iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59508

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Telugu Data Classification

Komal Parashar[1], V. Durga Bhavani[2], V. Keerthana[3], R. Narasimha[4]
[1]*Assistant Professor,* [2, 3, 4]*UG Students, Department of Computer Science & Engineering, CMR College Of Engineering & Technology, Hyderabad, Telangana*

Abstract: ***Telugu is considered as the difficult languages which is morphologically rich when it comes to Dravidian languages. There are many Telugu documents available on Internet, it is important to organize the data by automatically by assigning a collection of text with predefined categories. Here the Telugu data is classified into multiple areas like business, sports, entertainment, nation, editorial is the main goal throughout this research work. This research work provides up an efficient model by adopting some ML classifiers such as SVM, Naive Bayes, and Logistic regression to perform some areas of classification on Telugu data. The results obtained by various machine-learning models are compared and an efficient model is discovered, and it is observed that the Naive Bayes model outperformed with reference to accuracy, precision, recall, and F1-score.***

Keywords: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, ML classifiers

I. INTRODUCTION

These days, many people use online multimedia platforms such as blogs, online shopping review sites, feedback forums, and social networking sites. Share opinions and views in their native languages on Facebook, Twitter, WhatsApp, Instagram, LinkedIn about a specific topic. Here the text classification helps in the classification of the user opinions or views provided by the users into the domain it falls under and helps in organizing, identification of user preferences, personalized delivery of content, market analysis and facilitates the targeted users. Telugu, a really complex Dravidian language, poses challenges for organizing its abundant online content. To streamline this process, we aim to categorize Telugu documents into predefined topics like business, science, sports, etc. We'll employ modern techniques like Support Vector Machine (SVM); Naive Bayes; and Logistics Regression; backed by Natural Language Processing (NLP) in Machine Learning. By leveraging these methods, we seek to accurately classify Telugu text data and generate meaningful insights.

II. LITERATURE SURVEY

Deepu et al., n.d. [1] proposed a rule-based approach for opinion classification of Malayalam motion picture audits, tending to challenges stemming from client input containing spelling botches. Essentially, Sanjib et al.,[2] created a framework utilizing administered classification methods to classify Odia motion picture audit estimations as positive and negative. In differentiate, S.S. Mukku et al., [3] displayed a system for Telugu opinion investigation utilizing Doc2Vec models prepared with different ML procedures Moreover, J.Sultana et al., [4] compared conventional Profound Learning and ML approaches for opinion expectation on instructive information, finding MLP to abdicate the finest results. At long last, D Naga et al., [5] examined n-gram include selection's effect on news article content classification utilizing semi-supervised learning strategies, highlighting SVM's prevalence. The creators of the current think about propose to analyze Telugu news assumptions utilizing machine learning strategies to address the require for assumption investigation in Telugu news.

Kamal Sarkar et al., [6] and colleagues did some study about feeling analysis using a multinomial Naïve Bayes classifier enhanced with fancy determination characteristics. They focused on looking at feelings in tweets written in Bengali and Hindi, showing how their method can work in different languages. The Multinomial Naïve Bayes thing is great for handling jobs about text figures. The determination chat included deeper feeling analysis, underlining how smart their process is. The work by Sarkar and the gang gives us some good ideas for studying feelings in various languages. This study is super important in the international social media world because looking at feelings helps in knowing what people think across different languages.

Reddy Naidu et al., [7] presented a novel two-phase estimation research approach in their sentiment analysis investigation for Telugu e-News. This technique uses Telugu SentiWordNet, a lexical resource specific to the language, to categorize phrases found in Telugu e-News articles. The authors' use of a two-phase process points to a sophisticated and specialized method for handling the complexities of sentiment analysis in Telugu. Naidu et al. substantially advance sentiment analysis research by concentrating on regional language subtleties and offering a system that works with Telugu e-news information.

Their research contributes to our knowledge of sentiment dynamics in regional languages and may help tailor sentiment analysis techniques to a different type of linguistic circumstances. There is potential for improving sentiment analysis methods through more investigation of their methodology.

Samuel et al., [8] and his colleagues conducted a study that somehow connects to something about Coronavirus Tweets in the field of social media analytics. Their research on the Calculated Relapse and some other fancy Naïve Bayes techniques resulted in good precision, especially when they're dealing with the tiny Tweets. Because Twitter, is about short messages, so being brief is very important. This paper gives a guide for making tweet categorization methods better for the occurrences which happens like right now, you know, the epidemic.

III. EXISTING SYSTEM METHODOLOGY

In the existing implementation Telugu news is translated into English using the Google translation library available in Python. Then determined the sentiment scores using various tagging techniques and mark them as - positive/negative. Then, an attempt was made for classifying polarity value of Telugu news statements using several ML classifiers namely Naive Bayes, Random Forest, Passive-Aggressive Classifier, Perceptron, and SVM (Support Vector Machine). Here, the models are created for classification. One is a binary class and the other is a multiclass model. In binary classification, the system categorizes sentiment into positive polarity or negative polarity. Meanwhile, in the multiclass classification task, the system further categorizes the sentiment into business, editorial, entertainment, nation, and sports. Results were implemented using test data against performance parameters.

A. Implementation of Existing System

1) Telugu news is translated into English
2) Sentiment analysis is performed using different models
3) The classifiers SVM, Random Forests and Naïve Bayes used for the multiclass task
4) Training the Classifier Models
5) Testing the trained classifier models
6) Analysis of model performance

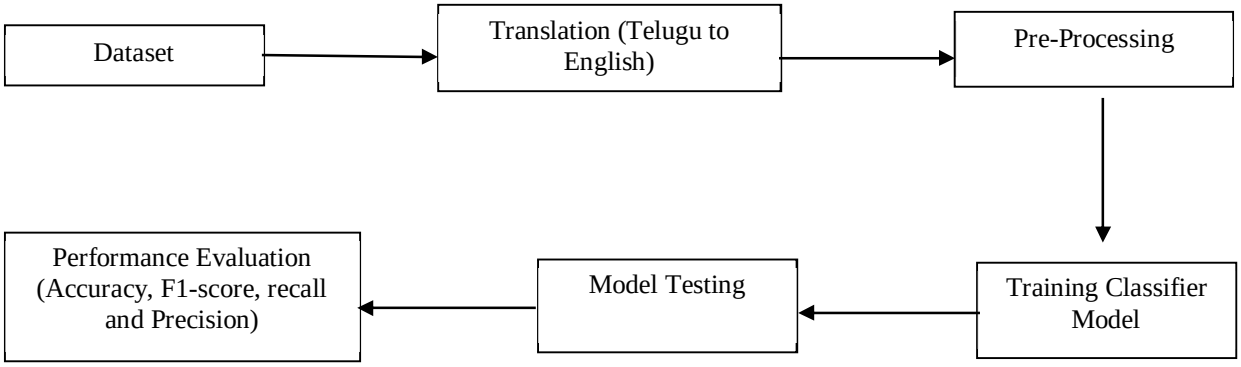


Fig. 1 Flow Diagram of Existing System

IV. PROPOSED METHODOLOGY

A. Dataset

Text classification is a crucial aspect and influences strongly on social practices. Most of the news articles generated in Telugu have not received much attention in the sentiment analysis community. These reasons led to choose the news dataset. The Telugu News dataset was collected by Kaggle (SRK, 2020). This Telugu news statements belonging to five areas and class labels used (Business, Editorial, National, Sports, Entertainment).

heading	body	topic
ఐడిబిఐపై ఆర్‌బిఐ నజర్	భారీ ఎత్తున మొండిబకాయిలు పెరిగిపోవడంతో ఐడిబిఐ ...	business
బ్యాంకింగ్ చీఫ్‌లతో నేడు జైట్లీ భేటీ	న్యూఢిల్లీ : ఆర్థిక మంత్రి అరుణ్ జైట్లీ సోమవా...	business
కీలక వికెట్ తీసిన జడేజా..	కటక్: ఇంగ్లండ్‌తో జరుగుతున్న సెకండ్ వన్డే మ్యా...	sports
మరో రెచ్చగొట్టే చర్యకు దిగిన పాకిస్థాన్	\nఇస్లామాబాద్ : పాకిస్థాన్ అంతర్జాతీయ ఉగ్రవాది...	nation
గోవాలో కొడుకుతో కలిసి అల్లు అర్జున్ స్విమ్మింగ్!	స్టార్ హీరోగా వరుస సినిమాలతో బిజీగా ఉన్నప్పటి...	entertainment

Fig. 2 Some samples of Telugu news dataset

B. Pre-Processing

Pre-Processing step is the first crucial step for removing the cleaning the data so that model can understand the effectively to yield a high-performance result. The pre-processing steps includes

1) Removes special characters
2) Sklearn: It is Label Encoder assigns numbered values to categories.
3) Tokenization: Breaking down the text into tokens and creating vocabulary.

C. Feature Extraction

Feature extraction: - Used in converting raw data into computer understandable that is into numerical format. The Feature Extraction includes

1) Vectorization: Converting the text into numerical values using TF-IDF (Term Frequency Inverse Document Frequency) and count vectorization.
2) Divides text into feature matrix using n-grams (uni-grams, bi-grams, tri-grams, 4-grams, 5-grams) which is the sequence of words.

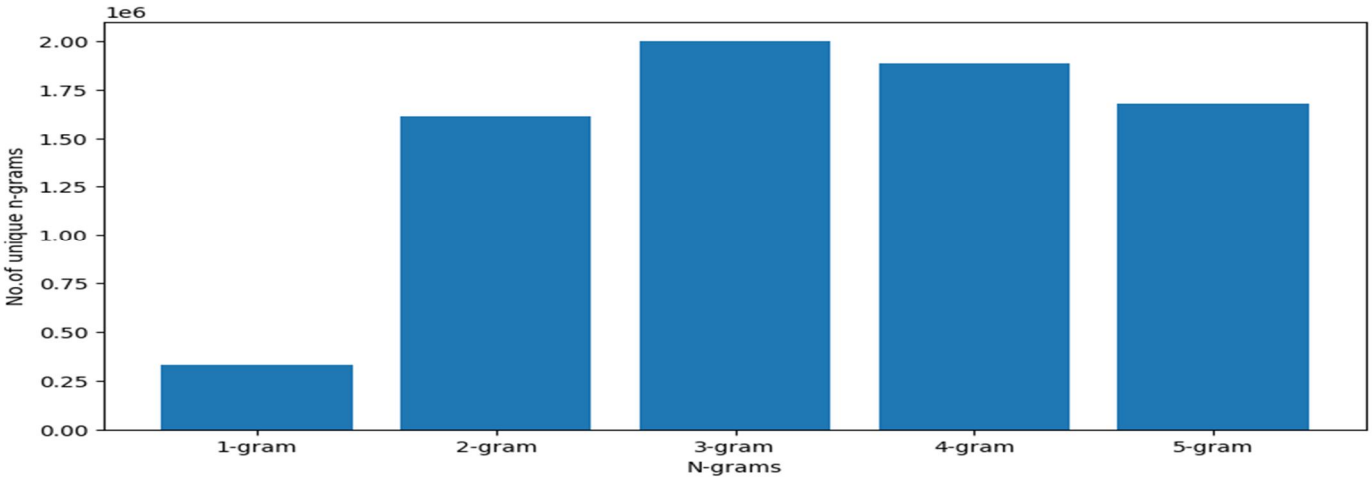


Fig. 3 Number unique N- grams

D. Classification Models

In our extend we used mainly three ML classification models those are Naïve Bayes, SVM, and Logistic Regression. A robust methodology which is simple and effective in classifying the Telugu text into different types of categories. Here, Naïve Bayes is very efficient classification technique for the small range of dataset and it calculates the probability for each category base on the input extracted features. Equally, SVM strikes to get the optimal hyperplane separating data points of various categories with the largest margin and efficiently captures the relationships in the high dimensional space. Logistic regression is also called as linear model which measures the probability of the binary output which can be further extended to classify multiple categories.

E. Implementation

The Flowchart starts with the importing of the packages such as indic-nlp library for the Indian Language understanding. Then, the dataset is extracted from the Kaggle for the Telugu data classification and processing data and giving a numerical value to each topic with the sklearn library- Label Encoder so that the model can understand and yield efficient results. After the Label encoding the feature extraction of N-gram using Count vectorization and TfidVectorization which is used for counting the frequency of occurrence of the n-gram (uni-gram, bigram, tri-gram, 4-gram, 5-gram). The vectors are generated in the feature extraction is kept for training the classification models Naïve bayes, SVM and Logistic Regression. After the models are trained the trained model is used for the prediction and performance of these models are being evaluated. The Evaluation of the trained is provided by the evaluation metrics Accuracy, F1-score, Precision and Recall.

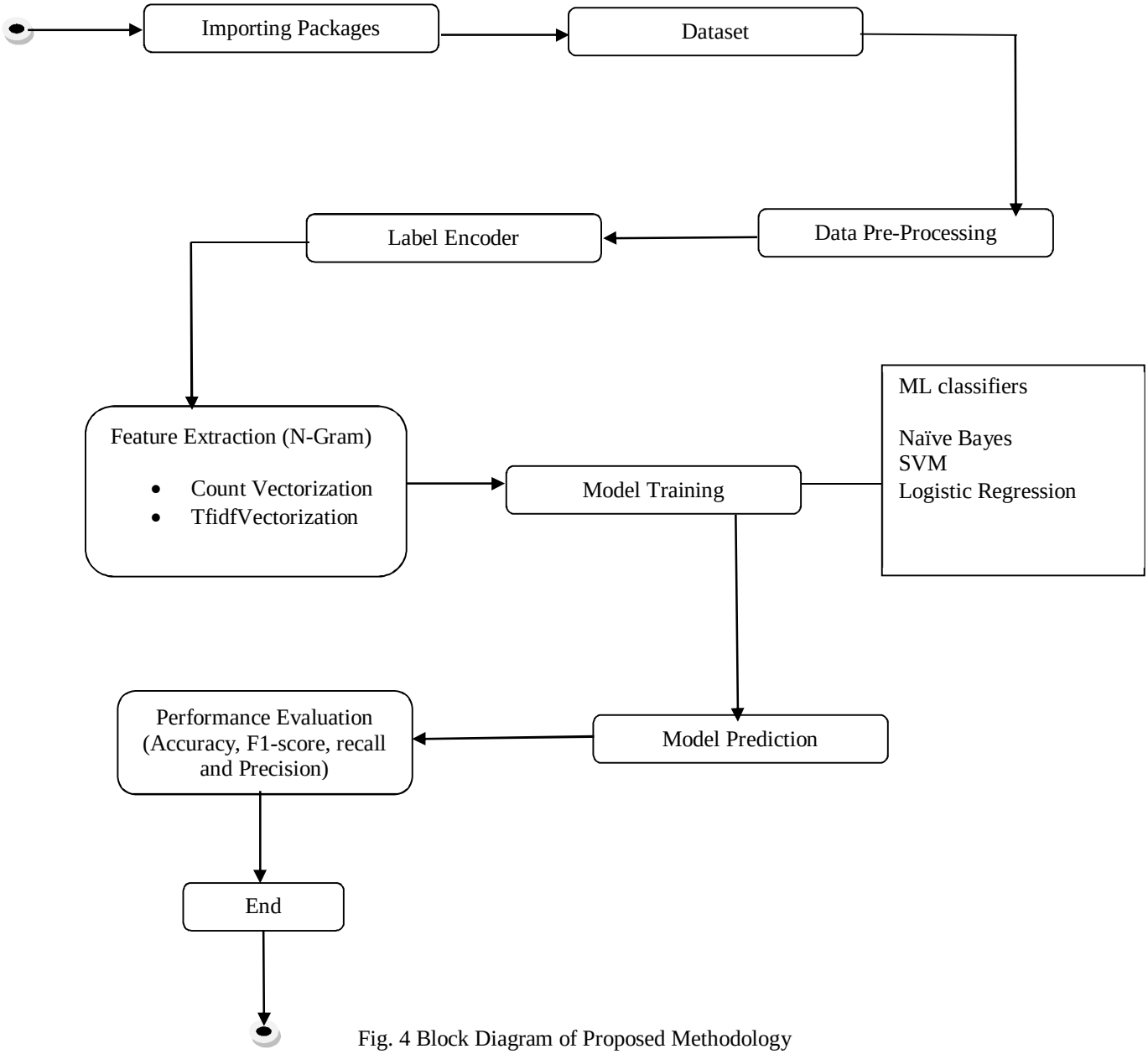


Fig. 4 Block Diagram of Proposed Methodology

V. PERFORMANCE EVALUATION

A. *Performance Evaluation of Naïve Bayes*

Table 1: Performance Evaluation metrics for Naïve Bayes

	Precision	Recall	F1-score	Support
Business	0.88	0.97	0.92	653
Sports	0.97	0.94	0.96	437
Nation	0.95	0.90	0.93	1673
Entertainment	0.96	0.98	0.97	1289
Editorial	0.80	0.83	0.82	277
accuracy			0.93	4329
macro avg	0.91	0.92	0.92	4329
weighted avg	0.94	0.93	0.93	4329

B. Performance Evaluation of SVM

Table 2: Performance Evaluation metrics for SVM

B	Precision	Recall	F1-score	Support
Business	0.89	0.83	0.86	534
Sports	0.97	0.85	0.91	380
Nation	0.85	0.93	0.89	1296
Entertainment	0.93	0.94	0.93	1058
Editorial	0.90	0.73	0.81	195
accuracy			0.90	3463
macro avg	0.91	0.86	0.88	3463
weighted avg	0.90	0.90	0.90	3463

C. Performance Evaluation of Logistic Regression

Table 3: Performance Evaluation metrics for Logistic Regression

	Precision	Recall	F1-score	Support
Business	0.91	0.82	0.86	534
Sports	0.97	0.82	0.89	380
Nation	0.84	0.93	0.88	1296
Entertainment	0.92	0.94	0.93	1058
Editorial	0.89	0.69	0.78	195
accuracy			0.89	3463
macro avg	0.91	0.84	0.87	3463
weighted avg	0.89	0.89	0.89	3463

VI. RESULTS AND DISCUSSION

The Table 4 interprets the data of comparison of accuracies between the Machine Learning Classification Models Naïve Bayes, SVM and Logistic Regression. We can observe that Naïve bayes is having high performance and outstanding accuracy than SVM and Logistic Regression.

Table 4: Accuracy of the classification models

	Accuracy
Naïve Bayes	93
SVM	90
Logistic Regression	89

Figures

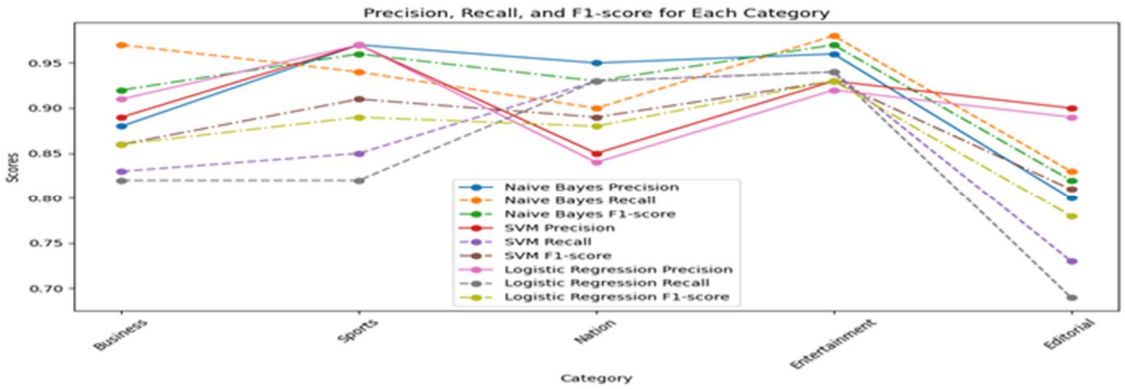


Fig. 5 graphical representation of precision, f1-score and recall of the Classification models on each of the category.

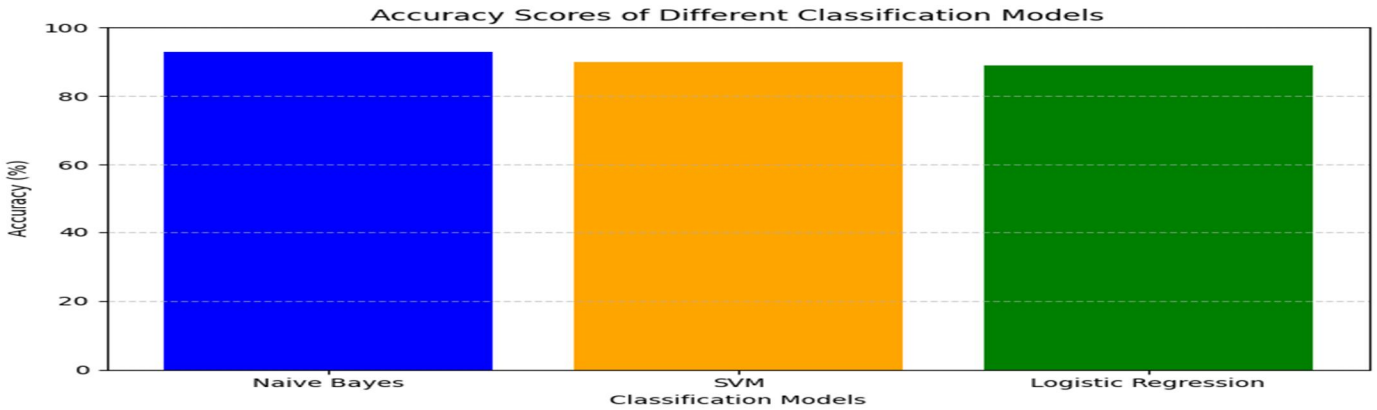


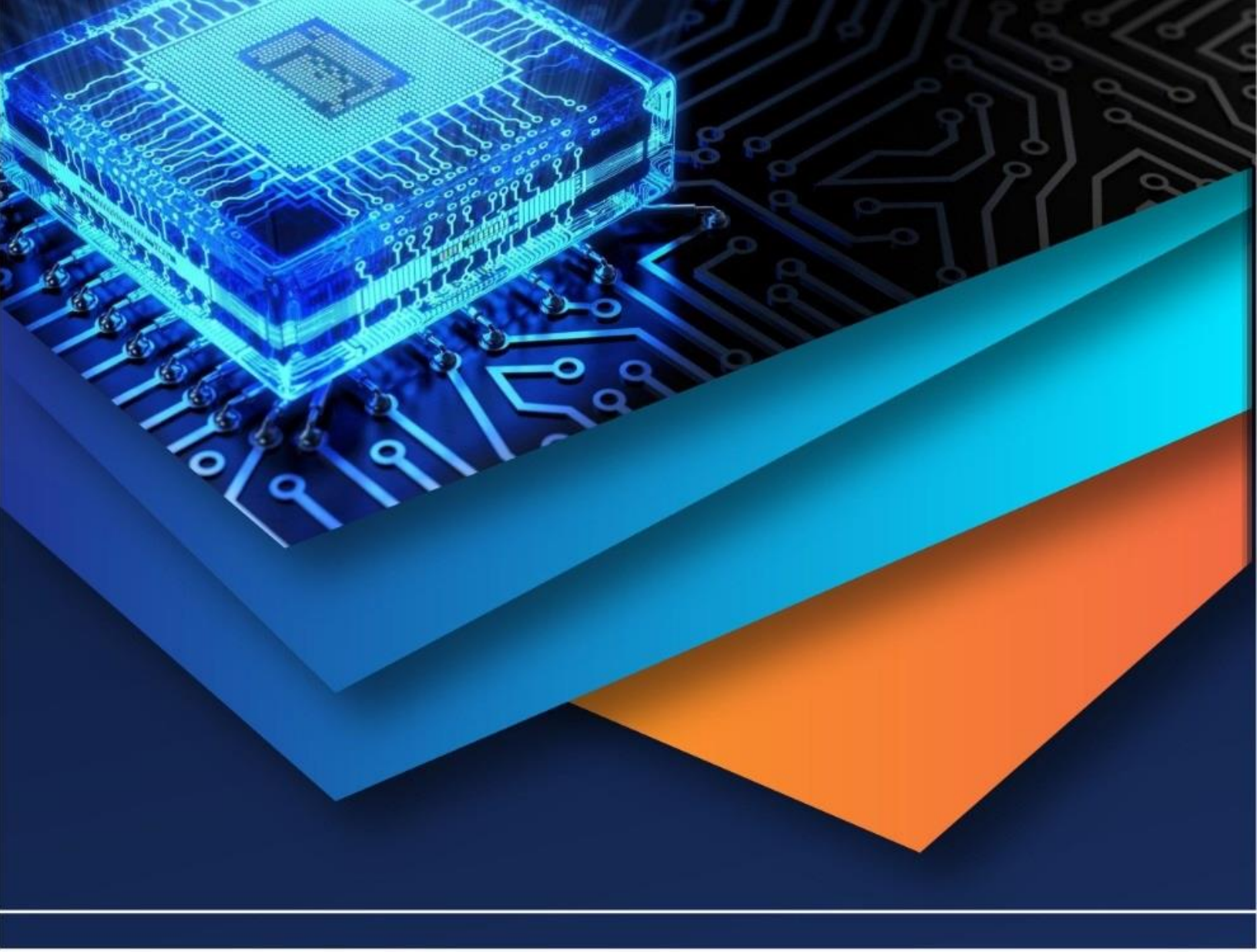
Fig. 6 Graphical representation of accuracy scores of the classification models SVM, Naïve Bayes, Logistic Regression

VII. CONCLUSION

Our research work helps in the classifying the Telugu text into the categories such as Business, sports, nation, Editorial and Entertainment. The Telugu news dataset is used for classification of categories using the ML classification models like Naïve Bayes, Logistic Regression and SVM. The Naïve Bayes performance outperforms Logistic Regression and SVM.

REFERENCES

[1] Nair, D. S., Jayan, J. P., Rajeev, R. R., & Sherly, E. (2014, September). SentiMa-sentiment extraction for Malayalam. In 2014 International conference on advances in computing, communications and informatics (ICACCI) (pp. 1719-1723). IEEE.
[2] Sahu, S. K., Behera, P., 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.
[3] Mukku, S. S., Choudhary, N., & Mamidi, R. (2016). Enhanced Sentiment Classification of Telugu Text using ML Techniques. SAAIP@ IJCAI, 2016, 29-34.
[4] Sultana, J., Sadaf, K., & Jilani, A. K. (2021). Classifying Student's Academic Performance using SVM. Journal of Engineering and Applied Sciences, 8(2), 61-61.
[5] Sudha, D. N. (2021). Semi Supervised Multi Text Classifications for Telugu Documents. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(12), 644-648.
[6] Sarkar, K. (2020). Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets. Sādhanā, 45(1), 196.
[7] Naidu, R., Bharti, S. K., Babu, K. S., & Mohapatra, R. K. (2017, March). Sentiment analysis using telugu sentiwordnet. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 666-670). IEEE.
[8] Samuel, J., Ali, G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. Information, 11(6), 314.
[9] Mukku, S. S. (2017). Sentiment Analysis for Telugu Language (Doctoral dissertation, PhD thesis, International Institute of Information Technology).
[10] Chattu, K., & Sumathi, D. (2023, July). Sentiment Classification of Low Resource Language Tweets Using Machine Learning Algorithms. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 1055-1061). IEEE.
[11] Sultana, J., Rani, M. U., Aslam, S. M., & AlMutairi, L. (2021). Predicting indian sentiments of COVID-19 using MLP and adaboost. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 706-714.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY