IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 4, May 2025

Impact Factor: 7.67

Fake News Detection Using Machine Learning with Cloud Deployment

Gayatri Dhumal, Dr. Sumit Hirve, Atharva Gaikwad, Rasika Mahakalkar , Aditya Midgule

Students, Department of Computer Science Engineering

Guide, Department of Computer Science Engineering

MIT-ADT University, Pune, India

Abstract: *The exponential growth of digital content has made online platforms the primary source of news consumption worldwide. However, this convenience has also led to the widespread dissemination of fake news, which poses significant threats to societal trust, political stability, and public safety. To address this challenge, this project presents a machine learning-based framework for automated fake news detection, seamlessly integrated with a cloud deployment infrastructure to ensure scalability, accessibility, and real-time performance.*

The proposed system employs advanced Natural Language Processing (NLP) techniques to analyze textual features and classify news articles as real or fake using supervised learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machines (SVM). The model is trained on benchmark datasets containing labeled news articles and fine-tuned to achieve optimal accuracy and generalization.

Keywords: Support Vector Machines (SVM), Logistic Regression

I. INTRODUCTION

In the digital age, the rapid proliferation of information through online platforms has revolutionized how people consume news. However, this revolution has come with a dark undercurrent — the alarming spread of **fake news**, which can distort public perception, influence elections, incite violence, and undermine trust in legitimate journalism. Traditional methods of content verification have proven insufficient in addressing the scale and speed at which misinformation propagates. Therefore, there is an urgent need for **automated, scalable, and intelligent systems** capable of detecting fake news with high accuracy and real-time efficiency.

This paper proposes an integrated framework that leverages machine learning algorithms for fake news detection, coupled with cloud-based deployment to ensure scalability, accessibility, and real-time processing. By harnessing the power of Natural Language Processing (NLP**)** and supervised learning models, the system can identify linguistic patterns and semantic cues commonly associated with deceptive content. The cloud deployment further enables continuous monitoring, distributed processing, and on-demand model updates, making it suitable for large-scale, real-world applications Unlike existing systems that are often limited to static, offline environments, this project emphasizes a dynamic, cloud-native architecture to deliver an end-to-end solution—from data ingestion and preprocessing to prediction and feedback loops. This not only improves performance and adaptability but also democratizes access by allowing integration with various news aggregation platforms, social media channels, and browser extensions.

Ultimately, the fusion of machine learning with cloud technology offers a promising pathway toward safeguarding digital information ecosystems, restoring public trust, and enhancing media literacy in an increasingly complex information landscape.

II. LITERATURE REVIEW

The problem of fake news detection has drawn significant attention from researchers in recent years, particularly with the growing influence of social media on public discourse. This section reviews key studies and technological

DOI: 10.48175/IJARSCT-26452

advancements in the domain of fake news detection, focusing on machine learning techniques, Natural Language Processing (NLP), and recent trends in cloud-based deployment.

Machine learning (ML) has become a cornerstone of fake news detection due to its ability to classify and learn from large datasets. Early studies by Rubin et al. (2016) and Conroy et al. (2015) explored the use of classical ML classifiers such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression to distinguish between real and fake articles based on linguistic and syntactic features. These models showed promising results but were limited in handling complex semantics.

Subsequent research introduced ensemble models and deep learning techniques to improve accuracy. Ahmed et al. (2018) employed Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to capture sequential dependencies in news content. Meanwhile, Shu et al. (2020) proposed hybrid frameworks that combine content-based and context-based features (e.g., user behavior and source credibility) for improved detection.

NLP is crucial in transforming unstructured textual data into machine-understandable formats. Common techniques include TF-IDF vectorization, n-gram modeling, lemmatization, and stop word removal. Recent advancements have shifted toward word embeddings like Word2Vec, GloVe, and FastText, which offer semantic understanding of text.

Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) have significantly enhanced fake news detection. BERT understands contextual meaning more effectively than traditional models, as demonstrated in studies by Vaswani et al. (2017) and Devlin et al. (2019). These models have set new benchmarks in text classification tasks, including misinformation detection.

While model accuracy is essential, practical deployment at scale is often overlooked in academic work. Cloud platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure offer infrastructure for deploying ML models in real-time environments. Zhou et al. (2019) emphasized the need for deployment pipelines that allow low-latency predictions, auto-scaling, and continuous model retraining.

Although significant progress has been made, challenges remain. Many studies focus on static, offline models with limited real-time capabilities. Few have integrated cloud deployment for continuous updates and real-world usability. Additionally, multilingual fake news detection and domain adaptation remain underexplored areas.

This project addresses these gaps by combining advanced ML techniques with real-time cloud deployment, offering a scalable, user-friendly system for fake news detection in dynamic environments.

III. METHODOLOGY

The methodology adopted for this project encompasses a complete pipeline from data collection to real-time cloud deployment, ensuring both technical robustness and practical applicability. The process begins with data collection from reliable and publicly available sources, primarily the Fake and Real News dataset from Kaggle and the LIAR dataset, which contains short political statements labeled with truth values. These datasets provide a mix of legitimate and deceptive content, including article headlines, full text, dates, and truth labels (real or fake), serving as the foundation for training and evaluating the model.

Once collected, the raw textual data undergoes thorough preprocessing to enhance quality and prepare it for machine learning. The text is converted to lowercase to maintain consistency, and common noise such as stop words, punctuation, and special characters is removed. Tokenization and lemmatization are then applied to break down sentences into meaningful words and reduce them to their root forms. This cleaned data is transformed into numerical vectors using techniques like Term Frequency–Inverse Document Frequency (TF-IDF) or advanced word embedding models such as Word2Vec or GloVe, depending on the model used.

The processed data is then used to train multiple machine learning algorithms, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Random Forest, and XGBoost. These models are trained using a standard 80-20 train-test split and evaluated using k-fold cross-validation to minimize the risk of overfitting and ensure generalization. Advanced deep learning models like Long Short-Term Memory (LSTM) networks or transformer-based architectures like BERT are also explored to capture deeper semantic meaning within the text. Each model's performance is assessed based on key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with the best-performing model selected for deployment.

Following model selection, the chosen model is serialized using tools like Pickle or Joblib and integrated into a cloud-based application via a RESTful API built using Flask. The entire application is containerized using Docker to ensure portability and consistent runtime behavior. The cloud deployment is handled through platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), or Microsoft Azure, which provide scalability, reliability, and real-time accessibility. The system is further enhanced with monitoring and logging capabilities to track performance and manage requests. An optional front-end interface or browser extension may also be developed to provide users with a user-friendly means of inputting news articles and receiving instant authenticity verification. This comprehensive methodology ensures that the fake news detection system is accurate, scalable, and practically deployable in real-world scenarios.

IV. SYSTEM DESIGN AND FEATURES

The system also includes a simple and intuitive user interface, either as a web-based dashboard or mobile-friendly page, where users can input news text or upload documents for verification.

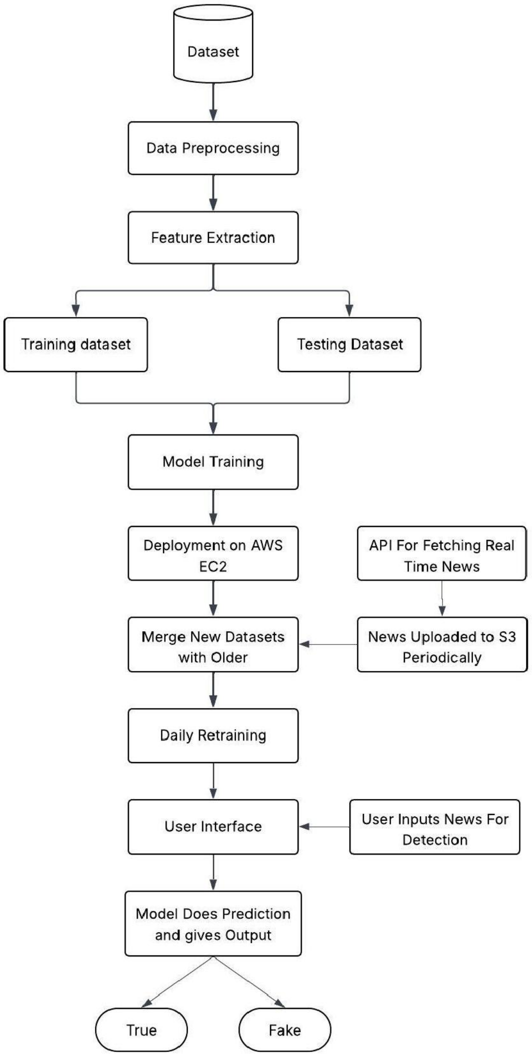

Dataset
Data Preprocessing
Feature Extraction
Training dataset
Testing Dataset
Model Training
Deployment on AWS EC2
API For Fetching Real Time News
Merge New Datasets with Older
News Uploaded to S3 Periodically
Daily Retraining
User Interface
User Inputs News For Detection
Model Does Prediction and gives Output
True
Fake

Fig. 1. Flowchart for Fake News Detection using Machine Learning with cloud deployment

DOI: 10.48175/IJARSCT-26452

For scalability and reliability, the entire system is containerized using Docker and deployed on cloud platforms like AWS EC2, Google Cloud Run, or Microsoft Azure. Load balancers and auto-scaling features ensure the system performs well under varying traffic conditions. Key features of the system include real-time detection, a responsive UI, cross-platform compatibility, logging and monitoring for prediction requests, and an easily updatable backend to allow future model improvements. Security features such as HTTPS communication and user input sanitization are also implemented to prevent misuse or system compromise.

Key Features of Fake News Detection:

- Automated Fake News Classification - The system automatically identifies whether a given news article is real or fake using machine learning models trained on real-world datasets. This reduces human effort and enables large-scale screening of digital content
- Natural Language Processing (NLP) - Based Analysis - The system uses NLP techniques such as tokenization, lemmatization, and vectorization (TF-IDF or word embeddings) to understand and process news content at a semantic level, improving classification accuracy.
- Train-Test Split for Model Validation - The dataset is divided into training and testing subsets, ensuring that the model is trained effectively and validated against unseen data for robust performance.
- High Accuracy and Reliability - The machine learning model is evaluated using key metrics like accuracy, precision, recall, and F1-score to ensure reliable performance in identifying misleading content.
- Scalable Cloud Deployment - The trained model is deployed to cloud platforms (such as AWS, GCP, or Azure), allowing for real-time, scalable access to the detection service through APIs or web interfaces.
- User-Friendly Interface - An optional user interface (web or mobile-based) is provided, enabling users to input news text or URLs and receive immediate classification feedback, promoting ease of use
- Real-Time Prediction - With cloud integration and optimized model inference, the system provides near real-time predictions, making it practical for use in content moderation tools or media platforms.
- Secure and Reliable Access - The API and web services are designed with secure protocols (HTTPS) and monitoring mechanisms to ensure safe, uninterrupted access and prevent abuse.

V. IMPLEMENTATION

The implementation of the proposed system is carried out in a series of well-defined phases, ensuring a structured approach from data collection to cloud deployment.

Phase 1: Data Acquisition

The project begins by sourcing data from public datasets available on platforms like Kaggle. These datasets contain labeled news articles categorized as "real" or "fake." The data is downloaded in CSV format and imported into the system using Python libraries such as Pandas. Initial analysis is performed to check for missing values, class imbalance, and basic statistics of the data.

Phase 2: Data Preprocessing

In this phase, raw news text is cleaned and standardized to make it suitable for machine learning. Techniques such as lowercasing, removal of stop words, punctuation, special characters, and HTML tags are applied. Lemmatization is performed to reduce words to their root form. This phase uses NLP libraries such as NLTK to prepare the dataset for vectorization.

Phase 3: Feature Extraction

Once preprocessing is complete, the text data is converted into numerical features using the TF-IDF (Term Frequency–Inverse Document Frequency) vectorizer. This method helps in identifying the most significant words in each document relative to the entire corpus. The result is a sparse matrix of features that can be input to a machine learning model.

DOI: 10.48175/IJARSCT-26452

Phase 4: Training Phase

In this phase, the preprocessed and vectorized dataset is used to train two classical machine learning algorithms Logistic Regression and Random Forest which are well-suited for binary classification tasks such as fake news detection. The dataset is first split into training and testing subsets, typically using an 80:20 ratio. This ensures that the models are trained on a significant portion of the data while being tested on unseen samples to evaluate their generalizability.

Phase 5: Testing and Evaluation

The trained model is evaluated using the testing dataset to assess its ability to generalize to unseen data. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to understand the effectiveness of the classifier.

Phase 6: Model Deployment

After achieving satisfactory results, the model is saved using Joblib or Pickle. A lightweight RESTful API is developed using Flask. This API serves as the interface between the user and the model, accepting user input and returning the classification result in real time.

Phase 7: Cloud Deployment

It is then deployed on cloud platforms like AWS EC2. This ensures that the fake news detection system is accessible globally and can handle real-time requests efficiently. Security measures such as HTTPS and input validation are added for safe operation.

VI. RESULT

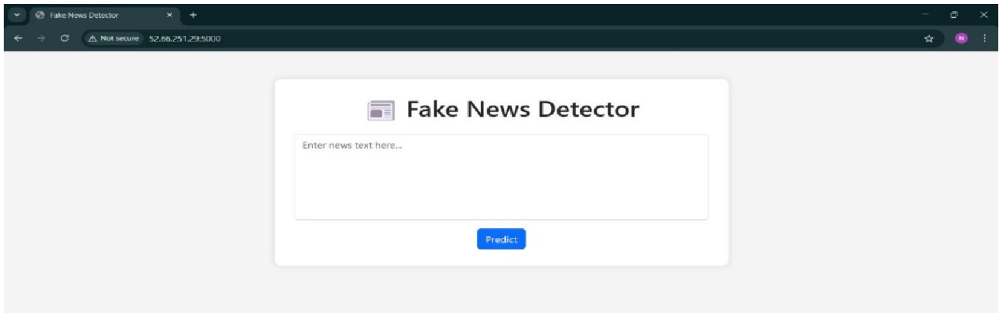
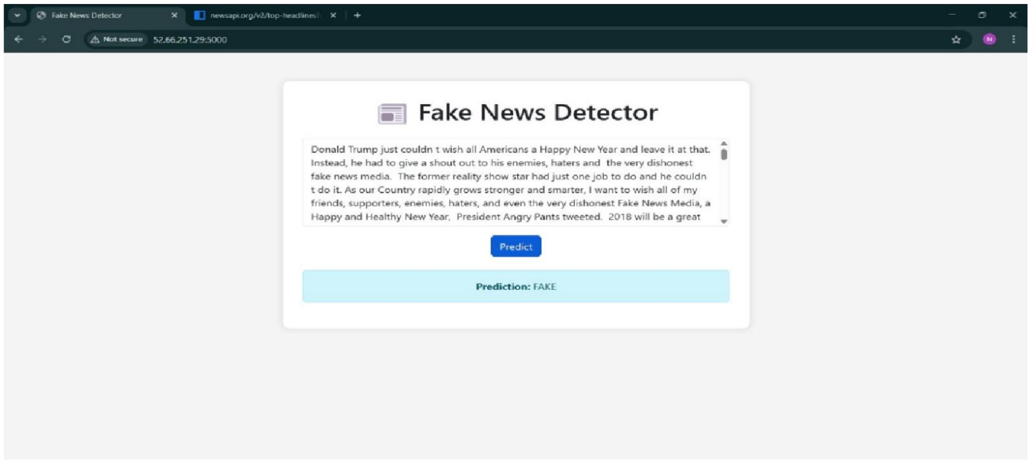


Fig. 1. User Interface

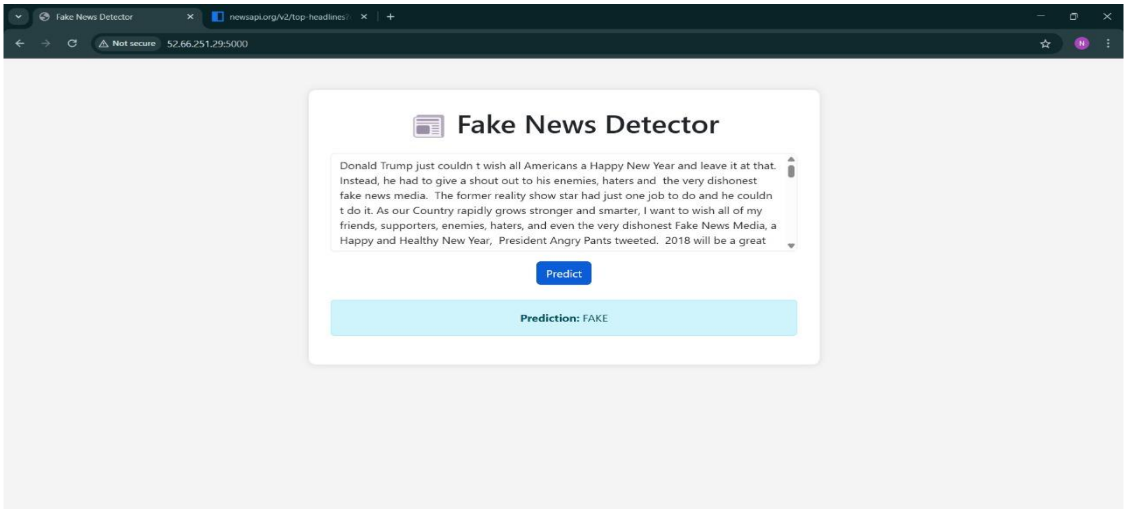


Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great

Prediction: FAKE

Fig. 2. Detects Real News Accurately

DOI: 10.48175/IJARSCT-26452



Fake News Detector

Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great

Prediction: FAKE

Fig.3. Detects Fake News Accurately

VII. CONCLUSION

In an era where misinformation spreads rapidly across digital platforms, the need for automated and accurate fake news detection systems has become critically important. This project successfully demonstrates the application of machine learning techniques specifically Logistic Regression and Random Forest classifiers to detect fake news with high accuracy and efficiency.

The system was developed through a systematic process involving data preprocessing, feature extraction, model training, evaluation, and cloud-based deployment. Both classifiers showed promising results, with Random Forest outperforming Logistic Regression in terms of accuracy and recall, making it well-suited for real-world use where missing fake news could have serious consequences.

By deploying the model via a Flask API and making it accessible through a cloud platform, the system offers scalability, accessibility, and real-time prediction capability. It highlights the practical viability of integrating machine learning and cloud technologies for solving real-world problems.

Despite certain challenges such as limited dataset scope and the difficulty in detecting nuanced language the system lays a solid foundation for future enhancements using deep learning, multilingual support, and real-time data streams. Overall, the project illustrates how machine learning, when combined with cloud computing, can serve as a powerful tool in the global fight against misinformation and fake news.

VIII. ACKNOWLEDGMENT

We would like to extend our heartfelt thanks to the faculty of MIT ADT University for their exceptional guidance, continuous support, and constructive feedback during the course of this Fake News Detection project. Their expertise and encouragement have played a vital role in the development and improvement of this project. We are especially appreciative of their insights into technical implementation, machine learning models, and the ethical implications of automated systems, which significantly contributed to the success of our work. Our sincere thanks to our mentors for their dedication, patience, and valuable input, which helped us address challenges and refine our approach. This project would not have been possible without their unwavering commitment to excellence in both innovation and education.

REFERENCES

[1]. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

[2]. Shu, K., Wang, S., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

[3]. Buntain, C., & Golbeck, J. (2017). Characterizing Fake News Users on Twitter. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 1-8. https://doi.org/10.1145/3123021.3123032

[4]. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 422-426. https://doi.org/10.18653/v1/P17-1051

[5]. Joulin, A., Grave, E., Mikolov, T., & Ranzato, M. A. (2017). Bag of Tricks for Efficient Text Classification. arXiv:1607.01759. https://arxiv.org/abs/1607.01759

[6]. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

[7]. Kwon, E., Cha, M., & Jung, K. (2017). A Study on the Role of Social Media in Fake News Detection. Proceedings of the International Conference on Big Data and Internet of Things, 185-192. https://doi.org/10.1109/BDIOT.2017.34

[8]. 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. https://doi.org/10.1145/3132847.3132885

[9]. Ahmed, A., Traore, I., & Saad, S. (2020). Fake News Detection on Social Media: A Data Mining Perspective. ACM Computing Surveys, 53(5), 1-35. https://doi.org/10.1145/3408381

[10]. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group.

[11]. Ruder, S. (2019). Transfer Learning for Natural Language Processing. O'Reilly Media.

[12]. Zhang, P., White, J., Schmidt, D. C. (2023). Blockchain for Secure Mental Health Data Sharing. IEEE Access (Q2).

[13]. Zhang, Y., & Wang, W. Y. (2020). Deep Learning for Fake News Detection. Springer.

[14]. How to Spot Fake News: 6 Steps". (2022). The New York Times. https://www.nytimes.com/guides/insights/how-to-spot-fake-news

[15]. The Battle Against Fake News in the Digital Age". (2021). The Guardian. https://www.theguardian.com/world/2021/feb/17/fake-news-and-the-battle-for-truth-in-the-digital-ageCummins, N., Scherer, S., Krajewski, J. (2024). Voice Analysis for Depression Screening. IEEE Transactions on Neural Systems and Rehabilitation Engineering (Q1).

[16]. Understanding the Impacts of Fake News on Public Health". (2023). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/fake-news

[17]. Fake News Dataset by George Washington University (2020). GWU Data Science Group. https://www.gwu.edu/fake-news-detection

[18]. FNSPID: Fake News Source Identification Dataset. (2022). Indian Institute of Technology Madras. https://fnsid.iitm.ac.in

[19]. Hugging Face Transformers. (2023). Hugging Face. https://huggingface.co/

[20]. PyTorch: An Open Source Machine Learning Framework. (2023). PyTorch. https://pytorch.org/

[21]. PyTorch: An Open Source Machine Learning Framework. (2023). PyTorch. https://pytorch.org/

[22]. Deep Learning for Fake News Detection using Text and Image Analysis" (2021). IEEE Access. https://ieeexplore.ieee.org/document/9001801

[23]. "An Overview of Fake News Detection Approaches and Methods" (2019). SpringerLink. https://link.springer.com/article/10.1007/s10207-019-00471-5

[24]. "A Review of Fake News Detection Using Machine Learning" (2022). IEEE Xplore. https://ieeexplore.ieee.org/document/9135358

DOI: 10.48175/IJARSCT-26452