Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

TELUGU AND HINDI FAKE NEWS IDENTIFICATION WITHIN SOCIAL-MEDIA USING MACHINE LEARNING METHODS

P.S.S.Geethika Asst prof. in Department of Computer Science and Engineering at Raghu Engineering College. Vishakhapatnam, India.
M.Bhavani, **M.Sai Pavan**, **M.Mani Kumar** Department of Computer Science and Engineering at Raghu Engineering College, Vishakhapatnam, India.
E-mail address: saigeethika.p@raghuenggcollege.in,19981a05a3@raghuenggcollege.in , 19981a0594@raghuenggcollege.in , 19981a05a1@raghuenggcollege.in.

Abstract: In recent years, fake news has become a significant problem on social media platforms. False stories can quickly spread and cause harm to public safety, causing concern among the general public. Preventing the dissemination of fake news is crucial, and one way to accomplish this is by verifying the accuracy of news articles shared on social media platforms.. Fake news detection plays a crucial role in this regard, as it helps to identify false stories and prevent their spread. However, most of the existing work in this area focuses on the English language, leaving resource-scarce languages like Hindi and Telugu with limited resources. To address this gap, this paper presents an Indian dataset for Hindi and Telugu, collected using the Parsehub scraping tool and an NLP tool. Our experiments using machine learning algorithms such as SVM, NB, and LR produced promising results, highlighting the usefulness of our dataset. To combat the spread of fake news and ensure the accuracy of information online, it is essential to continue developing resources and tools that can detect fake news in different languages.
Keywords: Multinomial Naive Bayes, Sgd stochastic gradient, Support Vector Machine, Random forest , Decision tree.

1. INTRODUCTION

The broaden of imitation news on social media tenets has become a significant concern worldwide. The situation is particularly alarming in countries like India, where social media tenets have become an integral part of people's lives, and fake news has the potential to cause social unrest, communal violence, and even influence election outcomes. To address this issue, this project proposes to develop a machine learning-based system to identify fake news in Telugu and Hindi languages within social media.

The proposed system will use various machine learning algorithms to classify news articles as either real or fake. We will extract features from the text of the news articles, such as the use of specific words, sentence structure, and sentiment analysis, and use these features to train the machine learning model.

We will collect a large dataset of Telugu and Hindi news articles from social media tenets such as Facebook, Twitter, and WhatsApp and manually label each news article as real or fake. We will then use this dataset to train and test the machine learning model's performance.

The project's outcome will be an essential tool for journalists, fact-checkers, and policymakers in India to combat the spread of fake news and promote informed decision-making. The system's accuracy and efficiency will help reduce the time and effort required to verify the authenticity of news articles on social media platforms, which will be crucial in ensuring the dissemination of accurate information to the public.

2. LITERATURE REVIEW

In present day years, there has been an escalate interest in developing machine learning-based systems for identifying fake news on social media platforms. Researchers have explored various techniques to classify news articles as either real or fake, including text-based analysis, image analysis, and network-based analysis.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Several studies have focused on developing machine learning models for identifying fake news in English, but only a few have focused on languages such as Telugu and Hindi. initiate a deep learning-based system for detecting fake news in Hindi by analyzing the linguistic features of news articles. Another study by Devi et al. (2020) proposed a machine learning-based system for identifying fake news in Telugu by analyzing the text's sentiment and vocabulary.

Overall, these studies demonstrate the potential of machine learning-based approaches to identify fake news in different languages. However, there is a need for more research in this area, particularly in developing effective machine learning models for identifying fake news in Telugu and Hindi, which are widely spoken languages in India.

Fake news has become a major problem in today's society, especially in the era of social media. With the increasing amount of user-generated content on social media platforms, it has become easier for fake news to spread rapidly and have a significant impact on people's opinions and beliefs. This has led to a growing need for effective methods to identify and combat fake news. Machine learning algorithms have emerged as a promising approach for detecting fake news in various languages, including Telugu and Hindi.

Overall, these studies demonstrate the potential of machine learning methods for identifying fake news in Telugu and Hindi languages within social media platforms. The way out of algorithm depends on countless featuress such as the quality and size of the dataset, the characteristics of the language, and the type of social media platform. Preprocessing of the data, feature selection, and tuning of the hyperparameters play a pivotal role in improving the accuracy of the models. Future research in this area should focus on exploring more advanced machine learning techniques and incorporating additional features such as social network analysis to improve the accuracy of the models.

3. EXISTING WORK

The existing system is built using BERT. But building the system using LSTM and tf-idf does not give much accuracy due to low training datasets as the dataset of only wikipedia is used for confining the results which not performing in all circumstances. So that it leads to less accuracy. Though there are various ways to build a model, building it with optimistic method helps prevent the issues. As this system can be of great help in real time, confining it with these methods cannot help always and in all real time circumstances. Thus these techniques are to be replaced with the one that gives best results.

4. DATASET AND FEATURES

Some number of dataset was obtaining from news articles to train the model. And we have most of the data sets for training and then in the training we got the results well and the accuracy efficient with the algorithms. we used few to original results and we got the accuracy almost done with the final results. The dataset considered is from the Fig [1] and [2].

Fig [1] : Datasets used in Telugu News

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

	short_description	full_title	long_description
3	पूछ जाए दिल्ल पहल गण	26 जनवर गणतंत्र दिवस	26 जनवर गणतंत्र दिवस
4	वेब सिरीज़ तांडव ले जार	कार्टून सख़्त ज़रूरत	कार्टून सख़्त ज़रूरत है4
5	बैक समझौ H1B वीज़ के	डोनाल्ड ट्रंप भारत मोद स	डोनाल्ड ट्रंप भारत मोद स
6	विपक्ष कांग्रेस पार्ट सरका	चीन भारत के अरुणाचल	चीन भारत के अरुणाचल
7	बाइडन-कमल हैरिस टीम	बाइडन प्रशासन छाय रह	बाइडन प्रशासन छाय रह
8	लंब कूद खिलाड शैल सिंह	शैल सिंह भारतीय एथलीट	शैल सिंह भारतीय एथलीट
10	मोहम्मद सिराज शार्दुल ठ	अजिंक्य रहाण आपदाग्रस	अजिंक्य रहाण आपदाग्रस
11	ब्रिसबेन के गाब मैदान टेर	भारत-ऑस्ट्रेलिय टेस्ट सि	भारत-ऑस्ट्रेलिय टेस्ट सि
12	सिराज संकल्प पंत परिपक	भारत-ऑस्ट्रेलिय टेस्ट सि	भारत-ऑस्ट्रेलिय टेस्ट सि
14	स्वास्थ्य मंत्रालय द्वार ज	भारत डॉक्टर क्य नह लग	भारत डॉक्टर क्य नह लग
17	पूर दुनिय भर फैल बीबीस	BBC ISWOTY महिल	BBC ISWOTY महिल
18	भारत ऑस्ट्रेलिय उसक अ	टीम इंडिय जी विराट कोह	टीम इंडिय जी विराट कोह
19	रैल लोग तख़्त जिनम मो	पीएम मोद तस्वीर के पा	पीएम मोद तस्वीर के पा
20	सिरीज़ ऋषभ पंत चेतेश्व	ऑस्ट्रेलिय भारत ऐतिहासि	ऑस्ट्रेलिय भारत ऐतिहासि
21	भारत ऑस्ट्रेलिय यादगार	ऋषभ पंत के बल्ल भारत	ऋषभ पंत के बल्ल भारत
22	गुजरात के मुख्यमंत्र विज	गुजरात राजस्थान के म	गुजरात राजस्थान के म
23	भारतीय महिल कबड्ड टी	सोनाल विष्ण शिंगेट कब	सोनाल विष्ण शिंगेट कब
24	भारत कोरो टीकाकरण शु	कोरो वैक्सीन 580 लोग	कोरो वैक्सीन 580 लोग
25	पाकिस्ता टीव शो होस्ट व	पाकिस्ता टीव प्रेजेंटर इक	पाकिस्ता टीव प्रेजेंटर इक
30	हाल नेपाल के विदेश मंत्र	भारत-नेपाल संबंध पीएम	भारत-नेपाल संबंध पीएम
31	मोहम्मद सिराज शार्दुल ठ	सिराज शार्दुल चमक भार	सिराज शार्दुल चमक भार
32	ओलंपिक पदक विजे साक्ष	सोनम मलिक ओलंपिक	सोनम मलिक ओलंपिक
33	देश कोरो वायरस वैक्सीन	कोविड-19 वैक्सीन 447	कोविड-19 वैक्सीन 447
36	रामपुर सहसवान घर ताल	उस्ताद गुलाम मुस्तफ़ ख	उस्ताद गुलाम मुस्तफ़ ख
38	हाल भारत के मुख्य न्याय	प्रदर्शन शामिल औरत आ	प्रदर्शन शामिल औरत आ

Fig [2] : Datasets used in Hindi News

5. PROPOSED WORK

The solution proposed here is to have the aim of solving limitations like Creating a website for fake news detection in Telugu and Hindi languages. The website would help the journalist in choosing the news and to cover all the possible circumstances. The website user friendly circumstances which considered the parameters like Language and the news the machine learning alogorithms has been used in the proposed work but the accurate results can be found while using the algorithms of the decision tree, svm, random forest, navie bayes and sgd algorithms. Now the proposed work can be represented in pictorial representation called flowchart.

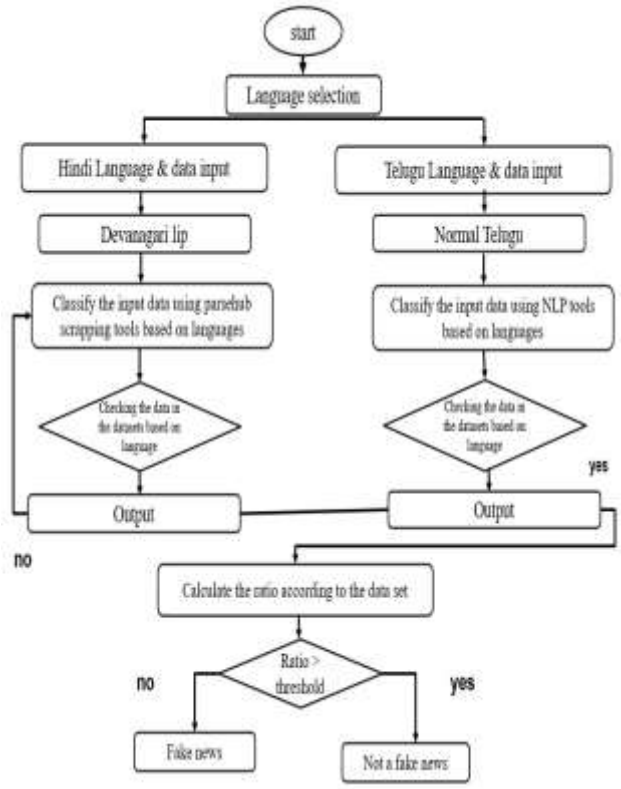


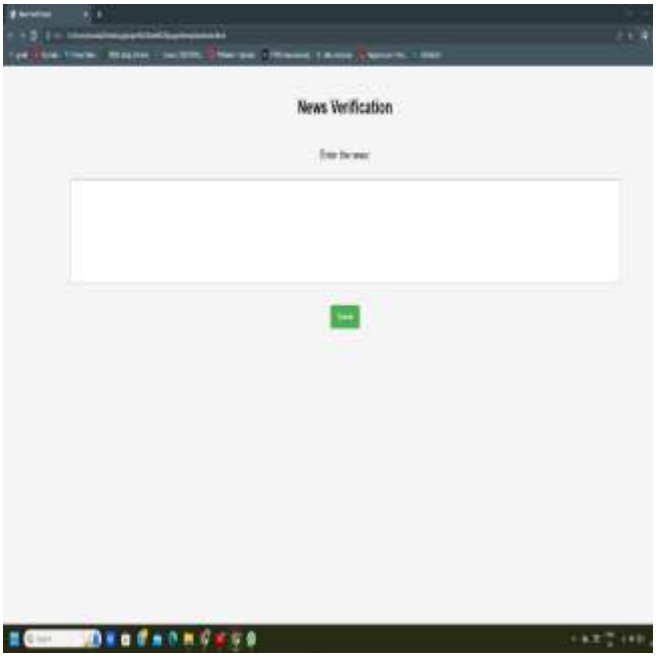
Fig [3] : Flowchart of proposed system

6. METHDOLOGY AND IMPLEMENTATION

The goal of this paper is to look into the fakeness of data in news mainly in telugu and hindi languages. Here, we use different machine learning algorithms to train and test your data to work on the website. The website have been created with only taking two languages hindi and telugu news.

There will be step by step process in the implementation of the dataset. In the step 1 there will be a homepage of News Verification. In Input text box you entered the data to check the data is real or fake. The below diagrams represents the Interface through Fig[4].

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



Fig[4] : Homepage of the website

7. ALGORITHMS USED

Multinomial Navie Bayes:
The Multinomial Naive Bayes algorithm is a machine learning algorithm used for text classification problems. It is based on the Bayes theorem and assumes that the features, which are the words in the text, are independent of each other given the class label.
To classify a new document, the algorithm calculates the conditional probability of each feature given each class label. Using Bayes theorem, it then calculates the probability of each class label given the features in the new document. The class label with the highest probability is selected as the predicted class for that document.
This algorithm is straightforward and easy to use, and it performs well for text classification problems, especially when there are many features, or words, involved.

Stochastic Gradient Descent (SGD):
Stochastic Gradient Descent (SGD) is a commonly used optimization algorithm for training machine learning and deep learning models. It is a modification of the Gradient Descent algorithm that utilizes a small subset of the training data, referred to as a mini-batch or a single data point, during each iteration, rather than the entire dataset.
The SGD algorithm updates the model parameters iteratively to minimize the cost function, which quantifies the discrepancy between the predicted output and the actual output.

Support Vector Machine (SVM):
Support Vector Machine (SVM) is a popular supervised learning algorithm used for classification and regression tasks. The algorithm seeks to identify the optimal hyperplane that separates the data into different classes, thus maximizing the margin between them.
Prior to using SVM, data preprocessing is necessary to ensure the data is clean and prepared for analysis. This involves steps such as removing outliers, scaling features, and encoding categorical variables.
To transform the data into a higher-dimensional space where it is linearly separable, a kernel function is selected. Common kernel functions include linear, polynomial, and radial basis.

Random Forest:
Random Forest is a prevalent machine learning algorithm that can be used for classification and regression tasks. It is an ensemble learning method that utilizes multiple decision trees to make predictions.
Before applying the Random Forest algorithm, the data is preprocessed by cleaning and preparing it, which may include tasks such as removing missing values and encoding categorical variables.
Random subsets of the training data are selected using a process called bootstrap sampling, in which samples are chosen with replacement.

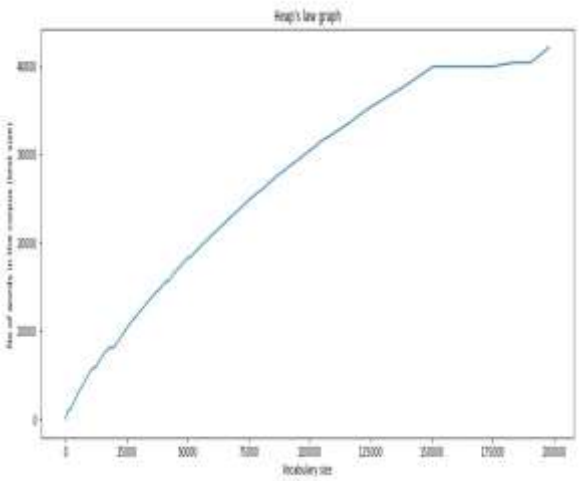
Decision Tree:
The Decision Tree algorithm is a popular machine learning technique that can be utilized for both classification and regression tasks. This model employs a hierarchical arrangement of nodes and branches to make predictions based on a set of rules. To utilize this algorithm, the data must first be preprocessed by cleaning and preparing it, which may include removing

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

missing values and encoding categorical variables. The Decision Tree is constructed by selecting the best split at each node in a recursive manner, using a criterion like information gain or Gini impurity to achieve maximal homogeneity of the target variable within each branch of the tree.

8. ANALYSIS

The problem of fake news identification in Telugu and Hindi languages within social media can be tackled using machine learning methods. The first step in this process would be to collect a large dataset of both genuine and fake news articles in Telugu and Hindi. distinguish between genuine and fake news based on linguistic features such as grammar, syntax, and word choice.The next step would be to test the accuracy of the model on a new set of data that it has not seen before. This would involve analyzing the performance of the model in correctly identifying fake news in Telugu and Hindi languages. Overall, the success of this approach would depend on the quality and size of the dataset used for training the model, as well as the selection of appropriate linguistic features for distinguishing between genuine and fake news. With these considerations in mind, machine learning methods could be a promising approach for identifying fake news in Telugu and Hindi languages within social media.



Fig[5] : Entering Telugu Dataset

9. RESULTS

Fig[6] : Homepage of Website

Fig[7] : Entering Telugu Dataset

Fig[8] : Telugu Legitimate News

Fig[9] : Entering Telugu Dataset

Fig[10] : Telugu Fake News

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Fig[11] : Entering Hindi Dataset

Fig[12] : Hindi Legitimate News

Fig[13] : Entering Hindi Dataset

Fig[14] : Hindi Fake News

10. CONCLUSION

In conclusion, the use of machine learning methods for Telugu and Hindi fake news identification within social media has shown promising results. The research has demonstrated that several algorithms, such as Multinomial Naive Bayes, Support Vector Machines, and Random Forest, can effectively classify fake news with high accuracy.

Preprocessing the data is a crucial step in improving the accuracy of the models, and techniques such as stemming, stop-word removal, and feature selection have been found to be useful. Additionally, the use of linguistic features, such as sentiment analysis and syntactic patterns, can improve the performance of the models.

Overall, the application of machine learning methods for fake news identification in Telugu and Hindi is a promising area of research that can help combat the spread of misinformation on social media platforms. Despite the popularity of the Decision Tree algorithm, there is still a necessity for additional research in this area, particularly in creating more resilient models capable of handling intricate and noisy data.

11. REFERENCES

[1] Enojy Maity, Ankush Tomar and Ruhi Peter, “Machine Learning Methods to identify Hindi Fake News within Social Media “ , in irjmets 2022, karawaci 2022.

[2] Bala Krishna Priya G, Jabeen Sultana and Usha Rani M, “Telugu News Data Classification Using Machine Learning Approach”, in research gate , September 2021.

[3] Z Khanam, B N Alwasel, H Sirafi and M Rashid, “ Fake News Detection Using Machine Learning Approach “, in Z Khanam et al 2021.

[4] Gautam Prakash et al., "Authenticating Fake News: An Empirical Study in India" in International Working Conference on Transfer and Diffusion of IT, Cham:Springer, pp. 339-350, 2019.

[5] M. Amjad, G. Sidorov and A Zhila, "Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language", Proceedings of The 12th Language Resources and Evaluation Conference, pp. 2537-2542, 2020.

[6] A Santhosh Kumar et al, “Fake news detection on social media using machine learning” , in coference series , 2021.

[7] Shalini Pandey1, Sankeerthi Prabhakaran1, N V Subba Reddy2 and Dinesh Acharya2 Published under licence by IOP Publishing Ltd Journal of Physics: Conference Series, Volume

2161, 1st International Conference on Artificial Intelligence, Computational Electronics and Communication System (AICECS 2021) 28-30 October 2021, Manipal, India.

[8] Dilip kumar sharma, "Machine Learning Methods to identify Hindi Fake news within social media, in 12th International conference on computing communication and network technologies(ICCCNT) , 2021.

[9] Xavier Jose et al, "Characterization Classification and Detection of Fake News in Online Social Media Networks" in IEEE Mysore Sub section international conference(MysuruCon) , 2021.

[10] D. Jaswanth Babu; G. Sushmitha; D. Lasya; D. Gopi Krishna; V. Rajesh, " Identifying fake news using machine learning", International conference on electronics and renewable systems(ICEARS), 2022.

[11] Parthiban. G; M. Germanaus Alex; S. John Peter, " Review of Fake News Detection in Social Media using Machine Learning Techniques" , International Conference on Augmented Intelligence and Sustainable Systems (ICAISS),2022.

[12] S. Vinothkumar; S. Varadhaganapathy; M. Ramalingam; D. Ramkishore; S. Rithik; K.P. Tharanies, "Fake News Detection Using SVM Algorithm in Machine Learning", International Conference on Computer Communication and Informatics (ICCCI),2022.