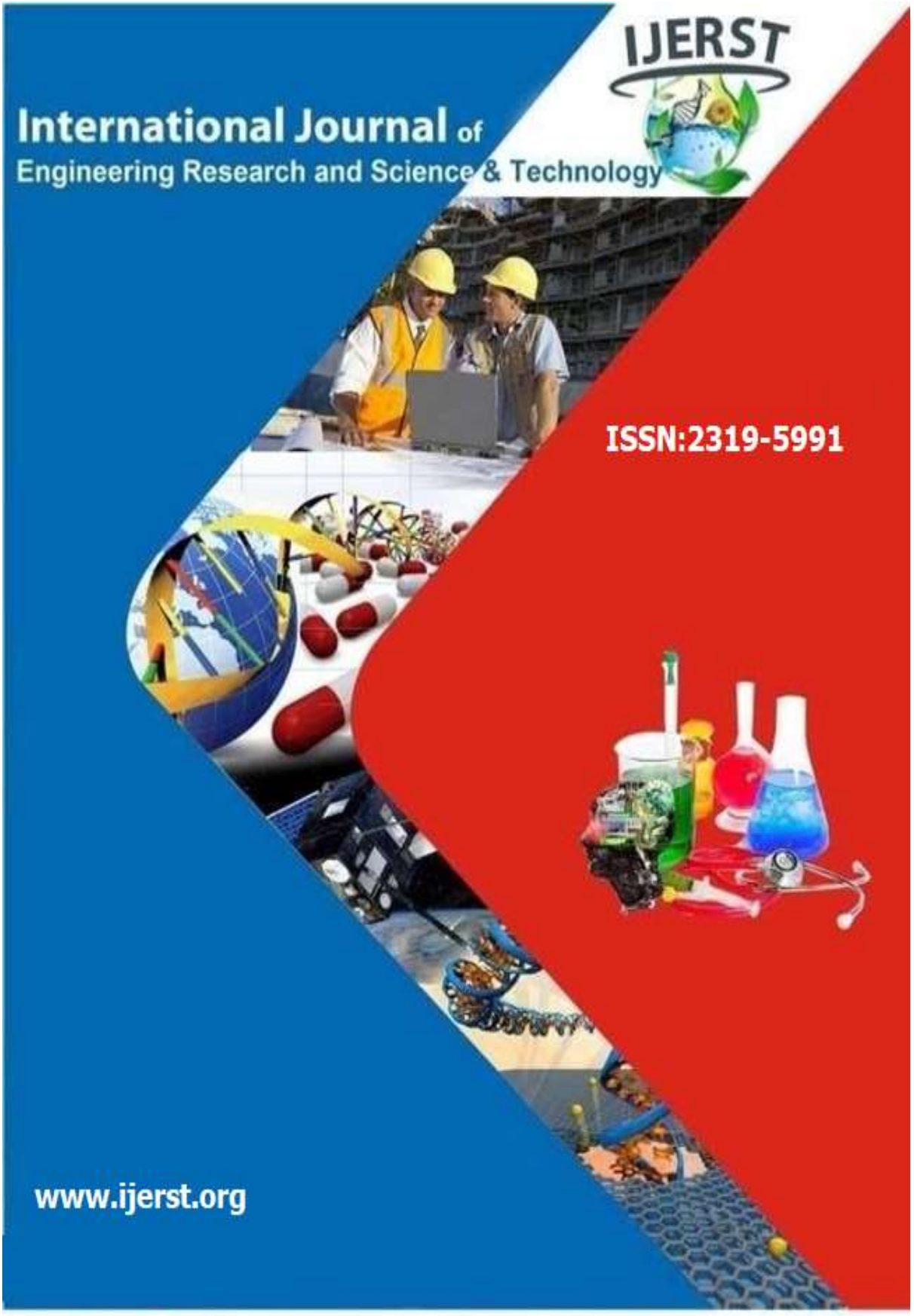

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

"A Hybrid Methodology for Telugu Sentiment Analysis"

V. Premalatha [1] **Dr. P. Dileep Kumar Reddy [2]**
Bharatiya Engineering Science and Technology Innovation University, Anantapur, Andhra Pradesh, 515731, India.
Narsimha Reddy Engineering College, Secunderabad, Telangana, 500100

Abstract- Sentiment Analysis is a subset of Natural Language Processing which is employed in wide range of business verticals to disentangle, analyze, distinguish and comprehend the general opinion of user reviews. This paper illustrates methodical approach which leverages lexicon-based approach and machine learning in the field of sentiment analysis to classify the opinions in Telugu language. Firstly, by employing Lexicon based approach - Telugu SentiWordNet identified the subjective sentences from the Telugu corpus. Secondly, by utilizing machine learning algorithms – SVM, Naïve Bayes and Random Forest we categorized the sentiment in the corpus. Our proposed methodology achieved highest accuracy of 85%.

***Index Terms* — Sentiment Analysis, Naïve Bayes, Random Forrest, Support Vector Machines, Machine learning, Lexicon based learning**.

INTRODUCTION

In 21st century, which is flaunted as the internet era, where the medium of eliciting opinions, expressing views has changed. Many people adopt digital media platforms like online feedback forums, blog posts, online retail product reviews, social media – Twitter, Facebook, Reddit, and so on to address their opinions and viewpoints on a specific entity. For some substantial degree, every choice we make is influenced by how our counterparts think and assess the world. Opinions from wide range of people can influence the decision making of an organization. From a single individual person to a big organization, everyone gets swayed by some extent on the choices they make from these opinions. As data generating in an exponential rate, it is becoming a widespread trend to capture these data silos from disparate data sources across web to disentangle and extrapolate key data sets to a Data-Driven decision making. Moreover, online platforms like Twitter and Facebook provides a window of opportunity for business leaders and researchers to access colossal amount of user reviews to comprehend the view of people. To understand the information resided in these huge data silos manually is an impossible task. The inception of machine learning and deep learning in the field of Natural Language Processing made this laborious and burdensome task of processing sentiment look easy and expediently. Since 2003, Opinion analysis or subjective analysis has become one of the most imperative field in Natural Language Processing. Sentiment analysis is the subset within the purview of NLP and computational linguistics which comprises of pre-processing, classification and apprehension of the mood and opinion of the customers expressed across the internet.

Many industries and companies proliferated since last decade focusing on NLP tasks like opinion mining, sentiment analysis, topic modelling, association analysis, review mining, and so on. Research in Natural Language Processing has been going on since 20th century but not much of research went through about knowing opinions until early 2000. Sentiment analysis has a wide range of applications like recommending products pertaining to the keenness of a customer towards an entity, estimation of customer satisfaction and retention rate, predicting result in political election depending on the user reviews on a political party generated from twitter, Facebook or any other social media. Sentiment analysis can be achieved in three different categories namely, Document level, Aspect or Feature level and Sentence level. Sentence level approach is to classify the sentence in the document or paragraph into

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

positive or negative polarity. Each sentence is considered as independent entity in this approach. Aspect or Feature level approach is a kind of approach where topics are extracted or filtered from the corpus using lexicon based approach and classify the sentiment for each of these aspects. Document level approach is to categorize the whole document into positive or negative polarity. Each document is regarded as a single entity in this approach.

Most of the research in this field mostly focused on English language. Models in deep learning algorithms like Recurrent Neural Network, Long Short-term Memory model achieved highest
accuracy while dealing with sentiment analysis in NLP. Whereas, very scant research went through in dealing with Indian regional languages like Telugu, Malayalam, Kannada, and so on. Indian languages are morphologically proficient and agglutinative in nature that makes task of generating efficient language specific tool problematic and onerous. In this paper, we are focusing on one of the regional language Telugu, which is fourth most spoken language after Hindi to analyze the sentiment in the Telugu corpus [1] and subsequently build a sentiment classifier. Even though English is an internally recognized language, less than 10% in India speak English. Greater than 90% speak languages such as Hindi, Gujarati, Tamil, Telugu, and so on. Telugu is a Dravidian language predominantly spoken in Andhra Pradesh and Telangana state [2]. 6.93% of the population in India speak Telugu [3]. The main challenges behind employing NLP in these regional languages are language ambiguity, lack of grammar resources, scarcity of documented standards. Even though millions speak these languages, there is a dearth of substantial resources about the literature and grammar. Developing Natural Language Processing models and algorithms without rudimentary lexicon resources is extremely taxing and laborious. Many NLP algorithms require considerable amount of dataset to train the model and extract the correlations, word patterns, etc. but data available for Indian languages are very scant compared to English language. It is highly intricate task to address these kinds of obscurities during development of NLP tasks.

In this paper, we gather Telugu reviews corpus from various sources available on the internet. Reviews extracted from these sources are teemed with objective and subjective sentences and few reviews have attached labels whereas, few comments are devoid of labels. To counter this problem, we leverage the use of lexicon based approach to assign labels to the comments using a lexical resource called SentiWordNet and also classify the review into objective sentence or subjective sentence. SentiWordNet is a sentiment lexicon acquired from the WordNet database where each element or word is accompanied with numerical scores indicating positive or negative polarity. SentiWordNet is open source and publicly available lexicon resource to tackle NLP tasks. After getting subjective sentences from the initial corpus, we employ machine learning algorithms namely, Naïve Bayes, Support Vector Machines and Random Forest to build the sentiment classifier which classifies the above mentioned reviews into positive or negative class. Then we test these results onto test dataset and evaluate these with performance metrics.

Section II explicates related word and associated background research dealing NLP tasks in Indian languages. Section III describes about the dataset. Section IV proposes a hybrid approach using lexicon based and machine learning for sentiment analysis and shows the experiment set up and describes machine learning models. Section V draws the results and evaluation metrics. Section VI present the conclusion and future work.

EXPERIMENTAL STUDY

The dataset is extracted from various online resources. The corpus consists of 3 different categories of reviews from wide range of sources. First category is extracted from amazon reviews dataset in English and translated into Telugu by using Google translation API. In second category, we leverage the usage of corpus ‘Sentiraama’ created by G. Rama Rohit Reddy at Language Technologies Research Centre, KCIS, IIIT Hyderabad [17]. This corpus has movie

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

reviews dataset annotated using a 2-scale value, distinguishing between positive and negative sentiment at document level. By adhering to the annotation procedure, each of the movie reviews were annotated by annotators meticulously. Third dataset is scraped from online Telugu blogs and Facebook comments.

	Positive	Neutral	Negative	Total
Amazon Reviews	936	52	951	1939
Sentiraama Movie Reviews	655	28	814	897
Telugu Blogs	91	30	96	219

Table. 1. Table shows summary statistics of dataset

Though there is a widespread predilection between the Telugu-speaking people and English language linguistically and at the same time culturally but there are far apart as any two languages can be. In English language, the systematic order is subject-verb-object. On a contrary, in Telugu language, the order of the sentence is subject-object-verb. For many English speakers the order of Telugu sentence seems inverted. Hence, With the help of proficient speakers of Telugu language from Andhra Pradesh, we analyzed each sentence and revamped into systematic Telugu sentence.

Review	Rating
ఇవి ఎక్కువసేపు ఉండవు.	0
ప్రచారం చేసిన విధంగా ఉత్పత్తి వేగంగా పంపిణీ.	1
మంచిది కాదు	0
మంచి ధర	1
3 నెలల్లోపు పనిచేయడం మానేసింది.	0

Review	Rating
These will not last long	0
Delivery was as fast as advertised	1
Not good	0
Good Price	1
Stopped working within 3 months	0

Fig. 1. Table shows few raw Telugu reviews collected from various sources and equivalent translation in English for reference

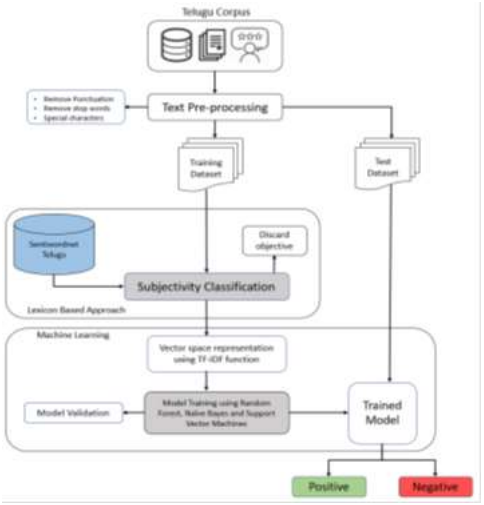

Fig. 2. Proposed architecture to classify sentiment in Telugu corpus using lexicon (Telugu SentiWordNet) based approach and machine learning

This section describes the approach in carrying out sentiment analysis on Telugu corpus. We collect a total of 3036 Telugu sentences from wide range of sources as shown in Table 1. As the dataset is in unstructured format with extra spaces, special characters, it is paramount to preprocess the dataset for standardization before feeding into our model. During preprocessing phase all the special characters like punctuations, emoticons and extra spaces are extricated. Data set is segregated into training data and testing data. We allocated 80% of data to training and 20% of the data to test the models. Next step is to filtered out Objective sentences from Subjective sentences from the corpus as they don't have any polarity. For this we make use of SentiWordNet developed by Amitava Das [18-20] et al. from IIIT Hyderabad. By comparing keywords to tokens from each sentence with SentiWordNet keywords, we extracted subjective sentences from the corpus and discarded objective or neutral sentences. Finally, we applied vector transformation method called TF-IDF (Term frequency – Inverse Document Frequency) onto subjective sentences and feed this vector matrix into our machine learning model. We train this dataset on various models namely, SVM, RF and NB and test this model on test dataset.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Consequently, we evaluated performance metrics of this model.

RESULTS AND DISCUSION

Below are the performance metrics while applying machine learning models namely, Support vector machine, Random Forest and Naïve Bayes on training dataset. The initial corpus of reviews is segregated into 80% training dataset and 20% into test dataset. Comparatively, Naïve Bayes algorithm achieved highest accuracy of 85.02% whereas, SVM scored 80.25% and Random Forest achieved 84.4%.

Classifier	Accuracy
Support Vector Machine	80.25
Random Forest	84.4
Naive Bayes	85.02

Table. 2. Table draws comparison of accuracy between various sentiment classifiers

COMPARISON OF PRECISION, RECALL AND F-1 SCORE FOR DIFFERENT MACHINE LEARNING METHODS ON TEST DATA SET

Classifier Model	Positive		
	Precision	Recall	F-1 Score
Support Vector Machine	0.67	0.83	0.74
Random Forest	0.89	0.80	0.84
Naive Bayes	0.85	0.85	0.86

Classifier Model	Negative		
	Precision	Recall	F-1 Score
Support Vector Machine	0.86	0.72	0.78
Random Forest	0.80	0.89	0.84
Naive Bayes	0.85	0.83	0.84

Table. 3. Table shows comparison between precision, recall and F-1 score for positive and negative polarity classifications among above-mentioned sentiment classifiers

CONCLUSION

Sentiment analysis is the subset within the scope of NLP and computational linguistics which comprises pre-processing, classification and apprehension of the mood and opinion of the customers expressed across the internet. This paper approaches the problem of understanding sentiment from huge corpus of reviews written in Telugu language by leveraging machine learning algorithms. Dataset is gathered from wide range of sources on the internet (Telugu blogs, Facebook comments, Sentiraama corpus and Amazon reviews. Three sentiment classifiers have been proposed along with cursory descriptions of each algorithm in section IV. We have achieved a highest accuracy of 85% for Naïve Bayes classifier.

REFERENCES

[1] P. Suryachandra and P. V. S. Reddy, "Statistical approaches in parsing for Telugu language," *2016 International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 2016, pp. 1-5.

[2] P. Suryachandra and P. V. S. Reddy, "Statistical approaches in parsing for Telugu language," *2016 International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 2016, pp. 1-5.

[3] R. Naidu, S. K. Bharti, K. S. Babu and R. K. Mohapatra, "Sentiment analysis using Telugu SentiWordNet," *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2017, pp. 666-670.

[4] Amitava Das in paper “Dr Sentiment knows everything” published in HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations Pages 50-55

[5] B. G. Patra, N. Mukherjee, A. Das, S. Mandal, D. Das and S. Bandyopadhyay, "Identifying Aspects and Analyzing Their Sentiments
from Reviews," 2014 13th Mexican International Conference on Artificial Intelligence, Tuxtla Gutierrez, 2014, pp. 9-15.

[6] M. Venugopalan and D. Gupta, "Exploring sentiment analysis on twitter data," 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, 2015, pp. 241-247.

[7] Pravarsha Jonnalagadda, Krishna Pratheek Hari, Sandeep Batha, Hashresh Boyina. “A rule based sentiment analysis in Telugu”, International Journal of Advance Research, Ideas and Innovations in Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[8] Naveen Bora, Hanisha balla "SentiPhraseNet: An extended sentiwordnet approach for Telugu sentiment analysis", International Journal of Advance Research, Ideas and Innovations in Technology
[9] Manukonda, Durga & Kodali, Rohith & Guduri, Dheeraj. (2019). Phrase-Based Heuristic Sentiment Analyzer for the Telugu Language. 6. 245-251.
[10] Mittal, N., Agarwal, B., Agarwal, S.G., Agarwal, S., & Gupta, P. (2015). A Hybrid Approach for Twitter Sentiment Analysis.
[11] Sharmista, Ramaswami "Rough set based opinion mining in Tamil" International Journal of Engineering Research and Development 2017
[12] Sahu, S.K., Behera, P.M., Mohapatra, D.P., & Balabantaray, R.C. (2016). Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative. CSI Transactions on ICT, 4, 111-115.
[13] Aditya Joshi and Balamurali A R and Pushpak Bhattacharyya "A fall-back strategy for sentiment analysis in hindi: a case study" in Proceedings of the 8th ICON 2010
[14] Rajesh Dongare, Lexicon Parser for syntactic and semantic analysis of Devnagari sentence using Hindi wordnet
[15] Meher Vijay Yeleti, Kalyan Deepak "Constraint based Hindi dependency parsing" in Language Technologies Research centre, IIIT Hyderbad, India
[16] B. M. Sagar, G. Shobha and P. Ramakanth Kumar, "Complete Kannada Optical Character Recognition with syntactical analysis of the script," 2008 International Conference on Computing, Communication and Networking, St. Thomas, VI, 2008, pp. 1-4.
[17] Gangula Rama Rohit Reddy, Radhika Mamidi "Multi domain corpus for sentiment analysis"
[18] A. Das and B. Gambäck. Sentimantics: The Conceptual Spaces for Lexical Sentiment Polarity Representation with Contextuality, In the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ACL 2012, Pages 38–46, Jeju, South Korea
[19] A. Das and S. Bandyopadhyay. Dr Sentiment Knows Everything! ACL/HLT 2011 Demo Session, Pages 50-55, June, Portland, Oregon, USA
[20] A. Das and S. Bandyopadhyay. SentiWordNet for Indian Languages, In the 8th Workshop on Asian Language Resources (ALR), COLING 2010, Pages 56-63, August, Beijing, China
[21] R. Bhargava, Y. Sharma and S. Sharma, "Sentiment analysis for mixed script Indic sentences," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 524-529.
[22] K. Shalini, A. Ravikurnar, R. C. Vineetha, R. D. Aravinda, K. M. Anand and K. P. Soman, "Sentiment Analysis of Indian Languages using Convolutional Neural Networks," 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2018, pp. 1-4.