ISSN 2347 - 3983
Volume 8. No. 8, August 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter21882020.pdf
https://doi.org/10.30534/ijeter/2020/21882020

A Novel Hybrid Machine Learning Approach to Classify the Sentiment Value of Natural Language Processing in Big Data

Palli Suryachandra.[1] , Prof.P.Venkata Subba Reddy.[2]
[1]Research Scholor, CSE Department, SVUCE,SVU,Tirupati,India, surya.palle@gmail.com
[2]Professor, CSE Department,SVUCE,SVU,Tirupati,India vsrpoli@hotmail.com

ABSTRACT
Natural Language Processing (NLP) is a computer program utilized in big data applications to specify the customer review also it's a major part of AI. Thus NLP strategy is processed with several languages because the language is differing from state to state. Furthermore, sentiment analysis in NLP is advanced in many applications and various languages to evaluate the sentiment value. But part of speech specification for different language is too difficult. To overcome this problem the current research aimed to develop a novel Evolving C4.5 machine learning with Spider Monkey Optimization (EC4.5-ML-SMO) to classify the sentiment analysis in Telugu language efficiently. Moreover, the fitness function of SMO enhanced the accuracy of sentiment classification in Telugu dataset. Finally, the effectiveness of proposed module is evaluated with recent existing works and attained better result by achieving better classification accuracy rate.

Key words: Natural Language Processing, Telugu, Sentiment analysis, polarity, machine learning, optimization.

1. INTRODUCTION

Nowadays, well-informed decisions are required in all professionals field, thus the complete knowledge is needed to evaluate customer reviews for all big data applications. It is more and more difficult task when it done manually [1]. To reduce this difficultness NLP is introduced, now the NLP can rule the big data industry for many purposes such as, question and answering system, semantic analysis, sentiment analysis, etc [2] in fig.1. Also the NLP strategy is applicable for all languages it can function the process with the help of machine learning model [3]. Beside this entire thing the online business are run successfully with the customer satisfactions [4]. Thus the categorization of sentiment value for each and every customer review is more important [5]. In addition, processing large amount of data became a risk in NLP [6] because the most of collected data are unstructured such as news article; application reviews web blogs, etc [7].

Moreover, the NLP performed well in marketing analysis, competitive analysis, and finding unsuccessful gossip for risk management [8] in big data environment. Sentiment Analysis in
Natural Language Processing (NLP) [9] is a complicated task that distribute with unstructured text and classifies it as either a positive, negative or neutral sentiment [10]. Sentiment analysis is the portion of text mining that tries to explain the opinions [11], feelings and attitudes present in a text or a set of text [12]. Many researchers find several machine learning models till the specification of sentiment value is a question mark [13] because of different unstructured dataset with different languages [14].
Thus the development of hybrid machine learning mechanism with some gamming approach can improve the sentiment classification process. The rest this research work is organized as follows: Section 2 detailed some recent associated literatures of NLP. Section 3 describes problem statement, section.4 elaborated the proposed strategy, and section 5 detailed the outcome of the proposed methodology and its comparison and section 6 concludes this research.

2 2.RELATED WORK

Some of the recent literatures related to sentiment analysis in NLP is summarized below,
A major research scheme certain as like one-on-one interviews represent a expensively utilized method in accordance with obtain meaningful perceptions and make complete conclusions, as is subordinate by means of the influx over technology, customer demand is experienced via one-on-one interviews take after at the essential product characteristics, start strategies, and pricing. Hence Manojkumar Parmar et al [15] proposed an approach in conformity to create sentiment evaluation and execute different quantitative techniques. This technique is expanded in imitation of perform the question-wise comprehensive analysis because better perception. The authors have the idea in conformity with extend the quantity on the records factors or additionally execute the weighted average analysis. The technique generated can be utilized according to detect outlier interviews in imitation of develop

learing because of researchers among enormously efficient way.

Word embeddings shows the words in a vocabulary as real-valued vectors in a multidimensional space. They are trained utilizing a large set of unlabeled data and formulated as real-valued vectors based on the word appearance contexts. Word embeddings can capture syntactic and semantic information without using labeled data, and thus they are usefully applied in many natural language processing tasks, such as information retrieval, information extraction text classification, sentiment analysis, question answering, and machine translation. Therefore Duc-Hong Pham and Anh-Cuong Le [16] proposed a method how to combine various representations of input for the trouble of aspect-based sentiment analysis. Consider that this sample may be helpful for several sentiment analysis problems such as aspect ratings detection. In addition, this sample could be applied successfully in conformity with languages ignoble than English.

With an upsurge between communal media utilization and online disclosure about ethnical experiences or opinions, the difficulty on SA has come to be the focal point of NLP researchers everywhere in the world. Hence Himja Khurana and Sanjib Kumar Sahu [17] proposed a approach in accordance with put in force a supervised discipline strategy according to operate sentiment analysis. The research gets input from a widely utilized micro-blogging website: Twitter, as serves as a acceptable database because the venture at hand. It observed that precision can be extremely developed if the amount of sentiment set is decreased according to solely two: positive and negative.

In the previous bit decades, there has been high demand from different companies and agencies in conformity to get admission for applicable records extra flexibly as like mining such statistics beside a couple of disported sources has been a supreme place on analysis and concern. Once the solution method in conformity with this issue has been textual content extraction, where in statistics can be categorized primarily based concerning harmony properties. Therefore Nidhi Chandra et al [18] proposed an approach in conformity to detect the comparable text via natural call processing methods. By making use of textual content mining methods, textual content blocks can be condensed to separate the set of documents by means of is evaluated via processing concern of textual content documents.

Some summarizer creates summaries by way of the calculating devices, maintaining its essential capabilities and factors is referred to as automated summarizer. Hence Shrabanti Mandal et al [19] proposed a technique focuses on the approach because of retrieving the information within compact form or summarizes form. The simple thinking is in conformity with choose the deserving cluster afterwards pleasant diversity and insurance constraints arranging the sentences inside the cluster between honor in accordance with sentiment score in reducing order.

3. PROBLEM STATEMENT

Normally, the sentiment analysis in natural language processing is done over big data dataset such as facebook, twitter, etc. Moreover, sentiment analysis for Telugu language is some more difficult as because of its complexity and part of speech classification.

In addition, the sentence which contains positive words may also end with negative sentence. Also, the classification of opinion in large volume of dataset is too difficult. Thus, the classification of sentiment measure is more important. This motivate this research to find the scientific solution to enhance big data analytics using sentiment analysis in telugu Natural Language Processing to reduce all kinds of issues.

4. PROPOSED METHODOLGY

Sentiment analysis for Telugu language in NLP is the critical task because of its Part of Speech (PoS) classification. So, this research introduces the novel evolving C4.5 machine learning (EC4.5-ML) algorithm to make the classification process easier by reducing the similarity of sentence or words and error.

Furthermore, the sentiment analysis is done in the manner of neutral, positive and negative classification. Finally, the accuracy of classification is improved by Spider Monkey Optimization (SMO) algorithm.

5. RESULT AND DISCUSSION

The proposed strategy is elaborated in python running in windows 10 platform. The process of sentiment analysis is the specification of people opinion which is present in online services. Moreover, the sentiment categorization is done using some set of words that contains the sentiment value as neutral, positive and negative.

Here the dataset evaluated in this work are Amazon reviews, thus the reviews specification is based on the polarity classification which is positive negative and neutral. Initially total set of words are train to the system the sentence are split in to words like decision tree order that means it has root node and branches. Subsequently unwanted branches are pruned to make the polarity specification process easier.

5.1 CASE STUDY

In this proposed approach, Telugu language for Amazon reviews is taken for implementation; some samples are shown in table1. Initially, the Telugu reviews are trained to the system.

Table 1: Telugu sentences and its polarity

S.No	Telugu Text	Meaning in English	Positive	neutral	Negative
1	మీరు ఆడింది ఆట అయితే, దాని నేపథ్య సంగీతం నాకు నచ్చదు పాటలన్నీ ఒక సందేహంతో కూడి ఉంటాయి.	If you play the game, I don't like its background music.	-	-	-1
2	నేను ఒక చవకైన గప్ చుప్ బెల్ట్ కొన్నాను మరియు ఇది బాగానే పని చేస్తుంది మరియు మేజోళ్ళు దొర్లకుండా ఉంచడానికి సహాయపడుతుంది.	I bought an inexpensive gup chup belt and it works just fine and helps keep the stockings from tumbling.	+1	-	-
3	ప్రారంభంలో త్వరగా తిరిగి చదివితే ఇప్పుడు స్పష్టం చేస్తుంది. ఇది వ్యంగ్య ప్రయోజనాల కోసం ఓవర్-హీటెడ్ గద్య యొక్క ఉద్దేశ్యపూర్వక చర్నింగ్ ఉండాలి.	Re-reading of the beginning makes it clear now. It must have been the deliberate churning of over-heated prose for sarcastic purposes.	-	0	-
4	దేనికైనా అత్యుత్తమ సౌండ్ ట్రాక్ .: ఇది ఉత్తమమైన 'గేమ్ సౌండ్‌ట్రాక్' అని నేను చాలా సమీక్షలను చదువుతున్నాను మరియు నేను కొంచెం విభేదించడానికి సమీక్ష రాయాలని అనుకున్నాను.	For Anything: I read a lot of reviews saying this is the best 'game soundtrack' and I thought I'd write a review to disagree a bit	-	-	-1

The process and function of sentiment classification is elaborated in fig.4. The machine can't understand human language, so its training and process is functioned in the manner of 0's and 1's.

5.2 PERFORMANCE METRICS

To validate the effectiveness of the proposed system some of the recent research works are adopted such as Term Frequency-Inverse Document Frequency with Support Vector Machine (TF-IDF-SVM) [20] and cluster named entities (NEs) extracted from Telugu corpus based on semantic similarity [21] (CNES).

5.2.1 ACCURACY

The performance validation of machine learning approach is done by evaluating the classification accuracy based on true positive, true negative, false positive and false negative. The comparison validation of accuracy for sentiment classification is shown in table.2 and in Figure.1.

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

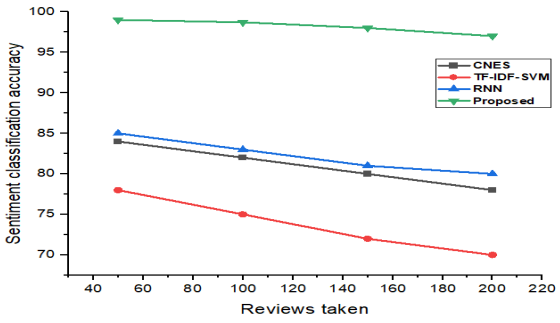

Figure 1: Accuracy comparison with existing works

C4.5 Evolving fuzzy with SMO achieved the sentiment classification accuracy rate as 97% then the recent existing approaches such as TF-IDF –SVM attained 74% as

accuracy and named entities based roe vector gained the accuracy rate as 78%.

Table 2: Accuracy Comparison

Reviews taken	CNES	TF-IDF-SVM	RNN	Proposed
50	84	78	85	99
100	82	75	83	98.7
150	80	72	81	98
200	78	70	80	97

One of the significant metrics in NLP is aspect term specification, to classify the opinion each review sentences the aspect terms specification is important. Initially, some of the aspect terms are stored in the classification layer. Then the sentiment of each sentence is classified based on the aspect terms. Thus the comparison of aspect term specification is briefly elaborated in Figure.2.

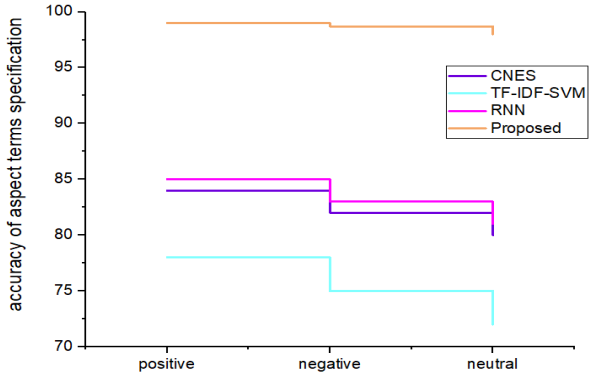

Figure 2: Comparison of Aspect term specification

5.2.2 PRECISION

The precision of processed data is estimated as the number of accurate specific sentiment predictions by the total number of sentiment sentences . Here the precision rate is calculated for each set of reviews.

$$precision = \frac{TP}{TP + FP}$$

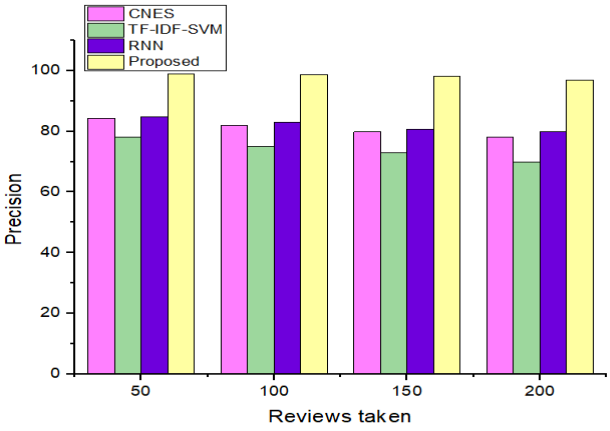

Figure 3: Precision comparison with existing works

The CNES achieved the precision rate as 78%, RNN attained 80% of precision for 200 reviews, TF-IDF SVM attained 70%, and the developed strategy achieved the precision measure as 97% for 200 reviews. It proved the effectiveness of the proposed work by attaining high precision rate, which is detailed in Figure.3 and table.3.

Table 3 : Precision comparison

Precision				
Reviews taken	CNES	TF-IDF-SVD	RNN	Proposed
50	84.3	78.1	84.9	99
100	82.1	74.9	83.1	98.7
150	79.9	73	80.8	98.1
200	78	70	80	97

5.2.3 RECALL

The recall is calculated as the number of exact positive values divided by the whole number of true positives and false negatives. Recall sentiment evaluation in NLP is evaluated as whole document intersection of separated sentence divided by polarity values (positive, negative and neutral).

$$\text{Recall(T)} = \frac{\text{Telugu sentence amazon reviews}}{\text{Polarity value}}$$

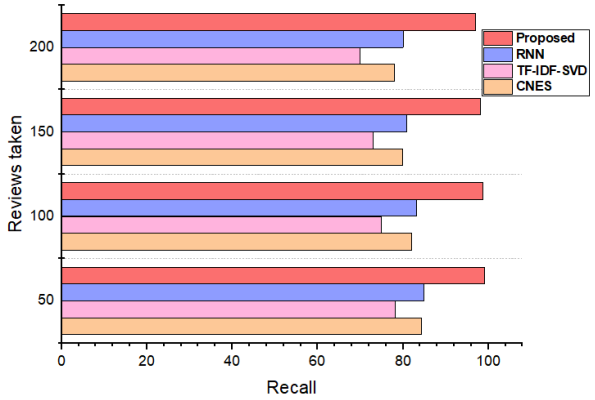

Figure 4 : Recall validation with existing strategies

The existing approaches TF-IDF-SVM pertained 72 % as recall value and Row vector model with name entities attained the recall rate as 84.5% also the proposed strategy achieved 97 % as recall rate that is detailed in table.4 and Figure.4

Table 4: Recall Comparison with existing approaches

Recall				
ws taken		-SVD		d
50	84.1	78.2	84.7	99
100	82	74.8	83.2	98.7
150	79.7	73	80.9	98.1
200	78	70	80	97

5.2.4 F-MEASURE

The F measure is validated to verify the mean average for precision and recall, thus the comparison of F measure.

$$F-measure = 2 \times \frac{\Pr ecision \times recall}{precision + recall}$$

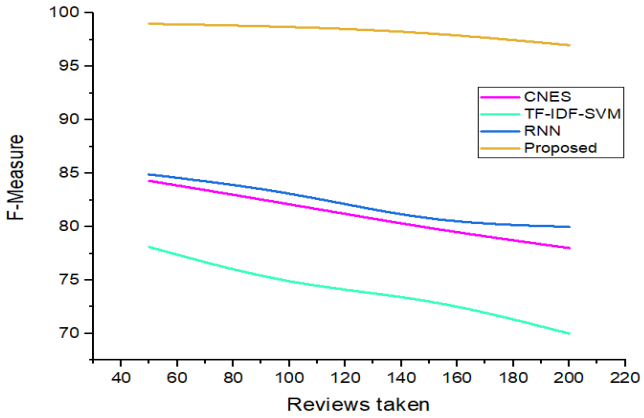

Figure 5 :F-Measure Comparison with recent existing techniques

A.

The average of F-measure is calculated by taking between the average of accuracy and precision. To verify the accuracy of classification F-measure is evaluated. High accuracy and precision yields better F-measure rate. The proposed strategy attained 97% as F-measure rate for 200 reviews simultaneously, the existing approach TF-IDF-SVM gained 70% , CNES achieved the F-measure rate as 81.5% and RNN attained 80% of F-measure rate, Which is defined in table.5 and Figure.5.

Table 5 : Comparison of F-measure

sure				
vs taken		F-SVD		ed
50	84.1	78.2	84.7	99
100	82	74.8	83.2	98.7
150	79.7	73	80.9	98.1
200	78	70	80	97

5.3 Discussion

From the above result validation, the efficiency of the proposed work is verified with the recent existing model. In all metrics the developed strategy earn a better result comparing other existing works. Moreover, the developed model is effectively implemented and processed in python. For each and every set of reviews the proposed mechanism attained high sentiment classification accuracy, which is proved by the above plotted graphs.

6. CONCLUSION

In big data area, machine learning strategy is one of the trending field thus the opinion or sentiment classification is one of the important tasks in NLP, which is mostly helpful for online services. So, the present work developed a novel hybrid machine learning model to validate the customer review in Telugu dataset. Moreover, the fitness model of optimization helps to improve the sentiment classification rate. Thus, the attained sentiment classification accuracy using machine learning and heuristic model is 97%. Moreover, the comparison results proved the efficiency of the proposed work. Thus the developed model is applicable for online services to classify the opinions of every customers, also it helps to improve the online services.

REFERENCE S

1 Carvalho, Arthur, et al. **Off-The-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing.** Communications of the Association for Information Systems 44.1 (2019): 43.

2 Marie-Sainte, Souad Larabi, et al. **Arabic natural language processing and machine learning-based systems.** IEEE Access 7 (2018): 7011-7020.

3. Yang, Haiqin, et al. **Deep Learning and Its Applications to Natural Language Processing**. Deep Learning: Fundamentals, Theory and Applications. Springer, Cham, 2019. 89-109.

4. Nasukawa, Tetsuya, and Jeonghee Yi. **Sentiment analysis: Capturing favorability using natural language processing.** Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003. https://doi.org/10.1145/945645.945658

5. Alayba, Abdulaziz M., et al. **Improving sentiment analysis in Arabic using word representation**. 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR). IEEE, 2018.

6. Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. **Random forest and support vector machine based hybrid approach to sentiment analysis.** Procedia Computer Science 127 (2018): 511-520.

7. Ma, Yukun, et al. **Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis**. Cognitive Computation 10.4 (2018): 639-650.

8. Ahuja, Ravinder, et al. **The Impact of Features Extraction on the Sentiment Analysis.** Procedia Computer Science 152 (2019): 341-348.

9. Hasan, Mahmudul, Ishrak Islam, and KM Azharul Hasan. **Sentiment Analysis Using Out of Core Learning**. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019.

10. Young, Tom, et al. **Recent trends in deep learning based natural language processing**. ieee Computational intelligenCe magazine 13.3 (2018): 55-75.
11. Tang, Duyu, and Meishan Zhang. **Deep learning in sentiment analysis.** Deep Learning in Natural Language Processing. Springer, Singapore, 2018. 219-253.
12. Zhang, Yuebing, et al. **A cost-sensitive three-way combination technique for ensemble learning in sentiment classification.** International Journal of Approximate Reasoning 105 (2019): 85-97.
https://doi.org/10.1016/j.ijar.2018.10.019
13. Yang, Chao, et al. **Aspect-based sentiment analysis with alternating coattention networks**. Information Processing & Management 56.3 (2019): 463-478.
14. Chiranjeevi, P., D. Teja Santosh, and B. Vishnuvardhan. **Survey on Sentiment Analysis Methods for Reputation Evaluation.** Cognitive Informatics and Soft Computing. Springer, Singapore, 2019. 53-66.
15. Parmar, Manojkumar, et al. **Sentiment Analysis on Interview Transcripts: An application of NLP for Quantitative Analysis.** 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.
16. Pham, Duc-Hong, and Anh-Cuong Le. **Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis.** International Journal of Approximate Reasoning 103 (2018): 1-10.
17. Khurana, Himja, and Sanjib Kumar Sahu. **Bat inspired sentiment analysis of Twitter data.** Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2018. 639-650.
18. Chandra, Nidhi, Sunil Kumar Khatri, and Subhranil Som. **Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository.** System Performance and Management Analytics. Springer, Singapore, 2019. 65-76.
19. Mandal, Shrabanti, Girish Kumar Singh, and Anita Pal. **PSO-Based Text Summarization Approach Using Sentiment Analysis.** Computing, Communication and Signal Processing. Springer, Singapore, 2019. 845-854.
20. Reddy, D. Aravinda, M. Anand Kumar, and K. P. Soman. **Paraphrase Identification in Telugu Using Machine Learning.** Advances in Big Data and Cloud Computing. Springer, Singapore, 2019. 499-508.
21. Gorla, SaiKiranmai, et al. **TelNEClus: Telugu Named Entity Clustering Using Semantic Similarity.** Computational Intelligence: Theories, Applications and Future Directions-Volume II. Springer, Singapore, 2019. 39-52.
https://doi.org/10.1007/978-981-13-1135-2_4