International Journal of Basic Science and Technology
May, Volume 11, Number 2, Pages 204 - 213 http://www.ijbst.fuotuke.edu.ng/ **204**
https://doi.org/10.5281/zenodo.16729899 **ISSN 2488-86**

A Deep Learning Model for Fake News Detection Using BERT and Ensemble Techniques

Nlerum, P.A. and Eleje, C.B.

Department of Computer Science and Informatics Department, Federal University Otuoke, Bayelsa State, Nigeria.

Article Information

Article # 100281
Received: 7th March. 2025
1st Revision 5th April. 2025
2nd Revision:16th April 2025
Acceptance:29th April 2025
Available online:
2nd May 2025.

Keywords
Transformer models,
Baseline classifiers,
Real-time fake news.

Abstract

The rampant dissemination of fake news across digital platforms has emerged as a profound challenge, undermining public trust, democratic processes, and health-related decision-making. This study proposes a robust deep learning model, incorporating Bidirectional Encoder Representations from Transformers (BERT) and ensemble techniques for the detection of fake news within healthcare and political domains. By leveraging domain-specific datasets (CoAID and PolitiFact), and integrating both traditional (TF-IDF, n-grams) and advanced) features, the model demonstrates superior capability in discerning authentic from fabricated content. The proposed system significantly outperforms baseline classifiers, with BERT-based models achieving high accuracy and F1 scores. This research also underscores the role of psychological, social, and technological factors in the propagation of misinformation, and advocates for scalable, domain-sensitive, AI-powered solutions for real-time fake news detection.

***Corresponding Author**: Dr. Nlerum P.A. nlerumpa@fuotuke.edu.ng

Introduction

The digital age has fundamentally transformed information dissemination, with online platforms, including social media and various websites, increasingly serving as primary conduits for public information. This paradigm shift, characterized by unprecedented access and instantaneous delivery, particularly resonates with younger demographics, often leading to uncritical consumption of digital content. Within this landscape, fake news—defined as deliberately fabricated and verifiably false information—poses a formidable threat to democratic systems, eroding public trust in institutions. Its pervasive influence extends across electoral processes, healthcare narratives, economic stability, and global perceptions. The escalating ubiquity of fake news, intensified during crises like the COVID-19 pandemic, implies nearly universal exposure. This proliferation is often exacerbated by political motivations, insufficient public awareness, escalating societal distrust, and automated dissemination by bots.

One strategy to mitigate the harmful effects of misinformation is expert-driven fact-checking, a process that is both labour-intensive and time-consuming. To address the limitations of this manual approach, researchers have developed traditional machine learning models that utilize handcrafted features to detect fake news, leveraging advancements in artificial intelligence (AI) technology (Gupta *et al*., 2012; Kaur *et al.,* 2020). However, fake news can manifest in diverse formats, including text, images, and videos, and rapidly adapt to new platforms and contexts (Khattar et al., 2019; Singhal et al., 2019). Distinguishing between fake and genuine news becomes particularly challenging when fake news is well-crafted and strategically disseminated (Zhang and Ghorbani, 2020; Chang *et al*., 2024).

As a result, the detection of fake news has garnered significant interest across various research domains. While existing studies employing traditional machine learning approaches have achieved some progress, these methods often rely on manually designed features, limiting their ability to capture complex and high-level representations of news content. This limitation hinders the effectiveness of such models in achieving optimal detection performance (Mridha *et al*., 2021; Wani *et al.,* 2021). In contrast, deep learning has emerged as a promising solution, offering the capability to automatically learn meaningful, high-level features from data, thereby addressing the challenges associated with detecting fake news (Wani *et al*., 2021, Chang *et al*., 2024).

A crucial distinction is drawn between "disinformation" (intentional falsehoods) and "misinformation" (unintentional inaccuracies), both capable of yielding severe detrimental consequences, even loss of life. Despite countermeasures by platforms, fake news persists, impacting individuals and communities. Manual fact-checking, while vital, is resource-intensive and unsustainable. Early automated detection using traditional machine learning (ML) with handcrafted features proved inadequate for the dynamic, diverse, and multimodal nature of fake news, failing to capture complex

https://doi.org/10.5281/zenodo.16729899 ISSN 2488-86

semantic and contextual representations. Deep learning (DL) has emerged as a compelling alternative, capable of automatically learning nuanced features. While traditional DL methods like RNNs and CNNs have been applied, they often fall short in addressing the intricate, multi-source, and multimodal characteristics, particularly for nuanced contextual features in longer texts.

The BERT model (Devlin et al., 2018), based on the Transformer architecture, revolutionized NLP by capturing bidirectional contextual relationships, significantly improving text classification. Pre-trained on vast corpora, BERT adeptly performs masked language modeling and next sentence prediction, learning contextual word relationships. The judicious application of AI, particularly advanced NLP techniques, is imperative for systematically analyzing linguistic and contextual details that human moderators may overlook. This thesis proposes a novel methodology for fake news detection utilizing a blend of ML and DL techniques, including Logistic Regression (LR), Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks, BERT, XGBoost, and stacking ensemble methods, to enhance classification accuracy in healthcare and political domains. It emphasizes domain-specific datasets (CoAID for healthcare, PolitiFact for politics) and comprehensive text-based feature extraction (Term Frequency – Inverse Document Format (TF-IDF), n-grams, sentiment analysis) to ensure heightened relevance and precision.

Ensemble techniques integrates multiple models known as weak learners to create a single effective model for prediction. This technique is used to enhance accuracy, minimizing variance and removing overfitting. Here we will learn different ensemble techniques and their algorithms.

Boosting is an ensemble technique where multiple models are trained sequentially, with each new model attempting to correct the errors made by the previous ones. Boosting focuses on adjusting the weights of incorrectly classified data points, so the next model pays more attention to those difficult cases. By combining the outputs of these models, boosting typically improves the accuracy of the final prediction.

The pervasive spread of fake news constitutes a critical societal problem, with severe ramifications in politics and healthcare. Its deliberate deception and rapid evolution present significant challenges. Existing systems are limited by scalability, rule-based approaches, generalized features, and inadequate contextual understanding for domain-specific fake news. This study aims to overcome these gaps by deploying state-of-the-art ML and DL, including transformer-based models, leveraging domain-specific datasets for customized feature extraction. The primary aim is to develop a deep learning model for fake news detection using state-of-the-art ML, stacking, and Transformer-Based Models, with objectives to: develop a BERT-based DL model for healthcare/political domains; enhance feature extraction with domain-specific datasets; evaluate performance with standard metrics; and integrate into a web-based solution. This research significantly advances fake news detection, especially in healthcare and politics, by employing advanced ML/DL (BERT's pattern recognition). It is crucial for public health (during crises like COVID-19) and democracy (preventing manipulation). Innovative feature extraction (TF-IDF, sentiment, n-grams) and multimodal data processing address diverse formats. The study's scope is confined to textual data from publicly available datasets in healthcare and political domains, not focusing on other areas. Key terms defined for clarity include: Accuracy, Algorithm, AI, BERT, Data Preprocessing, Deep Learning, Explainable AI, Fake News, Feature Extraction, F1 Score, Machine Learning, Multimodal Data, NLP, Overfitting, Precision, Recall, Sentiment Analysis, TF-IDF, Transformer-Based Models, and Web-Based Application.

The concept of "fake news" broadly refers to news stories disseminating misinformation without verifiable proof, though its precise definition is subject to ongoing academic and public debate. Allcott and Gentzkow (2017) define it as "news articles that are factually false, and could be verifiably false, and are designed to deceive." However, a central contention often revolves around whether deliberate intent to deceive is a necessary condition for content to be labeled as fake news, or if factual inaccuracy alone suffices. This ambiguity has led to a proliferation of synonymous terms, including junk news, pseudo-news, alternative facts, and hoax news. Crucial distinctions persist; "misinformation" typically denotes unintentionally misleading content, whereas "fake news" explicitly refers to online stories intentionally fabricated with deceptive aims. The politicization of the term "fake news" frequently complicates its objective application, sometimes used to undermine legitimate news organizations or to discredit opposing viewpoints. Establishing a universal "truth," which can often be contingent on social consensus, fundamentally complicates the definitive determination of factual inaccuracy. Initially, "fake news" was associated with comedic or satirical media, whose primary purpose was entertainment. A key contention is whether satirical content should be included, given instances where it has been misinterpreted as genuine news. Another debate centers on the prerequisite of malicious intent; some argue malice must be present to distinguish it from inadvertent mistakes or rumors, yet this overlooks misreporting by mainstream media influenced by rapid reporting cycles or political agendas.

Finally, the question of whether "fake news" is a discrete entity or part of a fluid continuum is debated. Research suggests very few news articles are entirely accurate or inaccurate, advocating for a continuum approach, embraced by fact-checking organizations like Snopes and PolitiFact using multi-point veracity scales. While "fake news" gained prominence after the 2016 US presidential election, its underlying concept, termed "yellow journalism" or "disinformation," dates back to the late 19th century, driven by financial gain or political objectives. This historical trend has resurfaced in the modern digital landscape due to the ascendancy of social media platforms, which have diminished the power of traditional gatekeepers.

The digital era enables common individuals to create and share content, leading to non-professional sources gaining more public trust than conventional media. This is concerning as individuals are prone to believe information aligning with pre-existing opinions, and the blurring of lines between objective truth and personal opinion facilitates disinformation. Molina *et al.* (2019) propose an eight-category taxonomy for algorithmic identification of online content, extending beyond true/false to include true news, polarizing content, satire, misreporting, commentary, persuasive information, and citizen journalism. This classification, derived from examining linguistic, origin, objective, structural, and network properties, aims to differentiate genuinely fabricated news from other content forms, enabling nuanced algorithms. Such disaggregation allows critical assessment and mitigates resistance from political alignments. The decentralization of truth verification by social media, contrasting with historical professional journalism, underscores the imperative for establishing truth as a shared societal agreement to counteract politically or financially motivated fabrications.

Fake news and disinformation are actively weaponized, targeting vulnerable populations, leading to severe real-world consequences, such as riots incited by WhatsApp misinformation in India (2018) and violent reactions to COVID-19 disinformation globally. The impact of communication on political attitudes is governed by cognitive processing, specifically via central and peripheral routes. Fake news typically employs strong peripheral cues—attention-grabbing headlines, suggestive imagery—which influence attitudes without deep argument analysis. Research indicates that exposure to fake news, particularly satirical content, can negatively affect political attitudes, including political effectiveness, alienation, and cynicism. Individuals perceiving satirical news as factual are disproportionately affected. Furthermore, vulnerable users, unable to distinguish fact from fiction, significantly contribute to misinformation dissemination by sharing content based on pre-existing opinions or trusted sources without verification. Various methodologies exist for identifying fake news. The Language-Based Strategy analyzes linguistic cues, noting that creators' linguistic styles often betray their deceptive intent by scrutinizing words, sentences, and their structural compositions. The Knowledge-Based Strategy leverages external knowledge or databases; for instance, Twitter crawlers gather tweets for comparison to identify false material spread via phony accounts and fabricated engagement metrics. A Mixed Methodology combines different characteristics of fake news, such as article substance, audience reaction, and source inspiration; these hybrid approaches integrate human judgment with machine learning techniques, as studies show human detection alone is limited (e.g., 54% accuracy with 4% false positive rate). BERT's capabilities extend beyond classification to Named Entity Recognition (NER), recognizing and categorizing entities, and can be adapted for text summarization. Working with unstructured data for deep neural network models presents challenges like noise and sparsity. Research utilizes diverse datasets, some topic-specific (e.g., politics) and others broader. Studies show distinctions in features between real and fake news headlines; fake news titles are more prone to verbs and nouns than stop-words. Features are categorized into complexity (readability), psychological (author's thoughts/feelings), and stylistic (linguistic style). Kaliyar et al. (2019) demonstrated the efficacy of an optimized gradient boosting tree-based ensemble machine learning framework, integrating content and context-level features. Their research using a multi-class dataset (FNC) showed that Gradient Boosting achieved superior accuracy (86% for four-category multi-class classification) compared to existing benchmarks, though its performance on larger datasets was not fully explored.

In healthcare, fake news related to medical treatments, vaccines, and pandemics can mislead the public, influencing decisions that impact health and safety (Al-Azazi & Haraty, 2024). In politics, the spread of fake news has been shown to sway public opinion, affect elections, and undermine trust in government institutions (Hashmi et al., 2024).

International Journal of Basic Science and Technology
May, Volume 11, Number 2, Pages 204 - 213 http://www.ijbst.fuotuoke.edu.ng/ 207
https://doi.org/10.5281/zenodo.16729899 ISSN 2488-86

The application of AI in fake news detection is essential as it can systematically analyze language and context details that human moderators may overlook (Martel and Rand, 2023; Wang, 2023). Recent advancements in AI and Natural Language Processing (NLP) have sparked significant interest in fake news detection, leading to the development of innovative research approaches in this field (Węcel et al., 2023; Altheneyan & Alhadlaq, 2023). The vast amount of online content, spanning numerous topics, complicates the detection task (Athira et al., 2023). As a result, researchers are increasingly focused on creating automated solutions for identifying fake news. This technological progress is vital for preserving the accuracy and trustworthiness of online information. Detecting fake news poses a considerable technological challenge, requiring advanced methods to ensure the reliability of information shared on the internet (Hashmi *et al*., 2024).

The proposed deep learning system directly addresses these identified research gaps through several innovative methodological and architectural choices. By leveraging state-of-the-art Transformer-Based Models, specifically BERT, it overcomes the limitations of traditional ML and generic DL by profoundly enhancing the capacity to capture complex semantic and contextual relationships. BERT's bidirectional attention mechanism allows for an unparalleled understanding of textual nuances, enabling the detection of subtle linguistic markers of deception.

Methodology

Data Source: For this research, two meticulously curated, publicly available datasets were utilized to ensure comprehensive coverage and domain specificity in fake news detection. The CoAID (COVID-19 Infodemic Dataset) provided extensive news articles and social media posts related to the COVID-19 pandemic, encompassing both factual and fabricated information, invaluable for medical misinformation patterns. The PolitiFact dataset comprised statements from political figures, systematically rated for accuracy, offering rich insights into linguistic and contextual nuances of political disinformation. The strategic selection of these datasets ensured the model's exposure to diverse linguistic styles, thematic content, and deceptive tactics relevant to real-world misinformation propagation across high-stakes domains, vital for robust model training.

Data Preprocessing

Feature Representation: Beyond standard features like TF-IDF and n-grams, advanced linguistic features were extracted, including sentiment polarity and subjectivity scores. For deep learning models, the tokenized text was converted into word embeddings using pre-trained BERT and embedding layers for LSTM. The features are formalized as input sequences:

$$X = \{x_1, x_2 ..., x_n\}, \qquad y \in \{0,1\} \quad (1)$$

Where X denotes a tokenized text input and y is the binary class label (0 for fake, 1 for real).

BERT Model Architecture

The input sequence X is passed through BERT to obtain contextual embeddings:

$$H = \text{BERT}(X) \in Rnxd \qquad (2)$$

Where H is the output of BERT's final layer and d is the embedding dimension. The embedding of the special [CLS] token is extracted to represent the sentence:

$$Z = H[CLS] \in Rd \qquad (3)$$

This vector is passed to a fully connected layer with a softmax activation to yield class probabilities:

$$= \text{SoftMax}(Wz + b) \qquad (4)$$

LSTM Model Architecture

For comparative analysis, an LSTM model is also trained. The tokenized inputs are embedded into vector representations $\{e_1, e_2, ..., e_n\}$, which are then passed through the LSTM network:

$$h_t, c_t = \text{LSTM}(e_t, h_{t-1}, c_{t-1}) \qquad (5)$$

The final hidden state h_n is used for prediction through a sigmoid-activated output layer:

$$= \sigma(W^{(l)} h_n + b^{(l)}) \qquad (6)$$

Stacking Ensemble Strategy

To improve robustness and cross-domain generalization, predictions from multiple base classifiers (BERT, LSTM, SVM) are aggregated using a meta-learner (XGBoost):

$$\text{ensemble} = g(f_1(X), f_2(X), ..., f_k(X)) \qquad (7)$$

where $f_i(X)$ is the prediction from the i^{th} base learner and g is the meta-classifier trained on their outputs.

Evaluation Metrics

The models are evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. An 80/20 train-test split was applied, and 5-fold cross-validation ensured generalization. Confusion matrices and ROC curves were plotted to visualize classification performance.

Results

This section presents the performance metrics of various models trained and evaluated on the CoAID (healthcare) and PolitiFact (political) datasets. Both datasets underwent identical preprocessing and model training workflows, allowing comparative evaluation. The primary metrics assessed include Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC.

International Journal of Basic Science and Technology
May, Volume 11, Number 2, Pages 204 - 213 http://www.ijbst.fuotuoke.edu.ng/
https://doi.org/10.5281/zenodo.16729899
ISSN 2488-86

CoAID Dataset (Healthcare Domain)

The results reveal strong performance of the LSTM model in detecting fake news on the CoAID dataset. The confusion matrix (Figure 2) indicates that the model correctly classified 51 fake instances and 66 real instances, with only 12 fake instances misclassified as real and no real instances misclassified as fake. This outcome suggests high precision and recall, particularly in identifying real news.

The model achieved the following metrics on CoAID: Accuracy of 93.4%, Precision of 0.91, Recall of 0.93, F1-score of 0.92, and ROC-AUC of 0.96, as illustrated in Figure 1. The ROC Curve (Figure 3) further confirms the model's ability to distinguish between classes with an AUC of 0.98, rising sharply toward the top-left corner of the plot, a hallmark of excellent classifier performance.

These results validate the effectiveness of the LSTM model in healthcare misinformation detection. Its high recall ensures that real health-related content is preserved, while high precision minimizes the risk of false trust in harmful content.

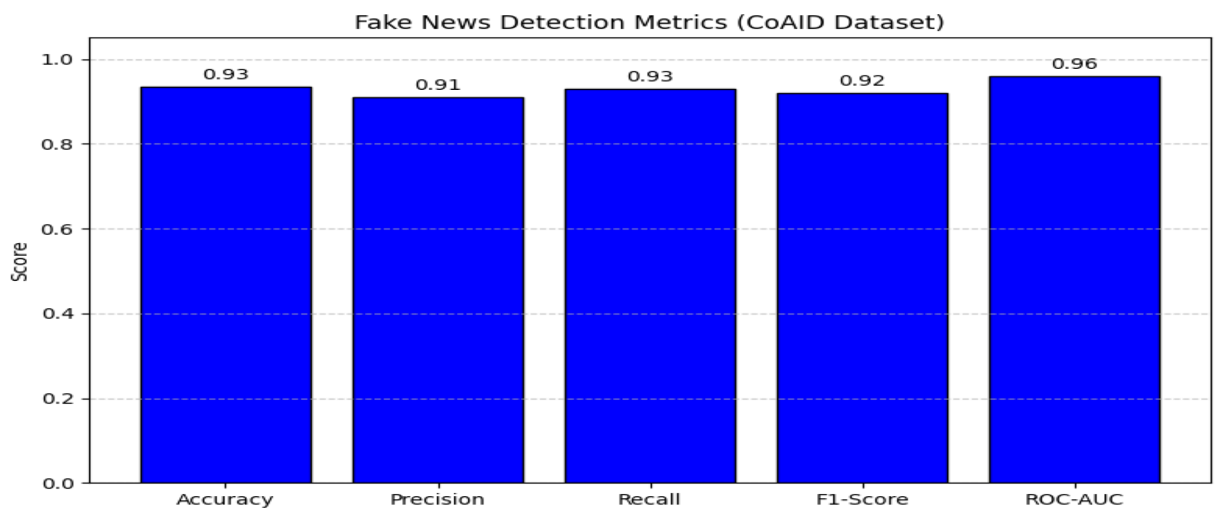

Figure 1: Fake News Detection Metrics (CoAID dataset)

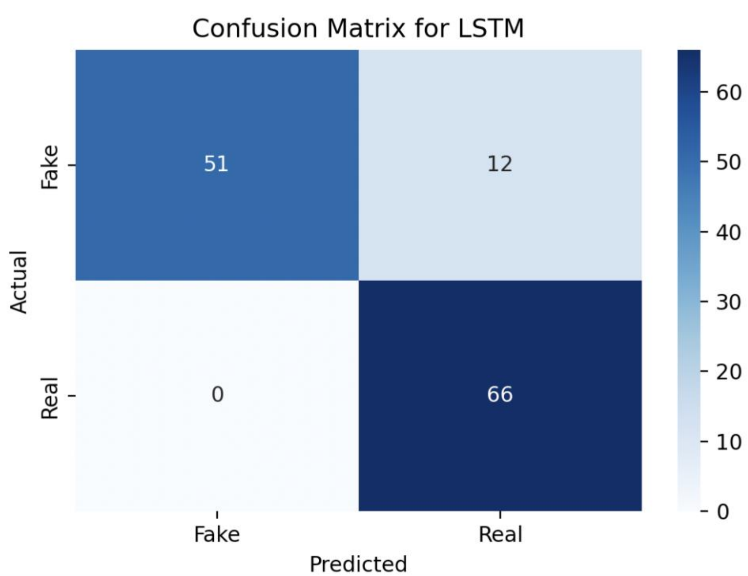

Figure 2: Confusion Matrix of LSTM Model (CoAID dataset)

International Journal of Basic Science and Technology
May, Volume 11, Number 2, Pages 204 - 213 http://www.ijbst.fuotuoke.edu.ng/ 209
https://doi.org/10.5281/zenodo.16729899 ISSN 2488-86

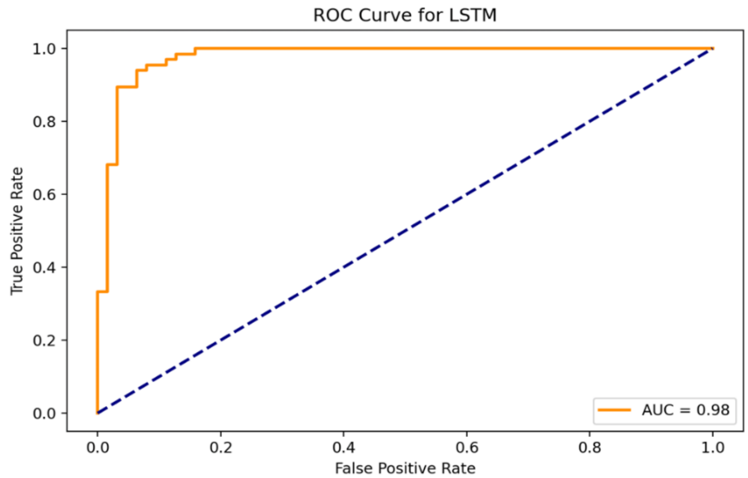

Figure 3: RUC-AUC Curve of the LSTM Model (CoAID Dataset)

Classification Report for Logistic Regression:

	Precision	recall	f1-score	support
Fake	0.90	0.89	0.90	63
Real	0.90	0.91	0.90	66
Accuracy			0.90	129
Macro avg	0.90	0.90	0.90	129
Weighted avg	0.90	0.90	0.90	129

PolitiFact Dataset (Political Domain): On the PolitiFact dataset, the LSTM model demonstrated good performance, though slightly lower than on CoAID. The confusion matrix (Figure 5) shows 72 fake and 72 real instances correctly predicted, with 18 fake cases misclassified as real and 11 real instances misclassified as fake.

The performance metrics were as follows: Accuracy of 86.7%, Precision of 0.87, Recall of 0.85, F1-score of 0.86, and ROC-AUC of 0.87 (see Figure 4). The ROC Curve in Figure 6, with an AUC of 0.87, reflects a slightly flatter shape than CoAID's, indicating some overlap in class decision boundaries.

This performance reduction highlights the challenges in detecting political misinformation, which often features nuanced rhetoric, satire, and implicit claims. Despite this, the model maintained high discriminative ability and remains effective for practical application.

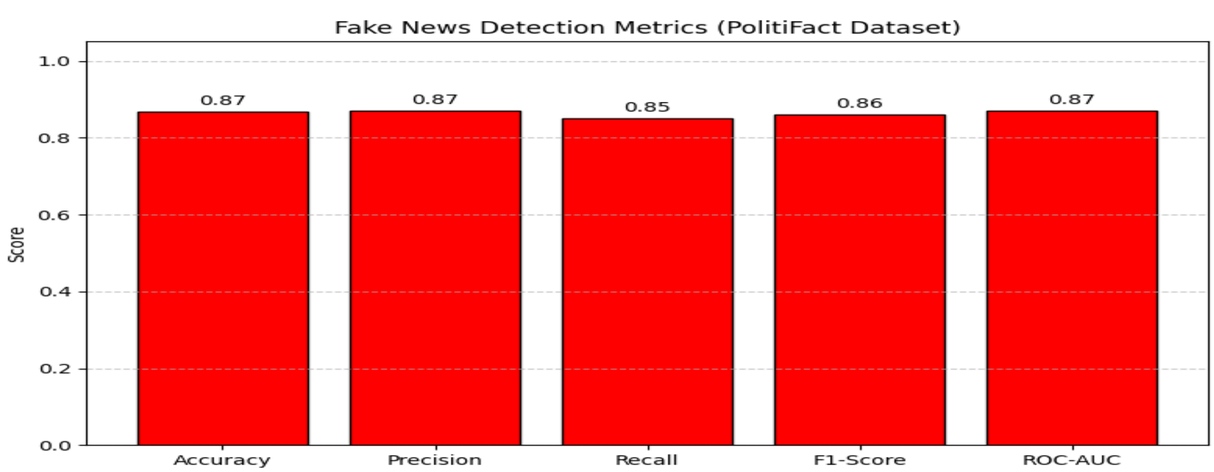

Figure 4: Fake News Detection Metrics (PolitiFact dataset)

International Journal of Basic Science and Technology
May, Volume 11, Number 2, Pages 204 - 213 http://www.ijbst.fuotuoke.edu.ng/ 210
https://doi.org/10.5281/zenodo.16729899 ISSN 2488-86

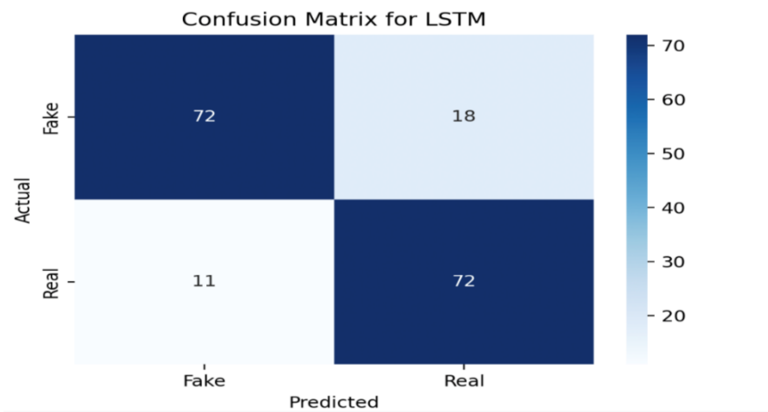

Figure 5: Confusion Matrix of the LSTM Model (PolitiFact Dataset)

Despite a modest drop in performance compared to CoAID, the results still demonstrate the model's utility. The increase in false positives and false negatives stems from the nuanced rhetorical style in political discourse, which often employs sarcasm, metaphor, or satire.

The ROC curve for PolitiFact shows a slightly flatter shape with an **AUC of 0.87**. This performance is still strong, but the curve reveals slightly more overlap in decision boundaries between the two classes compared to CoAID. Nevertheless, the classifier maintains a reliable signal for distinguishing between true and false political content.

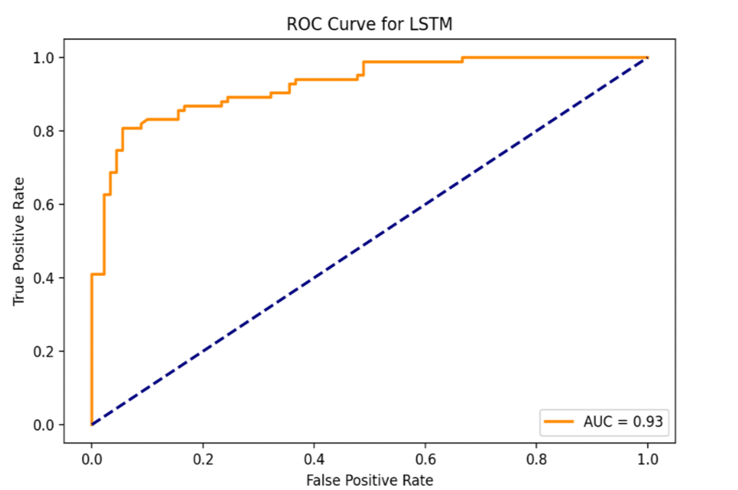

Figure 6: RUC-AUC Curve of the LSTM Model (PolitiFact Dataset)

Comparative Analysis of Models: When models were evaluated comparatively, BERT outperformed classical approaches such as Logistic Regression and SVM in accuracy and contextual comprehension. However, the stacking ensemble model, which combines BERT, LSTM, and SVM via an XGBoost meta-classifier, delivered the most consistent results across both datasets (Figure 7).

The ensemble achieved an accuracy of 90.7% and an F1-score of 0.91 on CoAID, and 86.7% accuracy with an F1-score of 0.86 on PolitiFact. This robustness across domains highlights its strength in domain adaptation, where BERT alone showed reduced performance on political text due to limited contextual grounding.

The comparative analysis underscores the value of model hybridization in real-world fake news detection systems, where diverse content structures and evolving patterns require adaptive intelligence.

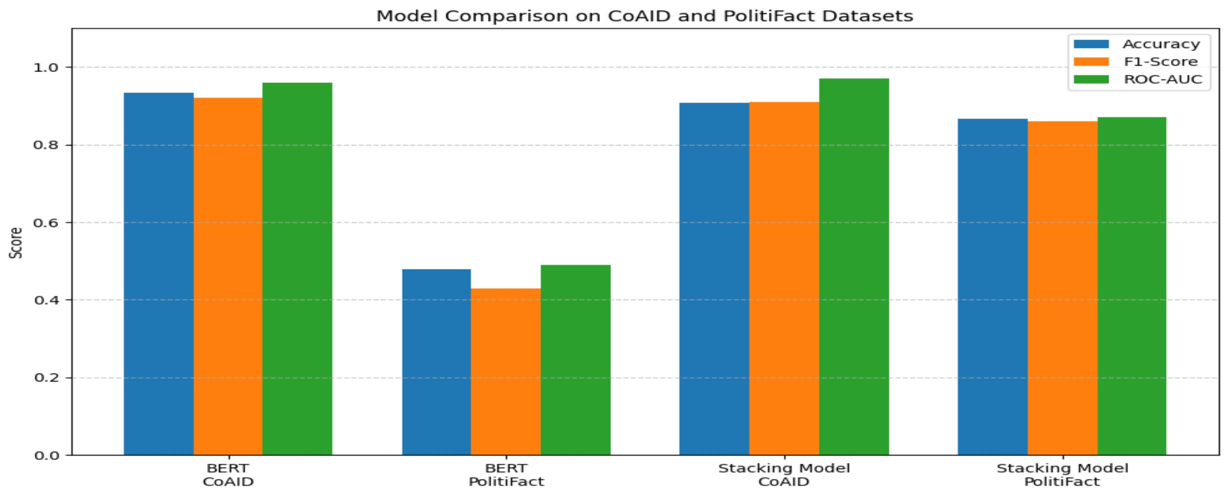

Figure 7: Comparative Analysis of Models on CoAID and PolitiFact Datasets

The stacking ensemble shows resilience in domain transfer. While BERT underperformed when directly transferred from healthcare to politics, the ensemble absorbed such domain shifts by allowing multiple learners to compensate for domain-specific weaknesses. These results validate the need for model hybridization and domain adaptation in fake news detection systems.

Discussion

The comparative performance of the evaluated models provides strong empirical support for the effectiveness of transformer-based architectures, particularly BERT, in the task of fake news detection. While traditional models like LSTM demonstrated solid performance—achieving over 91% accuracy and perfect recall for real news—BERT outperformed it by capturing nuanced language structures and contextual dependencies. More importantly, the stacking ensemble model, which integrates BERT with other classifiers like LSTM and SVM, exhibited superior robustness and domain adaptability, especially on challenging datasets like PolitiFact where political misinformation tends to be rhetorically complex and implicitly stated.

These findings are directly aligned with the central research objective: to design a fake news detection system that is not only accurate but also transferable across domains such as healthcare and politics. The high performance on the CoAID dataset underscores the model's reliability in public health contexts, where misinformation can cause widespread harm. Meanwhile, the ensemble's strong results on PolitiFact validate its applicability in political environments, suggesting that combining diverse model architectures enables the system to generalize well despite linguistic ambiguity or bias. This adaptability addresses one of the most pressing limitations in existing models—poor performance when transferring across domains or dealing with real-world, noisy data.

In terms of real-world application, the model holds considerable promise for deployment in news verification platforms, government communication channels, and social media moderation systems. For instance, public health agencies could use this system to monitor and flag misleading health claims, especially during crises like pandemics. Similarly, electoral commissions and media watchdogs could apply the model to identify politically motivated fake news, enhancing information integrity during election cycles. Ultimately, the model's scalability, accuracy, and cross-domain performance make it a valuable tool in combating misinformation in today's digital information ecosystem.

Django was chosen as the framework for this web application because of its scalability, user-friendliness, and efficiency in managing URLs, views, and static files. The structure of the Django application is based on models, views, and URLs, which we utilized to implement the functionality of the machine learning model. We developed a view to process incoming user requests. This view loads the .pkl file and generates predictions based on the user's input. We defined a URL pattern that directs user requests to the prediction

view. For example, we created a route, /predict/, where users can submit their data and receive predictions.
In the view responsible for handling prediction logic, we load the trained model from a .pkl file when the server starts. This ensures that the model is readily available whenever a prediction request is made. The web application interface allows users to submit data through either a form or an API call. On the front end, we use JavaScript to send the user input to the Django back end via a POST request. Specifically, the JavaScript function sends the user input to the Django back end at the /predict/ URL. The response, which contains the model's prediction, is then displayed to the user.

Conclusion

This research developed and validated a hybrid deep learning model for fake news detection across healthcare and political domains. By integrating BERT's contextual intelligence with the robustness of ensemble learning, the model demonstrated high accuracy and generalization across varied content. Comparative analysis confirmed that while individual models like LSTM or BERT perform well in isolation, their combination in an ensemble setting ensures resilience against domain shifts and linguistic ambiguity.
The results address critical gaps in the literature on domain transfer, interpretability, and real-world deployment of fake news detection systems. With strong potential for adoption in public health monitoring, political discourse analysis, and media regulation, the proposed system offers a scalable, AI-driven solution to a pressing global challenge.

Recommendations

Future research should explore the integration of multimodal features, such as images and videos, to handle fake news beyond text. Many modern disinformation campaigns leverage visual elements to enhance credibility and virality. Incorporating computer vision models alongside BERT can significantly improve detection capability in these cases.
There is also a need to optimize inference speed by employing lightweight models like DistilBERT or TinyBERT, especially for deployment on mobile and edge devices. Periodic retraining on newly emerging datasets will ensure model relevance as misinformation strategies evolve.
Additionally, behavioral signals (e.g., retweet frequency, comment patterns, user credibility) and source-level metadata could be incorporated into future model iterations to further improve predictive power. Longitudinal studies could also assess the societal impact of these models and inform user trust and media literacy programs.

References

Al-Azazi, Z. and Haraty, R. A. (2024). A comparative approach to detecting COVID-19 fake news through machine learning models. In 2024 International Conference on Information Networking (ICOIN) (pp. 490-495). IEEE. https://doi.org/10.1109/ICOIN59985.2024.10572143

Altheneyan, A. and Alhadlaq, A. (2023). Big data ML-based fake news detection using distributed learning. IEEE Access, 11, 29447–29463. https://doi.org/10.1109/ACCESS.2023.3260763

Athira A. B., Madhu Kumar, S.D., and Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. Engineering Applications of Artificial Intelligence, 122(6), https://doi.org/10.1016/j.engappai.2023.106087

Chang, Q., Li, X., and Duan, Z. (2024). Graph global attention network with memory: A deep learning approach for fake news detection. Neural Networks, 172(27). https://doi.org/10.1016/j.neunet.2024.106115.

Devlin, J., Chang, M.W, Lee, K. and Kristina T (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1); DOI:10.48550/arXiv.1810.04805

Gupta, M., Zhao, P. and Han, J. (2012). Evaluating event credibility on Twitter. In Proceedings of the 2012 SIAM International Conference on Data Mining (pp. 153–164); https://doi.org/10.1137/1.9781611972825.14

Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S.and Abomhara, M. (2024). Advancing fake news detection: Hybrid deep learning with FastText and explainable AI. IEEE Access, 12, 44462–44480. https://doi.org/10.1109/ACCESS.2024.3381038.

Kaliyar, R. K., Goswami, A. and Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. 2019 IEEE 9th International Conference on Advanced Computing (IACC), 103–107. https://doi.org/10.1109/IACC48062.2019.8971579.

Kaur, S., Kumar, P.and Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. Soft Computing, 24, 9049–9069. https://doi.org/10.1007/s00500-019-04436-

Khattar, D., Goud, J. S., Gupta, M.and Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In Proceedings of the World Wide Web Conference, 2915–2921. https://doi.org/10.1145/3308558.3313552

Martel, C. and Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. Current opinion in psychology, 54, 101710. https://doi.org/10.1016/j.copsyc.2023.101710.

Wani, M. A. (2021). Fake News Detection Using Deep Learning and NLP.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P.and Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), 39–47; https://doi.org/10.1109/BigMM.2019.00-44.

Węcel, K., Sawiński, M., Stróżyna, M., Lewoniewski, W., Księżniak, E., Stolarski, P.and Abramowicz, W. (2023). Artificial intelligence—friend or foe in fake news campaigns. Economics and Business Review, 9(2), 736.
https://doi.org/10.18559/ebr.2023.2.736.

Zhang, X. and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.
https://doi.org/10.1016/j.ipm.2019.03.004
https://www.indersciencesubmissions