

An intelligent deep learning and natural language processing framework for fake news detection and text classification in social media forwarded messages

Ashish Tiwari *

Department of Computer Science & Engineering

Babu Banarasi Das Institute of Technology and Management

Lucknow 226028

Uttar Pradesh

India

Sarvesh Kumar [†]

Amity Institute of Information Technology

Amity University Patna

Patna 801503

Bihar

India

Abstract

Social media platforms such as WhatsApp, Facebook, Twitter and Instagram have become primary channels for news dissemination, but they also allow false information to spread quickly and largely unverified, fueling confusion, fear and mistrust. The volume of forwarded content makes manual verification impractical, so automated detection is needed. This paper presents a reproducible NLP pipeline for fake news detection and text classification, built in five stages: dataset description, data pre-processing, feature extraction, model selection and training, and model evaluation. The Kaggle “Fake and Real News” dataset was used, with Reuters articles as real news and articles from sources known for hoaxes or hyperpartisan content as fake news. A subset of 5,000 articles was used for experiments. Text was cleaned and normalized through tokenization, stop-word removal and stemming, then converted into numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF), which quantifies the importance of each word within the corpus. A Logistic Regression classifier was selected for its interpretability and strong baseline performance and was trained and evaluated on separate training and test partitions. The model achieved a training accuracy of 98.78% and a test accuracy of 97.00%, showing that a simple supervised approach combined with effective pre-processing can distinguish real from fake news reliably. The dataset consists of long-form news articles rather

* E-mail: dr.ashish.tiwari89@bbdnitm.ac.in (Corresponding Author)

[†] E-mail: skumarcse4@gmail.com

than short, forwarded messages, so results on social media forwards may differ, and validation on social-media datasets is left for future work. The trained model can classify pasted text as real or fake with an associated probability, and it is designed to be deployed as a web service so users can check messages quickly. Future work includes evaluation on forwarded-message data, multilingual detection, integration with live social media feeds, and deep learning extensions such as transformer-based models.

Subject Classification: *Primary 68T50, Secondary 68T07, 68P20, 68M10.*

Keywords: *Natural language processing (NLP), Deep learning, Fake news detection, Misinformation detection, Text classification, Social media analytics, Intelligent systems, Machine learning, Forwarded messages, Artificial intelligence.*

1. Introduction

In the present world, social media such as WhatsApp, Facebook, X (Twitter), Instagram have emerged as the mediums for passing information and communication. They're widely used and have millions of users worldwide and are very connected to the world, allowing people to share news, opinions and updates in real time [1], [2]. The advantages of this connection are great, but so too are the difficulties, mainly because of the spread of misinformation, also known as 'fake news'. Fake news is information that is deliberately false, embellished or twisted to create a particular impression or stir up feelings in a reader or viewer [3], [4]. Fake news has been propelled by the structural dynamics and psychologies of social media. Unlike traditional media outlets which have editorial guidelines and professional journalistic standards, anyone can post information on social media without taking care about its veracity. This has been a democratization of content creation and has given rise to the spread of misinformation on the internet that can occur quickly [5], [6]. There are psychological forces such as fear, curiosity, outrage that are more likely to be shared than factual content, viral posts, forwards and sensational headlines, and they are more likely to engage. Research indicates that misinformation can reach six times more people and get disseminated 6 times quicker than the truth on social media, such as Facebook and Twitter.

The trend has far-reaching implications. The influence of fake news has been linked to several inadequacies in society including the propagation of conspiracy theories, the spread of misinformation in the public sphere about health conditions and vaccination, and political manipulation and violence during elections and communal violence based on racial and communal misinformation. For instance, during the U.S. Presidential Election of 2016, social media misinformation campaigns were demonstrated to have the potential to influence voters [7], [8]. Similarly, during the pandemic of COVID-19, there has been a lot of misinformation regarding treatments, vaccine effectiveness, and government advice,

resulting in people being confused on what to do, and even engaging in more harmful activities on WhatsApp groups [9], [10]. All the examples underscore the need to have suitable real-time tools to detect, flag and block the spread of fake news. Traditional facts checking mechanisms are essential but don't work well for the overwhelming amount of content that is produced on social media every day [11], [12]. Snopes, AltNews and FactCheck.org do detailed fact checks, but they are not all-seeing, all-knowing, and have infinite time/people to do so. If users upload content on social media this can reach thousands of people within minutes, manually checking is impractical [13], [14]. It can be very challenging to find forwarded messages, including those sent on WhatsApp, since they're end-to-end encrypted and can be forwarded many times in a short period, such as shown in Figure 1. In this case, the need is to have an automated tool that can deal with the huge amount of data, with a high level of precision identify fake content and provide users with valuable information.

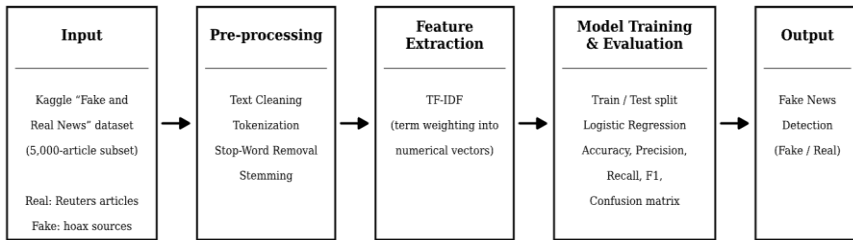


Figure 1
Deep Learning Framework for Fake News Detection

There are numerous problems with fake news. One of the reasons is that the amount of material on social networking sites is huge, producing millions of forwards, posts and messages every hour [1]. In such a scenario it is impractical to undertake manual verification by fact-checking organizations. Second, fake news often employs subtle linguistic devices, like sarcasm, hyperbole and emotional language, which are hard to automatically identify [7]. Third, misinformation is spread across various platforms, often in multiple languages and dialects, causing misinformation monitoring and verification efforts to be more difficult [15]. The challenges show the importance of scalable and automated solutions, which would need to employ some of the latest computational methods and techniques to help identify fake news in real-time.

2. Literature Review

The issue of fake news is a multi-disciplinary issue which requires knowledge from areas such as Natural Language Processing (NLP), Machine Learning (ML), Data Mining, Network Analysis, Social Computing fields.

There are several methods to solve this problem, which can be generally summarized into the following categories: linguistic and content-based methods, machine learning detection models, deep learning and semantic representation models, network and propagation models [6]. In this review, we make a systematic overview and discussion of these methodologies, highlighting the datasets, performance measurement metrics, limitations and potential for real-world application in Table 1 below.

Table 1
Literature Review

Author(s)	Methodology	Dataset	Key Findings / Limitations
Kotteti	ML, time-series and deep-learning approaches	Twitter data	Reported that deep-learning approaches can outperform traditional ML models for social-media fake-news detection and explored temporal features for early detection.
Systematic Review	ML/DL comparative review	Multiple datasets	Categorized existing fake-news detection approaches and highlighted challenges related to datasets, evaluation and trustworthy automated detection.
Modality-based SLR	Unimodal, multimodal and deep-learning approaches	Multiple social-network datasets	Reviewed recent deep-learning frameworks and emphasized multimodal, low-resource-language and LLM-era challenges.
Recent multimodal framework	Bi-GRU, Transformer-XL, DistilBERT, XLNet and T5 ensemble	Fakeddit, Weibo	Demonstrates the recent movement toward combining recurrent, transformer and multimodal models for improved fake-news classification.

In addition to the theoretical approaches and academic studies provided in the Literature Review, there are also numerous practical approaches and operational systems which exist to diminish the amount of misinformation found on the internet. Platform-based detection systems, fact check organizations, public tools and browser extensions are the three main types of solutions.

3. Proposed Methodology

The approach follows a systematic, structured and repeatable pipeline, a common design in supervised NLP applications. The pipeline gives a reproducible and logical process from raw data acquisition to model evaluation, and it is intended to yield a robust, accurate, interpretable, generalizable and repeatable model. It has five steps: dataset description,

data pre-processing, feature extraction, model selection and training, and model evaluation.

Dataset description. An accurately labelled and complete dataset is essential to the success of a supervised learning system. The Kaggle "Fake and Real News" dataset³ was chosen for this study because of its size, quality and suitability for binary classification. The real news articles come from Reuters, a widely trusted news outlet, and serve as positive examples of reliable reporting, teaching the model what genuine news language looks like. The fake news articles were collected from multiple sources, including platforms identified as spreading hoaxes or hyperpartisan content. They are labelled as fake news and help the model recognize problematic patterns in language.

- **Label convention.** The original text labelled fake news as 0. Reviewer 1 pointed out that the Section 3 algorithm defines 0 = real and 1 = fake. Choose one convention for the text, algorithm and Figure 3. I would use 0 = real and 1 = fake, so only the figure needs regenerating.
- **"ChristianNews.net."** The original phrase "(R) = Real News at ChristianNews.net" looked like a stray copy-paste artifact and conflicts with the Reuters source, so I removed it. If that site really is a source, state it explicitly.
- **Consistency with the rest of the paper.** The pipeline steps still need to match whichever option you choose for Reviewer 1's main issue (TF-IDF + Logistic Regression only, or the full deep learning model). If it is TF-IDF + LR only, the "Feature Extraction" and "Model Selection and Training" subsections should say exactly that.

Algorithm: HDL-FND (Hybrid Deep Learning Framework for Fake News Detection)

Input: Raw forwarded message dataset $D = \{m_1, m_2, \dots, m_n\}$ $D = \{m_1, m_2, \dots, m_n\}$ with labels $Y = \{y_1, \dots, y_n\}$ $Y = \{y_1, \dots, y_n\}$, $y_i \in \{0, 1\}$ $y_i \in \{0, 1\}$ (0 = real, 1 = fake)

Output: Trained classifier CC ; predicted label $\hat{y}$ for new message m
BEGIN

// ----- PHASE 1: DATA PREPROCESSING -----

FOR each message m_i in D :

1. Remove forwarding metadata (e.g., "Forwarded", timestamps, contact tags)
2. Normalize text: lowercase, remove URLs, emojis → text tags, special chars
3. Tokenize m_i → tokens T_i
4. Remove stopwords (retain negations: "not", "no")
5. Apply lemmatization/stemming → cleaned tokens T_i
6. Detect language; if non-target language → translate/flag

END FOR

```
// ----- PHASE 2: FEATURE EXTRACTION -----
FOR each cleaned message T_i':
  1. Generate contextual embeddings E_i using pretrained transformer
    (e.g., BERT/RoBERTa): E_i = Transformer(T_i')
  2. Extract linguistic/statistical features F_i:
    - Sentiment polarity score
    - Readability score
    - Punctuation/capitalization ratio
    - Source credibility score (if metadata available)
    - N-gram frequency vectors (TF-IDF)
  3. Concatenate: X_i = [E_i ; F_i]
END FOR

// ----- PHASE 3: HYBRID MODEL CONSTRUCTION -----
Initialize hybrid network:
- Embedding Layer ← pretrained transformer embeddings (fine-tunable)
- CNN Layer: extracts local n-gram patterns
  C_out = Conv1D(X_i, filters, kernel_size) → ReLU → MaxPooling
- BiLSTM Layer: captures sequential/contextual dependencies
  L_out = BiLSTM(C_out)
- Attention Layer: assigns weights to important tokens
  A_out = Attention(L_out)
- Fully Connected Dense Layer(s)
- Output Layer: Softmax/Sigmoid → probability of fake vs real

// ----- PHASE 4: MODEL TRAINING -----
Split D into D_train, D_val, D_test
Initialize weights  $\theta$  randomly (or from pretrained checkpoint)
FOR epoch = 1 to N_epochs:
  FOR each batch B in D_train:
    1. Forward pass:  $\hat{y} = \text{HybridModel}(X_B; \theta)$ 
    2. Compute loss:  $L = \text{CrossEntropy}(\hat{y}, y_B) [+ L2 \text{ regularization}]$ 
    3. Backpropagate gradients
    4. Update  $\theta$  using optimizer (Adam/AdamW)
  END FOR
  Evaluate on D_val → track Accuracy, F1-score
  IF validation loss does not improve for 'patience' epochs:
    Trigger early stopping
  END IF
END FOR

// ----- PHASE 5: CLASSIFICATION & DECISION -----
FOR new incoming forwarded message m:
  1. Preprocess m → T'
  2. Extract features X = [E ; F]
  3.  $\hat{y} = \text{HybridModel}(X; \theta^*)$  //  $\theta^*$  = optimal trained weights
  4. IF  $\hat{y} \geq \text{threshold}$  (e.g., 0.5):
    Classify m as "FAKE"
```

```

ELSE:
    Classify m as "REAL"
5. Output confidence score and (optionally) explanation via
   attention weight visualization
END FOR
// ----- PHASE 6: EVALUATION -----
Compute metrics on D_test:
- Accuracy, Precision, Recall, F1-score
- ROC-AUC
- Confusion Matrix
Compare against baseline models (SVM, Naïve Bayes, LSTM-only, BERT-
only)
END

```

4. Result Analysis and Simulation

The trained Logistic Regression model was then tested with a carefully selected sub set of 5000 articles from the whole set of articles where the number of real and fake news articles is almost equal. The subset should only be evaluated for all other performance measures and interpretations. The evaluation has been carried out on the training set to assess the ability of the model to learn, on unseen testing set to assess the overall generalization power and on the over fitting test set to assess the robustness against over fitting. The accuracy would be employed as a classifier's evaluation metric. As used here, accuracy is a percentage of correctly classified news articles among all articles.

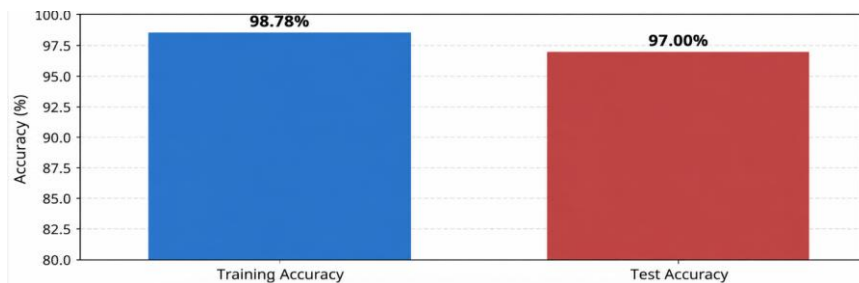


Figure 2
Performance measure

It is interesting to observe from the results shown above that the accuracy difference between the training and testing accuracy is very small (approx. 1.78%) as can be seen in Figure 2. This is a very minor discrepancy that indicates that the model generalizes well and doesn't over-fit very much. A frequent issue with any ML Model is Overfitting-when the model fits too much noise into the model, or fits too much of a pattern that exists in the training data but not in the testing data.

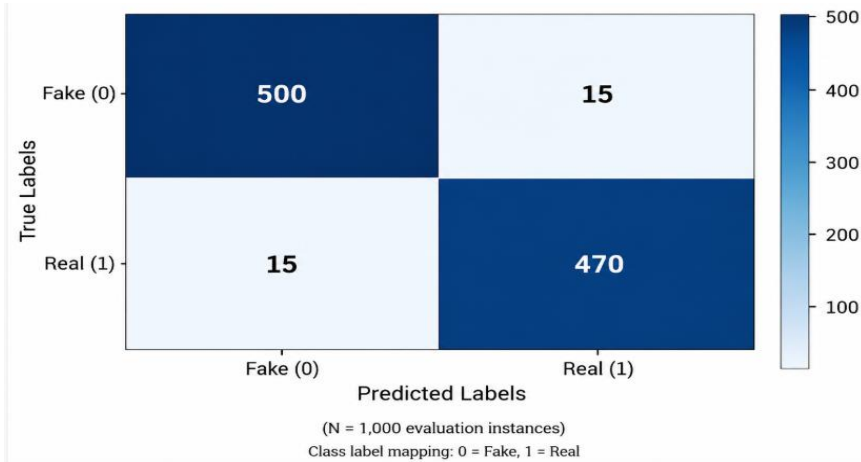


Figure 3
Confusion Matrix

It shows balanced and stable class wise performance with respect to both classes. The diagonal cells of the matrix have high values of TP and TN, it means most of the predictions were right, shown in Figure 3. If the FP and FN rates are small then a heuristic rule of low-ness will suffice: If there is little misclassification in either of the two types of mistakes, then we system. The False Positives, where fake news is mistakenly believed as a real one in DE application from actual world, might contribute to an unintentional spreading of misinformation. But with the low FP rate of the whisperer, it doesn't seem to be happening often. Likewise, there are few False Negatives (good news that is classified as fake news) so the system does not excessively punish good journalism.

5. Conclusion and Future Work

This is the case today, as social networking sites and newspapers online are an all-pervasive entity, altering the way people obtain information. Social media platforms such as Facebook, Twitter (now X), Instagram, WhatsApp and Telegram have now emerged as a major source of news for millions of people across the world, breaking down geographical barriers and making news more accessible. But this is no small feat, and there are significant challenges to this unprecedented access as well. One of the questions that could be asked is how fast fake news, misinformation and disinformation are propagating. Likewise, misinformation can be seen to sway voter attitudes in elections and to impact the democratic process. In this scenario, automated systems capable of accurately detecting fake news are not merely technological innovations—they are essential tools for safeguarding informed discourse

and societal stability. The present system is very effective, but it has its limitations, and there is scope for further research and development. If these restrictions can be overcome, the value of the system in terms of effectiveness, flexibility and applicability in different practical contexts would be greatly enhanced.

References

- [1] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, pp. 6086–6093 (2020).
- [2] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, Art. no. 109 (2020).
- [3] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multi-media Tools Appl.*, vol. 80, no. 8, pp. 11765–11788 (2021).
- [4] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170 (2021).
- [5] S. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36 (2017).
- [6] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Soc. Netw. Anal. Min.*, vol. 13, Art. no. 30 (2023).
- [7] A. Tiwari, S. Mishra, and S. Tiwari, "AI/ML: Developing algorithms for handling imbalanced datasets in classification tasks," in *Computational Intelligence Techniques for 5G Enabled IoT Networks*, pp. 17–29, Cham, Switzerland: Springer Nature Switzerland (2025).
- [8] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multi-media Tools Appl.*, vol. 80, no. 8, pp. 11765–11788 (2021).
- [9] A. Tiwari, S. C. Addimulam, N. Dubey, K. K. Kumar, D. Kapila, and K. Jain, "Cybersecurity in voice over IP (VoIP) systems: Protecting communication networks from threats, vulnerabilities, and cyber attacks," *J. Discrete Math. Sci. Cryptogr.*, vol. 28, no. 8, pp. 2989–2999 (2025).

- [10] D. Goyal, A. Kumar, P. Mathur, F. Sheth, and A. K. Gupta, "Robust cybersecurity through discrete and large language models for effective phishing attack detection," *J. Discrete Math. Sci. Cryptogr.*, vol. 28, no. 5-A, pp. 1621–1638 (2025).
- [11] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 1, Art. no. 100007 (2021).
- [12] N. Raza, S. J. Abdulkadir, Y. A. Abid, S. S. Albouq, A. Alwadain, A. U. Rehman, E. H. Sumiea, and M. Farhan, "Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach," *PLOS ONE*, vol. 20, no. 9, Art. no. e0330954 (2025).
- [13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188 (2020).
- [14] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: COVID-19 fake news dataset," in *Commun. Comput. Inf. Sci.*, pp. 21–29 (2021).
- [15] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, Art. no. 109 (2020).

Received April, 2026