AI-Based Fake News Detection Using Natural Language Processing: A Review and Analysis

Yash Sanap, Shivraj Thorat, Gaurav Sawant, Gaurav Gujar

D. Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune, India Department of Master of Computer Applications | Guide: Dr. Sanket Lodha

Abstract

Misleading news often reaches readers before its claims can be checked, which has encouraged research into artificial intelligence (AI) and natural language processing (NLP) as tools for flagging content that needs closer examination. This paper reviews 31 publications from 2011 to 2024 and organises their approaches into four groups: traditional machine learning, deep learning, transformer-based models, and multi-source or real-time systems. It then tests one concern that recurs in the literature, namely that scores measured on one dataset do not carry over to another. Four standard classifiers (Naive Bayes, Logistic Regression, Linear SVM, and Random Forest, all on TF-IDF features) were evaluated on two public datasets: Fake or Real News (2016 political articles) and CoAID (2020 COVID-19 articles).

Within Fake or Real News, five-fold macro-F1 ranges from 0.87 to 0.94. When a model trained on one dataset is tested on the other, macro-F1 falls to between 0.06 and 0.42. A model trained on an early CoAID snapshot also detects few fake articles in later snapshots, although the later test set holds only 15 fake articles, so that result is inconclusive. The highest-weighted terms of the linear models include source and formatting words, which suggests dataset-specific shortcuts. The paper also specifies CAMMS, a multilingual and multimodal framework with an explanation layer. CAMMS is described and given an evaluation plan, but it has not been implemented or evaluated here. The review finds no basis for naming one model family as best in every setting, and it points to cross-dataset testing, multilingual evaluation, transparent reporting, and human oversight as priorities.

Keywords: Fake news detection, natural language processing, machine learning, deep learning, transformers, BERT, explainability, multilingual detection, cross-dataset evaluation.

Introduction

Online platforms make it easy to publish information and share it with a large audience. The same convenience lets misleading stories circulate before anyone has examined the evidence behind them. Such stories can use familiar headlines, confident language, and a professional presentation, so reliability is hard to judge from appearance alone [26, 31].

AI-based detection systems can support the review process. Rather than replacing fact-checkers, they help identify articles or posts that deserve attention. Their role is to assist an investigation, not to issue a final judgement about whether a claim is true.

Many studies treat fake news detection as classification: a system receives a headline, article, or post and assigns a label, often “real” or “fake.” Traditional approaches combine numerical representations of text with classifiers such as Support Vector Machines (SVM), Naive Bayes, or Random Forest [11, 24]. Deep-learning methods learn patterns from training examples [8, 17], and transformer models such as BERT represent words in relation to their surrounding context [13, 16].

Published studies often report strong results, but success on one benchmark does not show that a detector will stay reliable in everyday use. New topics, unfamiliar sources, different languages, and changing misinformation tactics all create conditions that differ from the training data.

This paper does three things. It reviews 31 publications on AI- and NLP-based fake news detection and compares their reported strengths and limits. It measures, with experiments on two public datasets, how much the performance of four standard classifiers drops when the dataset, topic, or collection period changes. And it specifies a framework, CAMMS, aimed at the gaps that the review and the experiments identify. No new detection model is trained or evaluated.

Problem Statement

The volume and speed of online information make manual verification hard to carry out at scale. Automated tools may help, but their usefulness depends on more than a high accuracy score. A detector trained on one dataset may behave differently when it meets another language, topic, news source, or time period, because it can learn features that identify the dataset, such as a publisher's writing style, instead of features that identify misleading content.

The central problem addressed here is how to interpret the evidence from different detection studies without overlooking these differences, and how large the effect of a dataset change is for standard classifiers. A structured comparison is needed to understand what each approach offers, where its limits lie, and which questions need further investigation. Particular attention is given to dataset quality, generalisation, multilingual coverage, explainability, and practical deployment.

Objectives and Research Questions

This study has five objectives:

- Examine the AI and NLP methods used in the selected fake news detection studies.
- Organise these methods into broad approach families.
- Compare the reported datasets, text representations, evaluation measures, and findings.
- Measure how four standard classifiers perform within a dataset, across datasets, and across collection periods.
- Suggest research directions grounded in the gaps identified in the literature and in the experiments.

The study addresses the following research questions:

- RQ1: Which AI and NLP techniques are used in the selected studies?
- RQ2: How do traditional machine learning, deep learning, and transformer-based approaches differ in their reported performance and requirements?
- RQ3: Which datasets, languages, and evaluation measures are described?
- RQ4: How much does performance fall when a model meets a different dataset or a later collection period?
- RQ5: What research directions could improve transparency, robustness, multilingual support, and practical usefulness?

Three hypotheses are tested in the Results section. H1: classifiers that score high within one dataset score substantially lower in macro-F1 on a different dataset. H2: a classifier trained on an earlier CoAID snapshot detects fake articles in later snapshots less well than it does under in-snapshot cross-validation. H3: the highest-weighted terms of the linear models include source and formatting artefacts, not only claim content.

Review of Related Work

The selected literature covers several ways of approaching fake news detection. Some studies focus on conventional classifiers and carefully chosen text features. Others examine neural networks, contextual language models, or combinations of text and additional information. Survey papers give a broader view of these developments and the challenges that accompany them.

These approaches should not be seen as a simple progression in which every newer method replaces the earlier

ones. A conventional classifier may perform well on a particular task, while a more complex model may have advantages under other conditions. The summaries below therefore describe each study within its own reported setting. Several summaries are based on the abstracts and stated contributions of the papers, and they should be checked against the full texts before the paper is submitted.

Summary of Selected Studies

Surveys, taxonomies, and datasets.

Conroy, Rubin, and Chen (2015) [11]

This paper proposes an early taxonomy of automatic deception detection. It separates linguistic-cue approaches from network approaches and recommends combining them.

Shu et al. (2017) [26]

This survey reviews fake news detection on social media from a data-mining perspective, covering content, user, and propagation features and the datasets then available. It illustrates a recurring difficulty: studies use different datasets and evaluation procedures, so results are hard to compare directly.

Zhou and Zafarani (2020) [31]

This survey links detection methods to theories about fake news from the social sciences and sets out open problems.

Agrawal, Pandey, and Goyal (2021) [2]

This survey examines the use of machine learning and NLP to detect misinformation on social media. It gives an overview of common techniques, but the survey itself does not establish whether a model will transfer to another language or domain.

Alnabhan and Branco (2024) [5]

This systematic review covers deep-learning approaches, including CNNs, RNNs, graph neural networks, attention mechanisms, and BERT. It also examines datasets, transfer learning, and class imbalance, and draws attention to unresolved issues in evaluation and model selection.

Castillo, Mendoza, and Poblete (2011) [9]

This study assesses the credibility of information on Twitter using content and user features, and is an early example of feature-based classification of social-media content.

Wang (2017) [29]

This paper releases LIAR, a benchmark of short political statements labelled by a fact-checking site, and tests models that use text alone and models that add speaker metadata.

Karimi et al. (2018) [18]

This work builds a dataset from several sources with several label classes and proposes a method for multi-source, multi-class detection.

Pérez-Rosas et al. (2018) [24]

This paper collects two news datasets, analyses linguistic differences between real and fake articles, and tests classifiers across domains, where performance falls.

Cui and Lee (2020) [12]

This paper releases CoAID, which pairs COVID-19 news articles and social-media posts with fact-checking labels. It is one of the two datasets used in this study.

Traditional machine learning.

Yuslee and Abdullah (2021) [30]

This paper examines Naive Bayes with TF-IDF and count-vector features, with and without n-grams, and reports that TF-IDF gave higher precision on its data. It shows that a simple probabilistic classifier can work well when the text representation suits the dataset.

Ozbay and Alatas (2020) [23]

This study applies twenty-three supervised algorithms to text features extracted from social-media news, after a pre-processing and feature-extraction pipeline.

Reis et al. (2019) [25]

This paper builds classifiers on features describing the text, the source, and the surroundings of a story, and discusses which features drive the predictions. It is one of the few studies that treats explanation as part of the task.

Deep learning.

Ma et al. (2016) [20]

This study uses recurrent neural networks to detect rumours in microblog streams, treating a post and its replies as a sequence.

Bahad, Saxena, and Kamal (2019) [8]

This study evaluates a bidirectional LSTM recurrent network for news classification and reports high accuracy. Its findings support the use of recurrent models for sequence dependencies, but leave open how the model performs on data outside the evaluated distributions.

Ajao, Bhowmik, and Zargari (2018) [3]

This paper combines convolutional and recurrent layers to identify fake news in tweets.

Nasir, Khan, and Varlamis (2021) [22]

This paper proposes a hybrid CNN-RNN model, with the convolutional layers extracting local features and the recurrent layers modelling their order.

Kaliyar et al. (2020) [17]

This study presents FNDNet, a deep convolutional neural network for fake news detection. It shows that deep learning can extract complex patterns from text, although generalisation remains an open question.

Goldani, Safabakhsh, and Momtazi (2021) [14]

This study trains a convolutional network with a margin loss for fake news detection and evaluates it on several benchmark datasets.

Agarwal et al. (2020) [1]

This study explores a blend of neural networks for detection and shows how dataset characteristics and the choice of deep-learning model affect classification results.

Umer et al. (2020) [28]

This paper applies a CNN-LSTM design to fake news stance detection, a sub-task that compares a headline with the body of an article.

Monti et al. (2019) [21]

This paper applies geometric deep learning to propagation graphs on social media, so that the way a story spreads, and not only its text, informs the prediction.

Transformer-based models.

Kaliyar, Goswami, and Narang (2021) [16]

This study introduces FakeBERT, which combines BERT with convolutional layers. It shows how transformer-based contextual representations can capture the semantics of deceptive news.

Aljawarneh and Swedat (2024) [4]

This study presents an enhanced BERT approach with transfer learning for news-text classification. Whether the approach stays effective on unfamiliar sources or topics requires separate testing.

Azizah et al. (2023) [6]

This paper compares BERT, ALBERT, and RoBERTa for fake news detection on Indonesian news. It adds a non-English evaluation, although its findings cannot be extended to other languages without further evidence.

Multi-source and real-time systems.

Silva et al. (2021) [27]

This work uses multi-modal data to address domain differences in fake news and tests across domains. The extra inputs give context that text alone does not capture, but they raise the cost and complexity of data collection.

Babar et al. (2023) [7]

This study builds a real-time detection system on a big-data pipeline with a deep neural network. Judging such a system needs response time and reliability as well as classification performance.

Three further works supply methods used or proposed here: BERT [13], the multilingual model XLM-RoBERTa [10], and the SHAP method for explaining predictions [19].

Review Methodology

This paper combines a narrative review of selected literature with a small empirical study. It does not follow a PRISMA-style systematic protocol, because search dates and screening counts were not logged. The review should therefore be read as a review of selected literature and not as an exhaustive account of all relevant research.

The literature was selected purposively from peer-reviewed venues and public preprints that publish on fake news detection, including IEEE, ACM, Elsevier, Springer, AAAI, and ACL venues. Search terms included “fake news detection” and “misinformation detection,” combined with NLP, machine learning, deep learning, BERT, transformers, multilingual, and multimodal. A study was kept if it proposed, evaluated, or surveyed an algorithmic detection method or released a benchmark. Non-scholarly web material and blogs were excluded. The studies were grouped by approach and compared in terms of methods, datasets, representations, reported metrics, and limitations. Accuracy, precision, recall, and F1-score are treated as study-specific findings and are not combined into a single ranking.

Experimental Design

Two public English datasets were used (Table 1). Fake or Real News (FoR) holds news articles, mostly about the 2016 US election, labelled FAKE or REAL [15]. CoAID holds COVID-19 news articles and posts with fact-check-based labels [12]; only its news-article records were used, from three collection snapshots (1 May,

1 September, and 1 November 2020). FoR predates 2020, while CoAID is from 2020.

Table 1: Datasets used in the experiments

Dataset	Articles used	Fake	Real	Notes
FoR	6,060	3,071	2,989	6,335 raw records; duplicate texts removed; near-balanced; political topics
CoAID (articles)	2,298	126	2,172	Articles over 200 characters; duplicates removed; 5.5% fake; health topics

Each document is the headline followed by the body text. Text is lower-cased and English stop words are removed. A TF-IDF vectoriser (sublinear term frequency, minimum document frequency of 2, at most 50,000 features) is fitted on training data only. Four classifiers are compared: Multinomial Naive Bayes, Logistic Regression, Linear SVM, and Random Forest (200 trees). The last three use balanced class weights because CoAID is highly imbalanced. The random seed is 42, and the implementation uses Python 3 and scikit-learn 1.9.1.

Five protocols were run. E1 and E2 are stratified five-fold cross-validation on FoR and on CoAID. E3 trains on one dataset and tests on the other, in both directions. E4 trains on the 1 May CoAID snapshot and tests on the September and November snapshots. E5 lists the highest-weighted terms of the logistic regression on each full dataset. Precision, recall, and F1 are reported for the fake class, and macro-F1, the mean F1 of both classes, is the main summary because a large majority class does not inflate it. Cross-validation results are mean ± standard deviation over folds. E3 and E4 are single splits.

Comprehensive Comparative Analysis

Table 2: Approach families in the reviewed literature

Approach Family	Key Technologies Used	Main Strengths	Main Limitations
Traditional ML	SVM, Random Forest, Naive Bayes, TF-IDF, linguistic features	Computationally efficient; competitive when the features suit the dataset.	Relies on manual feature design; depends on dataset composition.
Deep Learning	CNN, LSTM, BiLSTM, hybrid CNN-RNN, graph models	Learns representations from text; models sequence dependencies.	Needs more training data; few studies test across datasets.
Transformers	BERT, ALBERT, RoBERTa, XLM-R	Contextual representations; transfer learning.	Higher computing cost; mostly English; sensitive to dataset bias.
Multi-source and Real-Time	Text with images or metadata; streaming pipelines	Added context; timely alerts.	More data collection and governance; response time and reliability must also be shown.

Results and Discussion

Within-Dataset Results

On FoR, the best five-fold macro-F1 is 0.939 (Linear SVM), and all four models score 0.87 or higher (Table 3). Naive Bayes is the weakest and the linear models are the strongest. This agrees with the high single-dataset scores that the reviewed studies report.

Table 3: Five-fold cross-validation on Fake or Real News (mean ± SD)

Model	Accuracy	Precision	Recall	F1 (fake)	Macro-F1
Naive Bayes	0.873 ± 0.010	0.968 ± 0.006	0.775 ± 0.022	0.861 ± 0.013	0.872 ± 0.011
Logistic Regression	0.922 ± 0.003	0.907 ± 0.007	0.943 ± 0.009	0.925 ± 0.003	0.922 ± 0.003
Linear SVM	0.939 ± 0.003	0.933 ± 0.006	0.949 ± 0.011	0.941 ± 0.003	0.939 ± 0.003
Random Forest	0.902 ± 0.003	0.906 ± 0.007	0.901 ± 0.008	0.903 ± 0.003	0.902 ± 0.003

CoAID gives a different picture (Table 4). Only 5.5% of its articles are fake, so accuracy says little: always predicting "real" would score 94.5%. Naive Bayes shows the effect. Its accuracy is 94.6%, but its fake-class F1 is 0.015, because it almost never predicts "fake." Logistic Regression has the best fake-class F1 (0.717).

Table 4: Five-fold cross-validation on CoAID articles (mean ± SD)

Model	Accuracy	Precision	Recall	F1 (fake)	Macro-F1
Naive Bayes	0.946 ± 0.001	0.200 ± 0.400	0.008 ± 0.016	0.015 ± 0.031	0.494 ± 0.016
Logistic Regression	0.969 ± 0.007	0.728 ± 0.077	0.715 ± 0.083	0.717 ± 0.059	0.850 ± 0.031
Linear SVM	0.973 ± 0.004	0.902 ± 0.043	0.572 ± 0.062	0.698 ± 0.051	0.842 ± 0.026
Random Forest	0.966 ± 0.003	0.822 ± 0.021	0.476 ± 0.070	0.600 ± 0.055	0.791 ± 0.028

Cross-Dataset Results

Performance falls sharply when the model meets the other dataset (Table 5, Figure 1). Trained on FoR and tested on CoAID, macro-F1 is between 0.055 and 0.207, and accuracy is between 5.7% and 22.4%, well below the 94.5% that a majority-class guess would give. These models have high fake recall and about 5% precision, which means they label most COVID-19 articles as fake. In the other direction, macro-F1 is between 0.337 and 0.418, close to chance for a balanced two-class task. These results support H1.

Table 5: Cross-dataset results (single train/test split)

Direction	Model	Acc.	Prec.	Recall	F1 (fake)	Macro-F1
FoR → CoAID	Naive Bayes	0.224	0.050	0.730	0.093	0.207
FoR → CoAID	Logistic Regression	0.065	0.052	0.937	0.099	0.063
FoR → CoAID	Linear SVM	0.116	0.050	0.841	0.094	0.115
FoR → CoAID	Random Forest	0.057	0.054	0.976	0.102	0.055
CoAID → FoR	Naive Bayes	0.494	0.606	0.007	0.013	0.337
CoAID → FoR	Logistic Regression	0.425	0.458	0.738	0.565	0.357
CoAID → FoR	Linear SVM	0.421	0.436	0.484	0.459	0.418
CoAID → FoR	Random Forest	0.437	0.463	0.699	0.557	0.393

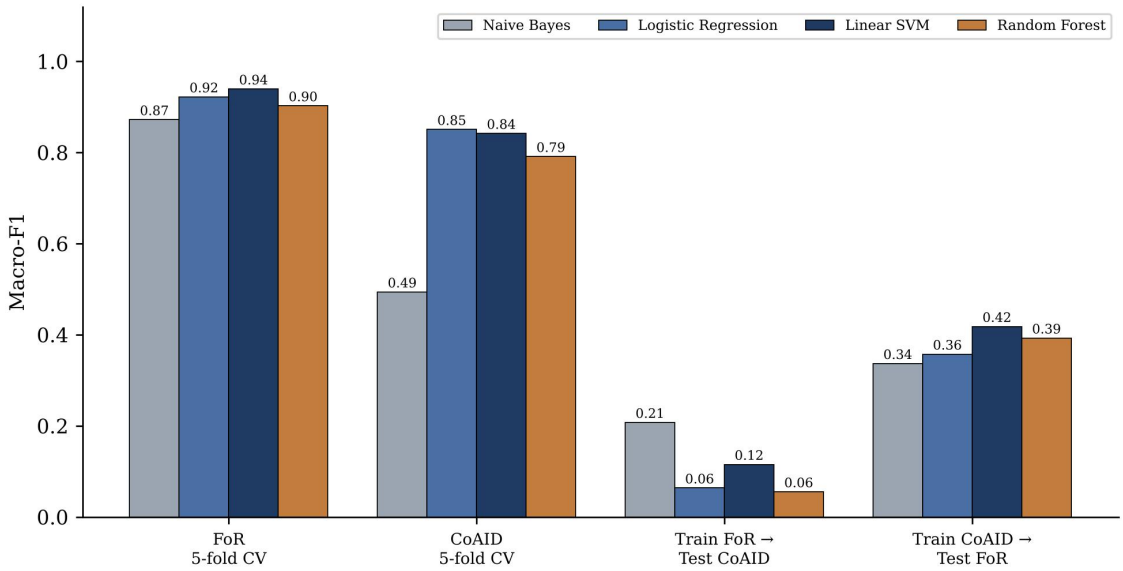


Figure 1: Macro-F1 of four classifiers within and across datasets

Two cautions apply. The datasets differ in topic (politics against health), in how labels were produced, and in class balance, so the drop combines several shifts that this design does not separate. The model trained on CoAID also sees few fake examples, which limits what it can learn. The result shows that in-dataset scores do not transfer. It does not identify a single cause.

Results on Later Collection Snapshots

The model trained on the 1 May 2020 snapshot (1,427 articles, 111 fake) was tested on the later snapshots (871 articles, 15 fake). Accuracy is high (97.1% to 98.3%) only because 98.3% of the test articles are real. Fake-class F1 is 0.000 to 0.214, against 0.717 for Logistic Regression in cross-validation (Table 6). The direction fits H2, but the test set has only 15 fake articles, so one or two predictions change the F1 considerably. H2 is neither confirmed nor rejected.

Table 6: Train on the 1 May 2020 CoAID snapshot, test on later snapshots

Model	Accuracy	Precision	Recall	F1 (fake)	Macro-F1
Naive Bayes	0.983	0.000	0.000	0.000	0.496
Logistic Regression	0.975	0.231	0.200	0.214	0.601
Linear SVM	0.978	0.167	0.067	0.095	0.542
Random Forest	0.971	0.188	0.200	0.194	0.589

What the Models Rely On

The terms with the largest weights in the logistic regression models are listed in Table 7. On FoR, the fake class is linked to “2016,” “october,” “november,” “share,” “article,” “com,” and “source,” and the real class to “said,” “gop,” and “tuesday.” Many of these describe how an article is written or where it appeared, not whether its claims are true. On CoAID, the real class carries “medscape,” “getty,” “images,” and “reading,” which look like site boilerplate. This supports H3 in a qualitative sense. No ablation was run to remove these terms, so how much of the score they account for is unknown.

Table 7: Highest-weighted terms in the logistic regression models

Dataset	Terms pointing to fake	Terms pointing to real
FoR	2016, october, hillary, election, november, article, share, wikileaks, russia, com, fbi, source	said, obama, gop, cruz, republican, debate, tuesday, state, fox, republicans, president, conservative
CoAID	coronavirus, president, china, chinese, government, natural, positive, want, trump, man, fatality, lab	19, covid, 2020, april, data, disease, experts, images, medscape, getty, based, reading

Comparison Across Approach Families

Table 8: Reported concerns in the literature compared with the evidence from the experiments

Concern in the literature	Related studies	Evidence from the experiments
Single-dataset scores overstate reliability	[26, 5]	Macro-F1 falls from 0.87–0.94 within FoR to 0.06–0.21 on CoAID
Cross-domain degradation	[24, 27]	Seen in both transfer directions
Class imbalance hides weak detection	[5]	High accuracy with near-zero fake F1 on CoAID (Naive Bayes, later snapshots)
Dataset artefacts and weak explanations	[25]	Source and format terms appear among the top features
Change over time	[5]	Direction consistent; evidence limited by 15 fake test articles

Pérez-Rosas et al. [24] reported lower performance in cross-domain tests, and Silva et al. [27] built their method around domain differences. The experiments here agree with both and add a measurement on two further datasets. Scores from the experiments are not compared numerically with scores in other papers, because datasets, label definitions, and splits differ.

Interpreting Reported Performance

The reviewed findings do not support a simple ranking based on accuracy. A larger percentage does not, by itself, show that one model is more capable than another. Studies differ in language, sample composition, label definitions, preprocessing, and evaluation procedure, and a meaningful head-to-head comparison needs shared conditions, including the same test data and metrics. The experiments show how large the gap can be: the same Linear SVM scores 0.94 macro-F1 on one test set and 0.12 on another.

Looking Beyond Accuracy

Where labels are imbalanced, an overall score can hide weak performance on the less common class. The temporal experiment is an example, with accuracy above 97% and almost no fake articles found. Precision, recall, F1-score, and class-level results make these differences visible, and cross-dataset testing shows whether a system has learned patterns that remain useful elsewhere.

Practical Use and Human Review

Text-based classification offers a starting point, while images, captions, and metadata can add context [27], at the cost of more demanding data collection. Real-time detection [7] adds further requirements: an alert must arrive quickly enough to be useful, but speed does not establish reliability. A practical system must also communicate uncertainty and let people examine questionable predictions. The most appropriate role for these tools is to support review and prioritisation. A classification label should not be presented as equivalent to evidence-based fact-checking.

Major Challenges

Multilingual and Mixed-Language Content

The Indonesian-language study [6] broadens the review beyond English, but one language-specific evaluation cannot establish wider multilingual capability. None of the reviewed studies evaluates a multilingual transformer on Indian languages, and both datasets used here are English. Reliability must be shown for the language and setting in which a system will be used.

Changes Across Topics and Time

A model may meet news that differs substantially from its training examples. Topics change, sources change,

and misinformation takes new forms. The cross-dataset results above show how large the loss can be, and testing on later material is one way to examine it.

Explainability

A prediction tells a reviewer which label the model chose, but not why. Useful explanations should help people inspect the basis of a decision and find possible errors, without implying that the model verified a claim because it can produce a convincing explanation. The term lists above show a simple case: the explanation reveals that the model uses source words.

Dataset Bias and Class Imbalance

Dataset construction influences what a model learns and how its performance is measured. Uneven representation of topics, sources, or classes can make a result less informative than it first appears. Clear documentation and evaluation across different datasets are needed to understand these effects.

Multimodal and Changing Misinformation

Combining text with other signals may improve context, but it also raises concerns about data quality, privacy, fairness, and complexity. None of the reviewed studies tests detectors against paraphrased or AI-generated misinformation, so existing benchmark results should not be assumed to carry over.

Inconsistent Evaluation

Differences in datasets, metrics, and experimental procedures make the literature hard to compare. Without shared evaluation conditions, claims that one approach is generally superior remain uncertain.

Directions for Future Work

The gaps identified in the reviewed literature and in the experiments suggest the following priorities:

- Evaluate models on datasets other than those used during development.
- Expand regional-language and mixed-language testing, with clear documentation of collection and labelling.
- Use time-based evaluation to examine performance on later news.
- Report precision, recall, F1-score, and class-level results alongside accuracy, and state the majority-class baseline.
- Inspect top features or token attributions and remove source and format artefacts where possible.
- Assess computing requirements, response time, and explanatory usefulness.
- Test resilience to paraphrasing and AI-generated misinformation.
- Explain how uncertain predictions are communicated and reviewed by people.
- Document search and selection procedures clearly enough for future reviews to be reproduced.

Proposed Work: Hybrid Multimodal Framework

To address the gaps identified, particularly generalisation, multilingual coverage, and explainability, this study proposes the Context-Aware Multilingual and Multimodal System (CAMMS). The framework has two pipelines (Figure 2). A Multilingual Text Module fine-tunes XLM-RoBERTa [10] so that one model can process English and regional-language articles. A Contextual Metadata Module takes information about the publisher and any attached media, such as the source's credibility history and features of the caption or image, where these are available.

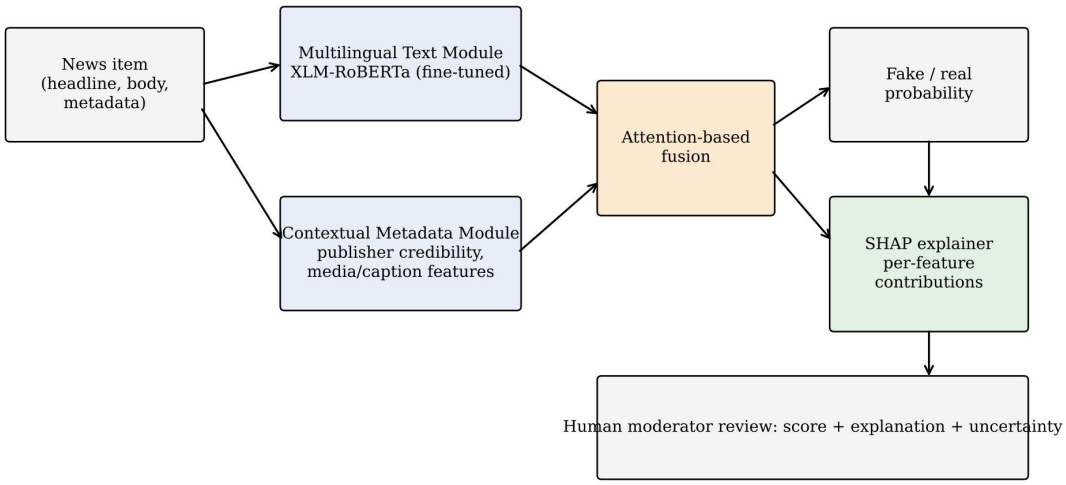


Figure 2: Architecture of the proposed CAMMS framework

An attention layer weights the two representations before a classifier outputs a probability. A SHAP-based explainer [19] attributes the prediction to input tokens and metadata features, so that a human moderator receives a score, the contributing features, and an indication of uncertainty instead of an opaque binary label.

CAMMS is a specification. It has not been implemented or evaluated in this paper, and no performance claim is made for it. It will be tested under the same five protocols as the baselines so that results are comparable, with an additional test that fine-tunes on English data and evaluates on a regional-language dataset whose collection and labelling are documented. The metadata module needs publisher and media fields that neither FoR nor CoAID provides in usable form, so a suitable dataset must be found or assembled first. The baseline results above set the reference point: a transformer result is informative only relative to what TF-IDF models achieve under each protocol.

Conclusion

This paper reviewed 31 publications on AI- and NLP-based fake news detection and tested one of the concerns they raise. The studies cover traditional machine learning, deep learning, transformer-based models, and multi-source or real-time systems, and different approaches perform well in particular settings. The reported results do not establish a universally best model.

The experiments give a measured example of why. Four standard classifiers reach macro-F1 of 0.87 to 0.94 within Fake or Real News, and between 0.06 and 0.42 when tested on a different dataset, so in-dataset scores should not be read as evidence of reliability in use. The models also lean partly on source and format terms. The evidence on change over time is inconclusive because of the small number of fake articles in the later CoAID snapshots.

The study has limits. Only classical models were run, both datasets are English, FoR dates from 2016, and the cross-dataset drop mixes topic, labelling, and class-balance shifts. Each model used one seed and one setting, with no tuning or significance tests, and the literature review is purposive and not exhaustive. CAMMS remains a specification. Future work should run it under the protocols described, add a regional-language dataset, and report cross-dataset and time-based results alongside in-dataset ones. No reported result from the literature was independently reproduced.

References

[1] Agarwal, A., Mittal, M., Pathak, A., & Goyal, L. M. (2020). Fake news detection using a blend of neural networks: An application of deep learning. SN Computer Science, 1(3), Article 143.

[2] Agrawal, C., Pandey, A., & Goyal, S. (2021). A survey on role of machine learning and NLP in fake news detection on social media. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 1–6.

[3] Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on Twitter with hybrid CNN and RNN models. Proceedings of the 9th International Conference on Social Media and Society, 226–230.

[4] Aljawarneh, S. A., & Swedat, S. A. (2024). Fake news detection using enhanced BERT. IEEE Transactions on Computational Social Systems, 11(4), 4843–4850.

[5] Alnabhan, M. Q., & Branco, P. (2024). Fake news detection using deep learning: A systematic literature review. IEEE Access, 12, 114435–114459.

[6] Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023). Performance analysis of transformer based models (BERT, ALBERT, and RoBERTa) in fake news detection. 2023 6th International Conference on Information and Communications Technology (ICOIACT), 425–430.

[7] Babar, M., Ahmad, A., Tariq, M. U., & Kaleem, S. (2023). Real-time fake news detection using big data analytics and deep neural network. IEEE Transactions on Computational Social Systems, 11(4), 5189–5198.

[8] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Computer Science, 165, 74–82.

[9] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web, 675–684.

[10] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

[11] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.

[12] Cui, L., & Lee, D. (2020). CoAID: COVID-19 healthcare misinformation dataset. arXiv:2006.00885.

[13] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[14] Goldani, M. H., Safabakhsh, R., & Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 58(1), Article 102418.

[15] Hamel, L. (n.d.). Fake news dataset [Data set]. GitHub. https://github.com/lutzhamel/fake-news

[16] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788.

[17] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet: A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32–44.

[18] Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. Proceedings of the 27th International Conference on Computational Linguistics, 1546–1557.

[19] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[20] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. Proceedings of the 25th International Joint

Conference on Artificial Intelligence, 3818–3824.

[21] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv:1902.06673.

[22] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), Article 100007.

[23] Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, Article 123174.

[24] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics, 3391–3401.

[25] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable machine learning for fake news detection. Proceedings of the 10th ACM Conference on Web Science, 17–26.

[26] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

[27] Silva, A., Luo, Y., Karunasekera, S., & Leckie, C. (2021). Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. Proceedings of the AAAI Conference on Artificial Intelligence, 35(1), 557–565.

[28] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B.-W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access, 8, 156695–156706.

[29] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 422–426.

[30] Yuslee, N. S., & Abdullah, N. A. S. (2021). Fake news detection using Naive Bayes. 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET).

[31] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), Article 109.