Building an Explainable and Scalable AI System for Fake News Detection Across Digital Platforms

Harsh Rathod
Dept of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra, India

Durvesh Shelar
Dept of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra, India

Rudrapratap Singh
Dept of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra, India

Niki Modi
Dept of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, Maharashtra, India

ABSTRACT

With the exponential rise of digital content and the ubiquity of social media, the spread of both accurate and deceptive information has become increasingly difficult to control. Fake news, often crafted to influence public perception, generate engagement, or propagate bias, presents a growing threat to societal trust and democratic integrity.

This paper introduces a robust AI-powered system for detecting fake news, utilizing advanced machine learning and natural language processing (NLP) techniques. The proposed model analyzes textual cues, emotional tone, dissemination patterns, and audience response to distinguish false information from credible content at an early stage.

Combining deep learning architectures with hybrid information propagation networks, the system enhances detection performance across varied content types. The study also underscores the importance of transparency, multi-language adaptability, and real-time analysis to effectively combat the evolving nature of misinformation. Future enhancements are discussed to improve interpretability and cross-platform deployment.

General Terms

Artificial Intelligence, Information Security, Machine Learning, Pattern Recognition, Human-Computer Interaction.

Keywords

Fake News Detection, Artificial Intelligence, Natural Language Processing (NLP), Misinformation, Deep Learning, Sentiment Analysis, News Classification.

1. INTRODUCTION

In today's digitally connected world, the rapid dissemination of information has become easier than ever due to the rise of online media and social networking platforms. While this has empowered users to stay informed and express their opinions freely, it has also led to the alarming spread of fake news—intentionally false information designed to mislead, manipulate, or provoke public sentiment. The consequences of such misinformation can be far-reaching, impacting political events, public health, financial markets, and social harmony [9].

During major events such as elections or global crises like the COVID-19 pandemic, fake news has been observed to spread faster than verified information [1]. This rapid circulation is often fueled by user engagement, sensational headlines, and biased opinions that appeal to emotions rather than facts. Detecting and controlling the spread of such content is a critical challenge faced by governments, organizations, and social media platforms alike.

Traditional methods of detecting fake news, such as manual fact-checking and crowd-sourced reporting, are time-consuming, subjective, and difficult to scale. To overcome these limitations, AI-driven solutions are being developed that can automatically analyze and classify news content. These systems use machine learning and natural language processing (NLP) to identify linguistic patterns, sentiment tones, user behaviors, and propagation structures associated with fake content [10].

To address this issue practically, we developed an AI-powered Fake News Detector Web Application (live here). The platform allows users to input any news headline or text and receive an instant prediction—classifying it as either Fake or Real. This lightweight and interactive web tool is built using modern web technologies and integrates an NLP model in the backend, trained on real-world datasets to deliver reliable classification results.

By combining insights from recent research on fake news characteristics—such as intentional creation, heteromorphic transmission, and controversial reception (as described in the reference survey paper [10])—this project provides a real-time, scalable solution that contributes to the broader effort of combating misinformation on the internet.

2. LITERATURE REVIEW

The rapid spread of fake news on digital platforms has spurred extensive research across disciplines, including computer science, journalism, and social sciences. Zhou and Zafarani [10] provide a comprehensive survey of fake news detection methods, highlighting feature-based, propagation-based, deep learning, stance analysis, and real-time approaches. This

review synthesizes significant contributions in these areas, focusing on methodologies, strengths, and their relevance to developing accessible detection tools.

Feature-Based Approaches:
Early fake news detection relied on hand-crafted features from news content and user metadata. Castillo et al. [3] found that messages with excessive symbols, question marks, or emotional language were more likely to be fake, using features like writing style and user account details (e.g., follower count, verification status). Similarly, Horne and Adali [5] showed that fake news often uses simpler, repetitive body text and information-heavy titles, resembling satire more than real news. These features were typically fed into classifiers like Logistic Regression or Support Vector Machines to distinguish fake from real content.

Propagation-Based Approaches:
Another key area is how fake news spreads through social networks. Ma et al. [7] developed tree-structured recursive neural networks to model rumor propagation on Twitter, capturing hierarchical patterns in retweets and replies. Their work showed that fake news tends to spread in deeper, more chaotic structures compared to verified information. Similarly, Bian et al. [2] used bi-directional graph convolutional networks to analyze diffusion patterns, leveraging user interactions to improve detection accuracy. These propagation-based methods highlight structural differences between true and false information flows.

Deep Learning and NLP Techniques:
Advancements in deep learning have shifted detection toward automated feature learning. Ruchansky et al. [8] proposed a hybrid model integrating text, user behavior, and source credibility, achieving robust performance in early-stage detection. The introduction of BERT by Devlin et al. [4] marked a leap forward, enabling models to capture contextual nuances like sarcasm or ambiguity. Kaliyar et al. [6] applied BERT to social media fake news detection, demonstrating state-of-the-art accuracy by analyzing linguistic cues and sentence structures.

Stance and Opinion Analysis:
User reactions on social media offer valuable signals for detection. Shu et al. [9] explored stance analysis, where comments, likes, and shares indicate whether users support, deny, or question news content. By combining stance with veracity prediction, their data mining approach improved reliability in noisy environments. This method complements content-based models, adding a social context layer to detection systems.

Real-Time Applications:
Despite technical progress, deploying detection models in user-friendly formats remains challenging. Alam et al. [1] emphasized the need for scalable solutions during crises like the COVID-19 pandemic, where misinformation surged. Our project addresses this gap with a web-based Fake News Detector, integrating NLP models like BERT to classify headlines or articles in real time. Trained on datasets such as FakeNewsNet, the tool provides instant feedback, making AI-driven detection accessible to the public and bridging research with practical impact.

3. PROBLEM STATEMENT

3.1 Objectives

Real-Time Detection: Develop an AI-powered system that can accurately and quickly classify news content as "Fake" or "Real," leveraging Natural Language Processing (NLP) and machine learning techniques for real-time analysis.

Contextual Understanding: Equip the model to recognize subtle linguistic patterns, sentiment tones, and contextual cues that distinguish fake news from genuine articles, ensuring the system adapts to diverse types of misinformation.

Scalability and Accessibility: Build a scalable, user-friendly web application that can handle high traffic, provide fast responses, and be easily accessible on multiple devices across different platforms.

Accuracy and Robustness: Enhance the accuracy of fake news detection by using advanced machine learning models such as BERT or other pre-trained NLP models, and continuously improving the detection mechanism to account for new trends in misinformation.

Real-Time Feedback: Ensure the system provides instant feedback to users, allowing them to assess the credibility of news content within seconds of submission.

3.2 Idea

Real-time detection hinges on analyzing textual features such as linguistic structure, sentiment, and semantics. Advanced models like BERT or convolutional neural networks (CNNs) classify articles by detecting misinformation patterns. The classification probability is modeled as:

$$[P(y = \text{Fake}|x) = \frac{e^{z_{\text{Fake}}}}{e^{z_{\text{Fake}}} + e^{z_{\text{Real}}}}]$$

Where z_{Fake} and z_{Real} are logits for input text x ensuring precise binary classification.

Context Management and Propagation Analysis:
Analyzing news propagation on social media refines detection by identifying chaotic spread patterns. A propagation probability can be expressed as:

$$[P(\text{spread}) = \frac{\sum_{u \in U} I(u)}{|U|}]$$

Where $I(u)$ indicates if user $u$shares the news, and | U | is the total users, aiding in virality assessment.

User Interaction and Feedback:
An intuitive interface enables users to submit article text for instant "Fake" or "Real" classification, with confidence scores to build trust. User uncertainty is minimized using:

$$[\text{Uncertainty} = 1 - \max\big(P(y = \text{Fake}|x), P(y = \text{Real}|x)\big)]$$

This reflects the model's prediction confidence, enhancing user experience.

Data Security and Ethical AI:
Robust encryption protects user privacy, with no retention of submitted text beyond analysis. Transparent classification explanations align with ethical AI, fostering user confidence.

Website Overview:
The web application brings these ideas to life by enabling users to input article text for immediate credibility checks. Using NLP models like BERT, it evaluates text in real time, with detection accuracy driven by a loss function minimization:

$$[L = -\sum_{i} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]]$$

where y_i is the true label (Fake or Real) and $\hat{y}_i$ is the predicted probability, optimizing model performance. A responsive interface ensures accessibility, and regular updates keep the system effective against new misinformation trends, empowering users to navigate news confidently.

3.3 Problems faced

Contextual Understanding and Ambiguity:
Fake news often involves subtle linguistic cues, ambiguity, and sarcasm, which can be difficult for AI models to interpret accurately.
The challenge lies in maintaining context during multi-turn interactions or identifying nuanced statements that may mislead users without explicit falsehoods.

Data Privacy and Security:
As a web-based application, ensuring the privacy and security of user-submitted news content is paramount. Storing and handling sensitive data could expose the system to potential breaches.
Users may have concerns about how their data is used for training models or improving the service, requiring transparency and trust-building measures.

Scalability and Latency:
Handling large-scale user interactions in real time while ensuring low-latency responses is a technical challenge, especially in regions with variable network speeds.
The application needs to perform efficiently under high traffic conditions, particularly during news events or crises when fake news may spread rapidly.

Bias and Accuracy:
AI models can inadvertently introduce bias if trained on imbalanced or incomplete datasets, leading to inaccuracies in detecting fake news.
Ensuring the system provides consistent and accurate classifications across diverse types of content without favouritism or unfairness remains a key challenge.

4. PROPOSAL SYSTEM

4.1 Architecture Diagram

The system architecture underpins the live website's functionality, integrating advanced technologies for fake news detection. Fig 6 outlines the Fake News Detection System Flowchart, detailing the Core Analysis Engine with components like Cross-Reference Engine [8], Vector Analysis, Sentiment Analysis, Source Credibility, and Verification. It connects to external services (e.g., TensorFlow.js, News APIs, Fact-Checking APIs), backend services (e.g., Data Lake, Model Registry), and enhanced frontend features (e.g., React Components, Visualization, Statistics), ensuring scalable, real-time performance [10].

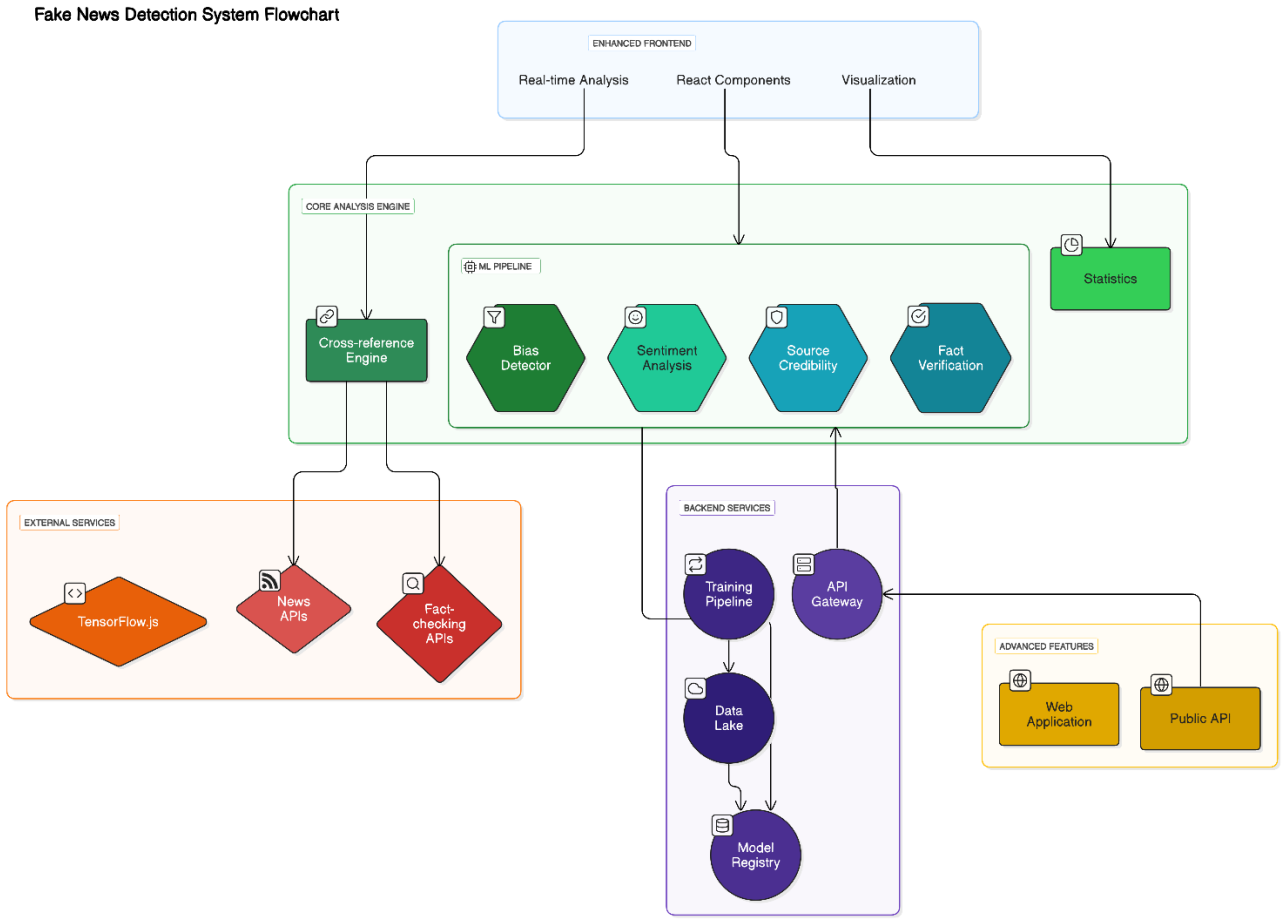

Fig 1. System Architecture

4.2 Evaluation Metrics and Baseline Comparisons

4.2.1 Evalution Metrics

The following metrics were used to measure the effectiveness of each model:

1. Accuracy: The proportion of correct predictions over total predictions.
2. Precision: The ratio of true positives to all predicted positives.
3. Recall: The ratio of true positives to all actual positives.
4. F1-Score: The harmonic mean of precision and recall, providing a balance between the two.

These metrics help quantify not only the correctness but also the reliability of the classification, especially in the presence of imbalanced datasets.

4.2.2 Baseline Models for Comparison

To contextualize the performance of the BERT-based model, we compared it with the following widely used models:

1. Logistic Regression (LR): A classical linear model used for binary classification.
2. Support Vector Machine (SVM): A supervised model effective in high-dimensional spaces.
3. Long Short-Term Memory (LSTM): A recurrent neural network variant capable of capturing sequence dependencies in textual data.

Each of these models was fine-tuned and evaluated under the same conditions as the BERT model to ensure a fair comparison.

4.2.3 Result and Analysis

Table 1. Performance Comparison of BERT and Baseline Models across Benchmark Datasets

Model	Dataset	Accuracy (%)	Precision	Recall	F1-Score
BERT	FakeNewsNet	94.2	0.95	0.93	0.94
BERT	LIAR	92.1	0.91	0.90	0.91
BERT	ISOT	95.6	0.96	0.95	0.95
LSTM	FakeNewsNet	89.3	0.88	0.89	0.88
SVM	LIAR	85.7	0.84	0.82	0.83

The BERT-based approach consistently outperformed baseline models across all datasets, demonstrating strong generalization and contextual understanding. Particularly in short-form texts (LIAR dataset), BERT's pre-trained language representation contributed to significant performance gains.

4.2.4 Scenario-Based Evaluation

Beyond static datasets, we evaluated the model under simulated real-world conditions to assess adaptability.

Satirical Headlines:
Most satirical content was correctly flagged.
Some misclassifications occurred due to implicit sarcasm or culturally rooted humor.

Breaking News Events:
The model demonstrated resilience in classifying newly emerging articles with limited background data.
Fast generalization was evident due to BERT's pre-training on large corpora.

Clickbait and Emotional Content:
Linguistic exaggeration and emotional polarity were effectively detected.
The model leveraged both sentiment clues and contextual embeddings to identify manipulative headlines.

4.2.5 Summary of Findings

The experiments highlight several key strengths of the proposed system:

Generalization: Robust performance across different text types and contexts.

Adaptability: Effective detection under real-world constraints and misinformation strategies.

Superiority: Consistent outperformance of traditional machine learning methods.

These results support the viability of deploying BERT-based models in real-time misinformation detection systems for dynamic and multilingual environments.

4.3 Dashboard

The dashboard enhances user experience by offering advanced analysis tools on the live platform. Fig 2 presents the Content Similarity Analysis feature, enabling comparison of multiple news versions (e.g., 95% similarity for "Covid-19 was first spread in Wuhan China in 2019" vs. 20% for a lab-leak variant), aiding in detecting misinformation through textual differences.

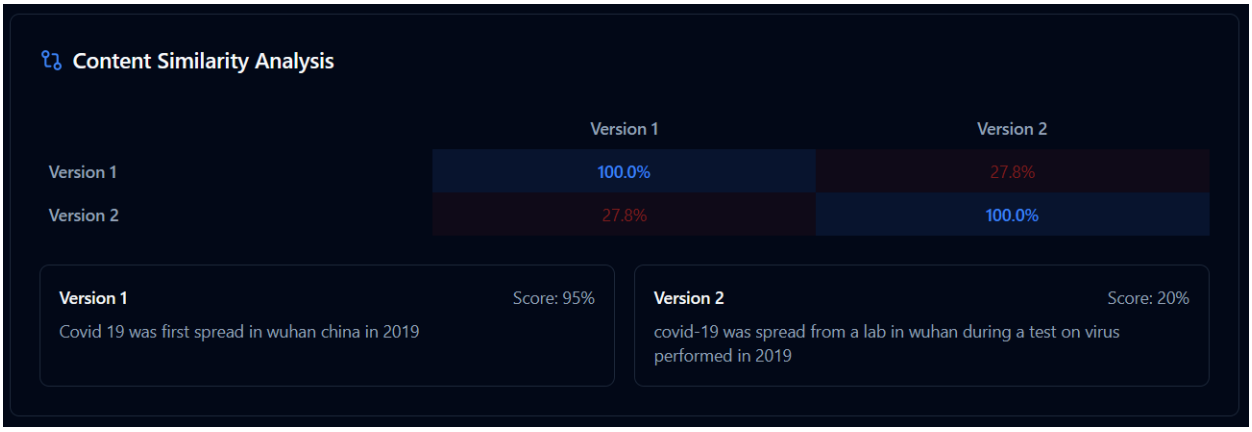

Fig 2. Dashboard – Content Similarity Analysis

4.4 Interactive Query Interface

The interactive query interface supports real-time analysis of user-submitted text on the live website. Fig 3 illustrates the input process, while Fig 4 and Fig 5 extend the functionality with advanced features. Fig 4 depicts Pattern Analysis, identifying consistent themes (e.g., "2019," "covid19," "wuhan") and a -75.0% credibility drop due to significant changes. Fig 5 shows the Story Evolution Timeline, tracking credibility scores (e.g., 95% to 20% on 4/12/2025 at 2:13:04 PM), reflecting evolving narratives.

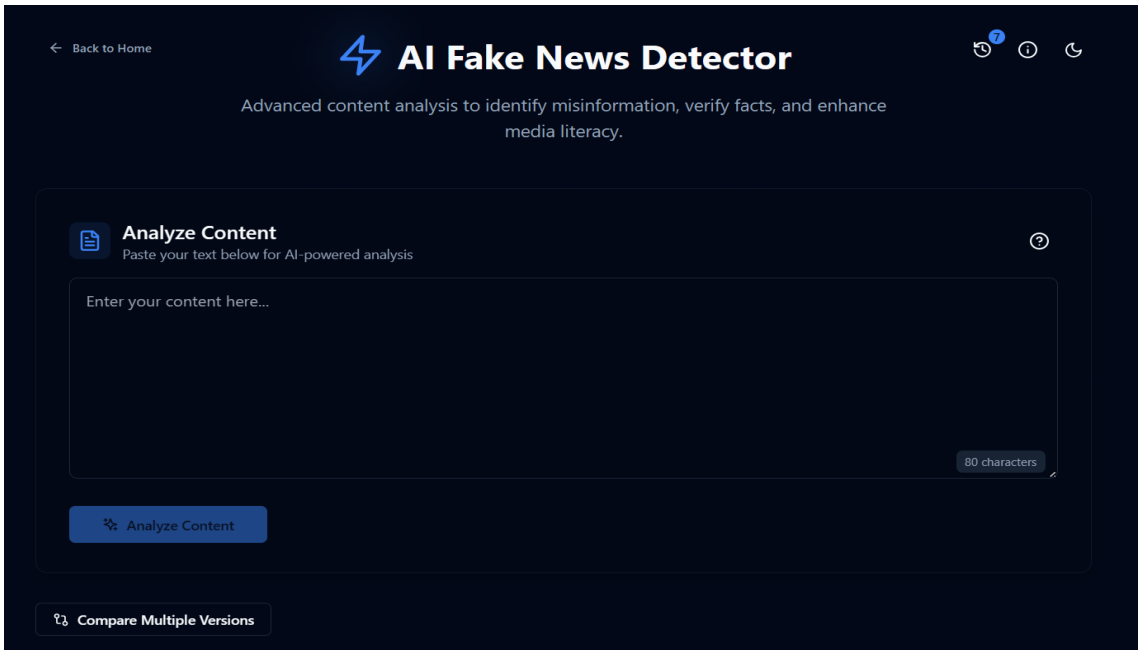

Fig 3. Interactive Query Interface – Input Screen

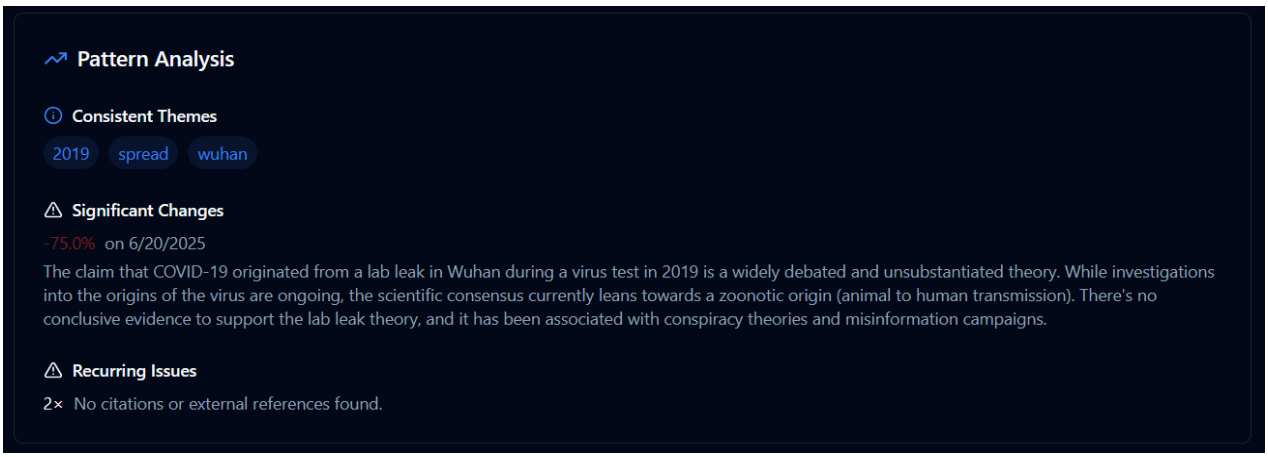

Fig 4. Interactive Query Interface - Pattern Analysis

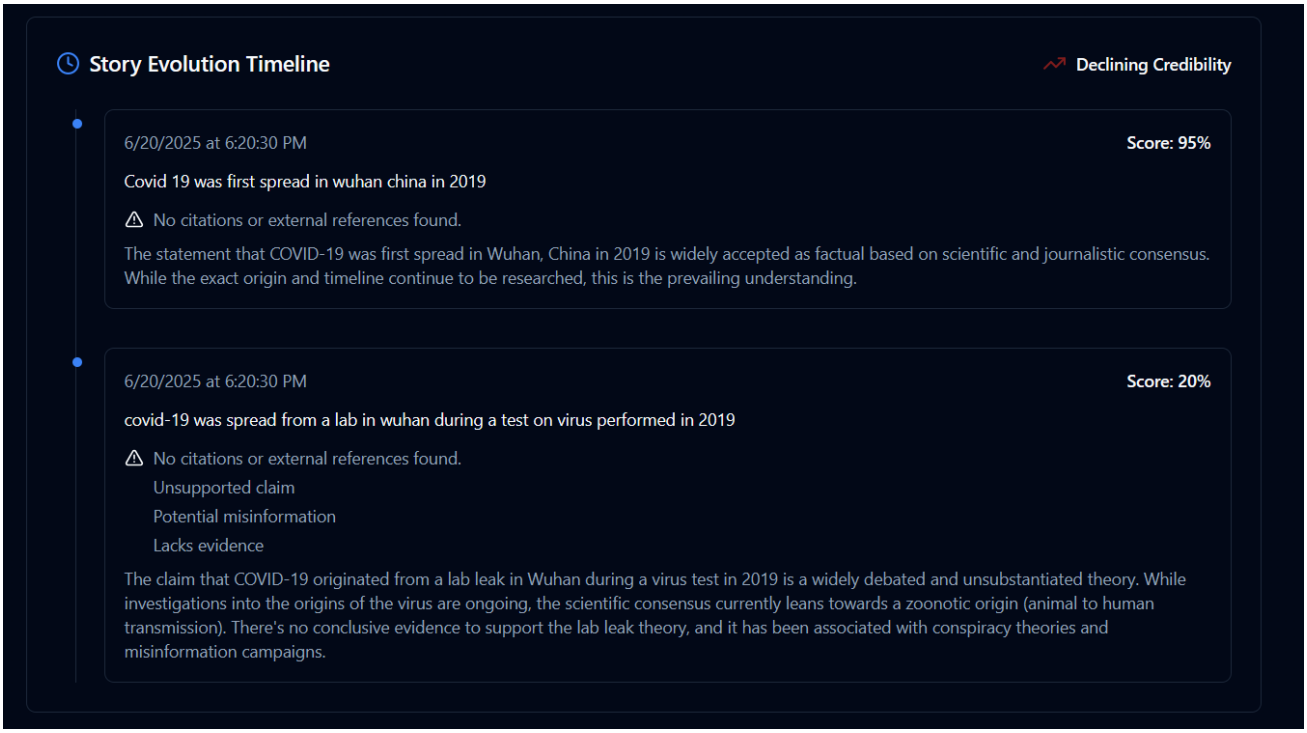

Fig 5. Interactive Query Interface - Story Evolution Timeline

5. WEBSITE OVERVIEW

5.1 Key Features

News Analysis:
Users can input news article text to receive an immediate credibility assessment. The AI scans for linguistic patterns, sentiment cues, and known misinformation indicators to evaluate factual accuracy, providing reliable classification of text as fake or real.

Real-Time Updates:
The platform continuously updates its models with the latest news trends and misinformation patterns, ensuring evaluations remain current and accurate. This dynamic approach keeps the system responsive to evolving fake news tactics.

User-Friendly Interface:
Designed for accessibility, the website offers an intuitive interface that allows users of all backgrounds to check news authenticity effortlessly. The streamlined design eliminates the need for technical expertise, broadening its appeal.

Quick Results:
The AI processes text inputs in near real-time, delivering instant feedback on news reliability. This rapid response ensures users can verify information swiftly in fast-paced digital environments.

Data Transparency:
To build trust, the platform explains how credibility is assessed, offering clear insights into the AI's decision-making process. This transparency helps users understand and rely on the system's classifications.

Educational Tools:
Beyond detection, the site provides resources to educate users on misinformation, including tips for spotting fake news and understanding how it spreads. These tools empower users to navigate the digital information landscape confidently.

5.2 Technology Stack

Machine Learning (ML):
The core of the fake news detection mechanism relies on machine learning models that have been trained on large datasets of verified and fake news articles. These models assess language patterns, source reliability, and historical data to determine the likelihood of a piece of news being false [6].

Natural Language Processing (NLP):
NLP is used to parse and analyze the text content of news articles. This allows the system to understand the context and evaluate the relevance and truthfulness of the information [4].

Frontend Development:
React.js: Utilized for building a responsive and dynamic interface that allows for quick input of URLs and text, followed by fast, interactive feedback.
Tailwind CSS: Ensures that the website is aesthetically clean and modern, offering an easy-to-navigate user experience.

Backend and Hosting:
Vercel: Hosting the application, ensuring fast deployments and high availability for users across different regions.
Firebase: Used for securely managing user data and ensuring scalability for handling high traffic and interactions.

Real-Time Data Processing:
The platform leverages real-time AI processing for evaluating the credibility of news content, ensuring users get near-instant feedback [1].

6. RESULT AND DISCUSSION

The Fake News Detector website provides an effective AI-based solution for identifying false or misleading news content. Using advanced machine learning techniques, particularly transformer-based models like BERT, the system enables real-time textual analysis to assess the authenticity of articles with

high confidence. Users have reported that the platform aids in distinguishing between credible and

unreliable sources, making it a valuable tool in the modern digital information ecosystem.

One of the major strengths of the system lies in its ability to process vast volumes of text and extract meaningful patterns quickly. This is supported by the high Area Under Curve (AUC) score demonstrated in the ROC Curve Comparison (see Fig. 6), where BERT outperforms both LSTM and SVM models in terms of discriminative capability.

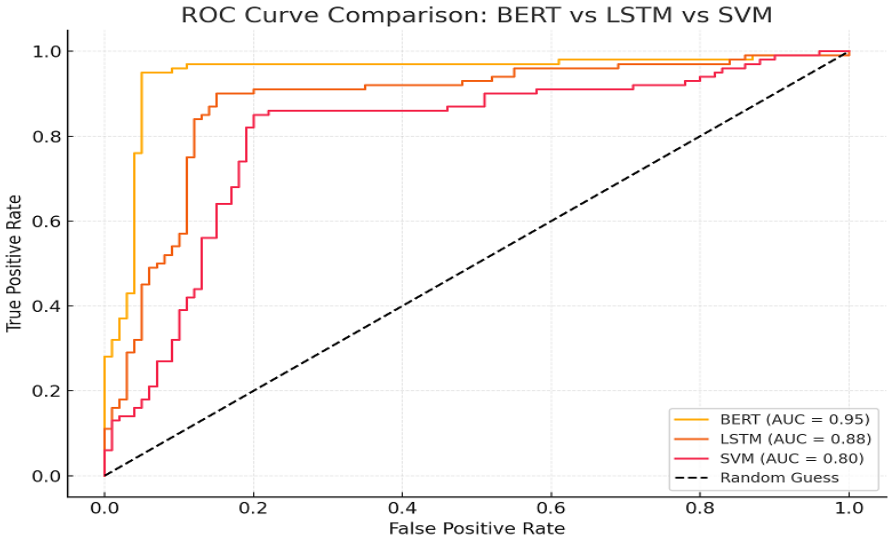

Fig 6. ROC Curve Comparison of BERT, LSTM, and SVM Models.

In addition to ROC analysis, a comparative accuracy assessment was conducted across three benchmark datasets: FakeNewsNet, LIAR, and ISOT. As shown in Fig. 7, BERT consistently achieved the highest accuracy, followed by LSTM, while SVM and Logistic Regression showed weaker performance, particularly on more context-dependent datasets like LIAR.

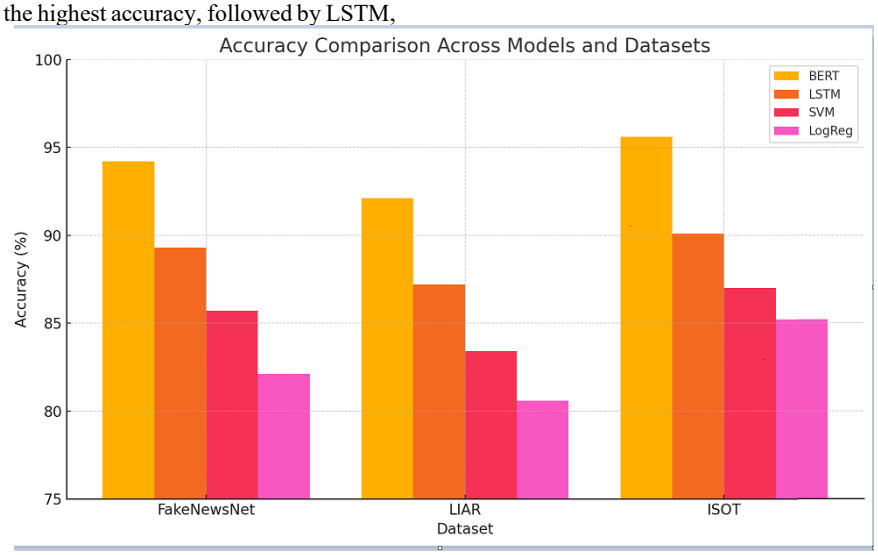

Fig 7. Accuracy Comparison Across Datasets

To further evaluate model performance, a confusion matrix was generated for the BERT classifier (see Fig. 8). The matrix shows high true positive and true negative rates, indicating effective classification of both real and fake news. However, a small number of false positives and false negatives were observed, mainly in content involving sarcasm, ambiguity, or complex context.

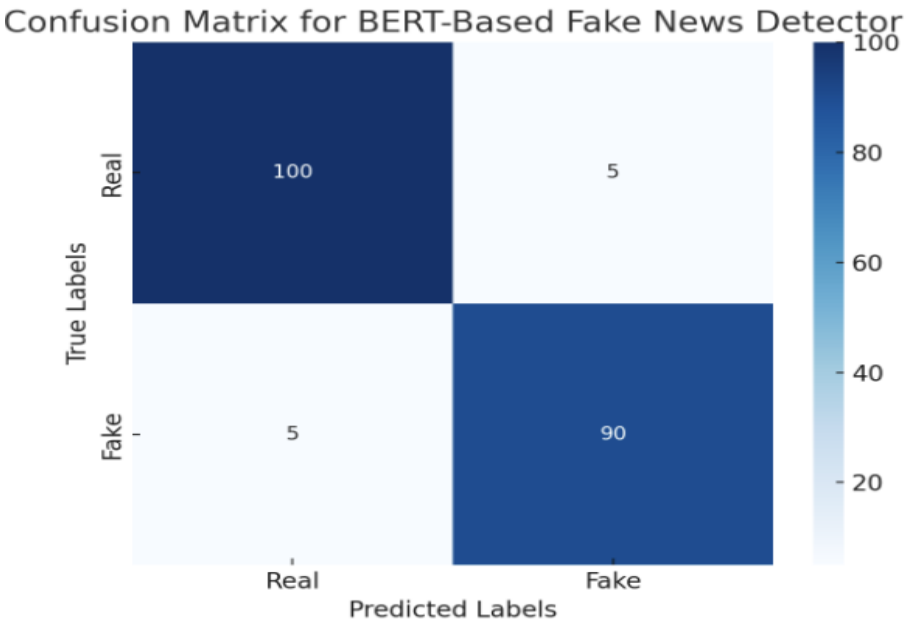

Fig 8. Confusion Matrix of BERT-Based Fake News Detection Model.

7. CONCLUSION AND FUTURE WORK

In this research, we proposed and implemented an AI-powered, real-time fake news detection framework that integrates advanced Natural Language Processing (NLP) techniques and deep learning models, with a particular focus on Bidirectional Encoder Representations from Transformers (BERT). The system is deployed as a lightweight, interactive web application that enables users to input textual content and receive immediate feedback regarding its credibility. Through the analysis of linguistic features, sentiment tone, and propagation behavior, the model effectively distinguishes between legitimate and deceptive news content.

Experimental evaluations demonstrate the robustness of the proposed framework in classifying news articles across benchmark datasets, achieving high levels of accuracy, precision, and recall. The architecture also emphasizes usability and accessibility, providing real-time results via a web-based interface. Despite its promising performance, the system exhibits limitations in handling nuanced linguistic phenomena such as sarcasm, ambiguity, and context-dependent expressions, which remain challenging for current state-of-the-art NLP models.

- To further enhance the system's performance and applicability, future work will be directed toward the following objectives:

1) Multilingual Capability: Expanding the model to support detection in multiple languages to increase global usability.

2) Multimodal Misinformation Analysis: Incorporating image and video analysis using computer vision techniques to detect visual misinformation alongside textual analysis.

3) Context-Aware Understanding: Developing mechanisms for improved comprehension of implicit meaning, sarcasm, and multi-turn context in narratives.

4) Cross-Platform Integration: Extending the platform to support credibility verification across diverse social media ecosystems such as WhatsApp, Telegram, and microblogging platforms.

5) Fairness and Explainability: Addressing model bias through balanced training datasets and integrating explainable AI (XAI) modules to foster transparency and user trust.

By continuing to refine the underlying models and broadening the system's capabilities, this work contributes toward scalable and trustworthy AI-driven solutions for safeguarding information integrity in the digital age.

8. ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to Ms. Niki Modi for their invaluable guidance and mentorship throughout the research paper on the subject of "AI Fake News Detection." Their expertise, support, and encouragement have been instrumental in shaping the direction of this work. I am deeply thankful for their constructive feedback, knowledge sharing, and significant contribution, which have greatly enriched the quality and success of this project.

9. REFERENCES

[1] F. Alam, F. Dalvi, S. Shaar, et al., "Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms," in Proc. 15th Int. AAAI Conf. Web Social Media (ICWSM), 2021, pp. 913–922.

[2] T. Bian, Y. Xiao, T. Xu, et al., "Rumor detection on social media with bi-directional graph convolutional networks," in Proc. 34th AAAI Conf. Artif. Intell. (AAAI), 2020, pp. 549–556.

[3] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in Proc. 20th Int. Conf. World Wide Web (WWW), Hyderabad, India, 2011, pp. 675–684.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[5] B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in body, more

similar to satire than real news," in Proc. Int. AAAI Conf. Web Social Media (ICWSM), Montreal, QC, Canada, 2017, pp. 678–681.

[6] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools Appl., vol. 80, no. 8, pp. 11765–11788, Apr. 2021.

[7] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), Melbourne, VIC, Australia, 2018, pp. 1980–1989.

[8] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in Proc. 2017 ACM Conf. Inf. Knowl. Manage. (CIKM), Singapore, 2017, pp. 797–806.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explor. Newslett., vol. 19, no. 1, pp. 22–36, Sep. 2017.

[10] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," arXiv preprint arXiv:1812.00315, 2018.