kumar** is received B.Tech in Computer Science and Engineering from JNTUH and M.Tech in Computer Science & Engineering from JNT University, Ananthapuramu. He has been awarded Ph.D. from JNTUA, India. Currently, he is working in the CSE Department as an Assistant Professor at JNTUA. He has around 13 Years of teaching experience. To his credit he has published 28 papers in various International Journals and conferences. His areas of research interest are Cross-Lingual Information retrieval and Natural Language Processing. He has received the award of outstanding faculty for the contribution and achievement in the field of Cross-Lingual Information Retrieval, at the Centre of Advanced Research and Design, VIF, Chennai.