Iris Journal of Mathematics

Iris Publishers

Research Article

DistilBERT-Based Binary Classification NLP-Framework to Enhance Misinformation and Fake News Detection Accuracy Performance

Hisham AbouGrad*, Fiza Riaz and Abdul Qadoos

Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London – UEL, London, United Kingdom

***Corresponding author:** Hisham AbouGrad, Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London – UEL, London, United Kingdom

Received Date: May 14, 2026

Published Date: May 21, 2026

Abstract

The mass growth in popularity of misinformation and fake news has become more common through social media platforms and communication channels. When fake information is presented as authentic news, it can deceive people and undermine their trust, which is a problem known as counterfeit news. Detecting all fake news is difficult because of varied writing styles and constantly evolving methods of misinformation. This becomes even more challenging when news articles are short, informal, noisy and come from various domains, such as politics, celebrity news and general newspapers. This research study has implemented a natural language processing (NLP) framework and several machine learning models to detect fake news. Logistic regression, support vector machine (SVM), random forest, long short-term memory (LSTM), and DistilBERT models are applied across three different datasets, which are FakeNewsDataset, WELFake, and GossipCop. Data preprocessing techniques, such as tokenizing, lemmatizing, and removing stop words, are applied, with the SMOTE data balancing technique to adjust class imbalances. TF–IDF features are used for traditional machine learning models, token sequences are applied for the LSTM model, and contextual token embeddings are used for DistilBERT. All models are tuned to ensure an optimal model configuration for detecting misinformation and fake news. Logistic regression and SVM models showed high accuracy performance, while LSTM has better results with longer news. Also, the DistilBERT model demonstrated the highest overall generalisation by achieving 96% classification accuracy on FakeNewsDataset, 98% on WELFake, and 85% on GossipCop. These research findings support that transformer-based NLP models have proven to deliver high classification accuracy for fake news detection, especially for complex and unstructured written text.

Keywords: NLP Detection Framework; Deep Learning Models; Binary Classification Algorithms; Fake News Detection Model; Misinformation Detection; DistilBERT Transformer Model; SMOTE Data Balancing Technique; Feature Preprocessing; Feature Engineering; TF-IDF Vectorization

Introduction

The way of accessing and exchanging information has changed intensely in recent years due to the rise of digital platforms, particularly social media. Such easy-to-assess platforms have become central to how people consume news, due to speed and convenience in the dissemination of information. This advancement has also introduced significant challenges, most notably the wide spread circulation of false or misleading content, which can lead to false information that causes people to make negative reactions and wrong decisions. As observed by Tian et.al. [1], the open nature of online platforms allows virtually anyone to post content, which results in the rapid spread of misinformation. Such content often goes without rigorous review, which influences public perception

and behaviour before factual corrections are revealed. The increasing presence of fake news presents a serious threat to societal trust, informed decision-making, and even public health, emphasising the need for effective detection strategies.

In the digital age, fake news, intentionally deceptive or incorrect information presented as legitimate news, spreads with alarming speed and reach. Its prevalence is exacerbated by algorithm-driven recommendation models that prioritise sensational or emotionally charged stories. Further, Gupta et.al. [2] proved the overwhelming volume of misinformation being generated across platforms, such as LinkedIn, Facebook, and X (Twitter), making fake news more difficult to recognise. Social media's capacity to amplify such content within seconds adds complexity to the challenge. The urgency to address fake news is reflected in ongoing efforts by governments, media organisations, and technology companies. Hence, this research study aims to address this issue by investigating how NLP and machine learning (ML) can be used to improve fake news detection. The research focuses on utilising the different models, such as logistic regression, SVM, random forest, LSTM and the DistilBERT transformer model, which is a lightweight alternative to BERT, known for its efficiency and contextual understanding [3]. It also addresses the problem of class imbalance by utilising the SMOTE technique. Thus, key questions explored in this research study include whether preprocessing techniques, such as tokenisation and stop-word removal, improve classification accuracy, how feature selection affects model performance, and how effective the chosen datasets are in supporting balanced learning. Overall, the study seeks to identify the most effective solution for detecting fake news with high accuracy and generalisability.

Related Work and State of the Art

Due to the quick dissemination of false information via social media and online news platforms, the detection of fake news has grown in importance. According to recent research studies, detecting false news requires more than just text categorisation; it also requires knowledge of writing style, context, source reliability, user behaviour, and occasionally visual material. While current research has shifted toward deep learning, transformer-based models, and multimodal approaches, earlier studies primarily used conventional machine learning and natural language processing techniques.

Machine Learning and NLP-Based Approaches

Recent research indicates that false news can manifest itself in a variety of formats, such as text, photos, captions, headlines, and social media engagement patterns [4]. As a result, multimodal techniques that integrate textual and visual data can enhance detection effectiveness, particularly on social media platforms where deceptive content is frequently accompanied by emotionally charged words or photos [5]. Language-specific preprocessing and suitable model selection are crucial for enhancing classification accuracy, as demonstrated by the application of NLP-based false news detection in non-English situations. Traditional machine learning models have limitations, even though they are helpful as baseline techniques [6]. They may find it difficult to convey deeper semantic meaning, sarcasm, emotional tone, informal writing, and contextual linkages between words since they frequently rely on surface-level qualities. Fake news content that is brief, loud, or gathered from social media makes this restriction more apparent.

Deep Learning and Transformer-Based Models

Deep learning methods have improved false news detection by learning sophisticated textual patterns directly from data. Hybrid models, such as CNN-BERT, have been implemented to capture both local and global semantic information, which boosts the ability of these models to understand bogus news content beyond a simple keyword pattern [7]. Word embeddings mixed with LSTM networks have also demonstrated significant results since LSTM models can process text sequentially and capture correlations between words in larger news items [8]. Multimodal and ensemble-based techniques have significantly increased detection accuracy by combining textual and visual features [9]. Conversely, Bi-LSTM models have been used to process text in both forward and backward directions, which allows the model to understand the whole context of a sentence more effectively than ordinary LSTM models [10]. Also, CNN-BiLSTM models have been applied to Arabic false news detection, which proves the effectiveness of hybrid neural networks (NNs) for binary classification tasks [11]. Likewise, graph-based models have introduced another direction by analysing the links between news content, users, and social media interaction patterns [12].

Further, other studies have incorporated NLP and deep learning techniques to automate fake news identification and improve classification performance [13]. Feature engineering and exploratory data analysis (EDA) also remain crucial because the quality of data preparation directly influences model accuracy [14]. Ensemble-based deep learning models, including Bi-GRU and Bi-LSTM architectures, have shown good performance in fake news detection tests [15,16]. More inventive methods have turned textual data into image-based representations to improve feature extraction and classification. Transformer-based models have become increasingly relevant because they provide higher contextual knowledge than typical machine learning models. BERT-LSTM models have been utilised for disinformation detection in mobile social media contexts [17]. Adversarial training approaches have also been utilised to improve model robustness against manipulated or deceptive inputs [18]. Deep ensemble models integrating CNN and Bi-LSTM have achieved strong F1-scores in online false news detection [19,20]. Sentence embeddings and deep learning have also been used to detect clickbait-style false information. Furthermore, temporal and textual elements have been implemented to improve rumour identification, suggesting that the timing and spread pattern of information might support fake news categorisation [21].

Multilingual and Emerging Detection Approaches

Multilingual, cross-lingual, and multimodal techniques have been the focus of recent research studies on fake news identification. The significance of language-specific models for non-English fake news detection has been demonstrated using Arabic transformer models to identify both human-written and generative ar-

tificial intelligence (GenAI) misinformation [22]. The usefulness of multilingual transformer designs has also been utilised by applying ensemble transformer models, such as ELECTRA, mBERT, and XLM-RoBERTa, to Urdu fake news detection [23]. Combining textual and visual data can further increase classification accuracy, as demonstrated by multimodal models that include BERT and R-CNN features [24]. Because model performance is directly impacted by text preparation quality, standard NLP data preprocessing techniques, such as tokenization, stemming, and vectorization, continue to be crucial to detect fake news and misinformation [25].

Several obstacles still exist despite these advancements, and many algorithms lose accuracy when dealing with brief, noisy, or informal information, but they do well on clean, structured datasets. Also, model fairness and biased predictions might be impacted by class imbalance. Further, a lot of studies only assess one model or one dataset, which makes it difficult to compare different areas fairly. Thus, logistic regression, random forest, SVM, LSTM, and DistilBERT are compared between three different datasets (i.e. FakeNewsDataset, WELFake, GossipCop) in this research study. Furthermore, the study uses SMOTE balance and preprocessing to provide an impartial and trustworthy assessment of transformer-based, deep learning, and conventional machine learning models for binary identification of fake news and misinformation.

Methodology and Methods

This research study implemented five different models instead of relying on one across three benchmark datasets to evaluate the effectiveness of various ML techniques for fake news detection, as shown in Figure 1. The selected models include logistic regression, random forest, SVM, LSTM and DistilBERT models.

Figure 1 exhibits the proposed model architecture, which has a combination that provides both traditional machine learning baselines and advanced deep learning architectures for a balanced comparison.

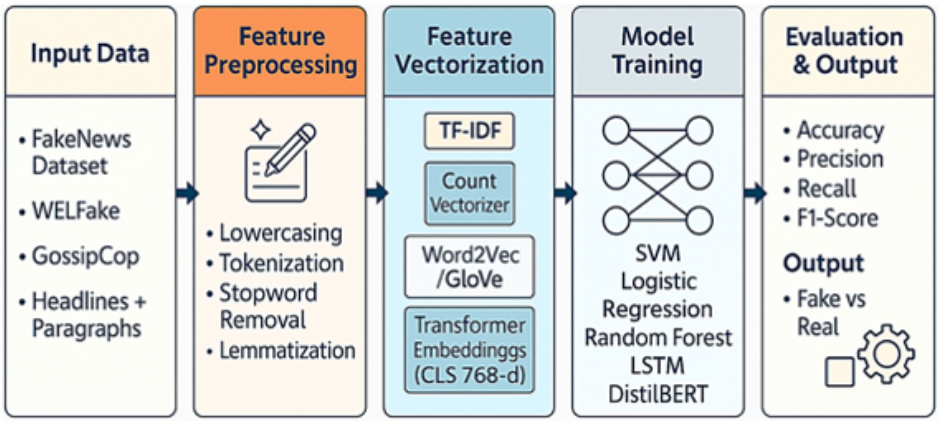

Figure 1: Architecture of Proposed Model.

Datasets and Data Preprocessing

The selected datasets for this research contain both real and fake news articles, labelled accordingly. The study datasets are: FakeNewsDataset, Mendeley data repository dataset containing short news articles and claims, balanced across real and fake classes [26]; WELFake, a large-scale Kaggle data repository dataset combining multiple sources with long-form articles and rich information [27]; and GossipCop, a GitHub platform dataset focused on celebrity news, known for being more challenging due to shorter and less structured writing [28]. The false information and fake news detection challenge is structured as a binary classification problem. The merged dataset can be expressed as follows:

$$D = \left\{ \left(x_i, y_i \right) \right\}_{\{i=1\}}^{\{N\}} \qquad (1)$$

where D denotes the dataset, x_i is the i^{th} news article, y_i is the corresponding class label, and N represents the total number of news samples. The classification label is defined as:

$$y_i = \begin{cases} 0 \ if \ the \ news \ article \ is \ real \\ 1 \ if \ the \ news \ article \ is \ fake \end{cases} \qquad (2)$$

This formulation treats each article as an input text sample, with each label serving as the intended output for binary classification. Since raw text often contains noise, a consistent data preprocessing pipeline is applied, which involves the following steps: (1) Deduplication and removal of non-informative entries; (2) Text normalisation through lowercasing, punctuation stripping, and stop word elimination; (3) Lemmatization to standardise word forms; and (4) Tokenization for model-ready formatting. The data preprocessing pipeline converts each raw news article into a clean text representation.

$$x_i^{'} = P\left(x_i \right) \qquad (3)$$

where $x_i^{'}$ represents the pre-processed version of the original article x_i and P (·) represents the preprocessing function, including lemmatization, tokenization, stop-word removal, punctuation removal, and text normalisation. To mitigate class imbalance, the SMOTE technique algorithm was applied to ensure equitable representation of both real and fake news instances during training, where SMOTE develops synthetic samples for the minority class by interpolating between a minority sample and one of its nearest neighbours [12].

$$x_{\{new\}} = x_i + \lambda\left(x_{\{nn\}} - x_i\right) \qquad (4)$$

Where $x_{\{new\}}$ is the synthetic sample, x_i is a minority class sample, and $x_{\{nn\}}$ is one of its nearest neighbours, while λ is a random number between 0 and 1.

Feature Engineering

Different feature engineering methods are utilised depending on the model's requirements. TF-IDF vectorization is used to convert cleaned text into numerical features for logistic regression, random forest, and SVM. TF-IDF prioritises phrases that are relevant in a document but are less common across the entire dataset. The TF-IDF score of a phrase t in document d can be calculated as:

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)+1}\right) \qquad (5)$$

where N is the total number of documents, $DF(t)$ is the number of documents containing term t, and $TF(t,d)$ is the frequency of term t in document d. To prevent division by zero, the +1 is utilised.

For the LSTM model, the cleaned text is padded to a predetermined length and transformed into token sequences. This enables the model to process the text in a sequential manner and identify word associations. The token sequence for each article can be represented as:

$$S_i = \left[w_1, w_2, w_3, ..., w_m\right] \qquad (6)$$

where S_i represents the token sequence of the i^{th} article, $w_1, w_2, w_3, ..., w_m$ are word tokens, and m represents the maximum sequence length after padding.

The DistilBERT model processes the text using the DistilBERT tokenizer, which has a fixed sequence length of 128 tokens. DistilBERT, unlike TF-IDF, utilises contextual embeddings, which determine the meaning of each word based on surrounding words. DistilBERT's tokenized input sequence formula is as follows:

$$T_i = \left[CLS, t_1, t_2, t_3, ..., t_m, SEP\right] \qquad (7)$$

where T_i represents the tokenised input sequence of the i^{th} article, $t_1, t_2, t_3, ..., t_m$ are sub-word tokens, and [CLS] and [SEP]. These are special tokens used for classification and sequence separation.

Model Training

Each of the five models is trained separately on the three datasets to ensure a fair comparison. The purpose of training is to learn the relationship between the pre-processed news article and its class label. This learning process can be represented as:

$$\hat{y}_i = f_{\theta\left(x_i^{'}\right)} \qquad (8)$$

where $x_i^{'}$ is the pre-processed news article, f_θ represents the trained classification model, θ represents the model parameters, and $\hat{y}_i$ is the predicted class label.

Logistic regression, random forest, and SVM models are trained with TF-IDF feature vectors. L2 regularisation is used in logistic regression training to minimise overfitting. An ensemble of 100 decision trees is used by the random forest model to increase stability and lower volatility. Since the SVM model works well with high-dimensional text classification data, it employs a linear kernel.

$$Objective = \left(\frac{1}{2}\right)\|w\|^2 + C\sum_{i=1}^{N}\xi_i \qquad (9)$$

Where w is the weight vector, C is the SVM penalty parameter, and ξ_i is the slack variable for classification errors or margin violations. A higher C penalizes classification errors more heavily, whereas a smaller C allows for more error tolerance and a wider margin. As a result, critical hyperparameters such as regularization strength, tree depth, and the SVM penalty parameter C are adjusted to improve model performance.

In the LSTM model, padded token sequences are processed through an embedding layer, then an LSTM layer, a dropout layer, and a dense output layer. The embedding layer translates tokens to numerical representations, whilst the LSTM layer detects sequential patterns in the text. Dropout is utilised to reduce overfitting, and early halting is done depending on validation loss. For the DistilBERT model, the pre-trained transformer model is run through the tokenized input. A classification layer receives the contextual representation produced by the classification token.

$$z_i = Wh_i + b \qquad (10)$$

where h_i represents the contextual output generated by DistilBERT, Wis the weight matrix, b is the bias term, and z_i is the output logit. The predicted probability of the fake news class is represented as:

$$p_i = P\left(y_i = 1 \mid x_i^{'}\right) \qquad (11)$$

where p_i is the probability that the news article belongs to the fake class. The neural network-based models are trained using a binary cross-entropy loss function:

$$L = -\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left[y_i(p_i) + (1-y_i)\log(1-p_i)\right] \qquad (12)$$

where L represents the loss value, N represents the total number of samples, and y_i is actually the label, and p_i is the probabil-

ity that is expected. This loss function helps the model distinguish between bogus and true news by penalising inaccurate predictions. Utilising the Hugging Face Transformers library, the DistilBERT model is optimised. The text is tokenized into up to 128-token sequences. The model is trained for three epochs using a weight decay of 0.01, a learning rate of 5×10−5, and a batch size of 16. The best performing model is saved, and batching, padding, training, and validation are all handled by the Hugging Face Trainer.

Evaluation Metrics

The performance of the final model was assessed using the independent testing subset. The testing subset was not used for TF-IDF fitting, SMOTE application, hyperparameter adjustment, early stoppage, or model choice. This guaranteed that the model's accuracy, precision, recall, and F1-score represented its performance on unseen data. To assess the efficiency of the proposed false news detection framework, various common classification measures were employed. These measures enable a fair comparison between deep learning models and conventional machine learning techniques and offer a thorough evaluation of model performance. Fake news detection is a binary classification job, and therefore, accuracy, precision, recall, and F1-score were used to assess each model's performance. Because they assess several aspects of classification quality, these metrics are frequently employed in research on text classification and natural language processing [29].

Accuracy measures the overall proportion of correctly classified instances among all predictions made by the model. It provides a general overview of model performance and is calculated using the following formula:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (13)$$

where true positive (TP) represents fake news articles correctly classified as fake, true negative (TN) represents real news correctly classified as real, false positive (FP) represents real news incorrectly classified as fake, and false negative (FN) represents fake news incorrectly classified as real [30]. While accuracy is useful for measuring overall performance, it may not always provide a complete picture when dealing with imbalanced datasets. Therefore, additional evaluation metrics such as precision and recall are used to better understand classification behaviour.

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. In the context of fake news detection, precision indicates how many articles predicted as fake are fake [31]. Precision can be calculated as:

$$\text{Precision} = TP / (TP + FP) \quad (14)$$

A high precision score indicates that the model produces fewer false alarms when identifying fake news. This is important in real-world practical applications where incorrectly labelling real news as fake may reduce trust in automated detection systems. Recall measures the proportion of actual positive instances that were correctly identified by the model. In fake news detection, recall represents how many of the actual fake news articles are successfully detected by the model. Recall is calculated as:

$$\text{Recall} = TP / (TP + FN) \quad (15)$$

A high recall value means that the model can detect most fake news instances, reducing the risk of misinformation spreading undetected. However, improving recall alone may sometimes increase false positives. To balance the trade-off between precision and recall, the F1-score is used. The F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification performance [32]. It is calculated as:

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (16)$$

The F1 score is particularly useful when evaluating models on datasets where both false positives and false negatives are important. In fake news detection, this metric helps determine whether a model can effectively identify misinformation while minimising incorrect classifications [33]. In addition to these evaluation metrics, statistical significance testing was examined to provide a more thorough comparison of models. This is significant because some models, such as logistic regression, SVM, and DistilBERT, produced remarkably similar findings on the FakeNewsDataset. In such instances, a slight variation in accuracy should not be seen as conclusive evidence that one model is superior. McNemar's test is appropriate for comparing two classifiers on the same test set since it looks at cases in which one model properly identifies a sample while the other incorrectly classifies it. As a result, McNemar's test is recommended for future validation of close model comparisons, particularly where stated accuracy and F1-score values are quite similar.

The evaluation results, as illustrated in Table 1, demonstrate that all models performed well overall, but their effectiveness differed among datasets. On the FakeNewsDataset and WELFake datasets, traditional ML models, including logistic regression, random forest, and SVM, proved high and consistent performance with accuracies between 95% and 97%, indicating their ability to handle structured and balanced data effectively. However, on the noisy and unstructured GossipCop dataset, their accuracy fell to about 80–82%. The LSTM model achieved 97% accuracy on WELFake but low on GossipCop, with just 76% accuracy and a poor F1-score of 43%, indicating difficulties handling short or informal statements. In contrast, DistilBERT performed consistently across all datasets, particularly on the difficult GossipCop dataset, where it attained the highest accuracy of 85%.

Table 1: Evaluation Metrics for Fake News Detection.

Models	Datasets	Accuracy	F1 Score	Precision	Recall
Logistic Regression	FakeNewsDataset	97%	97%	97%	96%
	WELFAKE	95%	95%	96%	95%
	GossipCop	81%	82%	83%	81%
Random Forest	FakeNewsDataset	96%	96%	96%	96%
	WELFAKE	95%	95%	95%	94%
	GossipCop	82%	76%	76%	77%
Support Vector Machine – SVM	FakeNewsDataset	97%	97%	97%	97%
	WELFAKE	96%	95%	96%	96%
	GossipCop	80%	75%	74%	77%
LSTM	FakeNewsDataset	77%	77%	82%	78%
	WELFAKE	97%	97%	97%	97%
	GossipCop	76%	43%	38%	50%
DistilBERT	FakeNewsDataset	96%	95%	95%	94%
	WELFAKE	98%	98%	98%	98%
	GossipCop	85%	85%	84%	85%

Overall, these results confirm that while classical models are fast, reliable, and effective for well-structured data, they lack the contextual learning power required for noisy real-world text. Deep learning models, especially DistilBERT, consistently deliver superior generalization and adaptability.

Discussions and Findings

The performance of logistic regression, random forest, SVM, LSTM, and DistilBERT models was evaluated on three benchmark datasets, which are summarised in Figure 2.

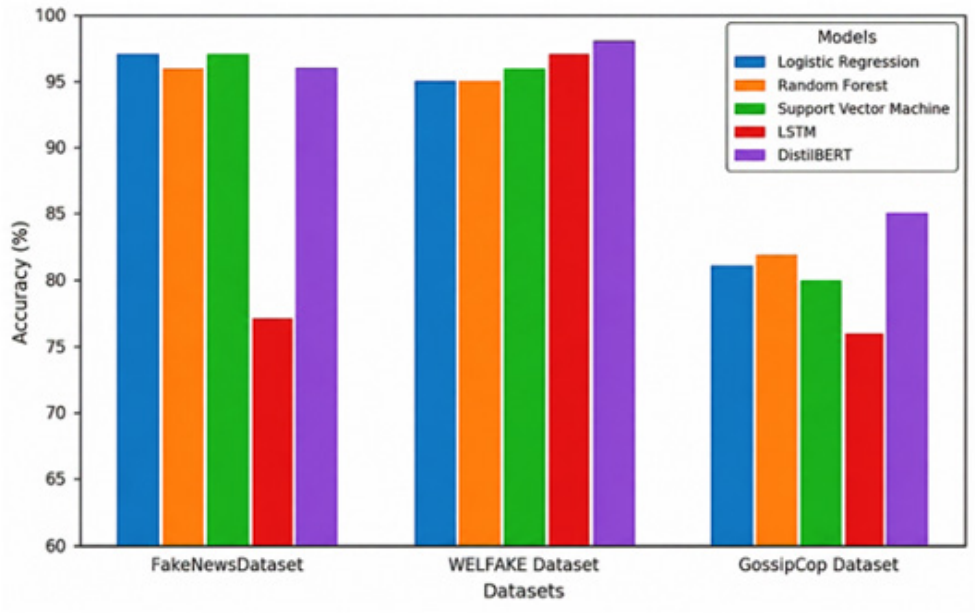

Figure 2: Comparison of different Models.

The conventional machine learning models performed well on the FakeNewsDataset. With 97% accuracy, logistic regression and SVM outperformed random forest and DistilBERT. The sequence-based model performed less well on this dataset, as evidenced by LSTM's 77%. This indicates that TF–IDF-based classical models can frequently manage structured datasets with distinct

lexical patterns. The DistilBERT model had the best accuracy of 98% on the WELFake dataset, followed by LSTM at 97%, SVM at 96%, logistic regression, and random forest at 95% accuracy. This suggests that additional contextual information is provided by longer and richer news stories, which helps transformer-based and deep learning models. The GossipCop dataset was the most challenging dataset. The performance of all models declined compared with the other datasets. The DistilBERT model achieved the strongest result with 85% accuracy and 85% F1-score. Logistic regression achieved 81% accuracy, random forest achieved 82%, SVM achieved 80%, and LSTM achieved 76%. The weaker performance on GossipCop appeared to be due to shorter text, informal writing, celebrity-focused content, and a more ambiguous language pattern.

Based on error analysis experiments, misclassifications are more likely to occur in short, informal, or confusing articles. This problem was most evident in the GossipCop dataset, where celebrity-related news frequently incorporates emotional terminology, clickbait-style phrasing, and insufficient contextual information. Traditional machine learning methods may misclassify such samples because TF-IDF features are primarily reliant on word-frequency patterns and cannot completely comprehend context. The LSTM model also performed poorly on GossipCop, possibly due to the limited sequential information given in shorter texts. The DistilBERT model outperformed this dataset because its attention mechanism is more adept at capturing contextual interactions between words. However, DistilBERT may misclassify articles that require external factual verification, background knowledge, or source credibility assessments.

Future evaluations should take out-of-sample testing circumstances into account to increase the proposed framework's practical applicability in people's everyday activities. While stratified training and testing divides provide a controlled experimental scenario, real-world fake news detection systems are frequently used to deal with novel topics, unexplored domains, and evolving patterns of disinformation. Cross-dataset evaluation, in which a model trained on one dataset is tested on another, is one practical case. For example, to investigate how effectively a model trained on FakeNewsDataset or WELFake generalises to noisier celebrity-related content, it may be tested on GossipCop. A time-based split is an additional situation in which more recent news articles are utilised for testing and older ones are used for training. Since misleading patterns evolve over time, this would more accurately reflect actual deployment. Determining where each model performs well or poorly can also be facilitated by domain-based testing, such as assessing political, health, entertainment, and general news independently.

Conclusion

The effectiveness of several machine learning models and deep learning approaches has been evaluated in this research study for fake news detection across well-trusted datasets with diverse textual characteristics. The research findings indicate that while classical machine learning models remain reliable baseline approaches, their effectiveness is largely dependent on the structure and quality of the data. These models perform well when textual content follows consistent patterns, but their performance declines when dealing with shorter, informal, or noisy text commonly found in online media sources. Deep learning models offer improved contextual understanding, particularly when analysing longer articles where relationships between words can be learned more effectively. The transformer-based DistilBERT model has outperformed the other models throughout all datasets processes, and in particular, the WELFake and GossipCop datasets. Traditional models, such as logistic regression and SVM, remained extremely competitive in the more structured FakeNewsDataset. Its contextual attention mechanism enables the model to capture semantic relationships within text more effectively than traditional feature-based methods, which allows it to better recognise complex false news and misinformation patterns.

The findings of this study highlight the increasing importance of transformer-based NLP models in addressing the challenges posed by misinformation in modern digital environments. In addition to model selection, the results emphasise the role of careful preprocessing and balanced datasets in improving the reliability of classification systems. While the proposed framework shows strong performance across multiple datasets, further advancements are necessary to improve the adaptability of fake news detection systems in real-world applications. Future research may focus on extending these models to multilingual environments, integrating multimodal information such as images and videos, and developing scalable real-time detection systems capable of analysing rapidly evolving online content. Such developments would contribute to more robust and practical misinformation detection frameworks that support trustworthy information dissemination in digital platforms.

Acknowledgements

The authors would like to express their sincere gratitude to the CDT staff of the Department of Computer Science and Digital Technologies (CDT) and ACE School Office Team of the School of Architecture, Computing and Engineering (ACE) at the University of East London – UEL for their invaluable support. In particular, the authors wish to thank Dr Aaron Kans, Dr Seyed Ali Ghorashi, Dr Nadeem Qazi, and Dr Mustansar Ghazanfar for their guidance, dedication, and professionalism in supporting this research study's conduct and dissemination.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

1. K Tian, G Rao, X Wang, M Yu, J Zhang, et al. (2025) CMFNThinker: A Novel Cross-source Multi-modal Fake News Detection Model, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE pp. 1–5.
2. A Gupta et al. (2022) Combating Fake News: Stakeholder Interventions and Potential Solutions. IEEE Access 10: 78268–78289.
3. H AbouGrad, S Santhosh, S Alsaid (2026) NLP Framework to Safeguard Youngsters Online Using Advanced Transformer-Based Models. Journal of Data Science and Intelligent Systems.

4. M Nasser et al. (2025) A systematic review of multimodal fake news detection on social media using deep learning models. Elsevier B.V.
5. A Matheven, BVD Kumar, (2022) Fake News Detection Using Deep Learning and Natural Language Processing, in 2022 9th International Conference on Soft Computing and Machine Intelligence, ISCMI 2022, Institute of Electrical and Electronics Engineers Inc pp. 11–14.
6. P Meesad (2021) Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. SN Comput Sci 2(6).
7. PK Verma, P Agrawal, V Madaan, R Prodan (2023) MCred: multi-modal message credibility for fake news detection using BERT and CNN. J Ambient Intell Humaniz Comput 14(8): 10617–10629.
8. H Abougrad, S Chakhar, A Abubahia. Decision Making by Applying Machine Learning Techniques to Mitigate Spam SMS Attacks. In Key Digital Trends in Artificial Intelligence and Robotics: Proceedings of 4th International Conference on Deep Learning, Artificial Intelligence and Robotics pp. 154–166.
9. R Baskar, S Sah, R Shyam, KS Kumar, H Patil, GN Reddy (2023) Advancements in Fake News Detection: Integrating NLP and Multi-Modal Approaches, in 2023 Intelligent Computing and Control for Engineering and Business Systems, ICCEBS 2023, Institute of Electrical and Electronics Engineers Inc.
10. A Chabukswar, PD Shenoy, KR Venugopal (2023) Fake News Detection Using Optimized Deep Learning Model Through Effective Feature Extraction, in 2023 International Conference on Recent Advances in Information Technology for Sustainable Development, ICRAIS 2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc pp. 118–123.
11. A Khalil, M Jarrah, M Aldwairi (2023) Hybrid Neural Network Models for Detecting Fake News Articles. Human-Centric Intelligent Systems 4(1): 136–146.
12. Q Chang, X Li, Z Duan (2024) Graph global attention network with memory: A deep learning approach for fake news detection. Neural Networks 172.
13. H AbouGrad, F Riaz (2026) Metadata-Enhanced Hybrid Fusion Architecture: Commercial Fake Reviews Detection Model Using Transformer Embeddings. FinTech and Sustainable Innovation pp. 1–8.
14. P Mittal, J Singh Saini, A Agarwal, RK Maheshwari, S Kumar, et al. (2024) Fake News Detection Using Machine Learning Techniques," in 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE), IEEE pp. 1374–1377.
15. ME Almandouh, MF Alrahmawy, M Eisa, M Elhoseny, AS Tolba (2024) Ensemble based high performance deep learning models for fake news detection," Sci Rep 14(1): 26591.
16. F Rustam, W aljedaani, AD Jurcut, S Alfarhood, M Safran et al. (2024) Fake news detection using enhanced features through text to image transformation with customized models. Discover Computing 27(1).
17. J Wang, X Wang, A Yu (2025) Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. Sci Rep 15(1).
18. S Maham, A Tariq, MUG Khan, FS Alamri, A Rehman, et al. (2024) ANN: adversarial news net for robust fake news classification. Sci Rep 14(1).
19. A Verma et al. (2025) ScrutNet: a deep ensemble network for detecting fake news in online text. Soc Netw Anal Min 15(1).
20. A Muqadas, HU Khan, M Ramzan, A Naz, T Alsahfi, et al. (2025) Deep learning and sentence embeddings for detection of clickbait news from online content. Sci Rep 15(1).
21. O Mairaj, SUR Khan (2025) Unveiling temporal patterns in information for improved rumor detection. Soc Netw Anal Min 15 (1).
22. H Himdi, N Zamzami, F Najar, M Alrehaili, N Bouguila (2025) Arabic fake news dataset development: humans and AI generated contributions. IEEE Access.
23. S Harris, HJ Hadi, N Ahmad, MA Alshara (2025) Multi-domain Urdu fake news detection using pre-trained ensemble model. Sci Rep 15(1).
24. P Zhu, J Hua, K Tang, J Tian, J Xu, et al. (2024) Multimodal fake news detection through intra-modality feature aggregation and inter-modality semantic fusion," Complex and Intelligent Systems 10(4): 5851–5863.
25. T Mahmud, T Akter, MT Aziz, M Kamal Uddin, MS Hossain, et al. (2024) Integration of NLP and Deep Learning for Automated Fake News Detection, in Proceedings - 2024 2nd International Conference on Inventive Computing and Informatics, ICICI 2024, Institute of Electrical and Electronics Engineers Inc pp. 398–404.
26. IK Sastrawan, IPA Bayupati, DMS Arsa (2021) Fake News Dataset. vol. 1.
27. (2025) Welfake dataset for fake news.
28. (2025) Fakenews-dataset/README.md at main · mbzuai-nlp/Fakenews-dataset.
29. R Sharma, V Sharma, TK Vashishth, Shashi, A Pandey, et al. (2025) Revealing the Reliability of Amazon Products via Innovative Fake Review Detection using Machine Learning, in Proceedings of 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2025, Institute of Electrical and Electronics Engineers Inc pp. 217–221.
30. S Akshara, S Shiva, S Kubireddy, T Arun, VVSL Kanthety (2023) A Small Comparative Study of Machine Learning Algorithms in the Detection of Fake Reviews of Amazon Products, in Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023, Institute of Electrical and Electronics Engineers Inc pp. 2258–2263.
31. H AbouGrad, A Shabarshov (2024) AI-Framework to Detect eCommerce Fake Reviews: A Hybrid Neural Network Machine Learning Model, Artificial Intelligence and Computational Technologies: Innovations, Usage Cases, and Ethical Considerations, Advances in Science, Technology & Innovation: IEREK Interdisciplinary Series for Sustainable Development.
32. İ Kulaksız, A Coşkunçay (2025) Fake News Detection on Mainstream Media Using Natural Language Processing," Black Sea Journal of Engineering and Science 8(1): 214–224.
33. M Tajrian, A Rahman, MA Kabir, MR Islam (2023) A Review of Methodologies for Fake News Analysis. IEEE Access 11: 73879–73893.