



July 2026

Comparative Evaluation of Transformer-Based Hybrid Models for Arabic and English Fake News Classification

Authored under supervision of:

Dr. Yasser Abdelhamid,

By:

Mohammed Al-jaaidi

Mohammed AbuSultan

Khalid Abbas Elzein

Master's Students, College of Informatics

Department of Artificial Intelligence

Midocean University

Contents

Abstract.....	3
Introduction	5
Objectives.....	7
Background	7
Fake News as a Text Classification Problem	7
Traditional Machine Learning Approaches	8
Deep Learning and Hybrid Models	8
Transformer-Based Language Models	8
Arabic Language Challenges.....	8
Dataset Bias and Generalisation	9
Review Scope and Study Selection	9
Related Work.....	10
Traditional Machine Learning and Linguistic Features.....	10
CNN and Recurrent Neural Networks	11
Transformer-Based Classification.....	11
Arabic Pretrained Language Models and Ensembles	12
Multimodal, Multilingual, and Explainable Approaches.....	13
Dataset Dependence and Generalisation.....	13
Comparative Synthesis of the Reviewed Studies	14
Research Trends Within the Reviewed Set	14
Interpretation of the Research Trends.....	21
Research Gap and Basis for the Proposed Study.....	21
Problem Statement.....	22
Research Questions and Hypotheses	23
Main Research Question	23
Research Hypotheses	23
Hypothesis Evaluation	24
Significance of the Study	24
Methodology	25
Research Design.....	25
Datasets and Task Definition	25
Data Preparation.....	26
Experimental Splits.....	26

Models.....	27
Training Procedure.....	28
Evaluation.....	28
Robustness and Error Analysis	29
Dataset Profile and Labelling Structure	29
Preprocessing Pipelines.....	31
Proposed Framework and Model Architecture	33
1. Text Input and Tokenisation	33
2. Transformer Encoder.....	33
3. CNN Branch.....	33
4. BiLSTM Branch	34
5. Feature Fusion	34
6. Classification Layer	34
Ablation Structure.....	36
Sequence Length and Long Articles	36
Experimental Design Matrix.....	36
Evaluation Tracks.....	37
Experimental Environment	38
Statistical Analysis.....	38
Ethical Considerations	39
Conclusion.....	39
References	41
 <i>Figure 1: Proposed Parallel Transformer-CNN-BiLSTM Architecture.....</i>	 35

Abstract

Fake news classification has progressed from traditional machine learning and manually designed linguistic features to deep learning, pretrained transformers, and hybrid neural architectures. Despite the strong performance reported by transformer-based models, results vary considerably across languages, datasets, and experimental settings. It also remains unclear whether adding convolutional and recurrent components to transformer representations provides a consistent improvement that justifies the additional computational cost.

This study reviews recent research on Arabic and English fake news classification and proposes a controlled comparative evaluation of transformer-based hybrid architectures. AFND will be used as the Arabic experimental case with a language-specific Arabic transformer, while WELFake will be used as the English case with BERT-base-uncased. The comparison will include TF-IDF-based Logistic Regression and Linear Support Vector Machine, CNN-BiLSTM with pretrained word embeddings, transformer-only classification, transformer-CNN, transformer-BiLSTM, and parallel transformer-CNN-BiLSTM.

All comparable models will be evaluated using consistent data partitions, preprocessing rules, training conditions, and repeated runs. Macro F1-score will serve as the primary evaluation measure, supported by accuracy, class-level precision, recall, weighted F1-score, ROC-AUC, and confusion matrices. Training time, inference time, parameter count, model size, and peak memory consumption will also be recorded. Duplicate content, class imbalance, article length, and source-related bias will be examined, including publisher-disjoint evaluation for AFND where the available metadata permits it.

The study treats the task as supervised fake news classification based on the labels provided by the selected datasets. It does not perform independent claim verification or use external evidence. The main contribution will be a controlled assessment of whether CNN and BiLSTM branches provide statistically stable and practically meaningful improvements over transformer-only baselines.

Keywords: Fake News Classification, Arabic Natural Language Processing, BERT, AraBERT, CNN-BiLSTM, Transformer-Based Hybrid Models, Comparative Evaluation.

Category	Keywords
Research Domain	Fake News Detection, Misinformation Detection, Text Classification
AI Techniques	Artificial Intelligence, Machine Learning, Deep Learning
NLP	Natural Language Processing, Arabic NLP
Proposed Model	BERT, CNN, BiLSTM, Transformer Models
Research Scope	Arabic Fake News Detection, English Fake News Detection

Introduction

Digital news platforms and social media have accelerated the publication and distribution of information. Although this development has improved access to news, it has also increased the speed at which false and misleading content reaches large audiences. The volume of online material makes continuous manual verification difficult, which has encouraged the development of automated methods for fake news classification.

Earlier classification systems relied mainly on traditional machine learning algorithms such as Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest. These approaches commonly represented text through bag-of-words, TF-IDF, or manually selected linguistic features. Deep learning models later reduced dependence on manual feature engineering by learning local and sequential textual patterns through CNN, LSTM, and BiLSTM architectures.

Pretrained transformer models introduced contextual language representations and achieved strong performance in many text classification tasks. BERT and Arabic-specific models such as AraBERT (Antoun et al., 2020) provide representations that account for the surrounding context of each token.

Recent fake news studies have therefore combined transformers with CNN, recurrent networks, ensemble methods, and multimodal components. The additional complexity of these models does not necessarily produce consistent improvement across datasets, languages, or evaluation settings.

Arabic fake news classification presents additional challenges related to morphology, spelling variation, dialectal diversity, and the limited availability of consistently labelled datasets. Dataset construction also affects model evaluation. When labels are associated with publisher credibility, or when duplicated and source-related content appears across training and testing partitions, a classifier may learn source identity and editorial style rather than general indicators of misleading content.

This study reviews recent Arabic and English fake news classification research and proposes a controlled comparison of traditional, deep learning, transformer-only, and transformer-based hybrid models. AFND and WELFake will be treated as separate Arabic and English experimental cases. The study will examine whether CNN and BiLSTM branches provide stable improvement over transformer-only baselines and whether any improvement justifies the additional computational cost.

The study treats the task as supervised fake news classification according to the labels provided by the selected datasets. It does not perform independent factual verification or retrieve external evidence. Its contribution lies in the controlled evaluation of model components, computational requirements, source-related bias, and generalisation under consistent experimental conditions.

The study makes four main contributions. First, it provides a structured synthesis of recent Arabic and English fake news classification research. Second, it applies the same ablation-based

comparison to separate Arabic and English experimental cases using language-appropriate transformers. Third, it evaluates predictive performance together with computational cost and repeated-run stability. Fourth, it examines whether duplicate content and publisher-related patterns produce optimistic estimates of generalisation

Objectives

The general objective of this study is to conduct a controlled evaluation of transformer-based hybrid architectures for Arabic and English fake news classification and determine whether their performance gains justify the additional computational cost.

The specific objectives are:

1. To examine the structure, labels, class distribution, missing values, duplicates, article lengths, and available source information within AFND and WELFake.
2. To prepare separate Arabic and English data-processing pipelines suitable for traditional machine learning, deep learning, and transformer-based models.
3. To implement TF-IDF-based traditional baselines, CNN-BiLSTM baselines, transformer-only models, and transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM architectures.
4. To conduct a controlled ablation comparison using consistent data partitions, training procedures, evaluation measures, and predefined random seeds.
5. To evaluate classification performance, statistical stability, and computational cost using macro F1-score as the primary measure, supported by class-level metrics, confidence intervals, training time, inference time, memory consumption, parameter count, and model size.
6. To examine the influence of duplicate content and source-related bias through deduplication, publisher-disjoint evaluation on AFND, and qualitative error analysis.
7. To determine which architecture provides the most defensible balance between predictive performance, computational cost, and generalisation.

Background

Fake News as a Text Classification Problem

Fake news is a broad term that refers to false or misleading content presented in a form resembling legitimate news. Computational datasets use the term for different categories, including fabricated articles, misleading reports, satire, rumours, manipulated content, and news published by sources classified as unreliable. The exact meaning of the label therefore depends on the collection and annotation procedure used by each dataset.

Most automated studies formulate the task as supervised text classification. A model learns patterns from labelled examples and assigns a class to previously unseen text. This differs from automatic fact verification, which requires external evidence and an assessment of the factual claims contained in an article.

The present study evaluates classification performance according to the labels provided by AFND and WELFake. Its results will therefore reflect the operational definitions of these datasets rather than independent verification of article content.

Traditional Machine Learning Approaches

Traditional fake news classification systems used algorithms such as Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest. Text was commonly represented through bag-of-words, TF-IDF, or manually selected linguistic features.

These approaches provide computationally efficient and interpretable baselines. Their performance, however, often depends on the selected features and the characteristics of the training dataset. Linguistic indicators associated with one publisher, topic, or period do not necessarily transfer to other sources or domains.

Deep Learning and Hybrid Models

Deep learning reduced dependence on manually designed features by learning textual representations directly from labelled data. CNN models identify local expressions and short token patterns, while LSTM and BiLSTM models represent sequential relationships across longer passages.

Hybrid CNN and recurrent architectures combine local and sequential information. Previous studies reported strong results on some benchmark datasets, but performance varied substantially across sources and domains. This variation indicates that architectural complexity alone does not guarantee reliable generalisation.

Transformer-Based Language Models

Transformer models produce contextual representations through self-attention, allowing each token to be represented in relation to the surrounding sequence (Devlin et al., 2019). BERT and related models have become strong baselines for fake news classification, while Arabic pretrained models provide representations adapted to Arabic vocabulary and morphology.

Adding CNN or BiLSTM layers after a transformer encoder provides additional local or sequential processing. These components also increase parameter count, training time, memory consumption, and inference cost. Their value must therefore be established through direct comparison with a transformer-only baseline.

Arabic Language Challenges

Arabic text classification is affected by rich morphology, prefixes and suffixes, derived word forms, spelling variation, and differences between Modern Standard Arabic and regional dialects. Unicode variation, optional diacritics, elongation characters, and inconsistent letter forms introduce further variation in written text.

Preprocessing therefore requires different treatment for traditional and transformer-based models. Traditional representations often benefit from normalisation and feature selection, while aggressive stemming or stop-word removal may remove information expected by a pretrained transformer tokenizer.

Arabic fake news datasets also differ in size, source composition, language variety, annotation procedure, and label definition. These differences limit direct comparison between reported results and increase the importance of documenting how each dataset defines article credibility.

Dataset Bias and Generalisation

Model performance depends partly on how a dataset was constructed and divided. A random article-level split may place articles from the same publisher, topic, or original collection in both the training and test sets. Under these conditions, a classifier may recognise recurring source characteristics rather than learn patterns that transfer to unseen publishers.

Exact duplicates and near-duplicate articles create an additional risk. When closely related versions of one article occur in different partitions, the reported test result may overestimate performance on genuinely unseen content.

Previous studies reported large differences across datasets such as ISOT, WELFake, LIAR, FA-KES, and ANS (Nasir et al., 2021; Merzah et al., 2026; Rout et al., 2025). These differences show that within-dataset performance and generalisation should be evaluated separately. Duplicate removal and publisher-disjoint evaluation are therefore included in the proposed experimental design.

Review Scope and Study Selection

This study uses a structured narrative review to examine recent developments in Arabic and English fake news classification. The review focuses on studies published between 2020 and 2026 that evaluate traditional machine learning, deep learning, transformer-based, hybrid, multilingual, or multimodal approaches relevant to the proposed comparison.

The initial study pool was assembled through broad academic web searches and targeted searches of official publisher repositories, including ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, Frontiers, MDPI, and PLOS.

The search focused on publications between 2020 and 2026 related to fake news classification, Arabic Natural Language Processing, transformer-based models, CNN and recurrent architectures, hybrid models, dataset bias, and generalisation. The titles, abstracts, full texts, publication details, and DOI records were reviewed to determine their relevance to the proposed study.

Studies were considered when they met the following criteria:

1. They addressed fake news, manipulated news, satire classification, news credibility, or closely related misinformation tasks.
2. They reported a machine learning, deep learning, or transformer-based method.
3. They included identifiable datasets, evaluation measures, or methodological findings.
4. They contributed directly to Arabic classification, English benchmark comparison, hybrid architecture design, dataset generalisation, or computational-cost analysis.
5. The full publication was available for review.

Studies were excluded when they provided insufficient methodological detail, did not evaluate a classification approach, duplicated another selected publication, or focused on a task with limited relevance to the proposed text-classification experiment.

Following the screening process, 15 studies were retained for detailed comparative analysis.

The selected studies were grouped according to their dominant technical contribution. Information was extracted on the research goal, language, dataset, preprocessing, model architecture, evaluation measures, main results, limitations, and relevance to the current study.

Related Work

The selected studies are organised according to their dominant technical contribution. The review begins with traditional machine learning and linguistic features, followed by CNN and recurrent architectures, transformer-based classification, Arabic pretrained models, and multimodal or multilingual systems. The final part examines dataset dependence and generalisation across sources and benchmarks.

This thematic structure supports comparison between studies that use different languages, datasets, labels, and evaluation procedures. Their reported scores are interpreted within the conditions of each experiment rather than treated as directly comparable results.

Traditional Machine Learning and Linguistic Features

Traditional machine learning remains relevant because it provides relatively simple and computationally efficient baselines. Himdi et al. (2022) examined Arabic fake news using emotion, polarity, parts of speech, negation, certainty, intensification, and reporting style. Random Forest produced the strongest performance among the evaluated classifiers. The study showed that deceptive Arabic content contains useful linguistic indicators, but the reliance on manually designed features limits transfer to other publishers, dialects, and topics.

Saadany et al. (2020) examined Arabic satirical news using lexical and grammatical features. Their models achieved high accuracy when distinguishing articles from selected satirical and

mainstream publishers. The result demonstrated that satire has identifiable stylistic characteristics. It also raised a methodological concern because the classifier may have learned the editorial styles of a small number of publishers rather than a general distinction between factual and misleading reporting.

Nagoudi et al. (2020) studied the generation and detection of manipulated Arabic news. Their work introduced an Arabic resource and evaluated both human readers and automatic classifiers. The strongest automatic result remained limited compared with results reported on some English benchmark datasets. This suggests that manipulated Arabic content presents a difficult classification problem and that performance depends on the way synthetic examples are produced.

These studies established the importance of lexical, stylistic, and grammatical signals. They also exposed the limits of fixed features. A feature associated with one source or type of misinformation may not retain the same meaning when the model encounters another topic, dialect, or publisher.

CNN and Recurrent Neural Networks

Deep learning methods reduced dependence on manual feature engineering. CNN models were used to detect local expressions and short token patterns, while LSTM and BiLSTM models were used to represent longer sequential relationships.

Nasir et al. (2021) combined CNN and recurrent layers using pretrained GloVe embeddings. The model achieved almost perfect performance on ISOT but considerably lower accuracy on FA-KES. The difference showed that architecture alone does not determine performance. Dataset composition, article style, source distribution, and task difficulty also have a major effect.

Merzah et al. (2026) evaluated a parallel CNN-BiLSTM architecture with FastText embeddings on Arabic and English data. The model achieved 98.85% accuracy on WELFake, 94.43% on AFND, and 71.63% on ANS. Ablation experiments indicated that the CNN and BiLSTM branches captured complementary information. At the same time, the difference between AFND and ANS showed that strong performance on one Arabic dataset did not guarantee similar performance on another.

The findings support the use of combined local and sequential features. They do not establish that CNN-BiLSTM architectures generalise consistently across sources and languages. Both studies also relied on static word embeddings rather than contextual transformer representations.

Transformer-Based Classification

Transformer models changed text classification by producing contextual representations in which the meaning of a token depends on the surrounding sequence. BERT and related models have therefore become common baselines in fake news research.

Raza et al. (2025) evaluated a progressively trained BERT model on WELFake. The model achieved 95.3% accuracy with balanced precision, recall, and F1-score. The study demonstrated that a carefully designed training procedure can improve performance without adding several architectural components. Its evaluation was limited to one English dataset, so the results did not establish cross-language or cross-domain generalisation.

Roumeliotis et al. (2025) compared CNN, BERT, GPT-4o, and GPT-4o-mini using a balanced sample from WELFake. The fine-tuned language models achieved the strongest results, while the basic CNN performed considerably worse. BERT remained close to the larger models while requiring a more accessible experimental setup. The use of a limited English sample restricts the extent to which the result can be transferred to other datasets or languages.

Essa et al. (2023) combined BERT representations with LightGBM. Their model achieved high performance on ISOT, TI-CNN, and the Fake News Corpus. The work showed that contextual transformer features can support an efficient external classifier. The architecture was not trained as one end-to-end neural model, and its near-perfect ISOT result must be interpreted in relation to the structure and source distribution of that dataset.

Rout et al. (2025) developed a compact transformer architecture and evaluated it on ISOT, FakeNewsNet, and LIAR. The model achieved 100% accuracy on ISOT, 80% on FakeNewsNet, and 59.56% on LIAR. The variation is more informative than the highest score. It shows that a model can appear highly effective on one benchmark while struggling with shorter, more ambiguous, or differently labelled content.

Together, these studies support transformers as strong contextual baselines. They also show that reported performance reflects both model capability and dataset characteristics. A strong transformer result should therefore be established before additional CNN or BiLSTM components are judged useful.

Arabic Pretrained Language Models and Ensembles

Arabic pretrained models have been developed to represent Arabic morphology, vocabulary, and language variation more effectively than models trained mainly on English content.

Al-Zahrani and Al-Yahya (2024) evaluated an ensemble of AraBERT, MARBERT, AraELECTRA, AraGPT2, and ARBERT. The strongest ensemble achieved an F1-score of 94% using weighted averaging. The result suggests that the participating models learned partly different decision boundaries. The study also found that adding more models did not always improve performance. Ensembles increase storage, inference time, and deployment complexity.

Touahri and Mazroui (2024) reviewed Arabic fake news classification methods, datasets, preprocessing procedures, linguistic features, deep learning models, and Arabic transformers. Their review found no single architecture that was consistently superior across all experiments. The authors identified limited public datasets, inconsistent annotation procedures, dialect variation, and domain dependence as continuing challenges.

Arabic transformer research therefore provides evidence that language-specific pretraining improves contextual representation. It does not remove the influence of data quality, annotation strategy, source distribution, or computational cost.

Multimodal, Multilingual, and Explainable Approaches

Recent studies have expanded the task beyond plain article text. Images, text embedded in images, propagation networks, social relationships, external knowledge, and explanation techniques have been included in more complex systems.

Al-alshaqi et al. (2025) used Optical Character Recognition to extract written content from images and combine it with article text. Their strongest result was obtained on TRUTHSEEKER, while performance declined on the merged dataset. The difference indicates that combining heterogeneous sources created a more difficult classification setting. The model processed textual information extracted from images, but it did not fully address visual manipulation or misleading relationships between images and articles.

Ganesh Karthik et al. (2025) combined a modified transformer with hybrid optimisation on Fakeddit and DGM4. The model outperformed several baselines, but the training process required multiple optimisation and fine-tuning stages. The results differed between the two datasets, which again shows that optimisation does not remove dataset dependence.

Jadhav et al. (2025) proposed a multilingual, multimodal, and explainable system using XLM-RoBERTa, CNN, BiLSTM, ResNet50, GraphSAGE, and attention-based fusion. The system improved on the transformer baseline and provided explanation outputs. Its architecture contained approximately 420 million parameters and required substantial GPU resources. Arabic was not included in the evaluated languages.

Lv et al. (2025) reviewed multimodal fake news detection across text, images, video, audio, social context, propagation structures, graphs, external knowledge, and large language models. Their survey identified annotation quality, heterogeneous fusion, interpretability, and computational cost as continuing problems.

These studies show the benefits of using information beyond article text. They also require datasets and infrastructure outside the present project scope. Their results are therefore relevant to understanding the direction of the field, but they are not direct baselines for the proposed text-only experiment.

Dataset Dependence and Generalisation

One repeated finding across the reviewed literature is the difference between within-dataset performance and generalisation. High results were frequently reported on ISOT and WELFake, while the same or similar architectures produced lower results on LIAR, FA-KES, ANS, merged datasets, and cross-domain settings.

These differences arise partly from the task itself. Some datasets contain complete articles, while others contain short claims. Some use binary labels, while others represent several levels

of truthfulness. Some labels result from individual fact verification, while others rely on the credibility of the publishing source.

AFND assigns article labels according to the credibility category of their publishers (Khalil et al., 2022). This allows a large Arabic dataset to be created, but it introduces a risk that a model will learn publisher vocabulary and editorial style. WELFake combines articles from several earlier datasets, which creates a separate risk of source-related patterns and duplicated or closely related content (Verma et al., 2021).

A random article-level split may place content from the same source in both training and testing data. The resulting score would measure performance on unseen articles but not necessarily on unseen publishers. Publisher-disjoint evaluation and duplicate analysis are therefore important when interpreting the reported results.

Comparative Synthesis of the Reviewed Studies

The reviewed studies differ in language, dataset construction, task definition, model architecture, and evaluation procedure. Their reported results should therefore be interpreted within the conditions of each experiment rather than ranked as if they were obtained from one shared benchmark. **Table I: Comparative Summary of the Reviewed Studies** summarises the datasets, approaches, reported results, methodological limitations, and relevance of each study to the proposed research.

Research Trends Within the Reviewed Set

The reviewed studies were grouped according to their dominant technical contribution. Some papers include methods belonging to more than one category, but each study was assigned to the direction that best represents its main research focus. **Table II: Dominant Research Trends Within the Reviewed Studies** summarises the distribution, strengths, and repeated limitations observed within the reviewed set.

Table I: Comparative Summary of the Reviewed Studies

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Himdi et al. (2022)	Arabic, crowdsourced news dataset	Linguistic features with Naive Bayes, SVM, and Random Forest	The dataset focused on a specific news context and relied heavily on manually designed textual features.	Narrow dataset and dependence on manually selected features	Supports the traditional Arabic baseline and linguistic error analysis
Saadany et al. (2020)	Arabic satire and genuine news from selected publishers	Lexical and grammatical features with machine learning	Accuracy reached up to 98.6%	The dataset was constructed from a limited set of satirical and mainstream publishers, creating a risk of publisher-style learning.	Demonstrates the risk of learning source style instead of general credibility
Nagoudi et al. (2020)	Arabic manipulated news, AraNews	Machine generation and automatic detection	Best automatic detector achieved 70.06% macro F1	The manipulated articles were generated through a defined synthetic procedure that may not represent all forms of real-world misinformation.	Supports discussion of machine-generated misinformation and Arabic data limitations

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Nasir et al. (2021)	English, ISOT and FA-KES	CNN-RNN with GloVe embeddings	The hybrid CNN-RNN achieved near-perfect performance on ISOT, while its performance on FA-KES was approximately 60%.	The large performance difference between the two datasets indicates severe dataset dependence, while the model relies on static GloVe embeddings.	Shows the value of hybrid features and the problem of dataset dependence
Merzah et al. (2026)	Arabic and English, AFND, ANS, and WELFake	Parallel CNN-BiLSTM with FastText	Accuracy reached 98.85% on WELFake, 94.43% on AFND, and 71.63% on ANS.	The architecture relied on static FastText embeddings and showed substantial performance variation across datasets.	Closest architectural basis for the proposed parallel branches
Raza et al. (2025)	English, WELFake	Progressively trained BERT	The model achieved 95.3% accuracy, 0.952 precision, 0.954 recall, and 0.953 F1-score.	The evaluation was limited to WELFake and did not examine cross-dataset or cross-language generalisation.	Supports the transformer-only baseline
Roumeliotis et al. (2025)	English, balanced WELFake sample	CNN, BERT, GPT-4o, and GPT-4o-mini	Fine-tuned GPT-4o models achieved approximately 98.6% accuracy, while the basic CNN achieved approximately 58.6%; BERT also produced strong results.	Small selected sample and one language	Shows the strength and practical relevance of a BERT baseline

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Essa et al. (2023)	English, ISOT, TI-CNN, and Fake News Corpus	BERT embeddings with LightGBM	BERT representations combined with LightGBM achieved 99.88% accuracy on ISOT, 96.94% on TI-CNN, and 99.06% on the Fake News Corpus.	The architecture was not trained as a single end-to-end neural model, and the results may reflect dataset-specific source patterns.	Supports comparison between contextual embeddings and hybrid classification
Rout et al. (2025)	English, ISOT, FakeNewsNet, and LIAR	Compact attention-based transformer	The lightweight transformer achieved 100% accuracy on ISOT, 80% on FakeNewsNet, and 59.56% on LIAR.	Performance varied severely across datasets despite the compact architecture.	Strong evidence for dataset sensitivity and the need for controlled evaluation
Al-Zahrani and Al-Yahya (2024)	Arabic multisource dataset	Ensemble of five Arabic pretrained language models	The weighted-average ensemble achieved a 94% F1-score.	The ensemble increases storage, inference cost, and deployment complexity, while adding more models produced only limited additional improvement.	Supports language-specific Arabic transformers and cost analysis

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Touahri and Mazroui (2024)	Arabic survey	Review of datasets, preprocessing, ML, DL, and transformers	The review found substantial variation across datasets, preprocessing procedures, and model families, with no universally superior approach.	The paper is a survey and does not provide a controlled common experimental benchmark.	Supports the Arabic research gap and the need for consistent evaluation
Al-alshaqi et al. (2025)	English, ISOT, WELFake, TRUTHSEEKER, and merged data	OCR text fusion with BERT	The strongest performance was reported on TRUTHSEEKER, while performance decreased substantially on the merged heterogeneous dataset.	The model extracts textual content from images but does not fully model visual manipulation or image-text inconsistency.	Shows the effect of combining sources and defines the text-only boundary
Ganesh Karthik et al. (2025)	Multimodal, Fakeddit and DGM4	Modified transformer with hybrid optimisation	The model achieved 90.42% accuracy on Fakeddit and 86.74% on DGM4.	The method requires a complex optimisation and multi-stage fine-tuning pipeline.	Supports computational-cost evaluation

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Jadhav et al. (2025)	Multilingual and multimodal data in Gujarati, Hindi, Marathi, and Telugu, with English included in selected cross-lingual analyses.	XLM-RoBERTa, CNN, BiLSTM, visual and graph models	The architecture contains approximately 420 million parameters and requires substantial GPU resources.	Approximately 420 million parameters and high GPU requirements	Shows potential value of hybrid branches and the cost of large architectures
lv et al. (2025)	Multimodal survey	Review of text, image, video, audio, graphs, and external knowledge	The survey identified continuing challenges in multimodal fusion, annotation quality, interpretability, and computational requirements.	The study is a survey and does not provide a common experimental comparison.	Positions multimodal analysis as future work outside the present scope

Table II: Dominant Research Trends Within the Reviewed Studies

Dominant Research Direction	Number of Studies	Share of Reviewed Set	Main Technical Focus	Repeated Strength	Repeated Limitation
Traditional machine learning and linguistic features	3	20.0%	Manual lexical, grammatical, and stylistic features	Lower computational cost and clearer feature interpretation	Limited transfer across sources, topics, and dialects
CNN, RNN, and hybrid deep learning	2	13.3%	Local and sequential feature learning using static embeddings	Complementary representation of short and long patterns	Strong dependence on dataset and embedding quality
Transformer-based text classification	4	26.7%	Contextual representations and transformer fine-tuning	Strong within-dataset classification performance	High variation across benchmarks and domains
Arabic pretrained models and ensembles	2	13.3%	Language-specific transformers and prediction combination	Better representation of Arabic language variation	Limited datasets and higher ensemble cost
Multimodal, multilingual, explainable, and optimisation-based systems	4	26.7%	Text-image fusion, graph learning, explainability, and optimisation	Richer evidence and broader task coverage	High computational demand and complex data requirements

Interpretation of the Research Trends

The reviewed literature shows a clear movement from manually designed linguistic features toward contextual transformer representations and more complex hybrid systems.

Transformer-centred approaches represent the largest direction within the reviewed set when standard transformer classification, Arabic pretrained models, and transformer-based multimodal systems are considered together.

Greater architectural complexity did not eliminate dataset dependence. Several studies reported very high results on ISOT and WELFake, while substantially lower performance appeared on LIAR, FA-KES, ANS, merged datasets, and cross-domain evaluations. These differences indicate that model performance is influenced by dataset structure, source composition, annotation procedure, and task difficulty.

Arabic research has progressed from linguistic features and conventional classifiers toward Arabic pretrained transformers and ensembles. Its main limitations remain the small number of public datasets, inconsistent annotation procedures, dialect variation, and source-related bias. The reviewed studies also provide limited controlled evidence about whether the same hybrid components produce comparable benefits when applied separately to Arabic and English data.

The main research opportunity is therefore not the simple combination of BERT, CNN, and BiLSTM. It is the controlled evaluation of each component under consistent experimental conditions, using language-appropriate transformers, repeated runs, computational measurements, and publisher-disjoint analysis.

Research Gap and Basis for the Proposed Study

The reviewed literature confirms that traditional machine learning, CNN, recurrent networks, transformers, ensembles, and multimodal systems have all been applied to fake news classification. It also shows that reported performance depends strongly on the selected language, dataset, annotation method, preprocessing procedure, and evaluation design.

Several research gaps remain.

First, many studies compare models developed under different experimental conditions. Differences in preprocessing, data partitions, hyperparameter tuning, and evaluation measures make it difficult to determine whether reported improvements result from the model architecture itself.

Second, hybrid architectures often combine several components without measuring the contribution of each component separately. A high result from a transformer-CNN-BiLSTM model does not show whether the improvement was produced by the transformer encoder, the CNN branch, the BiLSTM branch, or their interaction. A controlled ablation comparison is therefore required.

Third, limited research applies the same comparative structure separately to Arabic and English using language-appropriate transformer encoders. Existing bilingual studies often rely on static

word embeddings, while Arabic transformer research commonly focuses on one dataset or an ensemble of pretrained models.

Fourth, model selection is usually based mainly on accuracy or F1-score. Training time, inference time, memory consumption, parameter count, and model size receive less attention, even though a small performance improvement may not justify a substantial increase in computational cost.

Fifth, dataset construction may produce optimistic results. AFND associates article labels with publisher credibility, while WELFake combines records from several earlier collections. Models may therefore learn publisher style, duplicated content, topic patterns, or collection-specific characteristics rather than indicators that generalise to unseen sources.

The proposed study addresses these gaps through a controlled comparison of transformer-only, transformer-CNN, transformer-BiLSTM, and parallel transformer-CNN-BiLSTM architectures. AFND and WELFake will be treated as separate Arabic and English experimental cases. The models will use consistent preprocessing, data partitions, training rules, evaluation measures, and repeated runs within each language.

The contribution of the study will be an evidence-based assessment of whether CNN and BiLSTM branches provide stable and practically meaningful improvements over transformer-only baselines. A result favouring a simpler transformer model would be as relevant as a result favouring the complete hybrid architecture.

Problem Statement

Transformer-based models have achieved strong results in fake news classification, but reported performance varies substantially across datasets, languages, and experimental settings. Direct comparison between previous studies is difficult because they often use different preprocessing procedures, data partitions, model configurations, and evaluation measures.

Several studies have added CNN or BiLSTM components to transformer representations in order to capture local and sequential textual patterns. A higher score from a hybrid architecture does not establish that every added component contributes useful information. The performance gain must be measured against a transformer-only baseline under the same experimental conditions and interpreted alongside the additional computational cost.

The available evidence is also limited across Arabic and English. Arabic differs from English in morphology, spelling variation, language resources, and dataset construction. A hybrid architecture that improves performance on an English dataset does not necessarily provide the same benefit when applied to Arabic content with a language-specific transformer.

Dataset structure introduces an additional concern. AFND associates article labels with publisher credibility, which creates a risk that models learn source style rather than article-level

credibility. WELFake combines records from several earlier datasets and may contain duplicated content or collection-specific patterns. Random article-level splitting may therefore produce optimistic results when related content or recurring source characteristics appear in both training and testing data.

The research problem is the lack of controlled evidence showing whether CNN and BiLSTM branches provide stable, statistically supported, and computationally justified improvements over transformer-only baselines for Arabic and English fake news classification.

The study addresses this problem through a controlled ablation-based comparison using AFND and WELFake as separate experimental cases. Model performance, computational cost, duplicate content, and source-related bias will be examined without assuming in advance that the complete hybrid architecture is superior.

Research Questions and Hypotheses

The study is guided by the following main research question:

Main Research Question

Do CNN and BiLSTM branches provide stable and computationally justified improvements over transformer-only baselines for Arabic and English fake news classification?

The following research questions support the main question:

RQ1: Do transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models achieve higher mean macro F1-scores than the transformer-only baseline on AFND and WELFake?

RQ2: What contribution does each CNN and BiLSTM branch make when evaluated separately and in combination?

RQ3: Are the relative effects of the hybrid components consistent across the Arabic and English experimental cases?

RQ4: What additional training time, inference time, memory consumption, parameter count, and model size are introduced by each hybrid architecture?

RQ5: How do duplicate removal and publisher-disjoint evaluation affect the reported performance and interpretation of generalisation?

Research Hypotheses

The following hypotheses will be tested:

H1: The transformer-CNN-BiLSTM model will achieve a higher mean macro F1-score than the transformer-only baseline within each experimental case.

H2: The relative contributions of the CNN and BiLSTM branches will differ between AFND and WELFake.

H3: Publisher-disjoint evaluation on AFND will produce a lower mean macro F1-score than ordinary stratified random splitting.

Hypothesis Evaluation

The hypotheses will be evaluated using repeated experimental runs, mean macro F1-score, run-to-run variation, confidence intervals, and paired statistical comparisons.

A hypothesis will not be supported solely because one model produces a higher score in a single run. The size, stability, statistical evidence, class-level behaviour, and computational cost of the observed difference will also be considered.

Significance of the Study

The significance of this study lies in its controlled evaluation of model components rather than in architectural complexity alone. Previous fake news studies have reported strong results using traditional machine learning, deep learning, transformers, and hybrid architectures, but differences in datasets and experimental procedures often prevent direct comparison.

The study will compare transformer-only and hybrid models using consistent data partitions, preprocessing rules, training procedures, evaluation measures, and computational conditions. This design will help determine whether performance differences result from the CNN and BiLSTM branches rather than from unrelated experimental choices.

The research is also relevant to Arabic Natural Language Processing. Arabic fake news datasets remain more limited than English resources and differ in language variety, annotation procedure, and source composition. The Arabic experiment will use a language-specific pretrained transformer rather than applying an English model without linguistic adaptation.

The English experiment will provide a separate case for evaluating the same architectural question. Raw performance scores from AFND and WELFake will not be used to determine which language is easier. Instead, the study will examine whether the same architectural additions produce comparable patterns of improvement within each dataset.

A further contribution is the inclusion of computational cost and source-related robustness. A small increase in macro F1-score may have limited practical value if it requires substantially more training time, inference time, memory, or storage. Publisher-disjoint evaluation on AFND will also provide a stricter assessment of whether model performance extends beyond familiar sources.

The study will remain valuable regardless of which model achieves the highest score. Evidence that a simpler transformer model provides comparable performance at lower cost would be as relevant as evidence supporting the complete hybrid architecture.

Methodology

Research Design

The study will use a controlled comparative experimental design. Arabic and English will be treated as separate experimental cases rather than combined into one multilingual classification task. AFND will represent the Arabic case, while WELFake will represent the English case.

The same model comparison structure will be applied within both cases, but language-specific preprocessing procedures, tokenizers, pretrained transformers, and word embeddings will be used. Performance scores from the two datasets will not be compared directly because their labels, sources, collection procedures, and task characteristics differ.

The experiment is designed as an ablation-based comparison. Transformer-only, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models will be evaluated using consistent data partitions and training conditions. Traditional machine learning and CNN-BiLSTM models will provide additional reference baselines.

The purpose of the design is not to confirm that the most complex architecture is superior. It is to measure the contribution of each component and determine whether any observed gain is stable, statistically supported, and proportionate to the additional computational cost.

Datasets and Task Definition

AFND will be used for the Arabic experiment, while WELFake will be used for the English experiment. Both datasets will be inspected before model development to document their textual fields, label definitions, class distributions, missing records, article lengths, duplicate content, and available source information.

The main unit of analysis will be the news article. Where both title and body fields are available, they will be combined using a consistent separator. Records containing insufficient textual content or unusable labels will be removed according to documented criteria.

The primary experimental task will be binary text classification. WELFake already provides binary real and fake labels. For AFND, credible and not credible records will form the primary binary task, while undecided records will be excluded before data splitting. The number of excluded records and the resulting class distribution will be reported.

The task will measure agreement with the labels supplied by the datasets. It will not determine the factual accuracy of individual claims through external evidence. AFND results will therefore

be interpreted as classification according to publisher-associated credibility labels, while WELFake results will reflect the binary labels inherited from its combined source datasets.

All cleaning, label decisions, and duplicate removal will be completed before the final training, validation, and test partitions are generated.

Data Preparation

Data preparation will be completed separately for AFND and WELFake. The original datasets will be preserved, while all cleaning and transformation procedures will be applied to documented working copies.

Records with missing labels, empty textual fields, corrupted content, or insufficient article text will be removed according to predefined criteria. Where title and body fields are available, they will be combined using a consistent separator.

Exact duplicates will be identified using hashes of normalised text. Near-duplicate articles will be detected using a documented text-similarity method and threshold. Confirmed duplicate groups will be resolved before data partitioning to prevent related versions of the same article from appearing in different subsets.

Arabic preprocessing will include Unicode normalisation, removal of elongation characters, correction of inconsistent whitespace, and controlled normalisation of selected letter forms. Aggressive stemming, root extraction, and stop-word removal will not be applied to transformer inputs.

English preprocessing will remove invalid formatting, duplicated whitespace, and markup unrelated to article content. Lowercasing and token handling for transformer models will follow the selected BERT tokenizer.

Traditional machine learning models will use separately fitted TF-IDF representations. Transformer tokenisation, TF-IDF vocabulary construction, class weighting, and any feature selection will be derived from the training data only.

Experimental Splits

After cleaning and duplicate removal, each dataset will be divided into training, validation, and test partitions using an 80:10:10 ratio. The primary evaluation will use stratified random splitting to preserve the class distribution across the three subsets.

The partitions will be created once using fixed random seeds and saved for use by all comparable models. The test set will remain locked throughout model development. Hyperparameter tuning, checkpoint selection, preprocessing decisions, and model comparison during development will use only the training and validation data.

Near-duplicate records belonging to the same identified group will remain within a single partition. This prevents closely related articles from being distributed across training and testing data.

A second evaluation track will use publisher-disjoint splitting for AFND, provided that the publisher metadata supports reliable separation. Publishers assigned to the test set will not appear in the training or validation sets.

The publisher-disjoint result will be reported separately from the stratified random result because the two settings measure different forms of generalisation. Where WELFake metadata does not support reliable source separation, no publisher-disjoint result will be claimed.

Models

The study will evaluate seven model configurations:

1. Logistic Regression using TF-IDF features.
2. Linear Support Vector Machine using TF-IDF features.
3. CNN-BiLSTM using pretrained static word embeddings appropriate to each language.
4. Transformer-only classification.
5. Transformer-CNN.
6. Transformer-BiLSTM.
7. Parallel transformer-CNN-BiLSTM.

The traditional machine learning models will provide computationally efficient reference baselines. The CNN-BiLSTM model will represent a hybrid deep learning approach without contextual transformer representations.

The Arabic transformer configurations will use a selected AraBERT checkpoint, while the English configurations will use BERT-base-uncased. The exact AraBERT version will be reported with the final experimental configuration.

All transformer-based configurations within the same language will use the same pretrained encoder checkpoint, tokenizer, data partitions, and maximum sequence length. This will ensure that differences between the transformer-only and hybrid models result primarily from the added CNN and BiLSTM branches.

In the complete hybrid model, contextual token representations from the transformer encoder will be passed in parallel to CNN and BiLSTM branches. Their outputs will be pooled, concatenated, regularised through dropout, and passed to the final classification layer.

The transformer-CNN and transformer-BiLSTM models will provide direct ablations of the complete architecture.

Training Procedure

Model development will follow the same training and validation procedure within each experimental case. Comparable models will use the same saved data partitions and predefined random seeds.

Hyperparameters will be selected using the validation set only. The search will include learning rate, batch size, dropout rate, maximum sequence length, number of epochs, CNN kernel configuration, and BiLSTM hidden size where relevant. Comparable transformer models will receive the same tuning budget.

Early stopping will be based on validation macro F1-score. The checkpoint with the strongest validation performance will be retained for final testing. The test set will not be used to select hyperparameters, training duration, preprocessing rules, or model checkpoints.

Dropout and weight decay will be used to reduce overfitting. Gradient clipping will be applied to recurrent components. Training and validation loss, macro F1-score, and learning-rate progression will be recorded for each epoch.

Each neural architecture will be trained using at least three fixed random seeds. Results will be reported using the mean and standard deviation across repeated runs rather than the strongest individual run.

All directly compared models will be trained and evaluated using the same hardware environment. Any technical failure or interrupted run will be documented and repeated using the original configuration.

Evaluation

Macro F1-score will serve as the primary evaluation measure because it gives equal importance to both classes and provides a more balanced assessment when class distributions differ.

Supporting classification measures will include:

1. Accuracy.
2. Precision for each class.
3. Recall for each class.
4. F1-score for each class.
5. Weighted F1-score.
6. ROC-AUC.
7. Confusion matrix.

The study will also record the computational requirements of each neural architecture:

1. Training time.
2. Inference time per article and per batch.
3. Number of trainable parameters.
4. Peak GPU memory consumption.
5. Saved model size.

Timing and memory measurements will be collected using the same hardware and execution conditions. Initialisation time and one-time data loading will be separated from inference measurements where appropriate.

The main comparison will use mean results across repeated runs. Numerical improvement alone will not establish that one architecture is superior. Stability, class-level behaviour, confidence intervals, statistical comparison, robustness under publisher-disjoint evaluation, and computational cost will also inform the interpretation.

Robustness and Error Analysis

Robustness analysis will examine whether model performance depends on duplicate content, source composition, class distribution, and article length.

The primary stratified results will be compared with publisher-disjoint results on AFND. A substantial decrease under publisher-disjoint evaluation will indicate that part of the random-split performance was associated with source-specific characteristics.

The effect of duplicate removal will be documented by reporting the number of exact and near-duplicate records identified before splitting. Model results will be based on the cleaned and deduplicated datasets.

Error analysis will examine representative false positive and false negative predictions from the transformer-only and hybrid models. The analysis will consider:

1. Article length and truncation.
2. Publisher or source.
3. Topic and recurring named entities.
4. Repeated phrases and writing style.
5. Arabic spelling and morphological variation.
6. Prediction confidence.
7. Differences between the evaluated architectures.

Examples will be used to explain recurring error patterns rather than to make factual judgements about individual publishers or articles.

Dataset Profile and Labelling Structure

AFND and WELFake will serve as separate Arabic and English experimental cases. They differ in language, textual fields, label construction, source metadata, and collection procedure. Their raw performance scores will therefore not be compared to determine which language is easier to classify.

AFND contains credible, not credible, and undecided labels associated with the credibility classification of the publishing source. The primary AFND experiment will use credible and not

credible records as a binary classification task. Undecided records will be excluded before data partitioning, and the number of excluded records will be reported.

WELFake provides binary real and fake labels inherited from the datasets combined during its construction. Its collection history introduces possible duplication and source-specific patterns that require inspection before model training.

Table III: Profile of the Selected Datasets summarises the main characteristics, methodological risks, and planned controls for both datasets. The study will report the original record count, excluded records, removed duplicates, and final class distribution after data preparation.

Table III: Profile of the Selected Datasets

Characteristic	AFND	WELFake
Primary language	Arabic	English
Unit of analysis	News article	News article
Main textual fields	Article title and body, according to the downloaded dataset version	Article title and text
Original label structure	Credible, not credible, and undecided	Real and fake
Primary experimental task	Binary classification of credible and not credible articles	Binary classification of real and fake articles
Treatment of additional labels	Undecided records excluded before data splitting	Not applicable
Basis of labels	Publisher-associated credibility classification	Binary labels inherited from the combined source datasets
Source metadata	Publisher information available	Original source information may be incomplete
Main methodological risk	Learning publisher style, vocabulary, or topic patterns instead of article-level credibility	Learning duplicate, source-specific, or collection-specific patterns
Planned control	Duplicate removal and publisher-disjoint evaluation	Exact and near-duplicate removal with source-pattern analysis where metadata permits
Primary transformer	AraBERT, with the exact checkpoint reported in the final configuration	BERT-base-uncased

Interpretation of results	Independent Arabic experimental case	Independent English experimental case
----------------------------------	--------------------------------------	---------------------------------------

The final number of records used in each experiment will be determined after removing empty entries, unusable labels, exact duplicates, and confirmed near-duplicates. The study will report the record count at each cleaning stage rather than presenting only the final dataset size.

The original class distribution will also be documented before any balancing procedure is considered. Class weighting will be preferred over removing large numbers of valid records. Oversampling or undersampling will only be introduced if validation results show a clear need, and the procedure will be applied to the training data only.

The performance scores of AFND and WELFake will not be compared directly to determine which language is easier. Differences between the two datasets include label construction, publisher composition, article length, topic distribution, and collection history. Each dataset will instead be used to examine whether the same architectural change produces a similar pattern of improvement within its own experimental setting.

Preprocessing Pipelines

Separate preprocessing pipelines will be used for traditional machine learning and transformer-based models. Common cleaning and duplicate removal will be completed before data partitioning, while vocabulary fitting, tokenisation settings, class weighting, and feature selection will use training data only.

Arabic preprocessing will apply controlled Unicode and letter normalisation without aggressive stemming or stop-word removal for transformer inputs. English transformer inputs will follow the requirements of BERT-base-uncased. [*Table IV: Preprocessing Procedures for the Experimental Pipelines*](#) summarises the procedures applied to each experimental pipeline.

Table IV: Preprocessing Procedures for the Experimental Pipelines

Processing Stage	Arabic Traditional Models	Arabic Transformer Models	English Traditional Models	English Transformer Models
Remove empty or unusable records	Yes	Yes	Yes	Yes
Remove exact duplicates	Before splitting	Before splitting	Before splitting	Before splitting

Inspect and resolve near-duplicates	Before splitting	Before splitting	Before splitting	Before splitting
Join title and article body	Where both fields are available	Where both fields are available	Where both fields are available	Where both fields are available
Normalise whitespace and invalid formatting	Yes	Yes	Yes	Yes
Remove unrelated URLs and markup	Yes	Yes	Yes	Yes
Remove Arabic elongation characters	Yes	Yes	Not applicable	Not applicable
Remove optional Arabic diacritics	Applied consistently	Applied only when compatible with the selected checkpoint	Not applicable	Not applicable
Normalise selected Arabic letter forms	Yes	Limited and checkpoint-compatible	Not applicable	Not applicable
Lowercase English text	Not applicable	Not applicable	Yes	Managed by BERT-base-uncased
Stop-word removal	Evaluated during baseline validation	No	Evaluated during baseline validation	No
Stemming or lemmatisation	Not used by default	No	Not used by default	No
Text representation	TF-IDF	AraBERT tokenizer	TF-IDF	BERT-base-uncased tokenizer
Padding	Not applicable	Applied according to batch configuration	Not applicable	Applied according to

				batch configuration
Truncation	Not applicable	Based on Arabic token-length analysis	Not applicable	Based on English token-length analysis
Class weighting	Applied when justified by training-set imbalance	Applied when justified by training-set imbalance	Applied when justified by training-set imbalance	Applied when justified by training-set imbalance

Proposed Framework and Model Architecture

The proposed framework evaluates whether local and sequential processing branches add measurable value to contextual transformer representations. Arabic and English articles will pass through separate language-specific pipelines. AFND will use an Arabic transformer, while WELFake will use BERT-base-uncased. The datasets will not be combined during training.

The complete parallel architecture contains the following stages:

1. Text Input and Tokenisation

The article title and body will be combined using a consistent separator where both fields are available. Each article will be tokenised using the tokenizer associated with its language-specific transformer. Attention masks will distinguish valid tokens from padding positions.

2. Transformer Encoder

The tokenised sequence will pass through the selected pretrained transformer. The encoder will produce a contextual representation for every token.

The transformer-only baseline will use the pooled representation associated with the first classification token, followed by dropout and a linear classification layer. The hybrid models will retain the full sequence of contextual token representations for processing by the CNN and BiLSTM branches.

3. CNN Branch

The contextual token sequence will pass through one-dimensional convolutional filters with selected kernel sizes. The filters will extract local token patterns of different lengths.

Masked global max pooling will produce a fixed-length CNN representation while preventing padded positions from affecting the pooled output.

4. BiLSTM Branch

The same contextual token sequence will pass through a bidirectional LSTM. Valid sequence lengths derived from the attention masks will prevent padding from influencing the recurrent representation.

The forward and backward hidden representations will be combined to represent sequential information from both directions.

5. Feature Fusion

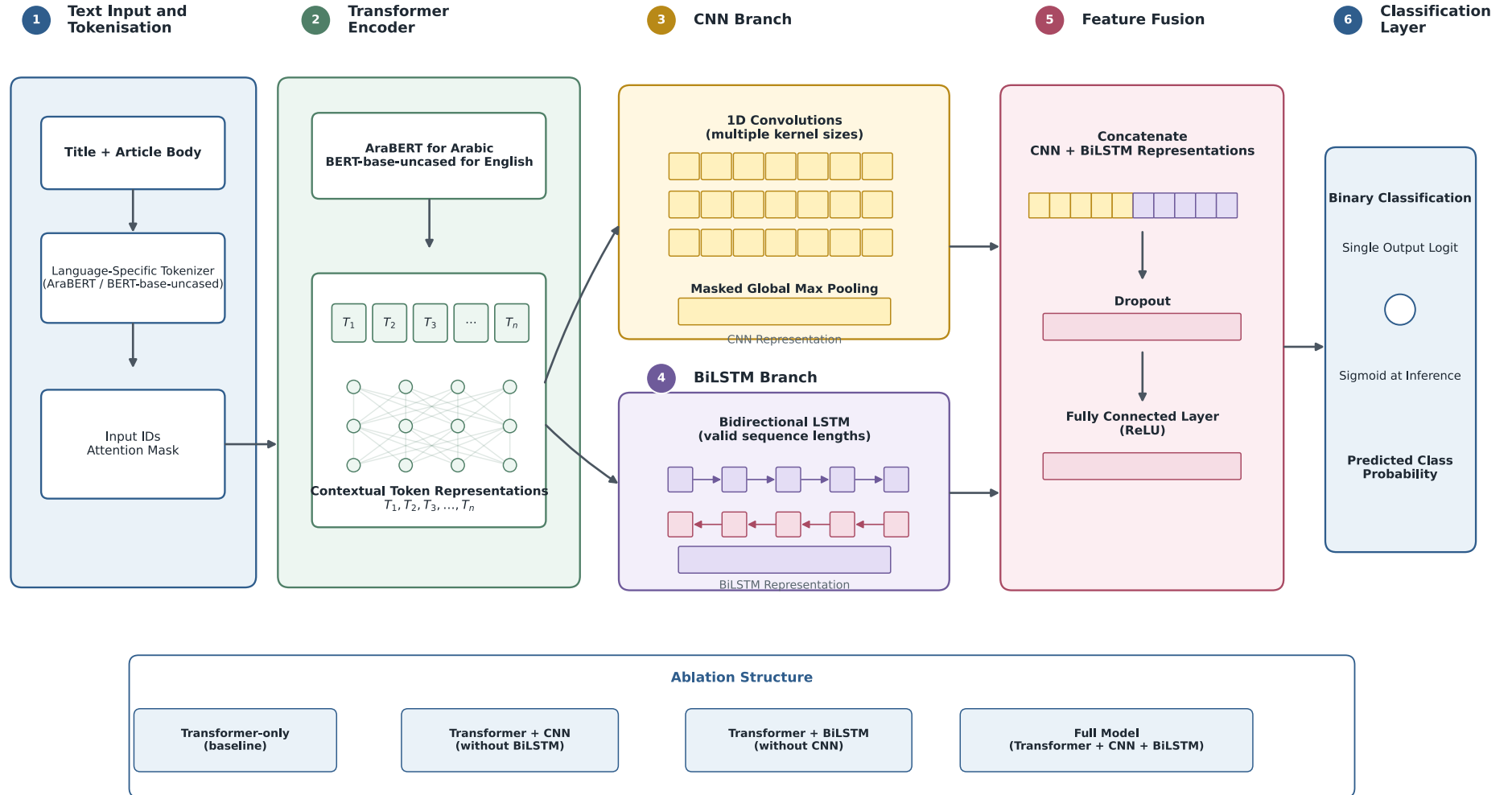
The pooled CNN representation and the BiLSTM representation will be concatenated into one feature vector. Dropout will be applied to the fused representation before it passes through a fully connected layer.

6. Classification Layer

The primary task is binary classification. The final layer will produce one output logit for each article. Binary cross-entropy with logits will serve as the classification loss.

Class weighting will be applied only when the training-set distribution justifies its use. The same weighting rule will be used across comparable models within each dataset.

Figure 1: Proposed Parallel Transformer-CNN-BiLSTM Architecture



Ablation Structure

Four transformer-based configurations will be evaluated using the same encoder checkpoint, tokenizer, data partitions, maximum sequence length, and training budget within each language:

1. **Transformer-only:** Uses the pooled transformer representation followed by the classification layer.
2. **Transformer-CNN:** Passes the contextual token sequence through the CNN branch before classification.
3. **Transformer-BiLSTM:** Passes the contextual token sequence through the BiLSTM branch before classification.
4. **Transformer-CNN-BiLSTM:** Processes the contextual token sequence through parallel CNN and BiLSTM branches and fuses their representations before classification.

This structure isolates the contribution of each added branch. The complete hybrid architecture will not be considered superior unless its improvement is stable across repeated runs and proportionate to its additional computational cost.

Sequence Length and Long Articles

Token-length distributions will be examined separately for AFND and WELFake. One maximum sequence length will be selected for all transformer configurations within each language.

The study will report the selected length, the percentage of articles exceeding it, and the average number of removed tokens. Head-only and combined head-tail truncation will be compared using validation data where resource limits permit. The test set will not influence the selected truncation rule.

Experimental Design Matrix

The experimental design matrix defines the models evaluated within the Arabic and English cases. All comparable models within the same language will use the same cleaned dataset, saved partitions, preprocessing rules, and evaluation records.

Traditional machine learning models will provide deterministic reference baselines. Neural architectures will be evaluated using at least three predefined random seeds to account for variation caused by initialisation and optimisation.

Table V: Experimental Design Matrix

Model	Arabic Case	English Case	Repeated Runs	Main Purpose
TF-IDF + Logistic Regression	AFND	WELFake	Deterministic baseline	Provide a simple and interpretable reference
TF-IDF + Linear SVM	AFND	WELFake	Deterministic baseline	Provide a strong traditional text-classification reference
CNN-BiLSTM	Arabic pretrained static embeddings	English pretrained static embeddings	At least 3 seeds	Evaluate hybrid deep learning without contextual transformers
Transformer-only	AraBERT	BERT-base-uncased	At least 3 seeds	Establish the main contextual baseline
Transformer-CNN	AraBERT-CNN	BERT-CNN	At least 3 seeds	Measure the contribution of local convolutional features
Transformer-BiLSTM	AraBERT-BiLSTM	BERT-BiLSTM	At least 3 seeds	Measure the contribution of sequential recurrent features
Transformer-CNN-BiLSTM	AraBERT-CNN-BiLSTM	BERT-CNN-BiLSTM	At least 3 seeds	Evaluate the complete parallel hybrid architecture

Evaluation Tracks

The primary evaluation track will use cleaned and deduplicated data divided into stratified training, validation, and test partitions. Every model within the same language case will use the same saved partitions.

A second evaluation track will use publisher-disjoint partitions for AFND, provided that the available metadata supports reliable source separation. Publishers included in the test set will not appear in the training or validation data.

The publisher-disjoint track will prioritise Logistic Regression, transformer-only, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM. It will examine whether performance remains stable when the models encounter publishers excluded from training.

The two evaluation tracks will be reported separately. Stratified evaluation measures performance on unseen articles drawn from the same dataset structure, while publisher-disjoint evaluation provides a stricter estimate of generalisation to unseen sources.

Experimental Environment

The experiments will be implemented in Python using PyTorch and Hugging Face Transformers for neural model development. Scikit-learn will be used for TF-IDF baselines, classification metrics, and supporting statistical procedures. Pandas and NumPy will support dataset preparation and analysis.

Training will be performed using a CUDA-compatible GPU environment. Directly compared neural models will be trained and timed on the same hardware to ensure that differences in training time, inference time, and memory consumption are not caused by changes in the execution platform.

The final report will document the GPU, CPU, system memory, operating system, Python version, PyTorch version, Transformers version, and CUDA version. Automatic mixed-precision training may be used where supported, but it will be applied consistently across comparable transformer configurations.

Maximum sequence length, batch size, and the number of repeated runs will be selected according to available memory while preserving a fair comparison between the evaluated architectures.

Statistical Analysis

The main analysis will compare mean macro F1-score across repeated runs. Accuracy, weighted F1-score, class-level precision, recall, and F1-score will support the primary comparison.

For each neural architecture, the study will report the mean and standard deviation across predefined random seeds. The highest individual run will not be used as the primary basis for model comparison.

Bootstrap resampling of the locked test predictions will be used to estimate 95% confidence intervals for macro F1-score. Paired bootstrap comparisons will estimate confidence intervals for differences between each hybrid model and its transformer-only baseline.

McNemar's test will be used to compare paired classification outcomes on the same test records. Because several hybrid architectures will be compared with the same transformer baseline, the Holm correction will be applied to control the family-wise error rate.

A significance level of 0.05 will be used. Statistical significance will be interpreted together with the size of the performance difference, confidence interval, class-level behaviour, stability across runs, publisher-disjoint results, and computational cost.

A hybrid architecture will be considered practically preferable only when its improvement is stable across repeated runs and proportionate to the additional training time, inference time, memory consumption, parameter count, and model size.

Negative or inconclusive results will remain part of the analysis. Failure to identify a meaningful improvement will indicate that the added CNN or BiLSTM components did not provide sufficient value under the selected experimental conditions.

Ethical Considerations

This study uses existing research datasets and does not involve direct interaction with human participants or the collection of new personal information. AFND and WELFake will be used according to their published access, citation, and redistribution conditions.

Dataset labels will not be treated as independent proof that every article is factually true or false. AFND associates article labels with publisher credibility, while WELFake inherits labels from several source datasets. Reported results will therefore be described as dataset-based classification outcomes rather than final judgements about publishers, journalists, or individuals.

Classification errors may create reputational or informational harm. The proposed models will be treated as research prototypes rather than autonomous moderation or fact-checking systems. Predictions would require human review before operational use.

Source metadata will be used only to study dataset bias and generalisation. Error-analysis examples will illustrate methodological patterns and will not be presented as factual accusations against individual publishers.

Cleaning decisions, excluded records, data partitions, model configurations, and unsuccessful results will be documented. Raw article content will not be redistributed unless permitted by the relevant dataset licence.

Conclusion

The reviewed literature shows that fake news classification has progressed from manually designed linguistic features and traditional machine learning to deep learning, pretrained transformers, hybrid architectures, ensembles, and multimodal systems. Despite this progress, reported results remain strongly influenced by dataset construction, source composition, language, annotation method, and evaluation procedure.

The literature does not establish that adding CNN and BiLSTM components to transformer representations always produces a meaningful improvement. Hybrid architectures increase model complexity and computational cost, while their reported benefits often vary across datasets and domains.

This study therefore proposes a controlled comparison of traditional machine learning, CNN-BiLSTM, transformer-only, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models. AFND and WELFake will be treated as separate Arabic and English experimental cases using language-appropriate transformer encoders.

The evaluation will consider predictive performance, repeated-run stability, statistical evidence, computational cost, duplicate content, and source-related generalisation. Macro F1-score will serve as the primary evaluation measure, supported by class-level metrics, confidence intervals, paired comparisons, and publisher-disjoint evaluation on AFND.

The contribution of the study will not depend on the complete hybrid model achieving the highest score. Evidence that a simpler transformer model provides comparable performance with lower computational requirements would represent an equally relevant research finding. The final conclusion will identify the architecture that provides the most defensible balance between performance, cost, and generalisation under the selected experimental conditions.

References

- Al-alshaqi, M., Rawat, D. B., & Liu, C. (2025). A BERT-based multimodal framework for enhanced fake news detection using text and image data fusion. *Computers*, 14(6), 237.
<https://doi.org/10.3390/computers14060237>
- Al-Zahrani, L., & Al-Yahya, M. (2024). Pre-trained language model ensemble for Arabic fake news detection. *Mathematics*, 12(18), 2941. <https://doi.org/10.3390/math12182941>
- Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (pp. 9–15). European Language Resources Association. <https://aclanthology.org/2020.osact-1.2/>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/N19-1423>
- Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, 9, 6581–6592.
<https://doi.org/10.1007/s40747-023-01098-0>

- Ganesh Karthik, M., Faizz Ahmad, K. S., Pamidimukkala, S. G., Sathe, A. P., Sirisha, G. N. V. G., Sitha Ram, M., & Koteswararao, C. (2025). Hybrid optimization driven fake news detection using reinforced transformer models. *Scientific Reports*, 15, 14782.
<https://doi.org/10.1038/s41598-025-99936-3>
- Himdi, H., Weir, G., Assiri, F., & Al-Barhamtoshy, H. (2022). Arabic fake news detection based on textual analysis. *Arabian Journal for Science and Engineering*, 47, 10453–10469.
<https://doi.org/10.1007/s13369-021-06449-y>
- Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., & Jadhav, S. (2025). Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation. *Frontiers in Artificial Intelligence*, 8, 1690616.
<https://doi.org/10.3389/frai.2025.1690616>
- Khalil, A., Jarrah, M., Aldwairi, M., & Jaradat, M. (2022). AFND: Arabic fake news dataset for the detection and classification of articles' credibility. *Data in Brief*, 42, 108141.
<https://doi.org/10.1016/j.dib.2022.108141>
- Lv, J., Gao, Y., Li, L., Shi, L., & Li, S. (2025). Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37, 306.
<https://doi.org/10.1007/s44443-025-00317-7>
- Merzah, B. M., Razmara, J., & Salmanian, Z. (2026). Hybrid deep learning models for fake news detection: Case study on Arabic and English languages. *Frontiers in Big Data*, 8, 1683786.
<https://doi.org/10.3389/fdata.2025.1683786>

- Nagoudi, E. M. B., Elmadany, A., Abdul-Mageed, M., Alhindi, T., & Cavusoglu, H. (2020). Machine generation and detection of Arabic manipulated and fake news. In Proceedings of the Fifth Arabic Natural Language Processing Workshop (pp. 69–84). Association for Computational Linguistics. <https://aclanthology.org/2020.wanlp-1.7/>
- 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), 100007. <https://doi.org/10.1016/j.ijime.2020.100007>
- Raza, N., Abdulkadir, S. J., Abid, Y. A., Albouq, S. S., Alwadain, A., Rehman, A. U., Sumiea, E. H., & Farhan, M. (2025). Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach. PLOS ONE, 20(9), e0330954. <https://doi.org/10.1371/journal.pone.0330954>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models. Future Internet, 17(1), 28. <https://doi.org/10.3390/fi17010028>
- Rout, J., Mishra, M., & Saikia, M. J. (2025). Towards reliable fake news detection: Enhanced attention-based transformer model. Journal of Cybersecurity and Privacy, 5(3), 43. <https://doi.org/10.3390/jcp5030043>
- Saadany, H., Orasan, C., & Mohamed, E. (2020). Fake or real? A study of Arabic satirical fake news. In Proceedings of the 3rd International Workshop on Rumours and Deception in

Social Media (pp. 70–80). Association for Computational Linguistics.

<https://aclanthology.org/2020.rdsm-1.7/>

Touahri, I., & Mazroui, A. (2024). Survey of machine learning techniques for Arabic fake news detection. *Artificial Intelligence Review*, 57, 157. <https://doi.org/10.1007/s10462-024-10778-3>

Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4), 881–893. <https://doi.org/10.1109/TCSS.2021.3068519>