International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 05 | May 2024 www.irjet.net p-ISSN: 2395-0072

Fake News Detection Using Machine Learning and Web scraping

Ayisha Nidha P P[1], Hridya Nanda K[2], Ayisha Marva A[3] ,Najna Nazir M K[4]

[123]Student, Dept. of Information Technology, KMCT College of Engineering, Kerala, India
[4]Assisstant Professor, Dept. of Information Technology, KMCT College of Engineering, Kerala, India

---

Abstract - *Technology which is getting updated day-by-day has made the work load of human more easier and effort less. Also, social media networks which were created for the purpose of recreation and socialization over internet has become an inevitable part of our life. Now a days, rather than entertainment, more often social media is used for checking out news feeds. But the trustworthiness of source and content of news that is being viewed and shared over the internet is a greater matter of concern. Because, pseudo news and hoaxes are spreading in a menacing manner like a pandemic which is yet to be sorted out. Without checking the reliability of source and without knowing whether the news is reel or real people used to share news which seems to be satisfactory, without realizing the fact that these incognizant acts may abate the power of fake news and may end up in a severe issue especially during the time of disaster. We have a set of trusted sites such as Google. In these sites, if we search for a real content relevant information will be provided. If we search for a fake content, mismatching and irrelevant information will be retrieved as output. In this application, the news to be checked is searched online and results are gathered through web scrolling i.e., from different sites or from different links of Google. Then, the content obtained through search and the content of the concerned news is compared. comparing and finding out similarities by understanding contents is not an easy task. For that, we use techniques like natural language processing and associated set of algorithms. If similarities are found then that would be a real news and if not, it would be a fake news. Thus, one can easily figure out fake news and break the chain of fake news from being shared over the internet globally.*

Key Words: Fake News, false information, analysis, Machine Learning, Web Scrapping.

1.INTRODUCTION.

The term "fake news" describes news that is manufactured, misleading, or false and in which the veracity of the comments, sources, or facts cited are not known. Throughout human history, misinformation, rumors, and gossip have all been forms of fake news. This fake news is disseminated over social media in an effort to maximize its efficacy. Social media is used by billions of people, but it also contains robots, or simply bots. These bots aid in the quicker spread of false information and increase its visibility on social media. The quick dissemination of false information has grown to be a global problem. The dissemination of inaccurate and deceptive information has had a substantial negative social and economic impact on a variety of industries, including healthcare and banking.

The need to recognize fake news has never been more urgent. The credibility of the source and quality of news that is being viewed and shared online, however, is a greater reason for concern. Because misleading information and hoaxes are alarmingly proliferating like an uncontrollably expanding illness. In the past, people would spread news that seemed credible without checking to see if the source was reliable or if it was phony. They didn't realize, though, that these thoughtless acts could accelerate the spread of false information and result in major issues, especially during emergencies.

As the United States got ready for Hurricane Irma in 2017, a lot of misleading information circulated online. This included a Facebook post that, according to the alert, incorrectly predicted the storm would approach Houston and provided a map with a 14-day forecast that was nine days longer than official forecasts. "The post had been shared over 36,000 times on Facebook when the national weather service publicly refuted the forecast on Twitter with in the same day."

These instances actually give a clear-cut idea about the danger hidden within fake news and the impact created by the same. So, it is high time to “think before click” i.e., to fact check the news and its reliability before sharing. For the above-mentioned purpose, design an application to check whether a news is fake or real so that users can share reliable information which ensures protection from critical false news disasters. Our system identifies a news posted on the application as real or not. A set of trusted sites such as Google. In these sites, if we search for a real content relevant information will be provided. If search for a fake content, mismatching and irrelevant information will be retrieved as output. In this application, the news to be checked is searched online and results are gathered through web scrolling i.e., from different sites or from different links of Google. Then, the content obtained through search and the content of the concerned news is compared. Comparing and finding out similarities by understanding contents is not an easy task. For that, we use techniques like natural language processing and associated set of algorithms. If similarities are found then that would be a real news and if not, it would be a fake news.

Fake news detection is used to avoid rumors from spreading across various platforms, such as social media and messaging platforms. The impetus for this work is to avoid the spread of Fake-news which can even lead to worse activities. There has been a rise in the news lately about riots that result in mass deaths; fake news detection aims to detect these and stop similar activities, thereby protecting society from these unwelcome violent acts. The proposed system helps to find the authenticity of the news. The news given by the user is classified as true or false based on the data collected using Web Scraping. This task uses five various classification models, including Random Forest, Logistic Regression, Decision Tree, KNN, and Gradient Booster. To improve prediction accuracy, a mixture of these models is tested. Further, the paper is structured as follows: section 2 takes a glance at previous work done in fake news detection. In the next section, data extraction, pre-processing. Thus, one can easily figure out fake news and break the chain of fake news from being shared over the internet globally.

In order to improve the trustworthiness of the online platforms such as online social networks and reduce the disastrous impacts on the society, it is essential to develop a reliable mechanism for early detection and containment of misleading contents like fake news. Many endeavours research based on supervised learning methods were made for the issue of detecting fake news. Many of those works suffer from some limitations of low accuracy. The basis for low accuracy can be caused by many reasons such as the mediocre feature selection, incompetent parameter tuning, non-availability of benchmarked and balanced datasets, etc. Some of the researchers propose comparatively accurate fake news detection and classification models using deep learning techniques to reduce the probability of these issues. One of the main reason of using deep learning techniques is that in traditional machine learning techniques, human intervention is used explicitly to do feature engineering. However feature engineering is not needed in deep learning because important features are automatically detected by deep neural networks in deep learning. Among the deep learning techniques, the convolutional neural network (CNN) is more popular in recent research because it automatically detects the important features without any human supervision. It has been evidenced in the literature that in many cases CNN models outperform than other contemporary models. Fake news detection is used to avoid rumors from spreading across various platforms, such as social media and messaging platforms. The impetus for this work is to avoid the spread of Fake-news which can even lead to worse activities. There has been a rise in the news lately about lynchings and riots that result in mass deaths; fake news detection aims to detect these and stop similar activities, thereby protecting society from these unwelcome violent acts [3]. The proposed system helps to find the authenticity of the news. The news given by the user is classified as true or false based on the data collected using Web Scraping. This task uses five various classification models, including Random Forest, Logistic Regression, Decision Tree, KNN, and Gradient Booster. To improve prediction accuracy, a mixture of these models is tested. Further, the paper is structured as follows: section 2 takes a glance at previous work done in fake news detection. In the next section, data extraction, pre-processing and classifiers are discussed.

2. LITERATURE SURVEY.

2.1 A Novel stacking approach for accurate detection of fake news.

With the increasing popularity of social media, people has changed the way they access news. News online has become the major source of information for people. However, much information appearing on the Internet is dubious and even intended to mislead. Some fake news are so similar to the real ones that it is difficult for human to identify them. Therefore, automated fake news detection tools like machine learning and deep learning models have become an essential requirement. In this paper, we evaluated the performance of five machine learning models and three deep learning models on two fake and real news datasets of different size with hold out cross validation. We also used term frequency, term frequency-inverse document frequency and embedding techniques to obtain text representation for machine learning and deep learning models respectively. To evaluate models' performance, used accuracy, precision, recall and F1-score as the evaluation metrics and a corrected version of McNemar's test to determine if models' performance is significantly different. Then, proposed the novel stacking model which achieved testing accuracy of 99.94% and 96.05 % respectively on the ISOT dataset and KDnugget dataset. Furthermore, the performance of our proposed method is high as compared to baseline methods. Thus, highly recommend it for fake news detection.

2.2 Machine learning-based approach for fake news detection.

In the modern era where the internet is found everywhere and there is rapid adoption of social media which has led to the spread of information that was never seen within human history before. This is due to the usage of social media platforms where consumers are creating and sharing more information where most of them are misleading with no relevance with reality. Classifying the text article automatically as misinformation is a bit challenging task. This development addresses how automated classification of text articles can be done. We use a machine learning approach for the classification of news articles. Our study involves exploring different textual properties that may be often used to distinguish fake contents from real ones. By using those properties, can train the model with different machine learning algorithms and evaluate their

performances. The classifier with the best performance is used to build the classification model which predicts the reliability of the news articles present in the dataset.

2.3 Big Data ML-Based Fake News Detection Using Distributed Learning.

Users rely heavily on social media to consume and share news, facilitating the mass dissemination of genuine and fake stories. The proliferation of misinformation on various social media platforms has serious consequences for society. The inability to differentiate between the several forms of false news on Twitter is a major obstacle to effective detection of fake news. Researchers have made progress toward a solution by emphasizing methods for identifying fake news. The dataset FNC-1, which includes four categories for identifying false news, will be used in this study. The state-of-the-art methods for spotting fake news are evaluated and compared using big data technology (Spark) and machine learning. The methodology of this study employed a decentralized Spark cluster to create a stacked ensemble model. Following feature extraction using N-grams, Hashing TF-IDF, and count vectorizer, we used the proposed stacked ensemble classification model. The results show that the suggested model has a superior classification performance of 92.45% in the F1 score compared to the 83.10 % F1 score of the baseline approach. The proposed model achieved an additional 9.35% F1 score compared to the state-of-the-art techniques.

3. PROBLEM STATEMENT.

3.1 Existing System

Social media networks which were created for the purpose of recreation and socialization over internet has become an inevitable part of our life. Now a days, rather than entertainment, more often social media is used for checking out news feeds. But the trustworthiness of source and content of news that is being viewed and shared over the internet is a greater matter of concern. Because, pseudo news and hoaxes are spreading in a menacing manner like a pandemic which is yet to be sorted out.

Without checking the reliability of source and without knowing whether the news is reel or real people used to share news which seems to be satisfactory. These blindly deeds make the one who shares the fake news as culprit without any acknowledgement and this incognizant act may upsurge the power of fake news and may end up in a severe issue especially during the time of disaster.

For example, in India, Misinformation related to coronavirus COVID-19 pandemic is in the form of social media messages related to home remedies that have not been verified, fake advisories and conspiracy theories. Also, fake news regarding economic crisis, monsoon floods and election is on its high range.

Therefore, it is high time to “think before click” That is, to fact check the news and its reliability before sharing. But differentiating fake and real news is quite a challenging task.

3.2 Problems in Existing System

Since the start of human civilization, fake news has often emerged. Though, the propagation of fake news can emerge through the utilization of the global media landscape and modern technologies. Fake news affects several fields including economic, political, and social environments . On the other hand, fake news and fake information have several faces. Fake news poses tremendous impacts as information molds the view of humans around the world, though critical decisions can be made through fake information, which also leads to wrong decision-making. Similarly, good decisions cannot be made by this fabricated, distorted, false, or fake information on the Internet. The major impacts of fake news affect health, innocent people, democratic impacts, and financial impacts.

4. PROPOSED SYSTEM

4.1 Introduction

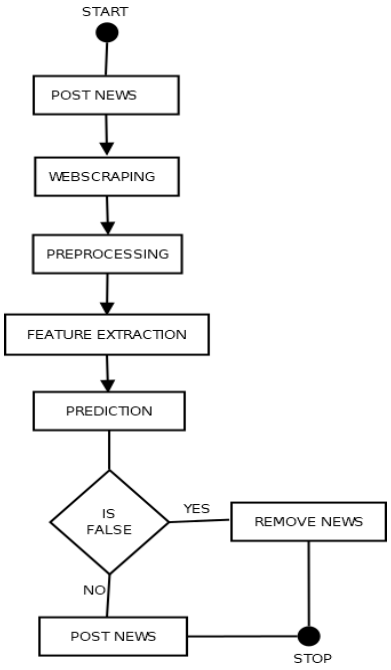
The motive of project is to identify a solution that may be used to detect and remove sites with fake news to help users to avoid being enticed by clickbait’s. It is our good fortune that such solution could be discovered and it would be much helpful for technical corporations and readers who are plagued by fake news. The proposed idea solves problems related to fake news which includes the tool usage that finds and filters out the fake news from the result provided by a search engine or a feed from social networking news. The aim is to use AI to identify articles and claim that are likely to contain false or highly biased information.

The challenge of fake news detection involves some of the following tasks: extracting and matching text-as it involves searching the several sources of data and then relating it to the input news, named entity recognition as it involves generation of token level fine-grained output, document, and entity-level sentiment analysis - as it involves the identification of the polarity, emotions negations, sarcasms, tone and bias of the sentences, document classification, stance classification, among others. The approach can be downloaded and attached to the browser or application used by the end user to receive news feeds. So that the application starts its work by using several methods like the methods related to the feature extraction so that we can verify the content we are seeing is valid one or false. This approach is implemented by integrating 2 modules: the web scraping and the LR module for detecting fake news.

4.2 Basic Working principle:

1. The news to be checked is searched online and results are gathered through web scrolling that is, from different sites or from different links of Google.
2. The content obtained through search and the content of the concerned news is compared.
3. Comparing and finding out similarities by understanding contents is not an easy work. For that, we use techniques like natural language processing and associated set of algorithms.
4. If similarities are found then that would be real news and if not, it would be a fake news.

Architecture diagram is given below:



- User can post news.
- Using web scraping process feature has been extracted through preprocessing.
- Predicting the posted news is fake or not.
- If it is fake it can't be posted.

Enabling chatbot as a feature helps in:

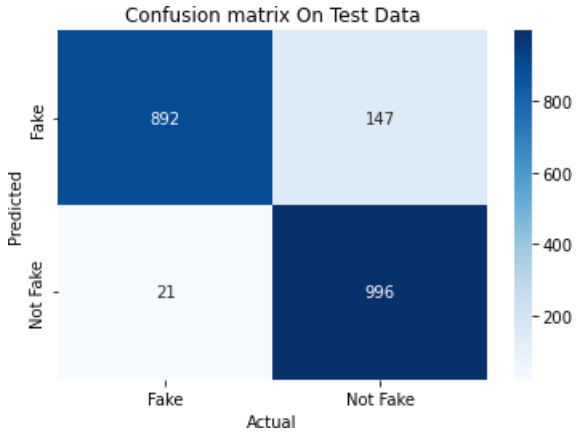
- Engaging large no. of people
- Limiting human interventions
- creating a more personalized , real-time and two-way experience
- Creating easier , friendly and instant to use communication platform.

4.3 Pre-processing

It involves the pre-processing steps needed to handle all the input texts and feeds. Initially, it reads the train, test and validation data and performs the processing such as tokenization and stemming. The exploratory data analysis such as response variable distribution and data quality checks like missing values or null values are then accomplished. NULL Value: start data pre-processing with checking the null values. We need to tackle with missing values during data analysis phase. So, handling the missing values is most important aspect before starting to work on data. However, before doing anything about missing values, we need to understand the pattern of missing values occurring. The techniques are useful in the early phases of the analysis of exploratory data.

4.4 Random Forest Algorithm

Among the ensemble learning techniques is Random Forest, a flexible and potent machine learning algorithm. To create forecasts, it synthesizes the results of several decision trees. Essentially, Random Forest makes use of decision trees, which produce predictions for each region by recursively dividing the input space into regions according to the values of input attributes. Feature randomness and Bootstrap Aggregating (Bagging) are two methods that provide randomness to the algorithm, guaranteeing variation within the decision trees and avoiding overfitting. Every decision tree undergoes independent training on a distinct bootstrapped dataset, and the final output is generated by averaging (for regression) or voting (for classification) the predictions made by the trees. Finding pertinent features in the dataset is made easier with the help of Random Forest, which offers a measure of feature relevance. It is widely utilized in many different sectors due to its scalability, robustness, and capacity to handle different forms of data.



Random forest classifier creates many trees by utilising features of subsets. It is extension of bagging and simply merges the output of multiple decision trees. It can be used for classification and regression issues as well as unsupervisedlearning. The accuracy score of 92%,92% F1-score , 98%precision and 86% recall of testing data set

4.5 Web scraping

It is one of the automated methods applied to obtain huge data from web URLs. Following Figure 4 shows web scraping. As we can see that, the data present on the web URLs is formless. It helps to get this formless data and store it in a proper form. Following some ways by which scraping to the websites can be done:

1. Writing your own code.
2. APIs.
3. Online services.

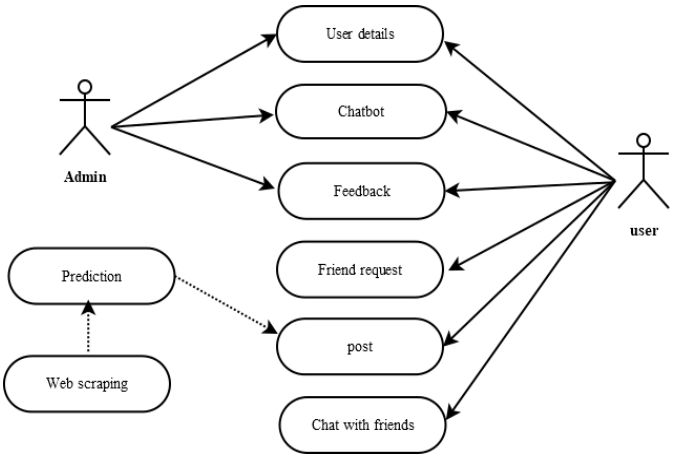
When open web scraper code is used and run to the URL that you have copied, a request is sent to the server. The sever then sends the data and grant you to read the HTML or XML page, in the response of the request made by us. The code then, finds the data and extracts it, parses the HTML or XML page. In this approach our own code is used for effective collection of data and its genuineness from several websites and social media platforms.

4.6 General Description

There is no common definition for describing fake news. Fake news can be spread over the world, which can be propagated in any field like COVID-19, politics, e-commerce, marketing, and so on. So, there is a need of analyzing fake news to understand the real news in any particular field. However, some of the writers, publishers, and vendors, posting non-authentic online comments or any third-party monitoring online comments will act as real customers and spreads fake news on online social media for increasing product sales. Similarly, a vast number of users on social media can broadcast fake news based on their opinions. In the case of the tourism field, tweets on social media may propagate fake news based on their imagination without spending at a destination (Das et al. 2021). It may lead to the loss of genuine consumers due to fake news on online platforms. In general, fake news is considered one of the huge threats to freedom of expression, journalism, and democracy. It also influences the political impacts, from which the fake news generation can be derived due to the comments, reactions, and shares posted on Facebook, WhatsApp, Instagram, and common websites (Brenes Peralta et al. 2021; Chang 2021). More specifically, fake news detection can be divided into four perspectives like source-based approaches, propagation-based approaches, style-based approaches, and knowledge-based approaches. Finally, recent advancements in “deep learning have been utilized for detecting” fake news from online social media platforms. Deep learning has several features over machine learning approaches, which are superior accuracy, the capability of extracting high-dimensional features, and “lightly dependent on data pre-processing”. Moreover, the recent broader “availability of data and programming schemes has increased the robustness and utilization of deep learning-based algorithms”. Thus, in the past years, various research articles on fake news detection models have been implemented based on deep learning techniques.

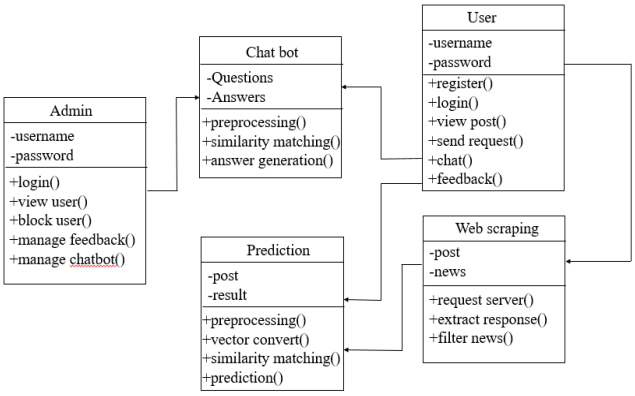
5. SYSTEM DESIGN AND DEVELOPMENT

5.1 Use Case Diagram



- Admin and User are the actors present in this use case Diagram.
- Admin can check User details, mange chatbot, analyze feedback.
- User can enter user details, use chatbot, give feedback, give friend request, post news, and also chat with friends.
- The news is only posted when it is not fake by predicted by web scraping method.

5.2 Class Diagram



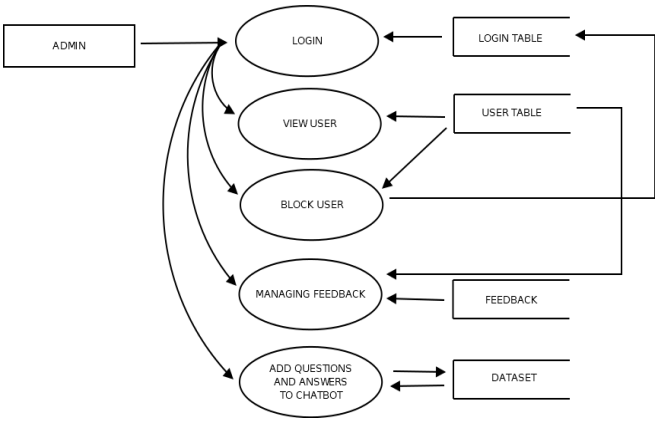
- There is Admin and user present in this diagram.
- There is also a prediction table for fake news detection
- The admin and user can perform their functions in their logins.
- Fake news is detected using Web scraping.

5.4 Data Flow Diagram

LEVEL 0

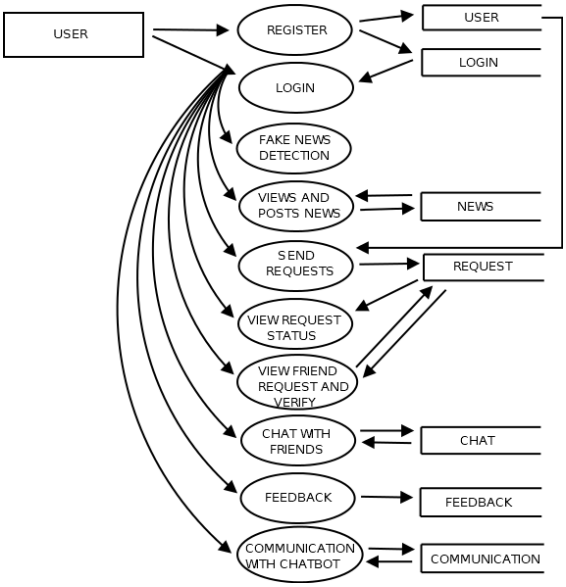
- It is the overall structure of the system
- There are 2 modules such as admin and User.
- These modules are connected to Fake news detection model.
- The model will predict whether the news is fake or not.

LEVEL 1.1



- Admin have login accesses to the platform.
- They can view user details.
- Can block user if there is any issues occurred.
- They can manage feedback from user.
- Also can question and answer to the chat bot.
- Each of the function will stored in respective tables in the database.

LEVEL 1.2



User can do following functions such as:

- Register and Login.
- View and post news.
- Send and accept friend request.
- They can chat with friends and chatbot.
- Can give feedbacks.

Above mentioned data are stored in database

6. RESULT AND DISCUSSION

Detecting fake news using web scraping involves collecting data from various online sources, analyzing content, and applying machine learning or natural language processing techniques to determine the credibility of the information. Here's how you can structure the result and conclusion section for a fake news detection project using web scraping.

Used web scraping techniques to gather information from a range of online sources, such as forums, social networking sites, and news websites. To assure the quality and consistency of the dataset, the acquired data underwent stringent preprocessing procedures, such as text cleaning, tokenization, and the removal of HTML tags and stop words. a variety of features were extracted from the text data, such as linguistic traits, word frequency, n-grams, and sentiment analysis scores.

Experiments revealed promising results, with the best-performing models achieving accuracies above 90% on the test dataset. Confusion matrices, ROC curves, and feature importance plots were utilized to visualize and interpret the

performance of the fake news detection models. These visualizations provided insights into the behavior of the models and highlighted key features contributing to the detection of fake news articles.

The experiments demonstrated varying performance among the different models, with some achieving higher accuracy and precision than others. Logistic regression exhibited simplicity and efficiency, while random forests showed robustness to overfitting. Support vector machines performed well in separating the classes but required careful parameter tuning. Ensemble methods such as stacking or boosting could potentially improve overall performance by combining the strengths of multiple models.Analysis of feature importance revealed that certain linguistic patterns, sentiment indicators, and metadata attributes played significant roles in discriminating between real and fake news.

Features related to sensationalism, inflammatory language, and unreliable sources often correlated with fake news articles, highlighting the importance of context and content analysis. Despite the promising results, our fake news detection approach faces several challenges and limitations. Biases in the training data, limited generalization across different languages and domains, and the dynamic nature of online misinformation pose significant obstacles. Future research efforts should address these challenges through more comprehensive data collection, robust model architectures, and cross-disciplinary collaborations.

Our study's conclusions have applications in the fields of digital citizenship, content control, and media literacy. Users that are able to analyze material critically and make well-informed judgments online can be enabled by effective false news detection systems. Scalable and sustainable solutions must be implemented in conjunction with fact-checking groups, social media companies, and legislators.

To sum up, our web scraping-based fake news detection study offers insightful information about the intricacies of online disinformation and the promise of computational methods to tackle this urgent social problem. Even if there are still obstacles to overcome, further study and development in this area could lead to the development of a more reliable and robust information ecosystem.

7.CONCLUSION

The study shows that machine learning and web scraping techniques can be used to detect bogus news. The models created for this project perform well when it comes to differentiating between authentic and fraudulent news articles on different websites. The research's conclusions have a big impact on content moderation, media literacy, and the battle against false information. We can enable consumers to make educated decisions and slow the spread of misleading information online by implementing efficient fake news detection technologies. Notwithstanding the encouraging outcomes, there are still several drawbacks to our fake news detection method, such as biases in the training set and difficulties identifying subtle types of false information. Subsequent investigations have to concentrate on resolving these constraints and investigating innovative methods to enhance the precision and dependability of systems designed to identify false news. Think of a number of directions that future study could go, such as including multimedia content analysis, utilizing user comments and social network data, and investigating group learning strategies. Misinformation identification and prevention can be further advanced through collaborations with academic researchers, social media platforms, and fact-checking organizations. When developing and implementing fake news detection systems, ethical factors like privacy, censorship, algorithmic bias, and transparency must be properly taken into account. Promote ethical AI practices and stress the value of maintaining moral principles in the fight against online disinformation.

REFERENCES

[1] A Novel Stacking Approach For Accurate Detection Of Fake News Tao Jiang , Jian Ping Li , Amin Ul Haq , Abdus Saboor , And Amjad Ali .

[2] Machine Learning-Based Approach for Fake News Detection H. L. Gururaj, H. Lakshmi, B. C. Soundarya , Francesco Flammini and V. Janhavi.

[3] Big Data Ml-Based Fake News Detection Using Distributed Learning Alaa Altheneyan And Aseel Alhadlaq.

[4] A Review Of Methodologies For Fake News Analysis Mehedi Tajrian , (Member, IEEE), Azizur Rahman, Muhammad Ashad Kabir , (Member, IEEE), And Md. Rafiqul Islam , (Senior Member, IEEE).

[5] OPCNN-FAKE: Optimized Convolutional Neural Network For Fake News Detection Hager Saleh, Abdullah Alharbi, And Saeed Hamood Alsamhi.

[6] Yimin Chen, Nadia K Conroy, and Victoria L Rubin. “News in an online world: The need for an “automatic crap detector””. In: Proceedings of the Association for Information Science and Technology 52.1 (2015), pp. 1–4.

[7] Mordechai Gordon and Andrea R English. John Dewey’s democracy and education in an era of globalization. 2016.

[8] Jorge J Palop, Lennart Mucke, and Erik D Roberson. “Quantifying biomarkers of cognitive dysfunction and neuronal network hyperexcitability in mouse models of Alzheimer’s disease: depletion of calcium-

dependent proteins and inhibitory hippocampal remodeling". In: Alzheimer's Disease and Frontotemporal Dementia. Springer, 2010, pp. 245–262.

[9] Praveen Kumar Donepudi et al. "Artificial Intelligence and Machine Learning in Treasury Management: A Systematic Literature Review". In: International Journal of Management (IJM) 11.11 (2020).