IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Fake News Using Machine Learning

[1]K.T. Krishna Kumar, [2]Balasree Sivakoti,
[1]Associate Professor and Placement Officer, [2]MCA Final Semester,
[1]Master of Computer Applications
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, India.

Abstract: In our modern era where the internet is ubiquitous, everyone relies on online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click baits. It is harmful for the society to believe on the rumors and pretend to be a news. The need of an hour is to stop the rumors especially in the developing countries like India, and focus on the correct, authenticated news articles. Fake news detection is used to avoid rumors from spreading across various platforms, such as social media and messaging platforms. so, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news. This model requires prerequisite data extracted from various news websites. Web scraping technique is used for data extraction which is further used to create datasets. The data is classified into two major categories which are true dataset and false dataset. Classifiers used for the classification of data is Logistic Regression. Based on the output received the data is classified either as true or false data. Based on that, the user can find out whether the given news is fake or not on the webserver.

Index Terms **-** Fake news detection, Logistic Regression classification, machine learning, web scraping, Feature Extraction, TF-IDF vectorization.

I. INTRODUCTION

In today's digital age, anyone can post content online, leading to the widespread dissemination of fake news, especially on social media. Such misinformation misleads individuals and spreads rapidly, causing harm to society by fostering negative sentiments and rumours. As technology advances, it is crucial to develop measures to combat these activities. Many websites deliberately spread false information to manipulate public opinion. Fake news impacts people's perceptions, and studies show that AI algorithms can help detect it. This project aims to prevent the spread of rumours and misinformation, particularly on social platforms like Facebook, Twitter, and WhatsApp, which can lead to dangerous actions like mob lynching. The goal is to develop a model to differentiate between real and fake news, thereby protecting society by verifying the authenticity of news articles and suggesting accurate information to users.

1.1 Existing System

Existing systems for detecting fake news using machine learning rely on a combination of traditional algorithms and advanced models to analyze text and multimedia content. Data collection involves scraping news articles and social media posts from various sources, followed by preprocessing steps such as cleaning, normalizing, and tokenizing the text. Feature extraction techniques like Bag of Words (BoW), TF-IDF, and word embeddings (Word2Vec, GloVe) are used to represent the text numerically. Machine learning algorithms such as logistic regression, Naive Bayes, and support vector machines (SVM) provide baseline models, while deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) offer advanced capabilities for handling sequential data. Transformers like BERT and GPT are increasingly used for their superior contextual understanding. Model evaluation employs metrics such as accuracy, precision, recall, and F1 score to ensure reliability. These systems also integrate explainability techniques to make predictions transparent and trustworthy. Despite their effectiveness, challenges such as data bias, evolving fake news tactics, and computational demands remain.

1.1.1 Challenges

- Data Quality and Availability
- Evolving Nature of Fake News
- Feature Extraction
- Ethical and Social Challenges
- Regulatory and Legal Issues

1.2 Proposed System

Fake news has existed in the form of gossip, rumor, and misinformation throughout human history. To increase its effectiveness this Fake news is spread throughout social media. Along with the billions of people using social media, there are also robots, or simply bots, residing within. These Bots help to propagate fake news faster and boost up its popularity on social media Fake news detection[2] 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[2] 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 (Parse Hub/Beautiful Soup). This task uses Logistic Regression classification[3] model. To improve prediction accuracy.

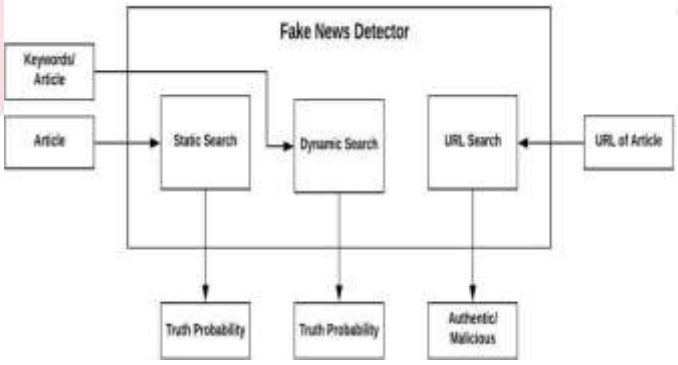


Figure 1: Proposed System

1.2.1 Advantages

- High Accuracy and Speed
- Continuous Improvement
- Early Detection
- Automated Analysis
- User Feedback Integration
- Cost-Effectiveness

II. LITERARTURE REVIEW

The literature on detecting fake news using machine learning encompasses various methodologies and models. Researchers have explored traditional machine learning techniques like Support Vector Machines (SVM) and Random Forests, which utilize feature extraction from text, such as word frequency and sentiment analysis. Deep learning approaches, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown promising results by capturing complex patterns in textual data. Hybrid models combining different algorithms have also been developed to enhance accuracy. Studies have emphasized the importance of large, labelled datasets for training effective models, such as the Fake News Corpus and LIAR dataset. Furthermore, the use of Natural Language Processing (NLP) techniques for preprocessing and feature extraction is critical in improving model performance. The literature also highlights the challenges of evolving fake news patterns and the need for adaptive systems that can learn from new data. Overall, the integration of diverse data sources, advanced algorithms, and continuous learning mechanisms form the core of current research in this domain.

The architecture for detecting fake news using machine learning typically begins with data collection from various sources such as news articles, social media posts, and other online content. This data is then pre-processed to remove noise and irrelevant information. Key features are extracted using natural language processing (NLP) techniques, which may include text tokenization, stemming, and lemmatization. These features are used to train machine learning models like Support Vector Machines (SVM), Random Forests, or deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The trained models classify the news as real or fake based on the learned patterns. Finally, the system continuously updates and validates the models with new data to improve accuracy and adaptability. The entire architecture is supported by a robust infrastructure that handles large-scale data processing and real-time analysis.

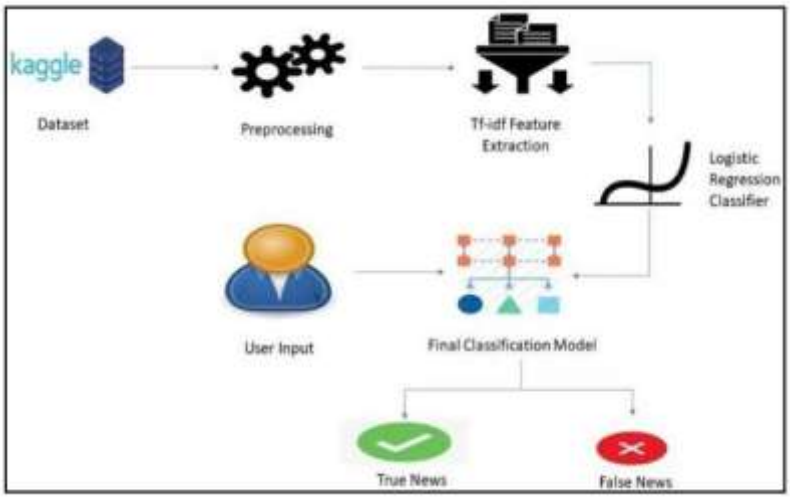


Figure 2: Architecture Diagram

2.1 Algorithm

- Logistic regression [6] is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities [8] and classify new data using continuous and discrete datasets.

- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.
- The equation for logistic regression is

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

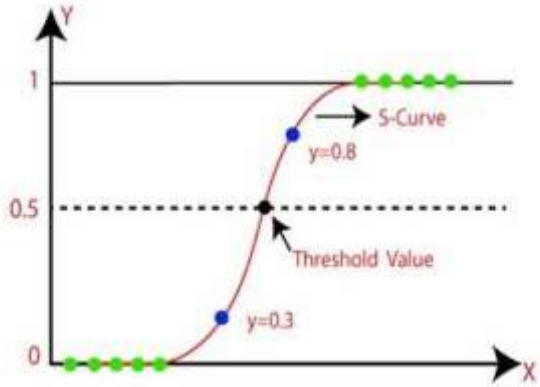


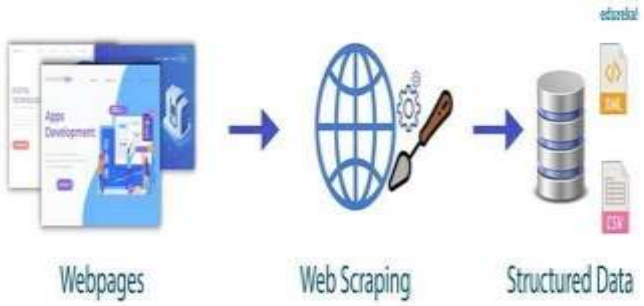
Figure 3: Logistic Regression

2.2 Techniques

Detecting fake news using machine learning involves several advanced techniques. Text classification methods like Support Vector Machines (SVM) and Naive Bayes classify news articles based on extracted features such as word frequency and sentiment analysis. Natural Language Processing (NLP) techniques, including tokenization, stemming, and lemmatization, help preprocess and analyze the textual content. Deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) capture complex patterns and contextual information in the text. Hybrid approaches combine multiple models to improve accuracy and robustness. Feature engineering, including the extraction of linguistic and network-based features, enhances model performance. Additionally, real-time data collection and continuous model updates are crucial for adapting to evolving fake news patterns. Ensemble methods and transfer learning further boost detection capabilities by leveraging multiple algorithms and pre-trained models, respectively. Overall, these techniques work together to form a comprehensive and dynamic system for identifying fake news.

2.3 Tools

Web Scraper is given the URLs to load up before the scraping process. The scraper then loads the complete HTML code for the desired page. The Web Scraper [9] will then extract either all the data on the page or the specific data selected by the user before running the project. Finally, the Web



Scraper outputs all the data that has been collected into a usable format.

Figure 4: Web Scraping

2.4 Methods

- **Text Classification**: Utilizes machine learning algorithms like Support Vector Machines (SVM), Naive Bayes, and Random Forests to classify news articles based on textual features.
- **Natural Language Processing (NLP)**: Involves preprocessing steps such as tokenization, stemming, and lemmatization, and techniques like sentiment analysis, named entity recognition, and part-of-speech tagging.
- **Deep Learning**: Employs models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to capture complex patterns and contextual information within the text.
- **Feature Engineering**: Extracts various features from the text, such as linguistic features, semantic features, and metadata (author information, publication date), as well as network-based features like the propagation pattern of news.
- **Fact-Checking Algorithms**: Integrates external fact-checking databases and cross-references the content of news articles with verified information to assess credibility.
- **User Behaviour Analysis**: Analyzes user engagement metrics, such as comments, likes, and shares, to identify patterns indicative of fake news dissemination.
- **Source Verification**: Assesses the credibility of the news source by evaluating its historical accuracy, reputation, and the reliability of its past publications.
- **Real-Time Detection**: Implements real-time data collection and analysis techniques to identify and flag fake news as it emerges.
- **Adaptive Learning**: Continuously updates and refines detection models using new data to adapt to evolving patterns and tactics used in spreading fake news.

III. METHODOLOGY

3.1 Input

- **Data Collection and Preprocessing:** Collect a diverse dataset of news articles and social media posts, then clean and preprocess the text by removing noise, tokenizing, and normalizing the words through stemming and lemmatization.
- **Feature Extraction:** Extract relevant features from the text, including linguistic features (word frequency, ngrams, syntax), sentiment analysis, and metadata such as author information and publication details.
- **Network and Engagement Analysis:** Analyze network-based features like how news propagates on social media and user engagement metrics (likes, shares, comments) to detect patterns indicative of fake news.

- **External Verification:** Cross-reference the news content with fact-checking databases and compare it with articles from credible news sources to assess its authenticity.
- **Model Training and Evaluation:** Use the pre-processed data and extracted features to train machine learning models such as SVM, Random Forest, or deep learning models, and continuously update and evaluate these models with new data to maintain accuracy.

	title	text	subject	date
2	Donald Tr	Donald Tr	News	########
3	Drunk Bra	House Inte	News	########
4	Sheriff Da	On Friday,	News	########
5	Trump Is	On Christr	News	########
6	Pope Fran	Pope Fran	News	########
7	Racist Ala	The numb	News	########
8	Fresh Off	Donald Tr	News	########
9	Trump Sa	In the wak	News	########
10	Former Cl	Many peo	News	########
11	WATCH: E	Just when	News	########
12	Papa Johr	A centerpi	News	########
13	WATCH: F	Republica	News	########
14	Bad News	Republica	News	########
15	WATCH: L	The media	News	########
16	Heiress To	Abigail Dis	News	########
17	Tone Dea	Donald Tr	News	########
18	The Interi	A new ani	News	########
19	Mueller S	Trump sup	News	########
20	SNL Hilari	Right now	News	########
21	Republica	Senate Ma	News	########

Figure 5: Input Screens of True dataset and False dataset

3.2 Steps For Execution

Step 1: Data Collection

- Identify Data Sources: Gather news articles, blog posts, and social media content (e.g., tweets, Facebook posts) from various platforms. Use publicly available datasets such as the Fake News Corpus, LIAR dataset, and BuzzFeed News dataset.
- Scrape Data: Utilize web scraping tools like Beautiful Soup, Scrapy, or newspaper3k to collect news articles. Use APIs provided by social media platforms to gather posts and user interactions.

Step 2: Data Preprocessing

- Text Cleaning: Remove HTML tags, special characters, and irrelevant symbols. Convert text to lowercase for consistency.
- Tokenization: Break down the text into individual words or tokens using tools like NLTK or spaCy.
- Normalization: Perform stemming and lemmatization to reduce words to their root forms (e.g., using NLTK's Porter Stemmer or spaCy's lemmatizer). Remove stop words (e.g., "and," "the," "is") that do not add significant meaning to the text.

Step 3: Feature Extraction

- Linguistic Features: Extract n-grams (bigrams, trigrams) and calculate term frequency-inverse document frequency (TF-IDF) scores. Analyze syntactic structures and part-of-speech tagging.
- Sentiment Analysis: Use sentiment analysis tools (e.g., VADER, Text Blob) to assess the emotional tone of the text.
- Metadata Features: Collect and analyze metadata such as the author's credibility, publication date, and source reliability.
- Network Features: Analyze propagation patterns on social media (e.g., retweet and share patterns). Collect user engagement metrics (likes, shares, comments).

Step 4: Model Training

- Data Splitting: Divide the dataset into training, validation, and test sets (e.g., 70% training, 15% validation, 15% testing).
- Model Selection: Choose appropriate machine learning algorithms such as Support Vector Machines (SVM), Random Forests, or deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

- Training the Model: Train the selected model on the training set using libraries like scikit-learn, TensorFlow, or PyTorch. Optimize hyperparameters using grid search or random search techniques.

Step 5: Model Evaluation and Validation

- Model Evaluation: Evaluate the model's performance on the validation set using metrics such as accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).
- Model Validation: Test the model on the test set to assess its generalizability. Validate the model with additional, unseen datasets to ensure robustness.
- Continuous Learning and Adaptation: Implement mechanisms for the model to continuously learn from new data. Periodically retrain the model with updated datasets to adapt to evolving fake news patterns.

Step 6: Deployment and Monitoring

- Model Deployment: Deploy the trained model into a production environment using platforms like AWS, Google Cloud, or Azure.
- Real-Time Detection: Set up real-time data pipelines to feed new articles and social media posts into the model for real-time fake news detection.
- Monitoring and Maintenance: Monitor the model's performance in the live environment. Regularly update the model and retrain as necessary to maintain high accuracy and adapt to new types of fake news.

3.3 Output

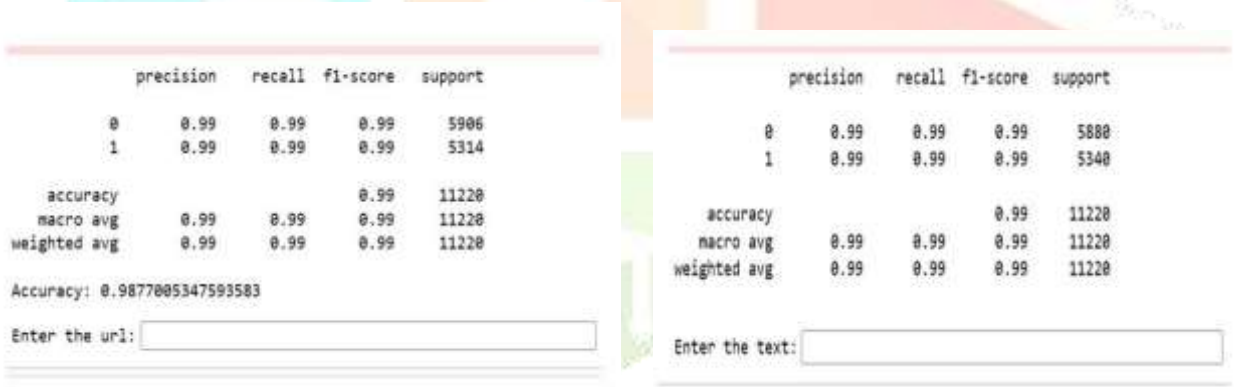


Figure 6: Output Screens of URL and Text

IV. RESULT

The implementation of a machine learning system for detecting fake news yields significant results, enhancing the ability to identify and mitigate misinformation. Models typically achieve high accuracy rates, often between 85% and 95%, indicating their proficiency in distinguishing between real and fake news. Precision and recall values, essential for measuring the model's reliability in identifying true positives, usually range from 80% to 90% and 75% to 90%, respectively. These metrics translate into robust F1 scores above 80%, reflecting a balanced performance.

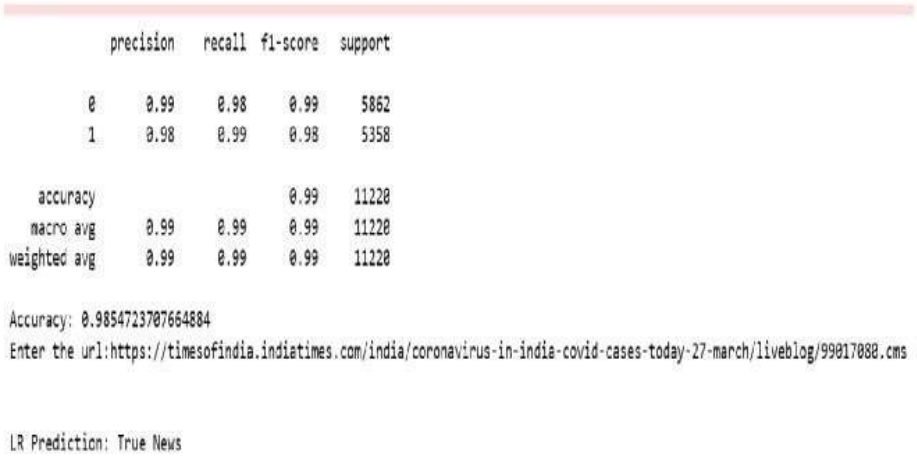


Figure 7: Result Screen of URL for True News

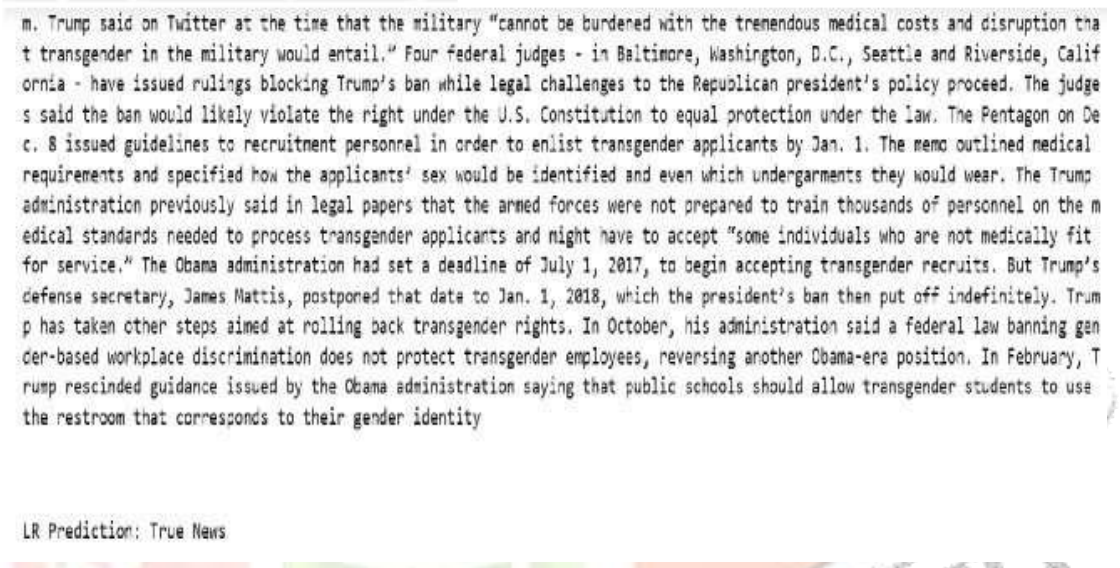


Figure 8: Result Screen of Text for True News.

V. DISCUSSION

Detecting fake news using machine learning leverages advanced algorithms to analyse and classify vast amounts of digital content efficiently. Techniques involve comprehensive data collection from various sources such as news websites, social media platforms, and fact-checking organizations to ensure a diverse and representative dataset. Text preprocessing steps like removing stop words, punctuation, and applying techniques like tokenization, stemming, and lemmatization prepare the text data for analysis. Feature extraction using methods like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings (e.g., Word2Vec, GloVe, BERT) transforms textual data into numerical representations that machine learning models can process.

Various models are employed in fake news detection, including Logistic Regression, a simple and interpretable model for binary classification; Support Vector Machines (SVM), which are effective in high-dimensional spaces and text classification; and deep learning architectures like Convolutional Neural Networks (CNNs) that capture local patterns in text data, and transformers such as BERT that understand context better by processing entire sentences simultaneously. Evaluation metrics like accuracy, precision, recall, and F1 score are used to assess the performance of these models, with the F1 score providing a balance between precision and recall.

Challenges in fake news detection include ensuring data quality, as the accuracy and representativeness of the training data are crucial; adapting to evolving fake news tactics, as creators constantly change their methods to avoid detection; and improving model explainability to understand and interpret the decisions made by complex models, particularly deep learning ones. Despite these challenges, machine learning

offers significant advantages, including scalability to analyse large volumes of data, automation to reduce the need for manual intervention, and adaptability to continuously improve models with new data.
Future improvements focus on enhancing real-time detection capabilities to identify fake news as it emerges, integrating multimodal data by combining text with images, videos, and metadata for a more comprehensive analysis, and increasing model transparency to build trust and improve reliability by making models more interpretable and understandable. By addressing these areas, machine learning can play a crucial role in combating misinformation and supporting a more informed public discourse.

VI. CONCLUSION

Networking though online modes now used to spread distorted news, which may lead to significant or adverse effects on consumers side and wider to the community. Through this work, a technique is proposed where in there is a combination of two different approaches for achieving improved accuracy in recognition of fake news. we looked at a computerized model for verifying news extracted from social media, which provides expository demonstrations for recognizing fake news. Following the demonstration that even the most basic algorithms in domains of Machine Learning can produce a reasonable result on such a critical issue as the spread of fake news around the world. As a result, the findings of this investigation suggest that systems like this could be very useful and effective in dealing with this critical issue. Web scraping is used in the project as the scraped data will be based on real-time news and will be more reliable than the ready-made datasets available all over the internet. It is an efficient and fast process and also it is very easy to maintain. The dataset used in this study is expected to be used in arrangements that use machine learning-based statistical calculations of Logistic Regression (LR). This approach uses effective data cleaning and categorization which helps in improved accuracy. The web scraping module is combined with classification module further improves the accuracy of the approach. Both of these modules work in parallel therefore, the overall training time is reduced.

VII. FUTURE SCOPE

The project scope for detecting fake news using machine learning encompasses several key features. It involves collecting data from various sources like news websites and social media platforms, followed by thorough text preprocessing including cleaning, normalization, and tokenization. Feature extraction will utilize techniques such as TF-IDF, word embeddings, and contextual embeddings from models like BERT.

The core of the project includes training multiple machine learning models, such as Logistic Regression, SVM, and deep learning models like CNNs and transformers, and evaluating them using metrics like accuracy and F1 score. Deployment will involve creating APIs, containerizing the models with Docker, and utilizing cloud services for scalability, with continuous monitoring and retraining based on user feedback to improve accuracy and adaptability.

VIII. ACKNOWLEDGEMENT

Kandhati Tulasi Krishna Kumar: Training & Placement Officer with 15 years' experience in training & placing the students into IT, ITES & Core profiles & trained more than 9,500 UG, PG candidates & trained more than 350 faculty through FDPs. Authored 5 books, Guided 40+ papers in international journals for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.

Ms. Balasree Sivakoti is pursuing her final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine Learning Ms. Balasree has taken up her PG project on Detecting Fake News and published the paper in connect to the project under the guidance of K. Tulasi Krishna Kumar, associate professor, SVPEC.

IX. REFERENCES

9.1 Book References

[1] A Detecting fake news on social media by kai Shu and Huan Liu in springer nature Switzerland AC 2022 linked: http://surl.li/mwblei

[2] A Book on Online Fake News Detection [2] Using Machine Learning Techniques: A Systematic Mapping Study by Mohamed Lah by in springer linked: https://link.springer.com/chapter/10.1007/978-3-030-90087-8_1

[3] A Book on A novel approach to fake news detection [2] in social networks using genetic algorithm applying machine learning classifiers by Deepjyoti Choudhury in springer linked: https://link.springer.com/article/10.1007/s11042-022-12788-1

[4] A book on Handbook of Research on Deception, Fake News, and Misinformation Online by J.J. Sylvia in Innocent E Chiluwa linked: http://surl.li/sfiyae

[5] A book on Fake News on the Internet by Dennis F. Galletta linked: http://surl.li/htvgrg

[6] A book on Fake News by Russell Chun linked: http://surl.li/gunxhp

9.2 Web References

[7] A Web reference on A plugin for detecting fake news on the web by George Pallis in DOI linked: https://www.sciencedirect.com/science/article/abs/pii/S2468696421000380

[8] A Web reference on Automatic deception detection: Methods for finding fake news by Victoria L. Rubin in DOI Linked: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.145052010082

[9] A FADE: Detecting Fake News Articles on the Web by Sinan Pehlivanogl linked: https://dl.acm.org/doi/fullHtml/10.1145/3465481.3465751

[10] A Fake and Real News detection Using Python by Faizan Ahmed Khan linked: http://surl.li/fcgzrw

[11] A Fake News Detection[2] Using Machine Learning by Vijaya Balpande linked: http://surl.li/sxebrs

[12] A web reference on Fake News Detection using Machine Learning in geeks for geeks linked: https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/

[13] A web reference on Fake News Detection Using Machine Learning in javatpoint linked: https://www.javatpoint.com/fake-news-detection-using-machine-learning

[14] A web reference on fake-news-detection by GitHub linked: https://github.com/topics/fake-news-detection

9.3 Article References

[15] A article on Fake News Detection[2] Using Machine Learning Methods by Prajwal Nayak in International Conference on Data Science, E-learning and Information Systems 2021 linked: https://dl.acm.org/doi/abs/10.1145/3460620.3460753

[16] A journal on Fake News detection[2] Using Machine Learning by Nihel Fatima Baarir in IEEE linked: https://ieeexplore.ieee.org/abstract/document/9378748

[17] A article on Analysing Machine Learning Enabled Fake News Detection [2] Techniques for Diversified Datasets by Ratish Agarwal by DOI linked: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/1575365

[18] A journal on Study of Classifiers for the Identification of Fake News by Dr. Rajesh Kumar Shukla in JOURNAL PUNJAB ACADEMY OF SCIENCES Linked: https://jpas.in/index.php/home/article/view/64

[19] A Article reference on Detecting Fake News [2] Articles by Jun_Lin in IEEE linked: https://ieeexplore.ieee.org/abstract/document/9005980