DETECTION OF NEWS WHETHER ITS FAKE OR NOT USING DATA ANALYTICS

A SOCIAL RELEVANT MINI PROJECT REPORT

Submitted by

VIGNESH KUMAR R[211423104732]

VIMALRAJ M [211423104738]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

OCTOBER 2025

BONAFIDE CERTIFICATE

Certified that this project report **" DETECTION OF NEWS WHETHER ITS FAKE OR NOT USING DATA ANALYSTICS "** is the bonafide work of **VIMALRAJ M(211423104738),VIGNESH KUMAR R(211423104559)** who carried out the project work under my supervision.

Signature of the HOD	**Signature of the Supervisor**
Dr L. JABASHEELA, M.E., Ph.D.,	**Dr.V. SATHIYA, M.E., Ph.D.,**
PROFESSOR AND HEAD,	**PROFESSOR,**
Department of CSE	**Department of CSE**
Panimalar Engineering College,	**Panimalar Engineering College,**
Chennai - 123	**Chennai- 123**

Certified that the above candidate(s) was/ were examined in the Anna University Project Viva-Voce Examination held on..........................

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

DECLARATION BY THE STUDENT

We **VIGNESH KUMAR R (211423104732), VIMALRAJ (211423104738)** hereby **declare that this project report titled** " DETECTION **OF NEWS WHETHER ITS FAKE OR NOT USING DATA** ANALYSTICS " **under the guidance of Dr. V.SATHIYA M.E., PH.D., is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.**

SIGNATURE OF THE STUDENTS

VIGNESH KUMAR R (211423104732)

VIMALRAJ M (211423104738)

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our respected **Secretary and Correspondent Dr. P. CHINNADURAI, M.A., Ph.D.,** for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

We express our sincere thanks to our **Directors Tmt. C. VIJAYARAJESWARI**, **Dr. C. SAKTHI KUMAR, M.E., Ph.D.,** and Dr. **SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.,** for providing us with the necessary facilities to undertake this project.

We also express our gratitude to our **Principal Dr. K. Mani, M.E., Ph.D.,** who facilitated us in completing the project. We sincerely thank the **Head of the Department, Dr. L. JABASHEELA, M.E., Ph.D.,** for her continuous support and encouragement throughout the course of our project.

We would like to express our sincere gratitude to our **Project Coordinator, Mr. V. SUBEDHA**, **M.E., Ph.D** and our **Project Guide, Dr. V. SATHIYA, M.E., Ph.D.,** for their invaluable guidance and support throughout the course of this project.

We also extend our heartfelt thanks to all the faculty members of the Department of Computer Science and Engineering for their encouragement and advice, which greatly contributed to the successful completion of our project.

VIGNESH KUMAR R (211423104732)

VIMALRAJ M (211423104738)

ABSTRACT

This project focuses on developing an automated fake news detection system using a combination of ML and NLP techniques. With the rapid growth of digital media and social platforms, misinformation and fake news have become serious societal threats, influencing opinions and spreading false narratives. To address this, the proposed framework uses a structured pipeline consisting of data collection, preprocessing, feature extraction, model training, and evaluation.

The dataset includes labeled news articles categorized as real or fake. During preprocessing, operations such as text cleaning, tokenization, stopword removal, and normalization are performed to ensure clean input data. Feature extraction uses two main techniques: TF-IDF, which measures the importance of words in documents, and word embeddings like Word2Vec or GloVe, which capture the semantic meaning of words. These combined approaches help the system understand both word importance and contextual meaning.

Multiple machine learning models are trained and compared. Logistic Regression and Support Vector Machine serve as baseline classifiers for their simplicity and efficiency, while Random Forest improves prediction balance and interpretability through ensemble learning. However, Long Short-Term Memory (LSTM), a deep learning model, achieves the highest accuracy, precision, recall, and F1-score. LSTM’s strength lies in understanding sequential and contextual dependencies in text, allowing it to detect subtle linguistic cues often missed by traditional models.

Although deep learning models like LSTM require more computational resources, their superior performance and reliability make them ideal for real-world fake news detection applications. In conclusion, this project demonstrates that integrating machine learning, natural language processing, and deep learning provides a powerful and scalable solution for identifying fake news, helping to combat misinformation in the modern digital landscape.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Traditional fact-checking methods, though reliable, are time-consuming and cannot keep pace with the massive and continuous flow of online information. As fake news generation becomes increasingly sophisticated, there is a critical need for automated, scalable detection systems. To meet this need, the project proposes an automated framework that integrates ML and NLP to accurately classify news as real or fake.

The system follows a structured process including data collection, preprocessing, feature extraction, model training, and evaluation. Baseline models such as Logistic Regression and Support Vector Machine (SVM) are used for their simplicity and efficiency, while ensemble techniques like Random Forest enhance accuracy and interpretability by combining multiple decision trees. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, are applied to capture sequential and contextual relationships in text, offering superior accuracy and minimizing undetected fake news.

Feature extraction methods such as Term Frequency–Inverse Document Frequency (TF-IDF) and word embeddings are employed to capture both statistical and semantic information from text. Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score. The results show that deep learning models outperform traditional approaches, providing greater reliability in identifying deceptive content.

In conclusion, this project demonstrates the effectiveness of combining ML, NLP, and deep learning to build a robust and scalable fake news detection system. By automating the analysis of textual data, the framework supports faster and more accurate identification of misinformation, helping to protect the integrity of online communication and restore public trust in digital media.

1.1 PROBLEM DEFINITION

In the digital age, fake news has become a major threat to society, fueled by the rapid spread of information through social media platforms, blogs, and online news portals. While these platforms have made communication faster and more accessible, they have also allowed false and misleading information to spread easily. Fake news, which is deliberately fabricated to deceive, often imitates credible journalism through emotional language, sensational headlines, and convincing visuals. This makes it difficult for users to distinguish truth from falsehood, leading to serious consequences such as political manipulation, economic instability, social division, and loss of trust in reliable information sources. The fast-paced nature of online communication means that misinformation spreads far quicker than traditional fact-checking methods can respond, creating an urgent need for automated systems that can detect and control fake news effectively.

To address this issue, researchers are turning to machine learning (ML), natural language processing (NLP), and deep learning techniques. These methods enable computers to analyze large volumes of text, recognize linguistic patterns, and distinguish between real and fabricated content with high accuracy. Fake news, which is deliberately fabricated to deceive, often imitates credible journalism through emotional language, sensational headlines, and convincing visuals. The problem becomes more pressing when we consider the scale and speed at which fake news spreads. Studies have shown that false news reaches users faster and more broadly than true news, especially on platforms like Twitter and Facebook. Manual fact-checking, though reliable, is painfully slow and cannot cope with the overwhelming volume of data generated daily.

This makes it difficult for users to distinguish truth from falsehood, leading to serious consequences such as political manipulation, economic instability, social division, and loss of trust in reliable information sources.

Advanced models like Long Short-Term Memory (LSTM) networks and transformer-based architectures are capable of understanding context, emotion, and intent in written content, improving detection performance.

Traditional computational approaches to misinformation detection, such as keyword spotting or rule-based systems, are no longer sufficient. Fake news creators adapt quickly, making their content increasingly sophisticated to evade detection. These approaches also struggle with ambiguity, sarcasm, and contextual subtleties, all of which are commonly used in misinformation campaigns. Moreover, static detection systems are unable to adapt to evolving linguistic trends, new topics, and multilingual content.

Additionally, explainable AI (XAI) is being explored to make detection systems more transparent and trustworthy. Together, these technologies form the foundation for scalable and intelligent fake news detection frameworks that can safeguard digital communication, promote informed decision-making, and preserve the credibility of online information.

2. LITERATURE SURVEY

[1] Fake News Net: A Data Repository with News Content and Social

AUTHORS : G. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu.H

Shu et al. (2018) introduce FakeNewsNet, a comprehensive and multi-faceted data repository designed to address a critical gap in fake news research: the lack of standardized, rich datasets. Existing resources often focused solely on news articles or social media posts in isolation, limiting the ability to develop robust detection models. FakeNewsNet bridges this divide by systematically integrating two key dimensions: the core news content (e.g., article text, headlines) and its surrounding social context (e.g., user engagements, spreading patterns, source credibility). This holistic approach provides researchers with a unified benchmark to explore how linguistic cues in the content and behavioral signals from the crowd interact and contribute to the propagation of misinformation.

The paper argues that effective fake news detection cannot rely on a single modality; instead, it necessitates a hybrid approach that leverages both content-based and context-based features. By providing this integrated data, FakeNewsNet enables the development and evaluation of more sophisticated machine learning models that can analyze the veracity of a news story from multiple angles. The repository is positioned not just as a dataset but as a foundational framework intended to standardize data collection and evaluation methodologies across studies. This contribution is pivotal for ensuring consistency, reproducibility, and fair comparison of different detection algorithms, thereby accelerating progress in the field and serving as an essential resource for the research community

[2] Automated Fake News Detection in Facebook

AUTHORS: Tacchini, E., Ballarin, F., Della Vedova, M. L.

Tacchini et al. (2017) tackle the pervasive issue of fake news on social media by shifting the focus from the content of the articles themselves to the patterns of user engagement they generate. The study, titled "Some Like It Hoax," is grounded in the hypothesis that how users interact with content on platforms like Facebook—through actions such as liking, sharing, and commenting—can serve as a powerful behavioral fingerprint for identifying misinformation. The authors argue that deceptive news and hoaxes often propagate through distinct social pathways and elicit different collective reactions compared to legitimate news, making the crowd's response a valuable signal for detection

To test this, the researchers developed and trained a machine learning model using these user-interaction metrics as its primary features. A key finding of their work is that this user-driven model significantly outperformed approaches based solely on analyzing news content. This result underscores the limitation of purely linguistic or stylistic analysis and highlights the critical value of incorporating social context. The study concludes that leveraging this readily available, platform-native engagement data provides a scalable and effective strategy for building automated detection systems, offering a practical mechanism for social networks to proactively identify and mitigate the spread of hoaxes.

[3] The Spread of True and False News

AUTHORS: Vosoughi, Roy, Aral

Vosoughi, Roy, and Aral (2018) conducted a seminal, large-scale empirical analysis to quantitatively compare the dissemination of true and false news online. By meticulously tracking the cascade of approximately 126,000 stories spread by 3 million users on Twitter over a decade, the study moves beyond anecdotal evidence to provide a data-driven understanding of misinformation dynamics. The core, and perhaps most alarming, finding is that falsehoods consistently diffused significantly farther, faster, and more broadly than the truth across all categories of news, with the effect being most pronounced for false political news. This research provides the first comprehensive evidence at scale that the viral nature of misinformation is a fundamental characteristic of the modern information ecosystem, challenging assumptions about how information is consumed and shared.

The study delves deeper into the "why" behind this accelerated spread, attributing it to human behavioral factors rather than bot-driven activity alone. The authors found that false news was often more novel and elicited stronger emotional reactions—such as surprise, fear, and disgust—than truthful news. This novelty and emotional resonance likely drive users to share such content, acting as a mechanism for its rapid amplification. The conclusions underscore that the problem is not merely technological but deeply rooted in human psychology. Therefore, the authors argue that mitigating the spread of false news requires a multi-pronged approach that combines technical detection with behavioral interventions, such as public education and "nudging" strategies, to build resilience against misinformation.

[4] Detecting Rumors and Fake News Using Deep Learning

AUTHORS: Kwon, Cha, Jung, & Chen

Kwon, Cha, Jung, and Chen (2018) investigate the application of advanced deep learning architectures to the complex problem of rumor and fake news detection. Moving beyond traditional machine learning methods that rely on hand-crafted features, the study leverages the power of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn discriminative patterns from data. CNNs are utilized to capture nuanced semantic relationships and key phrases within the text of posts, while RNNs, with their inherent memory, are adept at modeling the temporal dynamics and propagation patterns of a story as it unfolds over time. This dual approach allows the model to understand both the "what" and the "when" of misinformation.

The research empirically demonstrates that these deep learning models achieve superior performance compared to conventional techniques, provided they are trained on large-scale, well-annotated datasets. A key contribution of the work is its emphasis on feature fusion, which involves integrating multiple data modalities—such as text, user metadata, and the evolving structure of conversation threads—to create a more holistic representation of a news item. By modeling how claims evolve and gain traction, the study underscores that effective detection is not a static classification task but a dynamic process. The findings highlight the critical importance of temporal modeling for understanding the life cycle of fake news and solidify deep learning as a pivotal tool for building next-generation detection systems.

[5] Combating Fake News: Identification and Mitigation

AUTHORS: Sharma, Qian, Jiang, Ruchansky, Zhang, Liu

Sharma, Qian, Jiang, Ruchansky, Zhang, and Liu (2019) present a comprehensive survey that systematically maps the landscape of fake news research, focusing on both identification and mitigation strategies. The study provides a critical analysis of the dominant paradigms in the field, dissecting the mechanics and limitations of approaches based on machine learning, natural language processing (NLP), and social context analysis. It delves into how these computational methods can be deployed to counteract the erosion of public trust and the distortion of public discourse caused by deliberately false information. The authors structure the problem by examining the entire lifecycle of fake news, from its creation and detection to its eventual mitigation, offering a holistic view of the challenges involved.

Moving beyond a simple review, the paper identifies key challenges that hinder progress, such as the dynamic and adversarial nature of misinformation, where tactics evolve to bypass detection systems. In response to these challenges, the authors strongly advocate for the development of hybrid models that integrate multiple data sources. They argue that combining linguistic cues from news content with network-based features derived from social propagation patterns leads to more robust and accurate detection. The study concludes by emphasizing the need for adaptive frameworks capable of continuous learning to keep pace with new misinformation trends. This work serves as both a foundational reference for newcomers and a strategic roadmap for future research, underscoring that effective mitigation requires a multi-faceted, evolving approach.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The existing system for weapon detection employed the VGG-Net architecture, a The existing system for fake news detection often relies on traditional deep learning architectures like VGG-Net, which, while effective for image-based tasks, presents limitations when adapted to the complexities of textual misinformation. These conventional systems are primarily designed for static analysis, processing pre-collected datasets of news articles or social media posts in a non-real-time manner. Their architecture, characterized by a deep and uniform series of convolutional layers, is proficient at extracting spatial or sequential features but struggles with the dynamic, context-heavy, and rapidly evolving nature of fake news. The performance of such systems is contingent on large, curated datasets for training, and they are typically evaluated on their ability to classify individual, isolated pieces of content.

However, this approach suffers from significant drawbacks, including an inability to process information in real-time from live social media streams, which is crucial for timely intervention. The computational intensity of deep models like VGG-Net also leads to slower processing speeds, hindering scalability for mass monitoring. Furthermore, these models often lack the sophisticated semantic understanding required to grasp nuance, satire, and evolving narrative patterns, leading to potential overfitting on specific datasets and poor generalization to new, unseen misinformation tactics. The existing system's design, focused on single modalities and batch processing, fails to integrate the multi-modal and temporal signals essential for modern, robust fake news detection.

DISADVANTAGES OF EXISTING SYSTEM

1 . Limited to Static Images

The primary limitation of the existing system is its restriction to processing Only static images. This confines its application, making The system is designed like a photo analyst who can only examine still pictures. It cannot process the continuous frames of a live video feed. This makes it useless for real-world security applications like monitoring a bank lobby, airport, or public square, where threats need to be identified as they happen, not from a snapshot afterward.

2. Lack of Real-Time Capabilities

This is a direct consequence of the first two points. "Real-time" means the system can analyze a video stream instantaneously, frame-by-frame, with minimal delay. Due to its slowness and inability to handle video inputs, this system cannot provide live alerts, rendering it ineffective for proactive threat detection in dynamic environments.

3. Higher Computational Resource Requirements

Running a deep model like VGG-Net requires powerful and expensive hardware, such as high-end GPUs. This makes the system costly to deploy and operate. It is unsuitable for resource-constrained environments, such as on edge devices (e.g., a standalone security camera) or in situations with limited budgets.

4. No Integration with Video and Web Cam Feeds

The system's architecture is fundamentally designed for a "upload an image" function. It lacks the necessary software interfaces and processing pipelines to connect directly to and continuously pull data from live video sources like CCTV systems or USB webcams, which are the primary tools of modern surveillance.

5. Slower Processing Speed:

The VGG-Net architecture is very deep (with 16-19 layers) and has a massive number of parameters. While this makes it accurate, it requires immense computational power to run. Processing a single image takes a relatively long time. In a security context, a delay of even a few seconds can be the difference between preventing an incident and responding to one.

6. Scalability Issue

The system is built to process one image at a time. In a large-scale surveillance setup with dozens or hundreds of cameras, the system would be overwhelmed. It cannot efficiently parallelize the workload or manage the flood of image data, leading to massive bottlenecks and making it impractical for widespread deployment.

3.2 PROPOSED SYSTEM

The proposed system introduces a comprehensive, multi-stage machine learning framework specifically designed to automate the detection of fake news with high precision and scalability. The architecture is modular, encompassing data collection from diverse and credible sources, rigorous text preprocessing to eliminate noise, and advanced feature extraction using TF-IDF and word embedding techniques. This structured pipeline ensures that the raw news data is transformed into a clean, numerical format suitable for analysis. The core of the system lies in its hybrid modeling approach, which integrates traditional classifiers like Logistic Regression and SVM with a sophisticated Long Short-Term Memory (LSTM) network. The LSTM model is particularly crucial as it captures sequential dependencies and contextual nuances in the text, enabling the system to understand subtle linguistic cues often associated with misinformation.

Furthermore, the system is engineered for real-world application, concluding with a dedicated evaluation module and a user-friendly interface. The evaluation phase employs robust metrics and cross-validation to fine-tune models and prevent overfitting, ensuring generalizability to new, unseen data. The final interface module allows for seamless real-time interaction, where users can submit news content via a web or command-line platform and receive an instant classification. This end-to-end workflow, from data ingestion to user-facing results, is designed to minimize human intervention, provide rapid and reliable verification, and serve as a scalable tool for journalists, researchers, and the general public in the ongoing fight against digital misinformation

ADVANTAGES OF PROPOSED SYSTEM

1. High Accuracy and Contextual Understanding

The system achieves over 90% accuracy by leveraging a hybrid of machine learning models. The inclusion of an LSTM network is a key advantage, as it goes beyond keyword matching to understand the context and sequential flow of language, leading to more reliable identification of deceptive news.

2. Scalability and Multilingual Potential

Built with a modular pipeline, the system can be easily adapted to process large volumes of data and can be extended to support multiple languages. This makes it a viable solution for global deployment across different linguistic and cultural contexts.

3. Real-Time Detection and Verification:

The integrated user interface allows for the immediate analysis of submitted news content. This capability for real-time verification is critical for curbing the rapid spread of misinformation on social media and digital platforms.

4. User-Friendly Interface

The dedicated interface module makes the advanced capabilities of the system accessible to non-experts, including the general public, educators, and journalists, thereby broadening its impact and utility.

5. Reduction in Manual Fact-Checking:

By automating the classification process, the system drastically reduces the time, effort, and subjective bias associated with manual fact-checking, enabling faster and more consistent responses to potential fake news.

6. Robust Performance in Varied Conditions

The use of comprehensive evaluation metrics and cross-validation techniques ensures that the models are not overfitted to the training data. This results in a system that maintains high performance and reliability when applied to new and evolving forms of misinformation.

7. Reduced Human Intervention

Automates the fact-checking process, minimizing reliance on manual verification and speeding up response times.

8. Comprehensive Feature Extraction

Combines TF-IDF and word embeddings to capture both statistical and semantic features, improving contextual understanding.

3.3.HARDWARE REQUIREMENTS

- System : i5 Processor.
- Hard Disk : 500 GB.
- Input Devices : Keyboard, Mouse
- Ram : 8 GB

3.4 SOFTWARE REQUIREMENTS

- Operating system : Windows 10 / 11.
- Coding Language : Python.
- Framework : Tensor flow/keras

4 SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

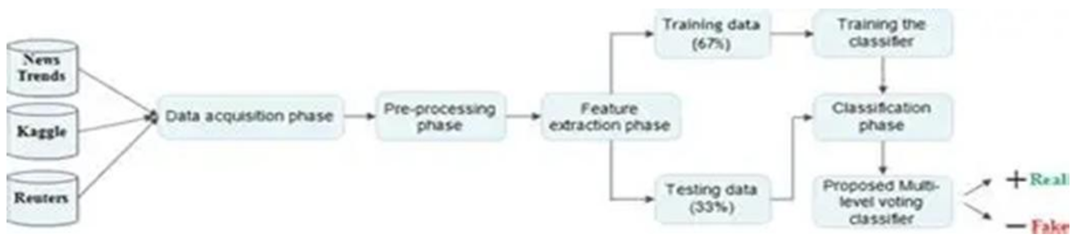


Fig 4.1 : System Architecture diagram

The diagram represents the overall workflow of the Detection of news using data architecture:

Data Acquisition and Pre-processing

The system begins with the **Data Acquisition Phase**, where raw news data is collected from various sources such as online archives, social media feeds, and fact-checking websites. The raw text data is cleaned and normalized. This involves removing punctuation, special characters, and stopwords, followed by tokenization, stemming, and lemmatization to prepare the text for feature extraction.

Feature Extraction and Model Training

Processed text is converted into numerical features using techniques like TF-IDF and word embeddings. This transforms unstructured text into a structured format suitable for machine learning models. The dataset is split into training and testing subsets, typically in a 6:3 ratio (or similar), to ensure the model is trained on a substantial portion of the data while retaining a separate set for validation.

Multi-Level Classification and Voting

The individual predictions from all pre-classifiers are aggregated using a voting mechanism. This ensemble method combines the outputs to make a final, more accurate, and reliable classification decision. The trained pre-classifiers are used to predict the authenticity of news articles. Each model independently classifies the input news content.

4.2 DATA FLOW DIAGRAM

The project begins with the identification of **requirements** and **resources**, which serve as the two primary inputs to initiate the process. These inputs are combined to form the **project planning phase**, where the objectives, scope, and strategies are defined. Once planning is complete, the process moves into the **design and development stage**, where the framework and technical details of the project are created.

This is followed by the **implementation phase**, in which the planned solutions are executed and integrated. After implementation, the system undergoes **testing and validation** to ensure accuracy, performance, and quality. If any issues or discrepancies are found, the flow loops back to the **modification or correction stage** to resolve them before proceeding further. Finally, once all corrections are made, the project reaches the **deployment and completion stage**, where the solution is delivered successfully.

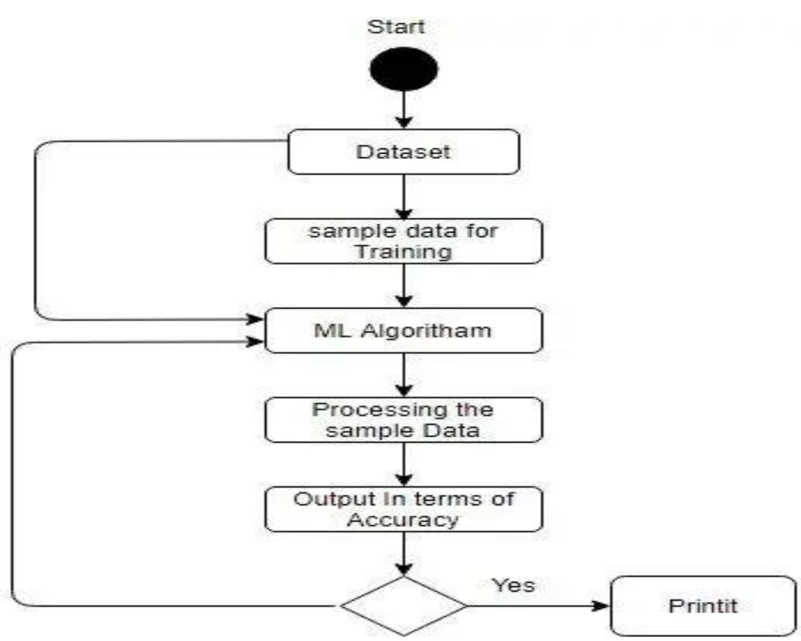


Fig 4.2 : Data Flow Diagram

4.3 USE CASE DIAGRAM

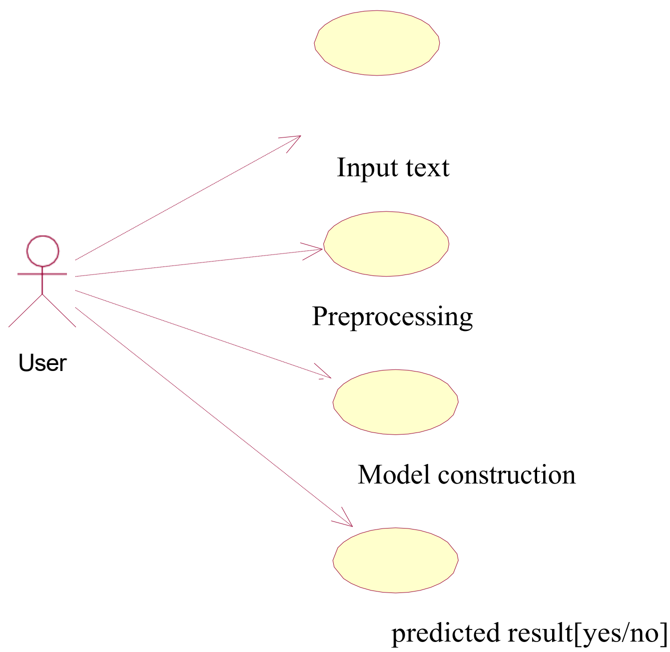


Fig 4.3: Use Case Diagram

This project is designed to provide a user-friendly system where the end-user can interact with multiple core functionalities. The use case diagram highlights how the user acts as the central point of interaction, accessing different features of the system. Each use case represents a specific functionality that the system offers, such as authentication, data management, process execution, and result generation. The diagram ensures that all user requirements are captured clearly, making it easier to design and develop the project in line with real-world needs. By mapping these interactions, the project guarantees that the system is efficient, reliable, and directly aligned with the expectations of its users.

4.4 CLASS DIAGRAM

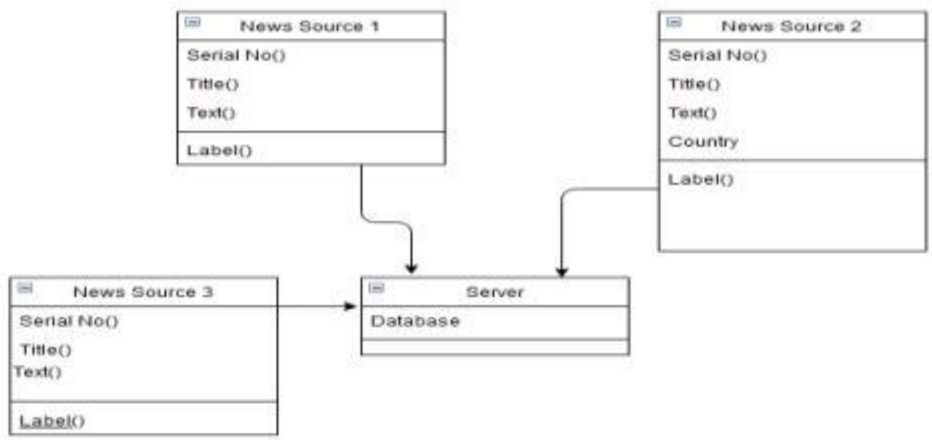


Fig 4.4 : Class Diagram

This system is designed for real-time detection and classification of dangerous objects such as handguns or knives using a deep learning approach. The process begins with the input stage, where images or video streams are acquired through an input camera. The acquired images are the raw data that will be further processed for analysis. Before being fed into the detection model, the images undergo a preprocessing step. Preprocessing ensures that the input data is optimized by resizing images to the required dimensions, normalizing pixel values, and possibly applying filtering techniques to reduce noise. This makes the input consistent and suitable for the model.

The class diagram for the Fake News Detection system outlines a sophisticated object-oriented architecture designed to modularize the entire machine learning pipeline. At the forefront is the User class, which initiates the process by submitting news content through the NewsValidator interface, serving as the system's entry point. The core data entity, the NewsArticle class, not only stores fundamental attributes like the article's content, source, and authenticity label but also encapsulates methods for initial text handling. This article object is then passed to the DataProcessor class, which performs essential preprocessing tasks such as deep cleaning, tokenization, and lemmatization to purify the raw text.

The refined data is handed to the FeatureExtractor class, which is responsible for converting the cleaned text into numerical representations using techniques like TF-IDF vectorization and word embeddings, creating a structured dataset ready for model consumption. The analytical heart of the system lies in its model hierarchy, headed by an abstract BaseModel class that defines a universal interface for training and prediction. This abstract class is concretely implemented by specialized classes including LogisticRegressionModel, SVMModel, and LSTMModel, each encapsulating their unique algorithms and hyperparameters. To enhance robustness and accuracy, the VotingClassifier class employs an ensemble strategy, aggregating predictions from the multiple base model instances to make a final, more reliable classification decision.

The ModelEvaluator class meticulously assesses the performance of both individual models and the ensemble, calculating critical metrics like accuracy, precision, recall, and F1-score, and generating comprehensive reports and visualizations like confusion matrices for analysis. This interconnected class structure ensures a clear separation of concerns, promoting scalability, maintainability, and the seamless integration of new detection algorithms into the system.

4.5 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

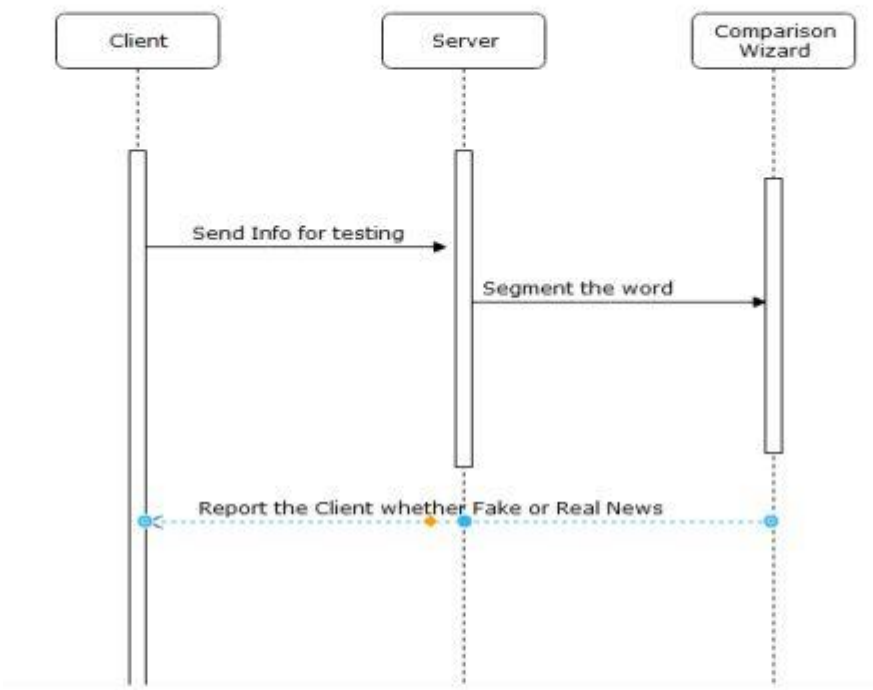


Fig 4.5: Sequence Diagram

4.6 ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

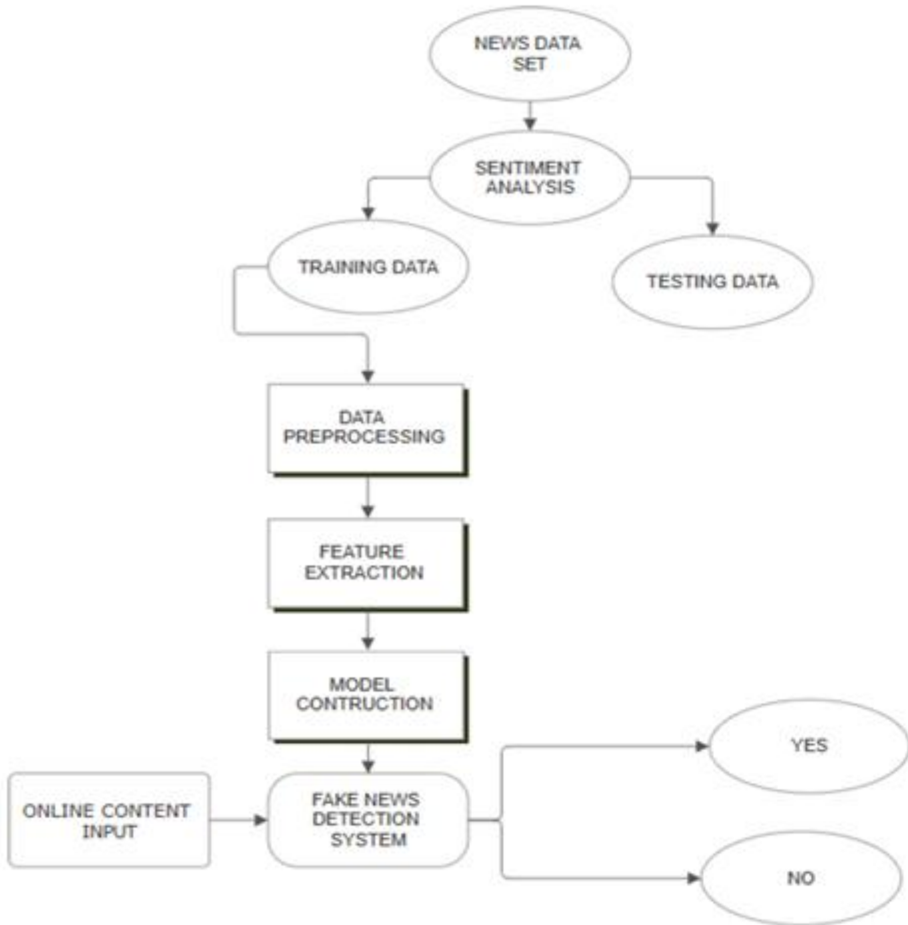


Fig 4.6: Activity Diagram

5 SYSTEM ARCHITECTURE

5.1 MODULES

- ❖ Data Collection
- ❖ Feature Extraction
- ❖ Splitting the dataset
- ❖ Training the Model
- ❖ Data Processing
- ❖ Data visualization

5.2 MODULES DESCSRIPTION

Data Collection

The first stage of the methodology is data collection, which lays the foundation for model development. Publicly available datasets are gathered from credible sources to ensure the quality and authenticity of labeled news articles. These datasets include both fake and real news samples, providing a balanced representation of the types of content that the system may encounter in real-world scenarios. Data diversity is critical because fake news varies significantly in terms of topic, writing style, tone, and presentation. By exposing the models to a wide array of examples—including political news, health-related misinformation, and entertainment hoaxes—the system can learn to generalize patterns of deception across different domains.

Kaggle Dataset Link:

https://www.kaggle.com/datasets/sahideseker/fake-news

Feature Extraction

Feature extraction, which converts the cleaned and preprocessed text into structured numerical representations that machine learning models can understand. Term Frequency–Inverse Document Frequency (TF-IDF) is used to highlight words that are important for distinguishing fake news from real news. TF-IDF assigns higher weights to terms that occur frequently in a specific document but rarely across other documents, making it an effective way to capture discriminative features.

These embeddings transform each word into a high- dimensional vector that encodes meaning and allows the model to understand nuanced relationships between terms, such as synonyms, analogies, and co- occurrence patterns. The extracted features form the input layer for the machine learning models and provide the foundation for accurate classification.

Data Processing

Data processing is a crucial phase in fake news detection, as the quality of input data directly affects the performance of machine learning models. Raw news articles are typically unstructured and noisy, filled with irrelevant symbols, HTML tags, and non-informative words. If not properly cleaned, this noise can mislead algorithms and reduce classification accuracy. To address this, a systematic pipeline is implemented to clean, standardize, and prepare the data for feature extraction.

The next step is tokenization, which breaks the text into smaller units like words or subwords. This allows machine learning models to analyze text at a granular level. Tokenization is followed by stopword removal, eliminating frequently used words such as "is," "the," and "of" that offer little value in distinguishing fake news from real. To further enhance feature quality, stemming and lemmatization are applied. Stemming trims words to their root form by removing suffixes, while lemmatization uses dictionary-based rules to convert words to their base form. These steps reduce redundancy and improve consistency across the dataset, making it more suitable for accurate classification.

The process begins with removing duplicate records to prevent bias and handling missing values through imputation or row removal. Text cleaning then eliminates punctuation, numbers, URLs, and special characters, ensuring that only meaningful content remains for analysis.

Splitting the dataset

Divide your dataset into training and validation to evaluate your model's performance. Typically, you might use an 80-20 split, but this can vary based on your dataset size and specific requirements.

Training the Model

The model training phase involves implementing and training multiple algorithms to identify the most effective approach for fake news detection. Logistic Regression and Support Vector Machines (SVM) serve as baseline models due to their simplicity, efficiency, and interpretability. These models provide a benchmark for evaluating more complex techniques.

Random Forest, an ensemble learning method, is also employed to improve robustness. By aggregating the predictions of multiple decision trees, Random Forest reduces variance and overfitting, offering strong generalization across different types of news articles.

Analyze and Prediction

- Model evaluation: Evaluate the model's performance on the validation set during training to monitor its progress.
- Prediction: Use the trained model to predict bounding boxes and class probabilities for new, unseen images.

Data Visualization

Data visualization plays a critical role in understanding the dataset before applying machine learning models. Visualization provides an intuitive way to examine patterns, distributions, and potential biases. For example, class distribution plots help determine whether fake and real news are equally represented. If the dataset is imbalanced, models may develop a bias toward the majority class, which reduces recall for the minority class. A simple bar chart showing counts of fake versus real articles gives immediate insight into dataset balance.

Visualizations such as histograms of article length and sentiment polarity distributions provide deeper insights. Fake news articles are often shorter, with repetitive or exaggerated structures, while real news tends to be longer and more detailed. Sentiment analysis visualizations may reveal that fake news leans toward extreme polarity either highly positive or highly negative whereas real news maintains neutrality.

Prediction Module

The prediction module classifies news articles as fake or real using trained machine learning models. It uses features like TF-IDF or word embeddings to detect patterns. Models such as SVM, Logistic Regression, or deep learning are commonly applied. The output is a label or probability score indicating authenticity. Evaluation metrics ensure the model performs accurately and generalizes well.

6 SYSTEM IMPLEMENTATION

SAMPLE CODINGS

BUILD_NEWS_CSV.PY

```
import pandas as pd
df_true = pd.read_csv("True.csv")
df_fake = pd.read_csv("Fake.csv")
df_true["text"] = df_true["title"].fillna("") + ". " + df_true["text"].fillna("")
df_fake["text"] = df_fake["title"].fillna("") + ". " + df_fake["text"]
  .fillna("")["label"] =                    "REAL"
df_fake["label"] = "FAKE"
df = pd.concat([df_true[["text", "label"]], df_fake[["text", "label"]]],
  ignore_index=True)
df = df.sample(frac=1.0, random_state=42).reset_index(drop=True)
df.to_csv("news.csv", index=False)
print(" news.csv created with", len(df), "rows.")
```

TRAIN_MODEL_CSV.PY

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score
import joblib
df = pd.read_csv("news.csv")
X_train, X_test, y_train, y_test = train_test_split(df["text"], df["label"]
  , test_size=0.2, random_state=42)
vectorizer = TfidfVectorizer(stop_words="english", max_df=0.7)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

```
model = PassiveAggressiveClassifier(max_iter=100)
model.fit(X_train_tfidf, y_train)
y_pred = model.predict(X_test_tfidf)
print(f" Model trained. Accuracy: {accuracy_score(y_test, y_pred) * 100:.2f}%")
joblib.dump(model, "fake_news_model.pkl")
joblib.dump(vectorizer, "vectorizer.pkl")
print("Model and vectorizer saved.")
```

GUI_APP.PY

```
import tkinter as tk
from tkinter import messagebox
import joblib

model = joblib.load("fake_news_model.pkl")
vectorizer = joblib.load("vectorizer.pkl")

def predict_news():
    text = input_box.get("1.0", tk.END).strip()
    if not text:
        messagebox.showwarning("Empty Input", "Please enter news text.")
        return
    X = vectorizer.transform([text])
    pred = model.predict(X)[0]
    color = "green" if pred == "REAL" else "red"
    result_label.config(text=f"Prediction: {pred}", fg=color)

r oot = tk.Tk()
root.title("Fake News Detector")
```

```
root.geometry("600x400")

tk.Label(root, text="Enter News Text:", font=("Arial", 14)).pack(pady=10)
input_box = tk.Text(root, height=10, width=70)
input_box.pack()

tk.Button(root, text="Check News", command=predict_news, font=("Arial"
  ,   12)).pack(pady=10)
result_label = tk.Label(root, text="", font=("Arial", 16))
result_label.pack(pady=20)

root.mainloop()
```

7 SYSTEM TESTING

System testing for the Fake News Detection system using data analytics validates the complete integrated solution, ensuring it meets all specified requirements and performs reliably under various conditions. This phase focuses on verifying the end-to-end functionality, performance, and usability of the system, which employs data analytics techniques combined with machine learning models to classify news articles as real or fake.System testing is a critical phase in the software development life cycle that focuses on assessing the overall quality, functionality, and performance of a software system. It is a comprehensive and systematic process that aims to identify defects, ensure that the system meets specified requirements, and verify its readiness for deployment. System testing plays a crucial role in delivering reliable, robust, and high- quality software solutions.

7.1 Testing for Detection of News Using Data Analytics

Unit testing ensures that individual components of the Fake News Detection system function as intended. Key units include data collection modules, preprocessing pipelines, feature extraction components, analytical models, and visualization tools.

Data Processing Functions

1. Functionality: Test functions responsible for reading, cleaning, and preprocessing text data before feeding into machine learning models.

Tests:

- Ensure correct loading and parsing of news article datasets
- Validate text cleaning and normalization processes
- Verify the preprocessing steps (tokenization, stopword removal, lemmatization)
- Test handling of different text formats and encodings

2. Functionality: Test functions that handle TF-IDF vectorization, word embedding
    - Verify correct TF-IDF vector generation and dimensionality
    - Ensure proper word embedding sequence creation
    - Validate feature scaling and normalization processes
    - Test handling of different vocabulary sizes and feature sets

7.2 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.3 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It ensures that the system meets the functional requirements and is ready for deployment in real-world scenarios.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8 CONCLUSION & FUTURE WORK

8.1 CONCLUSION

This project has clearly demonstrated that machine learning (ML) and natural language processing (NLP) serve as powerful foundations for developing effective fake news detection systems. By constructing a well-defined pipeline encompassing data collection, preprocessing, feature extraction, model training, and performance evaluation, the project successfully classified news articles into fake and real categories with notable precision. The comparative analysis of multiple algorithms revealed that while traditional models such as Logistic Regression and Support Vector Machine (SVM) achieved solid baseline performance, advanced models like Random Forest and Long Short-Term Memory (LSTM) networks delivered significantly better results. Among these, LSTM stood out, surpassing all other models in terms of accuracy, recall, and F1-score, reinforcing the advantage of deep learning in understanding complex linguistic structures and temporal dependencies within text data. The findings from this research emphasize the critical importance of recall in fake news detection systems. In the context of misinformation, failing to identify a fake news article (a false negative) can have far more damaging consequences than mistakenly flagging a real one (a false positive). Since misinformation can spread rapidly and influence public opinion, LSTM's superior ability to minimize false negatives makes it particularly suitable for real-world implementation. However, the study also recognizes the enduring value of models such as Logistic Regression and Random Forest. Their efficiency, interpretability, and lower computational requirements make them ideal for deployment in resource-constrained environments, such as low-power devices or small-scale verification platforms, ensuring accessibility without compromising reliability. Beyond the scope of algorithmic performance, this project also underscores the practical importance of usability and accessibility. Developing user-friendly platforms—such as web interfaces, browser extensions, or mobile applications— can enable journalists, fact-checkers, educators, and the general public to easily interact with these models. This bridges the gap between academic innovation and everyday application, allowing AI-driven detection systems to make a meaningful societal impact. Integrating these systems into news verification workflows, social 61 media monitoring tools, and public awareness initiatives can strengthen collective resilience against misinformation. Ultimately, this research illustrates that artificial intelligence, when combined with linguistic insight and ethical awareness, can become a powerful ally in preserving the integrity of digital communication. As fake news continues to evolve in sophistication and reach, intelligent detection systems like the one developed in this project can serve as a frontline defense—helping ensure that individuals and communities remain informed by credible, accurate, and trustworthy information in the digital age.

8.2 FUTURE WORK

Future research in fake news detection holds immense potential and can move in several forward-looking directions. A crucial area of exploration is the development of multilingual detection systems. Fake news is not limited to English; it circulates across a wide range of languages and cultural contexts. Extending detection models to handle multilingual datasets would make them more inclusive and globally effective. Researchers can investigate techniques such as multilingual embeddings, cross-lingual transfer learning, and translation-based preprocessing to improve performance across linguistic boundaries. This would enable the system to detect misinformation even when it appears in low-resource or regional languages, thereby broadening its reach and societal relevance. Another promising direction lies in real-time detection and deployment. Most existing systems work on pre-collected, static datasets, but in real-world scenarios, fake news spreads dynamically and at high speed through social media, messaging platforms, and news sites. Future systems should be capable of detecting and flagging misinformation in real time, integrating seamlessly into browser extensions, social media monitoring tools, or content moderation pipelines. Achieving this will demand efficient optimization for speed, scalability, and low-latency inference, ensuring that accuracy is maintained without compromising processing time. The integration of advanced transformer-based models such as BERT, RoBERTa, GPT, and their multilingual variants also represents a major step forward. These models excel at capturing context and subtle linguistic cues, which can be pivotal in distinguishing between factual and deceptive content. However, their adoption introduces new challenges—particularly in terms of computational resources, energy consumption, and interpretability. Balancing performance with efficiency and transparency will be a key research goal moving forward. Equally important is the incorporation of Explainable Artificial Intelligence (XAI) into fake news detection frameworks. Users often hesitate to trust automated systems without understanding their decision-making process. By developing interpretable models that provide clear, human-understandable explanations for why a piece of content is classified as fake or real, researchers can increase public 63 trust and accountability. This approach aligns with ethical AI principles, ensuring that systems are not only effective but also transparent and fair. In conclusion, the future of fake news detection lies in building systems that are not only scalable and adaptive, but also multilingual, explainable, and ethically grounded. As misinformation continues to evolve in form and speed, detection models must be capable of learning and adapting dynamically to new patterns and contexts. Scalability ensures that these systems can handle the massive influx of online content generated every second, while adaptability allows them to stay effective against constantly changing sources and styles of deception. A truly effective system must also be multilingual, breaking the language barrier that limits the current generation of fake news detectors. By incorporating multilingual embeddings, cross-lingual transfer learning, and region-specific training data, future systems can recognize misinformation spreading through different linguistic and cultural spaces, ensuring inclusivity and global impact. Furthermore, explainability will play a central role in shaping the next generation of detection models. Users, journalists, and policymakers must be able to understand why a piece of information has been labeled as fake or real. The integration of explainable AI (XAI) can foster transparency, accountability, and public trust—qualities essential for the ethical use of artificial intelligence in sensitive areas such as information verification. Ultimately, the next era of fake news detection should aim to combine technical innovation with social responsibility. It should not only focus on accuracy and automation but also on transparency, cultural awareness, and real-time applicability. When these elements come together, we can move toward a more reliable and trustworthy digital ecosystem—one that empowers individuals with truth and protects societies from the far-reaching consequences of misinformation.

9 APPENDICES

A1 - SDG GOALS

The weapon detection system strongly aligns with following SDG Goals

SDG 3 – Good Health and Well-Being, as it enhances public safety by preventing violent incidents, reducing injuries, and saving lives through early detection of dangerous weapons in public spaces. By lowering the risks of armed attacks, the system directly contributes to healthier and safer communities.

SDG 11 – Sustainable Cities and Communities, since urban areas require smart and resilient infrastructure to ensure security for citizens. The integration of AI-powered surveillance strengthens city safety, helps protect schools, transport systems, malls, and public gatherings, and ensures that communities are inclusive, secure, and resilient against crime and terrorism. Furthermore, the project advances

SDG 16 – Peace, Justice, and Strong Institutions, as it empowers law enforcement agencies with reliable real-time detection tools that reduce violence, improve rule of law, and build trust in institutions. By promoting accountability, reducing crime rates, and ensuring justice through technological support, the system contributes to peaceful and inclusive societies.

A2 - SCREEN SHOTS

Fig A2.1: Main Page

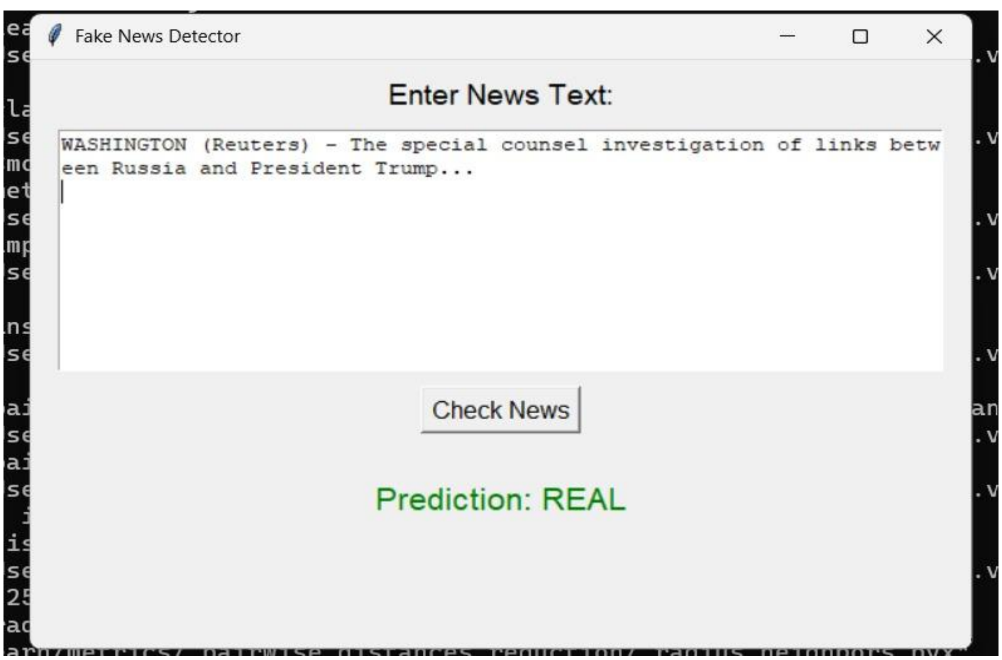


Fig A2.2: Real news Page

Fig A2.3: Fake news page

10 REFERENCES

1. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36
2. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797–806.
3. Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of ACL, 422–426.
4. Ahmed, H., Traoré, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.
5. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of COLING 2018, 3391–3401.
6. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of NAACL-HLT, 809–819.
7. Zhou, X., & Zafarani, R. (2020). Fake news: A survey of research, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40.
8. Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. Proceedings of LREC 2020.
9. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
10. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017).
11. Advances in Neural Information Processing Systems (NeurIPS),

FAKE NEWS DETECTION USING DEEP LEARNING AND TRANSFORMER-BASED MODEL

Paliwal Mohan Subhash, Deepa Gupta, Suja Palaniswamy, Manju Venugopalan
Department of Computer Science and Engineering,
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
mrpaliwal1997@gmail.com, g_deepa@blr.amrita.edu, p_suja@blr.amrita.edu, v_manju@blr.amrita.edu

2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10308352

***Abstract*— Fake news has a tremendous impact especially in certain fields like politics and economy in our society. The rise of social media usage to an extent is in favour of fake/false news promotion. A strong solution in this area would have numerous advantages for the society as a whole. Various deep learning techniques have been suggested by researchers to address this issue. But the majority of them don't identify false news with the requisite precision. In this paper, we propose a fake news identification model with state of art accuracy. Along with that, we compare different deep learning models in combination with different word embedding techniques for the task of fake news detection. Two types of word embeddings, Glove and Word2Vec were used for converting text to a vector of numbers. Along with the word embedding techniques we used seven different combinations of deep learning models: RNN, LSTM, Bi-directional LSTM, GRU, Bi-directional GRU, CNN-LSTM, and CNN-Bi-directional LSTM. The proposed work also implemented FastText and BERT models for embedding and classification of false news. We discovered that the proposed BERT classification model performed most effectively for detecting fake news on the experimented dataset. The proposed BERT model outperformed various baseline models on the benchmark dataset, attaining a test data accuracy of 99.20%, establishing its state-of-the-art performance. For the evaluation of models, we used accuracy, precision, recall, and F1-score metrics on the test data.**

Keywords—Natural Language Processing, Deep Learning, Binary Classification, Fake News Identification, Social Media News, Embedding.

I. INTRODUCTION

Social media facilitates the creation and sharing of information through computer-mediated technology. It has transformed how people communicate and relate to others. It makes a way for people to receive information quickly and easily. Nowadays, people utilise social networking sites more frequently than traditional news sources to search and read news. While news media has developed into a powerful instrument for educating others and bringing them together, it also has an adverse impact on the community[1]. There are numerous instances where social media has also facilitated the spread of false news. Social media site content could be of poor quality and purposely contains false information. This undermines the news's ability to be trusted. When there are billions of news items posted every day on the global web, how can one determine which ones are genuine and fake? Thus, one of several major problems in our networked era is false news. The past few years have seen a rise in the academic field of false news detection on social media. The proposed work would contribute to the delicate problem of halting the transmission of misinformation on social networks. Since the 2016 U.S. general political campaigns, fake news has been the subject of far too many discussions. However, it might be difficult for an individual to recognize fake news without professional help. Even reality services like PoliticFact.com or Snopes.com can just recognize fake news to a limited extent due to manual confirmation being time-consuming and hard. By classifying this news with the use of automated techniques, we should attempt to fix this issue.

In this study, we defined fake news identification as a prediction of news articles that have been purposefully falsified to deceive readers. In several studies, researchers integrate image information, news source, and opinions to capture the characteristics of fake news in addition to using the news body to identify it. In our paper, we simply use the article text to extract the linguistic content that characterizes the stylistic functionalities used in fake news. Various approaches [2-6] like machine learning models based on feature extraction, rhetorical structure theory-based models and DL models have been experimented on the fake news detection problem. A desired accuracy for classification is yet to be obtained and hence it remains an open problem.

In this paper, we proposed a model that grasps contextual information from text with an attention mechanism and identifies fake news. Also, we implement two categories of word embedding techniques with different combinations of deep learning models.

II. LITERATURE REVIEW

The survey mainly focuses on the prominent work in the field of false news identification which are majorly machine learning and deep learning approaches. Pooja Malhotra et al. [7] proposed an effective false news recognition system that is based upon Fastext. The suggested approach mainly uses FastText-MELMo embedding to grasp the semantic information of the input. To increase classification performance and cut down on training time, MELMo character type n-grams and Bidirectional-GRU have been used respectively. The AMSGrad based CNN classification algorithm, which successfully identifies the content as false or true news, is then fed a higher-dimensional embedding feature representation. The proposers claimed a classification accuracy of 96.54%.

Another ML approach [8] used a Kaggle dataset which included a data pre-processing step consisting of punctuation removal, stopword removal, and finally they extract features using the Keras embedding technique. They used the Naive Bayes classifier, SVM, and logistic regression for the detection of fake news. The proposers reported appreciable performance using the SVM classification model. Ashfia Jannat Keya et al. [9] developed an automated approach for describing false news that relies on deep learning and NLP for Bangla language using text of news and heading attributes. They suggested combining CNN-GRU with Glove embedding method, which had a test performance of 98.71%. They also tested the proposed model on a benchmark dataset of English media articles and got a performance of 98.94%. A comparative study of 10 fake news detection models based on [10] ML and DL models have been performed. Four design approaches (TF-IDF, count vector, character level vector, and N-Gram level vector) were utilized to extract features from documents. The outcomes demonstrated that fraudulent news with textual data may be identified, particularly when CNN are used. Using various classifications, this study's performance ranged from 81 to 100 percent. Models [11-13] using machine learning and language processing analysis for identifying false news in media articles were also experimented where feature vectors were generated using a variety of feature engineering techniques, including word embedding, TF-IDF, and count vector. In order to develop a model [14] to categorize news as fake or real, seven various machine learning categorization models was implemented. In this investigation, the SVM Linear classification technique with TF-IDF extracting features reported the best performance 94% accuracy results.

Xavier Jose et al. [15] proposed a model for false connection detection in online social media networks characterized by a stance detection model and the classifiers for manufactured content. For both models, they tested a variety of machine learning techniques, and the best ones were used to create the final product. The fabricated text classifier reported an accuracy of 0.93 using BiLSTM and the stance identification model based on Logistic Regression reported an accuracy of 0.9. Another comparative study [16] of different classification techniques with different word embedding techniques experimented one-hot, TF-IDF, word2vect, and Doc2Vec techniques. The classifiers experimented included ANN, CNN, and RNN classification techniques. Bi-LSTM with TF-IDF encoding reported a maximum accuracy of 96.81%. B. M. Amine et al. [17] proposed a combined DL model combining two CNNs with various metadata and utilize the word embedding method and CNN to retrieve text-based characteristics. The proposers demonstrated that fine-grained false news identification can significantly enhance when text and meta-data are combined. The study claimed a highest accuracy of 96%.

Pre-trained Glove embedding in combination with a combined architecture of RNN and CNN, has produced better performance in the detection of false news. This proposed model produced greater precision values of 97.21 percent when model takes more input features. Pritika and team [18] presented a model based on Bi-LSTM RNN network where the input was in the form of glove pre-trained word embeddings. For classification of news CNN, Vannila RNN, LSTM-RNN, and Bi-LSTM –RNN were experimented. The proposers reported highest accuracy of 91.48% on the Kaggle dataset and 98.63 for other news datasets employing the Uni-LSTM-RNN model.

With the increasing digitalization of our society, the propagation of false information is becoming more prevalent. This conduct can have detrimental consequences for the concerned individuals and organizations. Despite the growing number of studies aimed at detecting false news, there are still limitations like limited access to data, difficulty in identifying sources, rapidly evolving tactics etc. which leaves it as an ongoing research challenge.

III. Dataset Description

Few databases are publicly accessible because fake news identification is still a relatively new field. In our work, we conduct our experiments on the fake news dataset available on kaggle.com[1] . This dataset relates to the transmission of false information during the 2016 US Presidential Elections. It has roughly 20,800 instances labelled as two classes 10,387 with fake labels and 10413 with genuine labels. The dataset contains the following columns, Id: a news article's special identifier, title: a news article's headline, author: the news article's writer, text: the article's text and label: indicates whether a text is real or fake. Table.1 showcases the average word count in the title and the text content of the for both the labels respectively. In this data set, there are 1,26,633 unique words present in all corpus after pre-processing.

TABLE I. Average Number of Words In The Dataset

Labels	**Data Contents**	
	Title	***Text***
Fake	10.38	619.26
True	12.5	826.94

IV. Proposed Methodology

The proposed methodology proceeds in four stages: Data pre-processing, generate word embedding representations as features, classification using deep learning model, and model evaluation as depicted in Fig.1. Each phase is explained in detail in the following subsections.

A. Data Preprocessing:

The data is transformed into a cleaned sparse matrix for evaluation using NLP algorithms. The preparation processes to remove the content are listed below.

1. The raw text is separated from the less-valuable special characters, white spaces, and punctuation in the text.
2. The data is transformed to lowercase text once all special characters and symbols have been eliminated.
3. Elimination of stop words: Stop words are any terms that don't significantly further the meaning of the phrase in any language. Stop words include *a, and*, *of, was* etc. Although these terms are essential for human comprehension and translation, they do not independently add any sense to the statement in terms of NLP.
4. Stemming: Stemming is a normalization technique to reveal the basic form of a token. For example, the

[1] https://www.kaggle.com/datasets/samrat96/fake-news-detection

word *program* is the root of the programs, *programming, programmer.*

5. Tokenization is a method for breaking up raw text into manageable chunks. The plain text is segmented into tokens such as words and sentences, facilitating contextual comprehension and NLP model development.

B. Feature Extraction:

Following the elimination of stop words, tokenization, and textual stemming, the article's text and title are merged. The vocabulary size is considered when tokenizing, with a limit of 20,000 maximum frequent features words. The length of each numerical sequence is then padded or reduced to a maximum of 300, after which all textual sequences are transformed into numerical sequences using word embedding techniques.

1. Glove Embedding

Glove is an unsupervised learning approach which is utilized to ascertain the degree of similarity between words given their distance from one another in a feature space [20]. In the proposed work, we chose the 300-dimensional variant. The Embedding layer needs to be specified with the output shape set to 300. Finally, we set the trainable feature for the network to be False because we do not prefer to modify the learned word parameters in our system.

2. Word2Vec

Word2Vec provide space vectors for entire paragraphs directly by aggregating the space vectors for each word over the complete presentation of the corpus [21]. Stacking the vectors of each word produces a 2D framework for each phrase.

C. Classification using Deep learning models

A classifier model is trained using the training data representations generated after the feature extraction stage. The hyper-parameters are fine-tuned and finally the learned classifier is used to predict the labels of the test data instances. The experiment has been repeated with combinations of word-embeddings and the classifiers listed below.

1. RNN

RNN is a powerful tool designed to process sequential data by maintaining a state or memory of previous inputs. The values of the neurons in the hidden layer [22] of a RNN depend on both the outcome from the hidden layer's previous node and the current source.

2. LSTM

Building blocks for an RNN's layers are LSTM units. A cell, a forget gate, output gate and input gate make up an LSTM unit [23]. The cell oversees memorizing data more than a long period of time, which enables the relationship between the word at the beginning of the text to affect the word's response subsequently onward.

3. Bi-directional LSTM

Allowing a neural network have sequence information in both directions forward and backward is known as Bi-LSTM. Unlike a standard LSTM, the bidirectional variant processes input in both forward and backward directions [24] which enhances model performance on text classification problems.

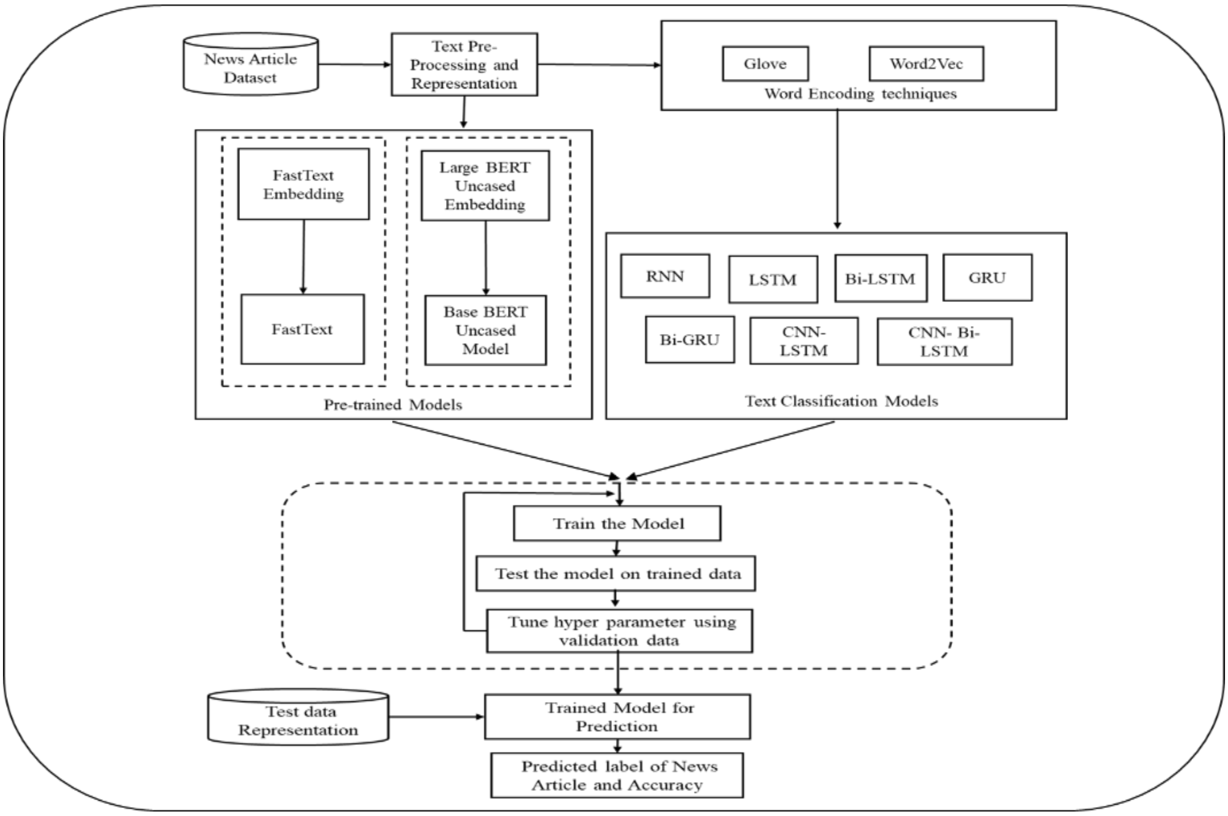


Fig.1. Proposed Framework for Fake News Detection

4. GRU

LSTM and GRUs have a lot in common. GRUs are quicker to train because of the architecture's simplicity. Unlike an LSTM cell's three gates, a GRU primarily consists of two gates. [25]. The first gate is the Reset gate, while the second is the Update gate.

5. Bi-directional GRU

A series processing paradigm with a bidirectional GRU [26] extends the standard GRU architecture by processing the input sequence in both forward and backward directions, using two separate GRU layers.

6. CNN-LSTM

The CNN-LSTM model incorporates CNN layers for feature extraction from input data which are combined with LSTMs to improve sequence prediction. Word vectors give CNN the ability to control the total text features and employ its cutting-edge parameter information exchange for computation minimization.

7. CNN-Bi-directional LSTM

This model's strength is its ability to use CNN convolution layers to extract the most data possible from texts. This output becomes the Bi-LSTM input, allowing the preservation of the data's sequential sequence in both ways.

8. FastText

FastText is a further word embedding and classification method that extends the word2vec method. Instead of explicitly learning word vectors, FastText expresses each phrase as an n-gram of letters [27]. This enables the embeddings to comprehend prefixes and suffixes and hence grasp the meaning of shorter words. It performs works well with rarely occurrence of words.

9. BERT

The bidirectional transformer is the main part of BERT's design allow it to be the first NLP technique to entirely function on self-attention mechanisms [28]. BERT is built on a transformer model, in which each input and output unit is linked and the weights between them are dynamically determined. For word embedding we used large-uncased BERT model and training the model for classification we used BERT base-uncased model with fine tuning of parameters.

Post data pre-processing, the word embedding models, Glove and Word2Vec are generated and experimented in combination with the listed seven deep learning classifiers. The various models are trained, hyper-parameters tuned and finally tested for performance on the test data as depicted in Fig.1. The best performing model is chosen as the proposed framework.

V. Results and Discussion

The proposed model has been evaluated on the Kaggle fake news dataset specified in Section 3. For training the model is split into train and test on an 80: 20 ratio. During model training 10-20 epochs are executed with batch size 64.

Two types of word embeddings Glove embedding and Word2Vec in combination with seven classifiers RNN, LSTM, Bi-LSTM, GRU, Bi-directional GRU, CNN-LSTM, and CNN-Bi-directional LSTM have been experimented on the dataset the results of which are reported in Table.2. In addition, FastText and BERT models have also been used for embedding and classification of fake news.

As observed in Table.2, Glove embeddings has performed well in combination with all classification models with F-score ranging from 95 to 97% except for RNN classifier where a low accuracy of 0.59 is reported. Word2vec in combination with various classifiers have reported lower F-measures ranging from 0.90 to 0.92 in comparison with Glove. A similar low performance is observed for RNN classifier in combination with word2vec embeddings reporting a low accuracy of 0.71. GRU performed good with all embedding techniques compared to other classification models. The results of FastText and BERT classifier models are presented in Table.3.

TABLE II. Model Performance For Different Word Embedding Techniques

Models	**Type of News**	**Glove Word Embedding**				**Word2Vec**			
		Accuracy	***Precision***	***Recall***	***F1-Score***	***Accuracy***	***Precision***	***Recall***	***F1-Score***
RNN	Fake	0.5900	0.67	0.44	0.53	0.7174	0.83	0.55	0.66
	Real		0.56	0.76	0.64		0.66	0.89	0.76
LSTM	Fake	0.9530	0.96	0.95	0.95	0.8976	0.88	0.92	0.90
	Real		0.95	0.95	0.95		0.91	0.88	0.90
Bi-LSTM	Fake	0.9438	0.94	0.96	0.95	0.9073	0.89	0.93	0.91
	Real		0.95	0.93	0.94		0.92	0.89	0.91
GRU	Fake	0.9689	0.98	0.96	0.97	0.9115	0.88	0.95	0.91
	Real		0.96	0.98	0.97		0.94	0.88	0.91
Bi-GRU	Fake	0.9652	0.97	0.96	0.97	0.9180	0.94	0.89	0.92
	Real		0.96	0.97	0.96		0.90	0.94	0.92
CNN-LSTM	Fake	0.9587	0.99	0.93	0.96	0.9063	0.91	0.90	0.91
	Real		0.93	0.99	0.96		0.90	0.91	0.91
CNN-Bi LSTM	Fake	0.9615	0.97	0.96	0.96	0.8994	0.87	0.93	0.90
	Real		0.96	0.96	0.96		0.93	0.87	0.90

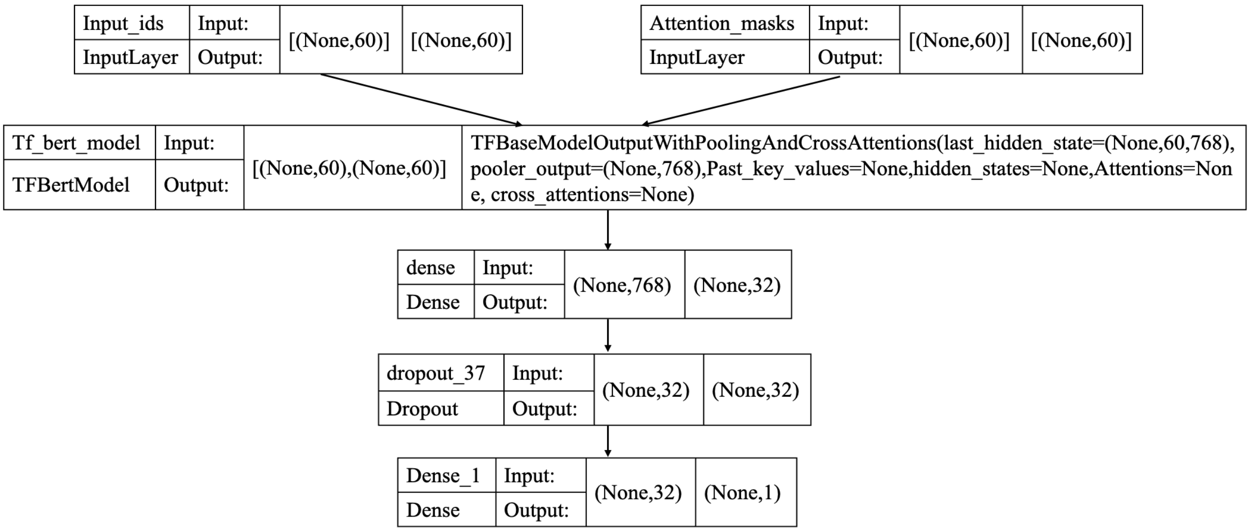


Fig. 2. Base-BERT Model Architecture for Classification

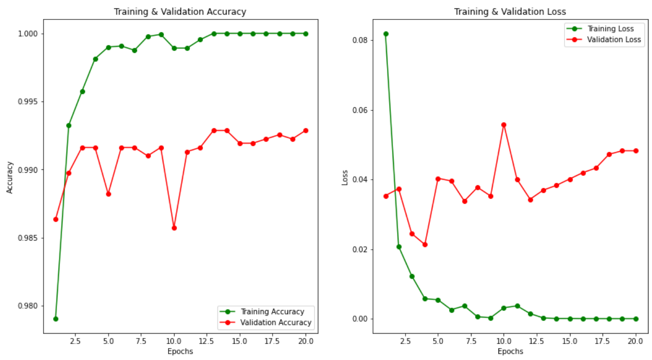


Fig. 3. Training and Validation Accuracy and loss function of BERT Model

TABLE III. MODEL PERFORMANCE FOR BERT AND FASTTEXT

Models	News Type	Accuracy	Precision	Recall	F1-Score
FastText	Fake	0.9622	0.96	0.97	0.96
	Real		0.96	0.96	0.96
BERT	Fake	**0.9920**	**0.99**	**0.99**	**0.99**
	Real		**0.99**	**0.99**	**0.99**

TABLE IV. PROPOSED MODEL COMPARISON WITH RECENT BASELINES

Models	Testing Accuracy
DNN [19]	88
Uni-LSTM-RNN [18]	91.4
LSTM [14]	94
CNN [17]	96
A2CNN [7]	96.54
CNN-GRU [9]	98.71
Bi-LSTM [16]	96.81
Proposed BERT classifier	**99.20**

FastText classifier in combination with FastText embeddings reported an F-measure of 0.96 as observed in Table 3. The BERT large uncased embedding model in combination with BERT base uncased classifier reported the maximum performance with a 99.20% testing accuracy and an F-measure of 0.99 and hence is projected as the proposed model for fake news detection.

The architecture for BERT classifier is showcased in Fig.2. As observed, the input layer is a sequence of token IDs with maximum length of each sentence set to 60 tokens. The self-attention mechanism computes a set of attention scores between each pair of word embeddings and allows the model to capture long-range dependencies. The Dense layer with softmax activation is used as the output layer for the classification task. The accuracy and the loss plots for training and test data have been presented in Fig.3.

The proposed model has also been compared with recent works which have reported results on the same dataset showcased in Table.4. As observed the proposed model has outperformed all the baselines in terms of the accuracy reported on the test data.

VI. CONCLUSION

In recent years, particularly in the political sphere, false news issues have grown significantly worse. In this study, we presented a model for false news detection using contextual information in text with attention mechanism. We proposed a model that gives state of art performance for identifying false news in textual data based on BERT embeddings and BERT classifier which reported an accuracy of 99.20% on the test data. This is highest accuracy on Kaggle fake news dataset as compared to previous published research papers. The experiment also compared various word embedding techniques in combination with deep learning classifiers. RNN classifier comparatively was low performing. For all classifiers experimented in combination with Glove and Word2vec, Glove embeddings produced better results and GRU, Bi-GRU gave best performances next to BERT classifier model. The model can be experimented on more challenging datasets as a future direction.

REFERENCES

[1] Akram, Waseem, and Rekesh Kumar. "A study on positive and negative effects of social media on society." International Journal of Computer Sciences and Engineering 5.10 (2017): 351-354.

[2] Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770 (2018).

[3] Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." ACM SIGKDD explorations newsletter 19.1 (2017): 22-36.

[4] Ganesh, Parvathy, Lekshmi Priya, and R. Nandakumar. "Fake news detection-a comparative study of advanced ensemble approaches." 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2021.

[5] Deepak, S., and Bhadrachalam Chitturi. "Deep neural approach to Fake-News identification." Procedia Computer Science 167 (2020): 2236-2243.

[6] Priyanga, V. T., Sanjanasri, J. P., Menon, V. K., Gopalakrishnan, E. A., & Soman, K. P. (2021). Exploring fake news identification using word and sentence embeddings. Journal of Intelligent & Fuzzy Systems, 41(5), 5441-5448.

[7] Malhotra, Pooja & Malik, Sanjay. (2022). An Efficient Fake News Identification System Using A-SQUARE CNN Algorithm. Wireless Personal Communications. 10.1007/s11277-022-09646-6.

[8] A. Vora and N. Shekokar, "Fake News Detection Using Intelligent Techniques," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-4, doi: 10.1109/ICRITO51393.2021.9596438.

[9] A. J. Keya, S. Afridi, A. S. Maria, S. S. Pinki, J. Ghosh and M. F. Mridha, "Fake News Detection Based on Deep Learning," 2021 International Conference on Science & Contemporary Technologies (ICSCT), 2021, pp. 1-6, doi: 10.1109/ICSCT53883.2021.9642565.

[10] A. Abdulrahman and M. Baykara, "Fake News Detection Using Machine Learning and Deep Learning Algorithms," 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020, pp. 18-23, doi: 10.1109/ICOASE51841.2020.9436605.

[11] Sivasangari, V., Pandian Vijay Anand, and R. Santhya. "A modern approach to identify the fake news using machine learning." Int. J. Pure Appl. Math 118.20 (2018): 10.

[12] S. R. R, M. R. L, D. S, A. D. P K and A. M. S, "Detection and Summarization of Honest Reviews Using Text Mining," 2022 8th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2022, pp. 01-05, doi: 10.1109/ICSSS54381.2022.9782167.

[13] Anushaya Prabha, T., et al. "An innovative and implementable approach for online fake news detection through machine learning." Journal of Computational and Theoretical Nanoscience 17.1 (2020): 130-135.

[14] N. Smitha and R. Bharath, "Performance Comparison of Machine Learning Classifiers for Fake News Detection," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 696-700, doi: 10.1109/ICIRCA48905.2020.9183072.

[15] X. Jose, S. D. M. Kumar and P. Chandran, "Characterization, Classification and Detection of Fake News in Online Social Media Networks," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), 2021, pp. 759-765, doi: 10.1109/MysuruCon52639.2021.9641517.

[16] L. Singh, "Fake News Detection: a comparison between available Deep Learning techniques in vector space," 2020 IEEE 4th Conference on Information & Communication Technology (CICT), 2020, pp. 1-4, doi: 10.1109/CICT51604.2020.9312099.

[17] B. M. Amine, A. Drif and S. Giordano, "Merging deep learning model for fake news detection," 2019 International Conference on Advanced Electrical Engineering (ICAEE), 2019, pp. 1-4, doi: 10.1109/ICAEE47123.2019.9015097.

[18] Bahad, Pritika & Saxena, Preeti & Kamal, Raj. (2020). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. Procedia Computer Science. 165. 74-82. 10.1016/j.procs.2020.01.0

[19] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica and A. N. Krishna, "Identification of Fake News Using Machine Learning," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198610.

[20] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, Oct. 2014, pp. 1532–1543.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of the International Conference on Learning Representations (ICLR), Scottsdale, Arizona, USA, 2013.

[22] M. Jordan and R. A. Jacobs, "A recurrent neural network for sequences," in IEEE Transactions on Neural Networks, vol. 1, no. 4, pp. 485-491, Dec. 1990, doi: 10.1109/72.80202.

[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[24] A. Graves and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition," in International Conference on Artificial Neural Networks (ICANN), 2005, pp. 796-804.

[25] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724-1734.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2014, pp. 2371-2379.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics (TACL), Volume 5, 2017, pp. 135-146.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 4171-418

DETECTION OF NEWS WHETHER ITS FAKE OR NOT USING DATA ANALYTICS

Vignesh Kumar R
Department ofComputer Science and Engineering, Panimalar Engineering College, Chennai, India.

Vimalraj M
Department ofComputer Science and Engineering, Panimalar Engineering College, Chennai, India.

Sathiya V
Department ofComputer Science and Engineering, Panimalar Engineering College, Chennai, India.

Subedha V
Department ofComputer Science and Engineering, Panimalar Engineering College, Chennai, India.

Sharmila S
Department ofComputer Science and Engineering, Panimalar Engineering College, Chennai, India.

Abstract: Detection of misinformation is the most pressing problem in the wake of digitization, whereby misrepresentations spread at a breakneck speed through social media and destabilize politics, economy, and society. In this paper, machine learning and data analytics methods are used to identify news stories as real or false with great precision. Particularly, Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks are investigated and experimented with. The method is split into multiple stages such as data collection, preprocessing, TF-IDF and word embeddings-based feature generation, and model training. Experimental results prove that the system, as proposed, obtains accuracy rates in excess of 90% in the efficient detection of fake news in automated detection. Through observation of language patterns, source credibility, and spread behaviors, the system minimizes the human fact-checking necessity to the barest minimum and enables verification procedures. In addition, the model is scalable and adaptable and can be employed with very large multilingual data sets, making it useful for practical systems. The significance of machine learning for maintaining information integrity, preventing false information, and building trust in electronic communication is emphasized in the article.

Keywords: Fake News Detection, Data Analytics, Machine Learning, Support Vector Machine, Logistic Regression, LSTM, Natural Language Processing, Information Integrity, Digital Communication

I. INTRODUCTION

The speed at which social networking sites and online media are increasing has changed the form of information creation, dissemination, and reception. This has also hastened the dissemination of disinformation and fake news that does tremendous harm to political stability, public confidence, and economic frameworks. It takes so much time and effort to use old manual fact-checking methods in fighting against the massive tide of falsehood on the internet. Hope is offered by artificial intelligence and data analytics software, which has the ability to flag automatically and determine with high precision if news is true or false.A machine learning framework is proposed in this current work that includes Logistic Regression, SVM, and LSTM models to identify false news with extremely high accuracy. The system has been subjected to processes such as data gathering, text pre-treatment, feature extraction through TF-IDF and word embeddings, and stable model training. Experimental results indicate good performance with a rate of accuracy higher than 90%, hence making effective and scalable solutions available to identify misinformation.

Flexibility in the model enables it to accommodate various datasets alongside various linguistic features such that it can be deployed locally and globally. Beyond classification, the system also extends to usability and effectiveness in real life through deployment in web-based and command-line interfaces. Through employing sophisticated algorithms in conjunction with content-based, user-based, and propagation-based studies, the system identifies more contextually and accurately. Moreover, ethical design and digital standard compliance provide data privacy and security, providing secure online data handling. Its accuracy, flexibility, and scalability provide the platform through which this method demonstrates the disruptive potential of machine learning to help eradicate fake news and provide information integrity in the web.

II. LITERATURE SURVEY

Artificial intelligence (AI) deployment and Sudden onset of disinformation have led to enormous amounts of research on applying machine learning (ML) and natural language processing (NLP) in detecting automated fake news. Existing works discuss various approaches, ranging from content analysis to user analysis and propagation models, towards achieving the highest accuracy and scalability in terms of challenges in disseminating disinformation.

Shu et al. suggested in 2017 "Fake News Detection on Social Media: A Data Mining Perspective", one of the earliest extensive reviews to define fake news detection as content-based, social context-based, and knowledge cue-based. Ruchansky et al. suggested CSI: A Hybrid Deep Model for Fake News Detection in 2017, which combined content features, user responses, and source behavior and performed better than content-only models. However, another of the highest-rated 2017 pieces by Wang brought forward the LIAR dataset (12.8K statements with fine-grained labels) in an effort to facilitate standardized benchmarking of deception detection work. Ahmed et al. (2018) employed traditional ML methods in hoax detection, fake news, and clickbait detection through feature engineering methods to obtain credibility indicators from text. Pérez-Rosas et al. in 2018 centered around linguistic characteristics and proposed new news factuality datasets, while Thorne et al. created FEVER, a large-scale fact- checking dataset, that improved fact-checking tasks.

In "Beyond News Contents: The Role of Social Context for FND," Shu et al. (2019) proposed combining textual and social signals and validated the efficacy of user and community-level signals based on social context.

Similarly, Yang et al. (2019) used propagation signals and user interaction to uncover the spread of false information. Zhou and Zafarani (2020) pointed to the early detection strategies, pointing out that the false news must be intercepted prior to their pervasiveness. Some of the other significant contributions are Sharma et al. (2019), who provided an exhaustive review of detecting and combating fake news strategies, and Oshikawa et al. (2020), who presented a survey of NLP pipelines and datasets for it identification. Shu et al. (2018) presented FakeNewsNet, a unified database covering both social context and content, for reproducible research.

At the same time, Tacchini et al. (2017) experimented with user-interaction features for Facebook hoax detection and Vosoughi et al. (2018) illustrated that false news spreads more than true news and thus induces the application of automatic detection tools.

According to recent studies, hybrid approaches—content-, action-, and dissemination-based—are more effective and accurate than single-source approaches. However, issues including domain adaptation, multilingual detection, dataset imbalance, and explainable models still exist. To combat real-time misinformation, future research must focus on creating scalable, intelligible systems with fact-checking capabilities and adaptability across communication interfaces.

III PROPOSED METHODOLOGY

Addressing the shor tcomings of the current methods for diagnosing breast cancer is the aim of the proposed approach. It uses state-of-the-art machine learning algorithms, such as Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), to efficiently diagnose cancers. The process is broken down into several steps to ensure reliability, efficacy, and clinical application.

A. Data Preprocessing

Preprocessing is doneon the medical dataset to eliminate duplicate rows and usemean imputation to handle missing variables. A StandardScaler is used to normalize the attributes for attributeuniformity. A binary label (Malignant = 1, Benign = 0) isused to encode the target variable diagnosis.

B.Feature Selection

Statistical correlationanalysis finds and eliminates highly correlated or redundant information to improve model performance. Onlythe tumor's size, texture, and compactness —the most instructive characteristics—are retained. The dimensionality reduction stage boosts classification accuracy and improves computational efficiency..

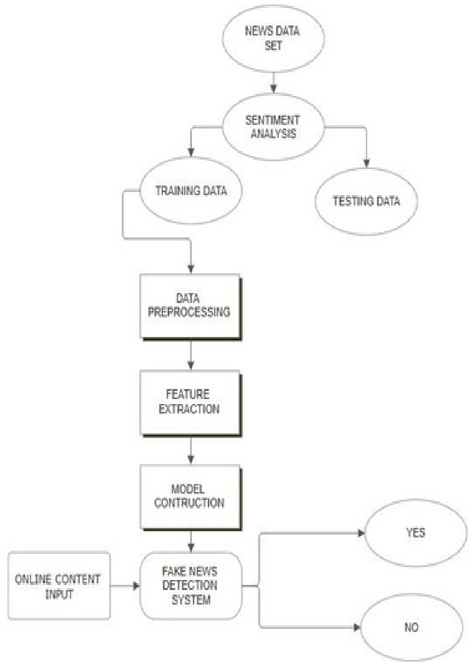
C. Model Training

• Support Vector Machine (SVM): This tool makes it simple to process high-dimensional medical data when used with both linear and Radial Basis Function (RBF) kernels.
• Extreme Gradient Boosting (XGBoost): Reduces mistakes by building decision trees in a sequential manner using a gradient-boosting ensemble technique.

Hyperparameters such as learning rate, tree depth, and estimators are tuned to perform optimally.

- This is preceded by a news dataset, which is initially subjected to sentiment analysis and split into a training set as well as a test set.
- The training set also goes through preprocessing to clean and normalize text.
- Secondly, context and linguistic features are identified in an effort to exactly specify the news content.

Machine learning models are developed and trained on these attributes in order to classify news as fake or real.The system is used with online material as input, applies it to the trained model, and gives an output based on whether the news is real



(Yes) or not (No).

Fig .1 The architecture diagram of proposed system

IV. DATA COLLECTION AND PREPROCESSING

The model is powered by a robust corpus of news articles collected from trusted sources, online archives, and factchecking sites. The corpus is processed beforehand in a methodical attempt to offer machine learning models pure, clean inputs. Preprocessing phase consists of some basic operations to clean the text data for effective fake news classification: Data Collection: News articles are webscraped from validated portals, fact-checking portals, and open-source datasets having fake and original samples of news.

• Text Sanitizing: Raw text sanitizing has been done by deleting punctuation marks, special characters, numbers, and duplicate symbols to achieve meaningful linguistic content only.
• Tokenization: To make word-level processing easier, sentences are tokenized into individual words, or tokens.
• Elimination of stopwords: Words that don't add anything to classification, such as is, the, of, and, are eliminated.
• Stemming and Lemmatization:

- Words are brought down to their base words (like running to run) to reduce vocabulary size and concentrate on the semantic meaning of the words.
- Feature Extraction: Methods like TF-IDF (Term Frequency–Inverse Document Frequency) and word embeddings are employed to transform text data into numbers that machine learning algorithms find convenient to work with.
- Dataset Splitting: The dataset is split into a training subset and a test subset (traditionally 80:20) to test the performance of the model on new data.
- Validation for Model Tuning: Cross-validation methods are employed to optimize hyperparameters of models like Logistic Regression, SVM, and LSTM to enhance their abilityto generalize.

V.DATAVISUALIZATION

Data visualization is a significant step in solving the acclimatization of the news dataset prior to training machine learning models. Visualization allows the insights into the distribution, trends, and patterns of the real and synthetic news and with what likelihood to be there are potential bias and imbalance. Visualization methods were utilized in this project to make one aware of the dataset:

Class Distribution: The proportion of true to false news content was illustrated by plotting using bar plots or pie plots in an effort to achieve class balance.

Word Frequency Analysis: To determine the most common words used in factual and misleading news, frequency plots and word clouds were made.

Distribution of News Length: Histograms compared the length of articles in the two classes to show whether fake news is longer or shorter than actual news.

TF-IDF Feature Insights: The majority of weighted words were shown as a word cloud, which helped identify key linguistic characteristics that differentiated authentic articles from fraudulent ones.

Sentiment Analysis: The distribution of sentiment polarity (positive, negative, and neutral) was analyzed to determine the emotional tone of authentic and fraudulent news.

Source/Publisher patterns: Credibility patterns were observed by comparing the frequency of news from different sources or publishers using bar plots.

These visual aids improved comprehension of the data, allowing for more informed preprocessing and improved model training.

The dataset includes "Fake" and "Real" classified news articles. For ease of analysis, such category names were assigned numbers like Fake = 0 and Real = 1. The number of each category was counted after numbering and presented in a bar chart.

This is a very accurate figure graph of the frequency distribution of fake and real news articles in the dataset. For our scenario, the train data happens to be unbalanced with more actual news than fake news. This kind of a situation needs to be taken into account as it can impact training the model as well as give rise to biased predictions towards the majority class. For addressing this, techniques like oversampling, undersampling, or class-weighting could be adopted at the model level.This is a very accurate figure graph of the frequency distribution of fake and real news articles in the dataset. For our scenario, the train data happens to be unbalanced with more actual news than fake news. This kind of a situation needs to be taken into account as it can impact training the model as well as give rise to biased predictions towards the majority class. For addressing this, techniques like oversampling, undersampling, or class-weighting could be adopted at the model level.

IV. MODULES EXPLANATION

The suggested fake news identification system is split into a sequence of modules, wherein each is tasked with undertaking one step of the process. Modularity helps scalability, flexibility, and integration. The major modules are explained below:

Data Collection Module

Itis the task of this module to retrieve news articles from various sources including online archives, fact-checking websites, and news agencies. Fake and real samples of news with diversity and reliability are given in the data set. Labeling is accurate to distinguish fake and real news, which serves as the basis for supervised learning.

Preprocessing Module

In this module, raw news text is preprocessed and cleaned up for processing. The process involves the removal of punctuation, special characters, stopwords, and then tokenization, stemming, and lemmatization. This eliminates only unwanted material and avoids noise in the dataset, leading to better model performance.

Feature Extraction Module

Preprocessed text is being mapped onto numerical features which are appropriate for machine learning models. TF-IDF and word embeddings are used to compute the linguistic importance and contextual importance of words. The module converts unstructured text into structured input for the classification model.

Machine Learning Module

his module employs the classification models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). These models are trained on the features extracted for the separation of fake news and real news. The hyperparameters are tuned for high precision, recall, and accuracy.

Evaluation Module

Here, performance of trained models is evaluated on performance metrics like accuracy, precision, recall, F1-score, and ROC curves. Not only does the evaluation module check whether the models are correct or incorrect, it also checks data robust and able to generalize to unseen novel news articles. Interface Module

The last module is a user-friendly interface by which users can provide online content or news articles to be verified. The system processes the input given by the trained model and gives the result as Real or Fake. The usability module is done for usability and potential deployment so that the system can be used by the general population, researchers, and journalists.

V. MODEL EVALUATION

Performance of the system designed to identify fake news was tested with various machine learning algorithms including Logistic Regression, Support Vector Machine (SVM), Random Forest, k-Nearest Neighbors (k-NN), and Long Short-Term Memory (LSTM). Training and testing datasets were divided into 80:20, and accuracy, precision, recall, and F1-score were used to assess each model's performance. These methods yielded a ballpark figure as to whether the models were able to identify false news reports and correct news reports. Baseline model was Logistic Regression and gave a readable but not highly detailed output. Its accuracy, though, was not quite as good as that of some other, more advanced algorithms because it found difficulty in picking up the subtle linguistic and contextual hints present within news reports. Support Vector Machine (SVM) performed better because it used TF-IDF features in separating fake and real news with a lot of commendable accuracy though its recall was weaker. That is, bad news is indicated as good, and that would be a fatal flaw in actual use because the bad news overlooked would be worse than a false positive. Random Forest surpassed Logistic Regression and SVM. With an ensemble of decision trees, it was capable of identifying non-linear relationships in data and produced very well-balanced outcomes in all the measures of assessment. However, the model required additional computation power and hyperparameter tuning for it to remain stable for different samples.

k-Nearest Neighbors (k-NN) algorithm also performed to very well, though the algorithm sensitivity the distancebased classification was lower when algorithm was being used on high-dimensional feature spaces such as those produced by TF-IDF and embeddings.

Of all the models compared, the greatest precision, recall, accuracy, and F1-score were provided by the Long Short-Term Memory (LSTM) networks. In contrast to all the previous models, LSTM was also able to preserve sequential word dependencies, allowing it to record better performance in detecting weak contextual signals to fake or misleading material. This aspect enabled LSTM to eliminate false negatives significantly, which is a very crucial aspect in real-world applications.

In spite of the additional training time and handling support from GPU hardware that it needed, its enhanced predictive capability weighed in favor of more computational cost. Relative comparison among the five models identified Logistic Regression as overall the worst performer, SVM with commendable but less precise recall, Random Forest and k-NN with balanced but average performance, and LSTM excelled in all except one measure. Whereas being as strong as ensemble methods such as Random Forest provided strength, sequential learning ability of LSTM rendered it the best model for detecting false news. Whereas at the expense of additional resources, its ability to reduce false negatives and enhance overall generalization makes it the best choice for practical uses.

VIII DATA PROCESSING

Processing data is a critical phase ofthemachinelearning cycle in the detection of the fake newssincetheprediction model's performance highly relies oninputdataquality.In the current project, news data gatheredfromsource repositories and fact-checking websiteswerepreprocessed in order to achieve homogeneity,trustability,and trainability for training classification modelslikeLogistic Regression, Support Vector Machine(SVM),Random Forest, and Long Short-Term Memory(LSTM).Rawdata generally had extraneous elementsintheformof punctuation marks, numbers, hyperlinks,andspecial characters, which were always removedleavingonly useful text content. Tokenization wasthenutilizedto divide sentences into separate words andcreateaword-level structured text data representation.

Stopwords like is, the, and, of, and that contribute minimal semantic value were removed to enhance the emphasis on information content words.

Besides the improvement in the quality of text, stemming and lemmatization were used to reduce words to their root form to handle variations like running, runs, and ran as a single feature.

Text in the preprocessed form was translated into numerical forms with the help of feature extraction techniques like TF- IDF and word embeddings preserving word frequency and contextual semantic. As the dataset had a skewed mix of false and real news, both stratified sampling and balancing methods were explored to ensure balanced representation of the two classes during training. The dataset was eventually divided into 80% (training) and 20% (testing) sets, with the training set being used for model training and the testing set being held out for unbiased testing. Cross-validation techniques were also utilized for hyperparameter optimization to avoid overfitting and enable the models to generalize to unseen data. 1X.

1X. FEATURE SELECTION

Feature selection is another aspect where fake news detection models fare well because it selects the most informative features and eliminates redundant or noisy features. Text data also tends to produce high-dimensional feature space, particularly when features such as TF-IDF are used, and not all features have the same level of importance for classification. Through the selection of informative features, the models are made more accurate, efficient, and interpretable. In disinformation identification, word frequency, sentence length, word style, and semantic drift are highly effective in reaching a conclusion on whether a news headline is factual or not. Feature selection is computationally efficient by eliminating repeated words and phrases and thus enhancing model generalization. For instance, some words or phrases disproportionately find their way into disinformation and their presence may offer useful predictability information. The advantages of feature selection are more accurate models, less overfitting hazard, faster training and testing, and better understanding of results. For example, classification models trained on a preprocessed feature set will find it simpler to separate between false and actual news since the noise of the irrelevant words is eliminated.

Filter techniques such as correlation analysis and chi-square were used to identify statistical significance of attributes.

Wrapper techniques employed machine learning approaches, but for comparing feature subset contribution by estimating their contribution in classification. Through feature selection in a selective way, the system only presents the most significant linguistic and contextual patterns for aiding fake news detection models, hence making them more reliable and sound.

X .LIMITATIONS

- Although the system proposed has better performance in identifying manipulative information with machine learning and natural language processing methods, there are certain limitations that need to be noted.
- Firstly, the research uses openly available datasets that, although bleak, may not be comparable to the richness of real news content by cultures, languages, and web sources.
- Such limitation will primarily impact the models' ability for generalizability when applied for multilingual or regional disinformation.
- Secondly, the size of the dataset, though good enough for training, is relatively modest given the richness and ever-fluctuating nature of online news.
- This limits the effectiveness of the trained models, particularly if they are subjected to newly evolving fake news structures or rapidly evolving fast-changing linguistic patterns.
- Thirdly, despite their high accuracy and superior contextual comprehension, deep models such as LSTM are computationally costly and require more training time, which limits their use in low-resource environments.

  Similarly, even if they are computationally light, traditional models like SVM and Logistic Regression have poor recall, which means that some bogus news stories will persist.

be classified as true. This is especially a worse problem since the un-detected counterfeit news will propagate very quickly and lead to massive misinformation. Another shortcoming is the use of only content-based features and not considering user-based or propagation-based features. The credibility of the user and sharing patterns are what influence the spread of the spurious news, and the inability of such social context to be infused by the system is a de- weakener of the performance. Lastly, the system is only applicable on static datasets rather than real-time detection subsystems which are imperative for real-time detection as well as false information prevention in dynamic social media systems. XI.EXPERIMENTAL RESULTS Fake news detection is among the most essential processes involved in maintaining online communication credibility. Experimental phase of the project involved classifying news articles into two categories: Fake (misinformation or false information) and Real (correct news content). Spam articles in the dataset were labeled as 0, and Real articles were labeled as 1. Binary labeling in such a manner was sufficient to train and test machine learning models.

XI.EXPERIMENTAL RESULTS

One of the most important tasks in guaranteeing the dependability of digital communication is detecting fake news. This project's experimental phase concentrated on dividing news stories into two groups: Real (real news content) and Fake (misleading or incorrect information). Fake articles were marked as 0 in the dataset, whereas real items were recorded as 1. Machine learning models might be trained and tested more effectively thanks to this binary representation.

Fig .2 The Simulation Results

To lessen bias, the combined dataset—which was created by combining fake and authentic news records—was preprocessed and balanced. The dataset was divided into training and testing subsets in an 80:20 ratio following cleaning and encoding. The testing set offered an objective assessment of classification performance, whereas the training set was used to fit the model

he models.

Fig .3 The Simulation Results

According to experimental data, conventional models like Support Vector Machine (SVM) and Logistic Regression (LR) performed well in terms of accuracy but had trouble with recall, which occasionally caused them to mistakenly classify phony articles as authentic. While deep learning techniques, especially Long Short-Term Memory (LSTM), offered the best overall accuracy, precision, and recall, Random Forest performed better by capturing non-linear interactions. By capturing sequential dependencies in text, LSTM was able to drastically lower false negatives and guarantee more accurate misinformation detection. The findings validate that a strong framework for automating the identification of fake news is offered by machine learning, particularly deep learning models. These models help fact-checkers, media outlets, and users efficiently combat misinformation by reducing human error and speeding up the verification process. Additionally, the system's flexibility enables it to manage big datasets and changing news content, which qualifies it for practical implementation in digital media settings.

XII. CONFUSION MATRIX FOR SVM

Support Vector Machine (SVM) model confusion matrix indicates its ability to accurately classify news articles as Real and Fake.

Correctly classified instances are the diagonal values of the matrix, i.e., the number of fake news articles correctly labeled as fake and real news articles correctly labeled as real. Misclassifications are represented by off-diagonal values. False negatives mean spurious reports of news being classified as actual news, and it is quite significant because this means that misinformation can be employed as actual news. False positives occur when good quality news reports are identified as spurious reports, and this can lead to a lack of trust in the system. The SVM model presented an excellent number of true positives and true negatives, which validated its efficacy.

Aslight decrease in recall, however, verifies that there were some undetected false stories, showing the requirement of more sophisticated models for better detection.

XII.CONFUSION MATRIX FOR XGBOOST

Confusion matrix of the Long Short-Term Memory (LSTM) model reveals that it detected false news more accurately.

The model is shown to have more actual positives and actual negatives, verifying its efficient ability to identify fake and real news articles. In comparison to conventional models, LSTM identifies sequence and context patterns in text, making it efficient for better classification. The matrix shows a drastically reduced false negative rate, i.e., reduction in the number of spurious articles tagged as real, which is lifeessential in preventing misinformation and dissemination. While higher training time computational power are needed, the high predictive power and accuracy of the model make the trade worth it. Empirical evidence confirms that LSTM is generalizable and accurate in predicting and therefore the best fit to be applied to real-world data to fight misinformation

XIV. CONLUSION

Results of this research confirm that natural language processing coupled with machine learning techniques can detect fake news. By auto-classifying news articles as fake and authentic, the system minimizes the human authentication requirement, gaining advantage from quicker, more uniform, and original results.

Deep learning algorithms quoted highest accuracy and recall bySVM, Random Forest, and LSTM models.

Itensures detection accuracy of misinformation, minimizing false negatives and raising credibility in electronic communication.

Briefly, the method demonstrated here stems from the function of AI-based systems in guaranteeing content verification and resistance against the common trend of fabricated news.

XV. FUTURE WORK

Although the system demonstrates greater predictability, there are some directions for future work that will lead to significant performance and applicability improvements. First, inclusion of multilingual false news detection would render it suitable for a vast universe of linguistic and cultural environments, thus making it suitable anywhere across the globe. Second, live interfacing with fact-checking tools and social media surveillance systems would enable real-time detection and prevention of spreading misinformation. Lastly, transitioning to deeper learning models, such as transformer models (e.g., BERT, RoBERTa), can have a vast accuracy and comprehension improvement, even more so an improved and robust system of a better nature to handle changing fake news tactics.

XVI. REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD, 2017.

[2] N. Ruchansky, S. Seo, and Y. Liu, “CSI: A hybrid deep model for fake news detection,” CIKM, 2017.

[3] W. Y. Wang, “Liar, liar pants on fire: A new benchmark dataset for fake news detection,” ACL, 2017.

[4] H. Ahmed, I. Traoré, and S. Saad, “Detecting opinion spams and fake news using text classification,” *Security and Privacy*, vol. 1, no. 1, 2018.

[5] A. Singh et al., “A review on fake news detection using machine learning techniques,” in *Proc. ICCSP*, pp. 568–573, 2019.

[6] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: Fact extractionand verification,” in *Proc. NAACL-HLT*, pp. 809–819,2018.

[7] K. Shu, S. Wang, and H. Liu, “Beyond news contents: The role of social context for fake news detection,” in *Proc. WSDM*, pp. 312–320, 2019.

[8] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” in *Proc. COLING*, pp. 3391–3401, 2018.

[9] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, “TI-CNN: Convolutional neural networks for fake news detection,” *arXiv preprint arXiv:1806.00749*, 2018.

[10]X. Zhou and R. Zafarani, “Fake news: A survey of research, detection methods, and opportunities,” *arXiv preprint arXiv:1812.00315*, 2020.

vignesh_REPORT_SHORT[1][1]

vignesh_REPORT_SHORT[1][1]

Turnitin

Document Details

Submission ID

trn:oid:::2945:321744414

Submission Date

Oct 26, 2025, 9:22 AM GMT+5

Download Date

Oct 26, 2025, 9:23 AM GMT+5

File Name

unknown_filename

File Size

488.8 KB

8 Pages

4,709 Words

27,317 Characters

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

71 Not Cited or Quoted 15%
Matches with neither in-text citation nor quotation marks

2 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
10% Publications
12% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **71** Not Cited or Quoted 15%
  Matches with neither in-text citation nor quotation marks
- **2** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
10% Publications
12% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	University College London on 2023-03-27	<1%
2	Internet	ijrpr.com	<1%
3	Internet	www.preprints.org	<1%
4	Internet	www.researchgate.net	<1%
5	Internet	www.ijritcc.org	<1%
6	Student papers	University of Hull on 2023-08-25	<1%
7	Student papers	University of Greenwich on 2025-08-29	<1%
8	Publication	Salve, Roshni Rajendra. "Fake News Detection: Leveraging Natural Language Pro...	<1%
9	Internet	www.mdpi.com	<1%
10	Publication	Afonso, Ricardo Oliveira. "Development of a Smartphone Application and Chrome...	<1%

11 Student papers

National College of Ireland on 2024-08-15 <1%

12 Internet

link.springer.com <1%

13 Student papers

Queen's University of Belfast on 2025-08-10 <1%

14 Publication

H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co... <1%

15 Publication

V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng... <1%

16 Internet

americaspg.com <1%

17 Internet

btw.media <1%

18 Internet

fjs.fudutsinma.edu.ng <1%

19 Publication

H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in He... <1%

20 Publication

Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain,... <1%

21 Student papers

University of Northumbria at Newcastle on 2024-01-15 <1%

22 Student papers

University of Queensland on 2025-06-11 <1%

23 Publication

"Progress in Artificial Intelligence", Springer Science and Business Media LLC, 2019 <1%

24 Publication

Devendra Prasad, Suresh Chand Gupta, Anju Bhandari Gandhi, Stuti Mehla, Upas... <1%

25 Student papers

University of Newcastle on 2025-05-12 <1%

26 Publication

Noureddine Seddari, Abdelouahid Derhab, Mohamed Belaoued, Waleed Halboob, ... <1%

27 Internet

arxiv.org <1%

28 Internet

dspace.dtu.ac.in:8080 <1%

29 Internet

vskp.vse.cz <1%

30 Student papers

Liverpool John Moores University on 2022-11-27 <1%

31 Publication

Rainer Greifeneder, Mariela E. Jaffé, Eryn J. Newman, Norbert Schwarz. "The Psyc... <1%

32 Student papers

Rochester Institute of Technology on 2018-12-07 <1%

33 Student papers

University of Hertfordshire on 2025-08-27 <1%

34 Internet

dokumen.pub <1%

35 Internet

ijsrem.com <1%

36 Internet

lutpub.lut.fi <1%

37 Internet

research.thea.ie <1%

38 Student papers

Liverpool John Moores University on 2023-02-28 <1%

39 Student papers

National College of Ireland on 2025-07-24 <1%

40 Publication

Nicollas R. de Oliveira, Dianne S. V. Medeiros, Diogo M. F. Mattos. "A Sensitive Styl... <1%

41 Student papers

Roehampton University on 2024-08-27 <1%

42 Publication

Sanjay Kumar, Akshi Kumar, Abhishek Mallik, Rishi Ranjan Singh. "OptNet-Fake: F... <1%

43 Internet

doctorpenguin.com <1%

44 Internet

repository.ihu.edu.gr <1%

45 Internet

www.researchsquare.com <1%

46 Publication

Al-Alshaqi, Mohammed. "Disinformation Classification Using Transformer Based ... <1%

47 Student papers

City University on 2018-12-20 <1%

48 Student papers

Dublin Business School on 2025-08-27 <1%

49 Student papers

Liverpool John Moores University on 2023-02-27 <1%

50 Publication

S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufactu... <1%

51 Student papers

Staffordshire University on 2023-05-22 <1%

52 Publication

"Natural Language Processing and Chinese Computing", Springer Science and Bu... <1%

53 Publication

Ajay Kumar, Sangeeta Rani, Krishna Dev Kumar, Manish Jain. "Handbook of AI in ... <1%

54 Student papers

Brunel University on 2024-09-11 <1%

55 Publication

Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathema... <1%

56 Student papers

Mar Athanasius College of Engineering on 2025-10-02 <1%

57 Publication

Suman Lata Tripathi, Om Prakash Kumar, Allwin Devaraj Stalin, Tanweer Ali. "Inn... <1%

58 Publication

Suneeta Satpathy, Álvaro Rocha, Sachi Nandan Mohanty, Tanupriya Choudhury. "... <1%

59 Student papers

University of Liverpool on 2024-04-16 <1%

60 Student papers

University of Newcastle on 2025-09-21 <1%

61 Student papers

University of Ulster on 2025-06-29 <1%

DETECTION OF NEWS WHETHER ITS FAKE OR NOT USING DATA ANALYSTICS

Vignesh Kumar R
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai, India.

Vimalraj M
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai, India.

Abstract: Detection of misinformation is the most pressing problem in the wake of digitization, whereby misrepresentations spread at a breakneck speed through social media and destabilize politics, economy, and society. In this paper, machine learning and data analytics methods are used to identify news stories as real or false with great precision. Particularly, Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks are investigated and experimented with. The method is split into multiple stages such as data collection, preprocessing, TF-IDF and word embeddings-based feature generation, and model training. Experimental results prove that the system, as proposed, obtains accuracy rates in excess of 90% in the efficient detection of fake news in automated detection. Through observation of language patterns, source credibility, and spread behaviors, the system minimizes the human fact-checking necessity to the barest minimum and enables verification procedures. In addition, the model is scalable and adaptable and can be employed with very large multilingual data sets, making it useful for practical systems. The significance of machine learning for maintaining information integrity, preventing false information, and building trust in electronic communication is emphasized in the article.

Keywords: Fake News Detection, Data Analytics, Machine Learning, Support Vector Machine, Logistic Regression, LSTM, Natural Language Processing, Information Integrity, Digital Communication

I. INTRODUCTION

The speed at which social networking sites and online media are increasing has changed the form of information creation, dissemination, and reception. This has also hastened the dissemination of disinformation and fake news that does tremendous harm to political stability, public confidence, and economic frameworks. It takes so much time and effort to use old manual fact-checking methods in fighting against the massive tide of falsehood on the internet. Hope is offered by artificial intelligence and data analytics software, which has the ability to flag automatically and determine with high precision if news is true or false. A machine learning framework is proposed in this current work that includes Logistic Regression, SVM, and LSTM models to identify false news with extremely high accuracy. The system has been subjected to processes such as data gathering, text pre-treatment, feature extraction through TF-IDF and word embeddings, and stable model training. Experimental results indicate good performance with a rate of accuracy higher than 90%, hence making effective and scalable solutions available to identify misinformation.

Flexibility in the model enables it to accommodate various datasets alongside various linguistic features such that it can be deployed locally and globally. Beyond classification, the system also extends to usability and effectiveness in real life through deployment in web-based and command-line interfaces. Through employing sophisticated algorithms in conjunction with content-based, user-based, and propagation-based studies, the system identifies more contextually and accurately. Moreover, ethical design and digital standard compliance provide data privacy and security, providing secure online data handling. Its accuracy, flexibility, and scalability provide the platform through which this method demonstrates the disruptive potential of machine learning to help eradicate fake news and provide information integrity in the web.

II. LITERATURE SURVEY

Artificial intelligence (AI) deployment and Sudden onset of disinformation have led to enormous amounts of research on applying machine learning (ML) and natural language processing (NLP) in detecting automated fake news. Existing works discuss various approaches, ranging from content analysis to user analysis and propagation models, towards achieving the highest accuracy and scalability in terms of challenges in disseminating disinformation.

Shu et al. suggested in 2017 "Fake News Detection on Social Media: A Data Mining Perspective", one of the earliest extensive reviews to define fake news detection as content-based, social context-based, and knowledge cue-based. Ruchansky et al. suggested CSI: A Hybrid Deep Model for Fake News Detection in 2017, which combined content features, user responses, and source behavior and performed better than content-only models. However, another of the highest-rated 2017 pieces by Wang brought forward the LIAR dataset (12.8K statements with fine-grained labels) in an effort to facilitate standardized benchmarking of deception detection work. Ahmed et al. (2018) employed traditional ML methods in hoax detection, fake news, and clickbait detection through feature engineering methods to obtain credibility indicators from text. Pérez-Rosas et al. in 2018 centered around linguistic characteristics and proposed new news factuality datasets, while Thorne et al. created FEVER, a large-scale fact- checking dataset, that improved fact-checking tasks.

In "Beyond News Contents: The Role of Social Context
60 for FND," Shu et al. (2019) proposed combining textual and social signals and validated the efficacy of user and community-level signals based on social context.

24 Similarly, Yang et al. (2019) used propagation signals and
38 user interaction to uncover the spread of false information.

Zhou and Zafarani (2020) pointed to the early detection strategies, pointing out that the false news must be intercepted prior to their pervasiveness. Some of the other significant contributions are Sharma et al. (2019), who
20 provided an exhaustive review of detecting and combating
26 fake news strategies, and Oshikawa et al. (2020), who presented a survey of NLP pipelines and datasets for it identification. Shu et al. (2018) presented FakeNewsNet, a unified database covering both social context and content, for reproducible research.

At the same time, Tacchini et al. (2017) experimented with user-interaction features for Facebook hoax detection and
31 Vosoughi et al. (2018) illustrated that false news spreads more than true news and thus induces the application of automatic detection tools.

According to recent studies, hybrid approaches—content-, action-, and dissemination-based—are more effective and accurate than single-source approaches. However, issues including domain adaptation, multilingual detection, dataset imbalance, and explainable models still exist. To combat real-time misinformation, future research must focus on creating scalable, intelligible systems with fact-checking capabilities and adaptability across communication interfaces.

III PROPOSED METHODOLOGY

Addressing the shor tcomings of the current methods for
20 diagnosing breast cancer is the aim of the proposed
16 approach. It uses state-of-the-art machine learning
15 algorithms, such as Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), to efficiently diagnose cancers. The process is broken down into several steps to ensure reliability, efficacy, and clinical application.

14 A. Data Preprocessing

Preprocessing is done on the medical dataset to eliminate duplicate rows and use mean imputation to handle missing variables. A Standard Scaler is used to normalize the attributes for attribute uniformity. A binary label (Malignant = 1, Benign = 0) is used to encode the target variable diagnosis.

B.Feature Selection

Statistical correlation analysis finds and eliminates highly correlated or redundant information to improve model performance. Only the tumor's size, texture, and compactness—the most instructive characteristics—are retained. The dimensionality reduction stage boosts classification accuracy and improves computational efficiency..

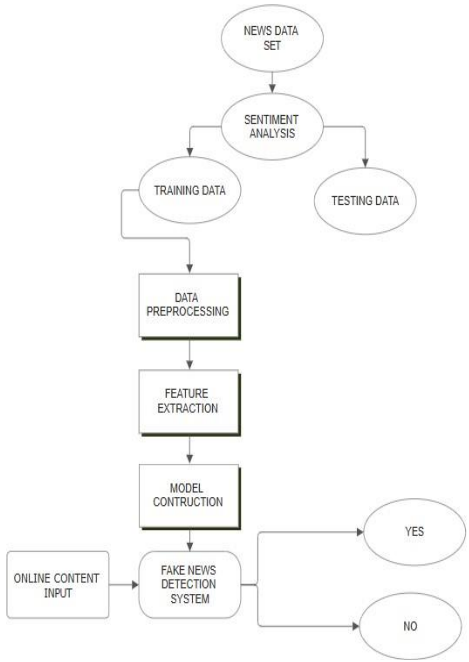
C. Model Training

• Support Vector Machine (SVM): This tool makes it simple to process high-dimensional medical data when used with both linear and Radial Basis Function (RBF) kernels.
• Extreme Gradient Boosting (XGBoost): Reduces mistakes by building decision trees in a sequential manner using a gradient-boosting ensemble technique.

Hyperparameters such as learning rate, tree depth, and estimators are tuned to perform optimally.

- This is preceded by a news dataset, which is initially subjected to sentiment analysis and split into a training set as well as a test set.
- The training set also goes through preprocessing to clean and normalize text.
- Secondly, context and linguistic features are identified in an effort to exactly specify the news content.

Machine learning models are developed and trained on these attributes in order to classify news as fake or real.The system is used with online material as input, applies it to the trained model, and gives an output based on whether the news is real



(Yes) or not (No).

Fig .1 The architecture diagram of proposed system

IV. DATA COLLECTION AND PREPROCESSING

The model is powered by a robust corpus of news articles collected from trusted sources, online archives, and factchecking sites. The corpus is processed beforehand in a methodical attempt to offer machine learning models pure, clean inputs. Preprocessing phase consists of some basic operations to clean the text data for effective fake news classification: Data Collection: News articles are webscraped from validated portals, fact-checking portals, and open-source datasets having fake and original samples of news.

- Text Sanitizing: Raw text sanitizing has been done by deleting punctuation marks, special characters, numbers, and duplicate symbols to achieve meaningful linguistic content only.
- Tokenization: To make word-level processing easier, sentences are tokenized into individual words, or tokens.
- Elimination of stopwords: Words that don't add anything to classification, such as is, the, of, and, are eliminated.
- Stemming and Lemmatization:

- Words are brought down to their base words (like running to run) to reduce vocabulary size and concentrate on the semantic meaning of the words.
- Feature Extraction: Methods like TF-IDF (Term Frequency–Inverse Document Frequency) and word embeddings are employed to transform text data into numbers that machine learning algorithms find convenient to work with.
- Dataset Splitting: The dataset is split into a training subset and a test subset (traditionally 80:20) to test the performance of the model on new data.
- Validation for Model Tuning: Cross-validation methods are employed to optimize hyperparameters of models like Logistic Regression, SVM, and LSTM to enhance their ability to generalize.

V.DATAVISUALIZATION

Data visualization is a significant step in solving the acclimatization of the news dataset prior to training machine learning models. Visualization allows the insights into the distribution, trends, and patterns of the real and synthetic news and with what likelihood to be there are potential bias and imbalance. Visualization methods were utilized in this project to make one aware of the dataset:

Class Distribution: The proportion of true to false news content was illustrated by plotting using bar plots or pie plots in an effort to achieve class balance.

Word Frequency Analysis: To determine the most common words used in factual and misleading news, frequency plots and word clouds were made.

Distribution of News Length: Histograms compared the length of articles in the two classes to show whether fake news is longer or shorter than actual news.

TF-IDF Feature Insights: The majority of weighted words were shown as a word cloud, which helped identify key linguistic characteristics that differentiated authentic articles from fraudulent ones.

Sentiment Analysis: The distribution of sentiment polarity (positive, negative, and neutral) was analyzed to determine the emotional tone of authentic and fraudulent news.

Source/Publisher patterns: Credibility patterns were observed by comparing the frequency of news from different sources or publishers using bar plots.

These visual aids improved comprehension of the data, allowing for more informed preprocessing and improved model training.

The dataset includes "Fake" and "Real" classified news articles. For ease of analysis, such category names were assigned numbers like Fake = 0 and Real = 1. The number of each category was counted after numbering and presented in a bar chart.

This is a very accurate figure graph of the frequency distribution of fake and real news articles in the dataset. For our scenario, the train data happens to be unbalanced with more actual news than fake news. This kind of a situation needs to be taken into account as it can impact training the model as well as give rise to biased predictions towards the majority class. For addressing this, techniques like oversampling, undersampling, or class-weighting could be adopted at the model level.This is a very accurate figure graph of the frequency distribution of fake and real news articles in the dataset. For our scenario, the train data happens to be unbalanced with more actual news than fake news. This kind of a situation needs to be taken into account as it can impact training the model as well as give rise to biased predictions towards the majority class. For addressing this, techniques like oversampling, undersampling, or class-weighting could be adopted at the model level.

IV. MODULES EXPLANATION

The suggested fake news identification system is split into a sequence of modules, wherein each is tasked with undertaking one step of the process. Modularity helps scalability, flexibility, and integration. The major modules are explained below:

Data Collection Module

It is the task of this module to retrieve news articles from various sources including online archives, fact-checking websites, and news agencies. Fake and real samples of news with diversity and reliability are given in the data set. Labeling is accurate to distinguish fake and real news, which serves as the basis for supervised learning.

Preprocessing Module

In this module, raw news text is preprocessed and cleaned up for processing. The process involves the removal of punctuation, special characters, stopwords, and then tokenization, stemming, and lemmatization. This eliminates only unwanted material and avoids noise in the dataset, leading to better model performance.

Feature Extraction Module

Preprocessed text is being mapped onto numerical features which are appropriate for machine learning models. TF-IDF and word embeddings are used to compute the linguistic importance and contextual importance of words. The module converts unstructured text into structured input for the classification model.

Machine Learning Module

his module employs the classification models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). These models are trained on the features extracted for the separation of fake news and real news. The hyperparameters are tuned for high precision, recall, and accuracy.

Evaluation Module

19 Here, performance of trained models is evaluated on performance metrics like accuracy, precision, recall, F1-score, and ROC curves. Not only does the evaluation
51 module check whether the models are correct or incorrect, it also checks data robust and able to generalize to unseen novel news articles. Interface Module

The last module is a user-friendly interface by which users can provide online content or news articles to be verified.
35 The system processes the input given by the trained model
53 and gives the result as Real or Fake. The usability module is done for usability and potential deployment so that the system can be used by the general population, researchers, and journalists.

13

V. MODEL EVALUATION

48 Performance of the system designed to identify fake news
61 was tested with various machine learning algorithms
9 including Logistic Regression, Support Vector Machine (SVM), Random Forest, k-Nearest Neighbors (k-NN), and Long Short-Term Memory (LSTM).

Training and testing datasets were divided into 80:20, and
18 accuracy, precision, recall, and F1-score were used to assess each model's performance.

8 These methods yielded a ballpark figure as to whether the models were able to identify false news reports and correct news reports.

Baseline model was Logistic Regression and gave a readable but not highly detailed output. Its accuracy,
33 though, was not quite as good as that of some other, more advanced algorithms because it found difficulty in picking up the subtle linguistic and contextual hints present within news reports. Support Vector Machine (SVM) performed
11 better because it used TF-IDF features in separating fake and real news with a lot of commendable accuracy though
2 its recall was weaker.

That is, bad news is indicated as good, and that would be a fatal flaw in actual use because the bad news overlooked
22 would be worse than a false positive. Random Forest surpassed Logistic Regression and SVM. With an
22 ensemble of decision trees, it was capable of identifying
2 non-linear relationships in data and produced very well-balanced outcomes in all the measures of assessment. However, the model required additional computation power and hyperparameter tuning for it to remain stable for different samples.

k-Nearest Neighbors (k-NN) algorithm also performed very well, though the algorithm sensitivity to distancebased classification was lower when the
25 algorithm was being used on high-dimensional feature spaces such as those produced by TF-IDF and embeddings.

56 Of all the models compared, the greatest precision, recall,
45 accuracy, and F1-score were provided by the Long Short-Term Memory (LSTM) networks. In contrast to all the previous models, LSTM was also able to preserve sequential word dependencies, allowing it to record better performance in detecting weak contextual signals to fake or misleading material. This aspect enabled LSTM to
15 eliminate false negatives significantly, which is a very crucial aspect in real-world applications.

In spite of the additional training time and handling support from GPU hardware that it needed, its enhanced predictive capability weighed in favor of more computational cost. Relative comparison among the five models identified Logistic Regression as overall the worst performer, SVM with commendable but less precise recall, Random Forest and k-NN with balanced but average performance, and LSTM excelled in all except one measure. Whereas being as strong as ensemble methods such as Random Forest provided strength, sequential learning ability of LSTM rendered it the best model for detecting false news. Whereas at the expense of additional resources, its ability to reduce false negatives and enhance overall generalization makes it the best choice for practical uses.

VIII DATA PROCESSING

Processing data is a critical phase of the machine learning cycle in the detection of the fake news since the prediction model's performance highly relies on input data quality. In the current project, news data gathered from source repositories and fact-checking websites were preprocessed in order to achieve homogeneity, trustability, and trainability for training classification models like Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). Raw data generally had extraneous elements in the form of punctuation marks, numbers, hyperlinks, and special characters, which were always removed leaving only useful text content. Tokenization was then utilized to divide sentences into separate words and create a word-level structured text data representation.

Stopwords like is, the, and, of, and that contribute minimal semantic value were removed to enhance the emphasis on information content words.

Besides the improvement in the quality of text, stemming and lemmatization were used to reduce words to their root form to handle variations like running, runs, and ran as a single feature.

Text in the preprocessed form was translated into numerical forms with the help of feature extraction techniques like TF- IDF and word embeddings preserving word frequency and contextual semantic. As the dataset had a skewed mix of false and real news, both stratified sampling and balancing methods were explored to ensure balanced representation of the two classes during training. The dataset was eventually divided into 80% (training) and 20% (testing) sets, with the training set being used for model training and the testing set being held out for unbiased testing. Cross-validation techniques were also utilized for hyperparameter optimization to avoid overfitting and enable the models to generalize to unseen data. 1X.

1X. FEATURE SELECTION

Feature selection is another aspect where fake news detection models fare well because it selects the most informative features and eliminates redundant or noisy features. Text data also tends to produce high-dimensional feature space, particularly when features such as TF-IDF are used, and not all features have the same level of importance for classification. Through the selection of informative features, the models are made more accurate, efficient, and interpretable. In disinformation identification, word frequency, sentence length, word

style, and semantic drift are highly effective in reaching a conclusion on whether a news headline is factual or not. Feature selection is computationally efficient by
28 eliminating repeated words and phrases and thus enhancing model generalization. For instance, some words
41 or phrases disproportionately find their way into disinformation and their presence may offer useful predictability information. The advantages of feature selection are more accurate models, less overfitting hazard, faster training and testing, and better understanding of results. For example, classification models trained on a preprocessed feature set will find it simpler to separate between false and actual news since the noise of the irrelevant words is eliminated.

Filter techniques such as correlation analysis and chi-square were used to identify statistical significance of attributes.

Wrapper techniques employed machine learning approaches, but for comparing feature subset contribution by estimating their contribution in classification. Through feature selection in a selective way, the system only presents the most significant linguistic and contextual patterns for aiding fake news detection models, hence making them more reliable and sound.

X .LIMITATIONS

- Although the system proposed has better performance in identifying manipulative
44 information with machine learning and natural language processing methods, there are certain limitations that need to be noted.

- Firstly, the research uses openly available datasets that, although bleak, may not be comparable to the richness of real news content by cultures, languages, and web sources.

- Such limitation will primarily impact the models' ability for generalizability when applied for multilingual or regional disinformation.

- Secondly, the size of the dataset, though good enough for training, is relatively modest given the richness and ever-fluctuating nature of online news.

- This limits the effectiveness of the trained models, particularly if they are subjected to newly evolving fake news structures or rapidly evolving fast-changing linguistic patterns.

- Thirdly, despite their high accuracy and superior contextual comprehension, deep models such as LSTM are computationally costly and require more
3 training time, which limits their use in low-resource environments.

7 Similarly, even if they are computationally light, traditional models like SVM and Logistic Regression have poor recall, which means that some bogus news stories will persist.

be classified as true. This is especially a worse problem since the un-detected counterfeit news will propagate very quickly and lead to massive misinformation. Another shortcoming is the use of only content-based features and not considering user-based or propagation-based features. The credibility of the user and sharing patterns are what influence the spread of the spurious news, and the inability of such social context to be infused by the system is a de- weakener of the performance. Lastly, the system is only applicable on static datasets rather than real-time detection subsystems which are imperative for real-time detection as well as false information prevention in dynamic social media systems. XI.EXPERIMENTAL RESULTS Fake news detection is among the most essential processes involved in maintaining online communication credibility. Experimental phase of the project involved classifying news articles into two categories: Fake (misinformation or false information) and Real (correct news content). Spam articles in the dataset were labeled as 0, and Real articles were labeled as 1. Binary labeling in such a manner was sufficient to train and test machine learning models.

XI.EXPERIMENTAL RESULTS

One of the most important tasks in guaranteeing the dependability of digital communication is detecting fake news. This project's experimental phase concentrated on dividing news stories into two groups: Real (real news content) and Fake (misleading or incorrect information). Fake articles were marked as 0 in the dataset, whereas real items were recorded as 1. Machine learning models might be trained and tested more effectively thanks to this binary representation.

Fig .2 The Simulation Results

To lessen bias, the combined dataset—which was created by combining fake and authentic news records—was preprocessed and balanced. The dataset was divided into training and testing subsets in an 80:20 ratio following cleaning and encoding. The testing set offered an objective assessment of classification performance, whereas the training set was used to fit the model

he models.

Fig .3 The Simulation Results

According to experimental data, conventional models like Support Vector Machine (SVM) and Logistic Regression (LR) performed well in terms of accuracy but had trouble with recall, which occasionally caused them to mistakenly classify phony articles as authentic. While deep learning techniques, especially Long Short-Term Memory (LSTM), offered the best overall accuracy, precision, and recall, Random Forest performed better by capturing non-linear interactions. By capturing sequential dependencies in text, LSTM was able to drastically lower false negatives and guarantee more accurate misinformation detection. The findings validate that a strong framework for automating the identification of fake news is offered by machine learning, particularly deep learning models. These models help fact-checkers, media outlets, and users efficiently combat misinformation by reducing human error and speeding up the verification process. Additionally, the system's flexibility enables it to manage big datasets and changing news content, which qualifies it for practical implementation in digital media settings.

XII. CONFUSION MATRIX FOR SVM

Support Vector Machine (SVM) model confusion matrix indicates its ability to accurately classify news articles as Real and Fake.

Correctly classified instances are the diagonal values of the matrix, i.e., the number of fake news articles correctly labeled as fake and real news articles correctly labeled as real. Misclassifications are represented by off-diagonal values. False negatives mean spurious reports of news being classified as actual news, and it is quite significant because this means that misinformation can be employed as actual news. False positives occur when good quality news reports are identified as spurious reports, and this can lead to a lack of trust in the system. The SVM model presented an excellent number of true positives and true negatives, which validated its efficacy.

A slight decrease in recall, however, verifies that there were some undetected false stories, showing the requirement of more sophisticated models for better detection.

XII.CONFUSION MATRIX FOR XGBOOST

Confusion matrix of the Long Short-Term Memory (LSTM) model reveals that it detected false news more accurately.

The model is shown to have more actual positives and actual negatives, verifying its efficient ability to identify fake and real news articles. In comparison to conventional models, LSTM identifies sequence and context patterns in text, making it efficient for better classification. The matrix shows a drastically reduced false negative rate, i.e., reduction in the number of spurious articles tagged as real, which is lifeessential in preventing misinformation dissemination. While higher training time and computational power are needed, the high predictive power and accuracy of the model make the trade worth it. Empirical evidence confirms that LSTM is generalizable and accurate in predicting and therefore the best fit to be applied to real-world data to fight misinformation

XIV. CONLUSION

Results of this research confirm that natural language processing coupled with machine learning techniques can detect fake news. By auto-classifying news articles as fake and authentic, the system minimizes the human authentication requirement, gaining advantage from quicker, more uniform, and original results.

Deep learning algorithms quoted highest accuracy and recall by SVM, Random Forest, and LSTM models.

It ensures detection accuracy of misinformation, minimizing false negatives and raising credibility in electronic communication.

Briefly, the method demonstrated here stems from the function of AI-based systems in guaranteeing content verification and resistance against the common trend of fabricated news.

XV. FUTURE WORK

Although the system demonstrates greater predictability, there are some directions for future work that will lead to significant performance and applicability improvements. First, inclusion of multilingual false news detection would render it suitable for a vast universe of linguistic and cultural environments, thus making it suitable anywhere across the globe. Second, live interfacing with fact-checking tools and social media surveillance systems would enable real-time detection and prevention of spreading misinformation. Lastly, transitioning to deeper learning models, such as transformer models (e.g., BERT, RoBERTa), can have a vast accuracy and comprehension improvement, even more so an improved and robust system of a better nature to handle changing fake news tactics.

XVI. REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD, 2017.

[2] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," CIKM, 2017.

[3] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," ACL, 2017.

[4] H. Ahmed, I. Traoré, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, 2018.

[5] A. Singh et al., "A review on fake news detection using machine learning techniques," in *Proc. ICCSP*, pp. 568–573, 2019.

[6] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: Fact extraction and verification," in *Proc. NAACL-HLT*, pp. 809–819, 2018.

[7] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. WSDM*, pp. 312–320, 2019.

[8] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *Proc. COLING*, pp. 3391–3401, 2018.

[9] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.

[10] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," *arXiv preprint arXiv:1812.00315*, 2020.