



COMPARATIVE EVALUATION OF TRANSFORMER-BASED HYBRID MODELS FOR ARABIC AND ENGLISH FAKE NEWS CLASSIFICATION

Prepared By:

- **Mohammed Al-jaaidi**
- **Mohammed AbuSultan**
- **Khalid Abbas Elzein**

Thesis submitted in partial fulfillment of the requirements for the degree of:
Master of Science in Artificial Intelligence

At: College of Informatics, Midocean University, UAE

Name of the Supervisor: Dr. Yasser Abdelhamid, College of Informatics, Midocean University, Comoros

Date: July 2026

Acknowledgements

Our special appreciation goes out to our supervisor, **Dr. Yasser Abdelhamid**, for his invaluable support and guidance throughout the process of formulating and conducting the research. His vast knowledge and experience in the field of artificial intelligence and natural language processing have greatly contributed to this thesis.

Moreover, we wish to thank the College of Informatics, Midocean University, for providing us with the necessary environment and resources for conducting this research. Lastly, we appreciate our families and friends for their patience and encouragement.

Abstract

Classification of fake news has evolved from classical machine learning with manually engineered language features to deep learning with pretrained transformers and hybrid neural architectures. While promising results have been presented by transformer-based approaches, performance varies greatly depending on the language, the dataset, and other aspects of experiments. Furthermore, it has yet to be confirmed whether the incorporation of convolutional and recurrent structures to transformers produces consistently better outcomes which justify additional computational costs.

The purpose of this paper is to review recent literature on fake news classification in Arabic and English languages and perform a comparative evaluation of transformer-based hybrid architectures. For Arabic cases, the Arabic Fake News Dataset (AFND) is proposed with Arabic transformer (AraBERT). English cases will utilize WELFake dataset with BERT-base-uncased transformer. The comparison involves traditional baselines (TF-IDF-based Logistic Regression and Linear Support Vector Machine), deep learning baselines (CNN-BiLSTM with pretrained static embeddings), transformer-only classification and three hybrid approaches: transformer-CNN, transformer-BiLSTM and parallel transformer-CNN-BiLSTM.

All models will be trained under the same conditions regarding data partitions, preprocessing pipelines, training parameters and random-seed runs. As an evaluation metric, macro F1-score is used together with accuracy, precision, recall per class, weighted F1-score, ROC-AUC and confusion matrix. In addition, computational overhead is measured through training time, inference latency, number of parameters, model size and peak memory usage. Moreover, we study

particular weaknesses of each dataset, including duplicate content, class imbalance and publisher-bias, through publisher-disjoint evaluation on AFND, if the metadata allows. We treat the problem of fake news classification as strictly supervised one with no external verification of facts provided by the dataset labels. The main contribution of this thesis lies in a controlled and empirical evaluation of the benefit of adding CNN and BiLSTM branches to transformers.

Keywords: Fake News Classification, Arabic NLP, BERT, AraBERT, CNN-BiLSTM, Transformer-Based Hybrid Architectures, Comparative Evaluation.

Table of Contents

1. [Chapter 1: Introduction and Overview](#)
 - [1.1 Overview](#)
 - [1.2 Problem Definition](#)
 - [1.3 Thesis Objectives](#)
 - [1.4 Thesis Contribution](#)
 - [1.5 Significance of the Study](#)
 - [1.6 Thesis Organization](#)
2. [Chapter 2: Background and Related Work](#)
 - [2.1 Fake News as a Text Classification Problem](#)
 - [2.2 Traditional Machine Learning Approaches](#)
 - [2.3 Deep Learning and Hybrid Models](#)
 - [2.4 Transformer-Based Language Models](#)
 - [2.5 Arabic Language Challenges](#)
 - [2.6 Dataset Bias and Generalisation](#)
 - [2.7 Review Scope and Study Selection](#)
 - [2.8 Detailed Literature Review](#)
 - [2.9 Comparative Synthesis of the Reviewed Studies](#)
 - [2.10 Research Trends Within the Reviewed Set](#)
 - [2.11 Interpretation of the Research Trends](#)
 - [2.12 Research Gap and Basis for the Proposed Study](#)
 - [2.13 Research Questions and Hypotheses](#)
3. [Chapter 3: Proposed Methodology \(Approach\)](#)
 - [3.1 Research Design](#)

- [3.2 Datasets and Task Definition](#)
 - [3.3 Data Preparation & Preprocessing Pipelines](#)
 - [3.4 Proposed Framework and Model Architecture](#)
 - [3.5 Ablation Structure](#)
 - [3.6 Sequence Length and Long Articles](#)
4. [Chapter 4: Experimental Setup, Results, and Discussion](#)
- [4.1 Experimental Design Matrix](#)
 - [4.2 Evaluation Tracks & Splits](#)
 - [4.3 Experimental Environment](#)
 - [4.4 Statistical Analysis & Evaluation Metrics](#)
 - [4.5 Ethical Considerations](#)
5. [Chapter 5: Conclusion and Future Work](#)
- [5.1 Important Features & Findings](#)
 - [5.2 Future Work](#)
6. [References](#)
-

List of Abbreviations

Abbreviation	Full Form
Acc.	Accuracy
AFND	Arabic Fake News Dataset
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network

DL	Deep Learning
F1	F1-Score (Harmonic Mean of Precision and Recall)
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

List of Figures

Figure No.	Title	Proposed Page
Figure 3.1	Proposed Parallel Transformer-CNN-BiLSTM Architecture	35

List of Tables

Table No.	Title	Proposed Page
Table 2.1	Comparative Summary of the Reviewed Studies	14
Table 2.2	Dominant Research	21

	Trends Within the Reviewed Studies	
Table 3.1	Profile of the Selected Datasets (AFND vs. WELFake)	29
Table 3.2	Preprocessing Procedures for the Experimental Pipelines	31
Table 4.1	Experimental Design Matrix	36

Chapter 1: Introduction and Overview

1.1 Overview

Digitization of news channels, social media sites, and other messaging apps has increased the speed and scale of publishing and distributing information. On one hand, this revolution has made the access to information more democratized, but, on the other hand, it has greatly amplified the speed and scale of distribution of misleading and fraudulent content. The scale of user-generated content on the Internet is too vast and makes any manual check of the information impossible without significant delays and complications. As a result, automated computational systems to classify fake news became an important domain for research in fields of Artificial Intelligence and NLP.

The classical methods to design systems for fake news classification involved machine learning algorithms such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. These systems analyzed the text in the form of matrices TF-IDF, bags of words, or some linguistics/stylometric/manual feature set. Even though such classical systems were computationally efficient, they lacked the ability to detect semantic information and sequence dependencies.

Deep learning models significantly decreased the need to use manual feature extraction. Convolutional Neural Network was used to detect the pattern of n-gram tokens, while Recurrent Neural Networks and, more specifically, Long Short-Term

Memory (LSTM) and Bi-directional Long Short-Term Memory (BiLSTM) models were used to analyze sequence dependencies.

Pretrained transformer models, first used in BERT, introduced a breakthrough innovation. By using self-attention, transformers generate bidirectional and contextual representations of input texts, vastly increasing the models' performance for NLP tasks. In the case of non-English languages, specialized models like AraBERT have been created to deal with morphological and syntactic peculiarities of these languages.

Recent research has introduced hybrid architectures that incorporate additional CNN, BiLSTM, or attention pooling layers on the output of pretrained transformers. The intuition behind such models is that although the transformer learns deep contextual and semantic representation of texts, the CNN and BiLSTM layers can learn additional localized and sequential features. However, such complicated architecture adds greatly to the number of parameters, memory usage, training time, and inference time. In this thesis, a tightly controlled empirical study will be conducted to prove whether the mentioned hybrid extensions give consistent and statistically significant improvements over simpler transformer models.

1.2 Problem Definition

Although the literature demonstrates high performance of hybrid models based on transformers, there are some methodological issues that need to be solved:

1. **Absence of Controlled Baselines:** Hybrid architectures (Transformer + CNN + BiLSTM) are said to surpass regular models. At the same time, many researchers do not conduct their experiments against controlled baselines and use results from the published benchmarks that were performed with different preprocessing and hyperparameters.
2. **Contribution of Specific Components:** There is a lack of ablation studies that evaluate the contribution of each individual component of hybrid architectures (CNN and BiLSTM).
3. **Cross-lingual Generalizability:** All hybrid architectures are evaluated using only English benchmarks. There is no proof yet whether such architectural improvements would yield the same level of effectiveness when applied to morphologically complex languages such as Arabic and with specialized models such as AraBERT.
4. **Computational vs. Performance Metrics:** Commonly used performance metrics (accuracy, F1-Score etc.) are not compared together with computation metrics (training/inference time, memory usage, model size), which is essential for deployment.

5. Datasets **and their Bias**: Current datasets have inherent bias in them. For example, the Arabic Fake News Dataset (AFND) uses publisher reputation to generate labels, which increases the likelihood of classifiers picking up publisher specific writing styles instead of generic features related to misinformation. WELFake dataset has duplications, which can lead to leakage of train-set patterns in random-split test-sets.

1.3 Thesis Objectives

The main goal of this thesis is to conduct an ablation-based comparative analysis of transformer-based hybrid architectures for Arabic and English fake news classification, assessing whether the performance improvement justifies the computational cost of doing so.

The specific objectives are:

- **Objective 1**: To prepare, deduplicate and clean Arabic Fake News Dataset (AFND) and English WELFake dataset by removing label errors and text duplications.
- **Objective 2**: The creation of standardized and dedicated data preprocessing and tokenization pipelines for both traditional models and contextual transformer-based models in both languages.
- **Objective 3**: The implementation of an integrated experimental setup involving seven setups that include TF-IDF+Logistic Regression, TF-IDF+Linear SVM, CNN-BiLSTM (static embeddings), Transformer only, Transformer-CNN, Transformer-BiLSTM, and Transformer-CNN-BiLSTM in parallel.
- **Objective 4**: Conducting a systematic ablation study on an equal budget of hyperparameters while employing fixed random seeds.
- **Objective 5**: Evaluating the predictive performance of the models in terms of macro F1 score, accuracy, class-wise metrics, and ROC-AUC score along with their computational cost in terms of training time, inference time, memory, and number of parameters.
- **Objective 6**: Model robustness against the publishing bias in terms of a dedicated publisher-disjoint experiment using AFND dataset.

1.4 Thesis Contribution

The research makes the following four main contributions to the field of automated misinformation detection via AI:

- **A Controlled Benchmark for Comparative Evaluation**: The study creates a fully controlled benchmark in which conventional, deep learning-based,

transformer-only, and hybrid architectures are compared on an equal footing in both Arabic and English.

- **Rigorous Layer Ablation** Study: The research analyzes the effect of CNN and BiLSTM layers in case of transformation representations, proving whether multi-layer fusion actually helps.
- **Bilingual** Comparison of Architectures: In order to compare the effect of hybrid layers in different linguistic settings, we conduct a bilingual experiment in Arabic and English using AraBERT and BERT-base-uncased, respectively.
- **Performance vs. Efficiency** Trade-off Matrix: The study maps prediction improvements (in terms of macro F1) against deployment costs (latency and memory).
- **Mitigation of Publisher Bias:** It shows the reduction of classification accuracy under unseen source evaluation (disjoint split by publisher), revealing limitations of the standard random split experiments.

1.5 Significance of the Study

This study is important due to its comparative analysis of different parts of the architecture. Previously, fake news detection has shown impressive results in both traditional machine learning, deep learning, transformers and hybrid architectures, however the variations in datasets and experiment settings make comparisons between different works difficult. The study will analyze transformer-based and hybrid models using identical data splitting, preprocessing, training, evaluation and computational setup. It will reveal whether differences in performance are caused by the CNN and BiLSTM branches of the architecture or due to other experimental differences.

The research is also applicable for Arabic Natural Language Processing. Currently there are fewer Arabic datasets for fake news analysis and they differ in language diversity, labeling procedure and content. The Arabic experiment will use language-specific pretrained transformer model instead of applying English model without adaptation to language features. For the English language experiment the problem of architectural design will be investigated independently. Performance scores on AFND and WELFake datasets will not be used to evaluate which language is easier for model analysis. Instead it will investigate whether the same architectural improvements lead to similar patterns in each dataset separately.

1.6 Thesis Organization

The rest of this thesis is organized as follows:

- **Chapter 2 (Background and Related Work)** covers the theoretical background of text classification, describes traditional, deep learning and transformers methods, discusses issues of Arabic Natural Language Processing and synthesizes 15 recent peer-reviewed articles defining the research gap.
 - **Chapter 3 (Proposed Methodology)** explains the design methodology, the properties of the dataset, preprocessing techniques used, and the mathematical structure of the proposed parallel Transformer-CNN-BiLSTM and its variations.
 - **Chapter 4 (Experimental Setup, Results, and Discussion)** explains the experimental setup, data splits, hardware configuration, statistical analyses including McNemar's test and bootstrap confidence interval testing, and discusses the results and error analysis.
 - **Chapter 5 (Conclusion and Future Work)** concludes the paper and explains future directions for research.
-

Chapter 2: Background and Related Work

2.1 Fake News as a Text Classification Problem

In computational linguistics, fake news detection is formulated as a supervised text classification task. Given a corpus of document-label pairs

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i represents the textual sequence of a news article and $y_i \in \{0, 1\}$ represents the binary credibility label (e.g., 0 for real/credible, 1 for fake/not credible), the objective is to learn a mapping function $f(x) \rightarrow y$ that minimizes classification loss over unseen data.

While fact verification systems check against an external database or knowledge graph, text classification is completely dependent on the linguistic style, semantic connection, syntax, and vocabulary of the source document.

2.2 Traditional Machine Learning Approaches

Early automatic detection systems were based on feature extraction methods. Unprocessed text was transformed into vector space using one of the two weighting schemes such as BoW or TF-IDF:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{|D|}{1 + |\{d \in D: t \in d\}|}\right)$$

where t represents a term, d represents a document, and D represents the corpus. Classifiers such as Logistic Regression and Support Vector Machines (SVM) were trained on these sparse, high-dimensional vectors. While these models are highly interpretable and computationally inexpensive, they lack semantic awareness, treat tokens as independent entities, and fail to capture sequence-based context.

2.3 Deep Learning and Hybrid Models

In order to learn both local and sequential aspects without the need for explicit feature engineering, deep learning models were used. CNN employs 1D filters on sequences of word vectors for learning local n-gram patterns:

$$c_i = f(w \cdot x_{i:i+k-1} + b)$$

where w is the filter weight, k is the kernel size, and b is a bias term. Conversely, Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, capture long-range sequential dependencies using input, forget, and output gates to manage information flow across time steps.

Hybrid CNN-LSTM model learns both local spatial as well as global temporal features but it uses the word embeddings obtained through pre-trained algorithms such as Word2Vec, GloVe or FastText, which cannot distinguish between ambiguous lexicon.

2.4 Transformer-Based Language Models

Pre-trained language models based on self-attention mechanism were able to overcome the drawbacks of the pre-trained models. Attention-based language model is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q , K , and V represent the Query, Key, and Value matrices, and d_k is the dimensionality of the keys. By stack-pooling multi-head self-attention layers, BERT models compute bidirectional, context-dependent token representations. In Arabic NLP, models like AraBERT are pre-trained on massive Arabic corpora, utilizing specific pre-segmentation tools to process the language's complex morphological structure.

2.5 Arabic Language Challenges

The following aspects of Arabic language make it very difficult for text classification task:

1. **Morphological** complexity of Arabic language: Arabic is a templatic and very inflectional language, where words are formed from three or four letter roots using some patterns.
2. **Clitics and Affixes:** Conjunctions, prepositions, and possessive pronouns are often attached directly to nouns and verbs in form of prefixes and suffixes.
3. Ambiguous Orthography: Lack of short vowels (diacritics/Harakat) and different orthographies of letters like Alif (ا, إ, ؤ, ة) cause extremely high lexical ambiguity.
4. **Dialectal** Diversification: News content may incorporate Modern Standard Arabic (MSA) and different dialects, and this variety may be misclassified by models trained on classical texts.

2.6 Dataset Bias and Generalisation

[Fake news detection: A hybrid CNN-RNN based deep learning approach](#)[Hybrid deep learning models for fake news detection: Case study on Arabic and English languages](#)[Towards reliable fake news detection: Enhanced attention-based transformer model](#)

The performance of a classification model depends, among other factors, on how a particular dataset is collected and partitioned. In case of random article-level partitioning, articles from one publisher, covering the same topics, or originally belonging to one corpus can end up in training and test splits. In this situation, a classifier will detect certain patterns in the source rather than develop transferrable patterns. Presence of exact duplicates or near-duplicates increases this threat: related versions of one article in different partitions may bias reported test results. Previous works demonstrated considerable variability in fake news detection tasks across datasets, such as ISOT, WELFake, LIAR, FA-KES, and ANS. Differences like this point to separate evaluation of within-dataset performance and generalisation capabilities. For this reason, removal of exact duplicates and publisher-independent testing is incorporated into the experiment setup.

2.7 Review Scope and Study Selection

The current study employs structured narrative review to investigate recent trends in Arabic and English fake news classification. The review covers the research papers published between 2020 and 2026 and evaluates the approaches in terms of traditional machine learning, deep learning, transformer-based, hybrid, multilingual, or multimodal models. The starting pool of papers was formed using

targeted search in official repositories, such as ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, Frontiers, MDPI, and PLOS. The selection criteria were the presence of a research question on fake news, machine learning/deep learning/transformer-based methods, datasets and metrics. As a result, fifteen studies were chosen for comparative analysis.

2.8 Detailed Literature Review

Traditional Machine Learning and Linguistic Features

[Arabic fake news detection based on textual analysis](#)[Fake or real? A study of Arabic satirical fake news](#)[Machine generation and detection of Arabic manipulated and fake news](#)Traditional machine learning approaches are still relevant due to the simplicity and efficiency of the baselines they provide. Himdi et al. Arabic fake news detection based on textual analysis used emotion, polarity, parts of speech, negation, certainty, intensification, and reporting style features to analyse Arabic fake news. Random Forest was the best-performing classifier in the paper. The study showed that deception in Arabic involves informative linguistic signals, but the use of manually engineered features prevents the approach from generalizing to new publishers, dialects and topics. Saadany et al. Fake or real? A study of Arabic satirical fake news examined Arabic satirical news based on lexical and grammatical features. The classifiers built by the authors performed very well in distinguishing the articles of selected satirical and mainstream publishers. The results proved that satire had distinctive stylistic properties. On the other hand, the methodology raises concerns since the classifier may have simply learned the writing style of certain publishers rather than the distinction between factual and misleading reporting. Nagoudi et al. Machine generation and detection of Arabic manipulated and fake news worked on Arabic news generation and detection. The paper introduced the Arabic resource and evaluated both human readers and automatic classifiers. The strongest classifier had limitations compared to the best-performing results achieved on some English benchmark datasets.

CNN and Recurrent Neural Networks

[Fake news detection: A hybrid CNN-RNN based deep learning approach](#)[Hybrid deep learning models for fake news detection: Case study on Arabic and English languages](#)Deep learning models decreased the dependency on manual feature engineering. CNN was used for local expressions and short token sequences, while LSTM and BiLSTM represented longer sequential patterns. Nasir et al. Fake news detection: A hybrid CNN-RNN based deep learning approach used CNN with recurrent layer and pretrained GloVe embeddings. The model showed nearly perfect

performance on ISOT and considerably lower accuracy on FA-KES. The results showed that the choice of the architecture did not guarantee the good performance of the model; the effect of the dataset composition, article style, source distribution and task difficulty was significant. Merzah et al. Hybrid deep learning models for fake news detection: Case study on Arabic and English languages used parallel CNN and BiLSTM and evaluated it with FastText embeddings for Arabic and English data. The model achieved 98.85% accuracy on WELFake, 94.43% on AFND and 71.63% on ANS. The ablation experiments revealed that the CNN and BiLSTM branches captured complementary features. Both papers relied on static word embeddings rather than contextual transformer representations.

Transformer-Based Classification

[Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach](#)[Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models](#)[Fake news detection based on a hybrid BERT and LightGBM models](#)[Towards reliable fake news detection: Enhanced attention-based transformer model](#)Transformer models revolutionized the field of text classification providing the means of contextual representation in which the meaning of a token is dependent on the context of the whole sequence. Raza et al. Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach evaluated the progressively trained BERT model on WELFake. The model showed 95.3% accuracy with balanced precision, recall and F1-score. Roumeliotis et al. Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models used CNN, BERT, GPT-4o and GPT-4o-mini on the balanced sample of WELFake. The fine-tuned language models had the best performance, while the basic CNN model performed considerably worse. BERT was close to the larger models in terms of accuracy but provided an accessible experimental setup. Essa et al. Fake news detection based on a hybrid BERT and LightGBM models combined BERT embeddings with LightGBM. Their model showed strong performance on ISOT, TI-CNN and the Fake News Corpus. Rout et al. Towards reliable fake news detection: Enhanced attention-based transformer model 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 inconsistency shows that a model can show outstanding results on one benchmark but poor performance on another, shorter, more ambiguous or differently labelled dataset.

Arabic Pretrained Language Models and Ensembles

[Pre-trained language model ensemble for Arabic fake news detection](#)[Survey of machine learning techniques for Arabic fake news detection](#)The pretrained models for Arabic were created to improve the representation of Arabic morphology, vocabulary and language variation. Al-Zahrani and Al-Yahya Pre-trained language model ensemble for Arabic fake news detection evaluated an ensemble of AraBERT, MARBERT, AraELECTRA, AraGPT2 and ARBERT. The best-performing ensemble showed the F1-score of 94% with the weighted averaging. The ensembles increase the storage, inference time and deployment complexity of the system. Touahri and Survey of machine learning techniques for Arabic fake news detection reviewed the classification of Arabic fake news, datasets, preprocessing steps, linguistic features, deep learning models and Arabic transformers. The review revealed that there was no architecture that was consistently better than others in all experiments; limited public datasets, annotation procedure inconsistencies, dialect variation and domain specificity remained the problems.

Multimodal, Multilingual, and Explainable Approaches

[A BERT-based multimodal framework for enhanced fake news detection using text and image data fusion](#)[Hybrid optimization driven fake news detection using reinforced transformer models](#)[Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation](#)[Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges](#)Al-alshaqi et al. A BERT-based multimodal framework for enhanced fake news detection using text and image data fusion used Optical Character Recognition to extract the written part of the image and fuse it with the article text. The best-performing result was obtained on TRUTHSEEKER, while the model performed worse on the fused dataset. Ganesh Karthik et al. Hybrid optimization driven fake news detection using reinforced transformer models used the modified transformer with hybrid optimization on Fakeddit and DGM4. Jadhav et al. Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation proposed the multilingual, multimodal and explainable system combining XLM-RoBERTa, CNN, BiLSTM, ResNet50, GraphSAGE and attention-based fusion. The system showed better performance than the transformer baseline, but it contained about 420 million parameters and required significant amount of GPU resources. Lv et al. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges reviewed multimodal fake news detection based on text, images, video, audio, social context and large language models, revealing the

problems with annotation quality, heterogeneous fusion, interpretability and computational complexity.

2.9 Comparative Synthesis of the Reviewed Studies

However, the studied papers vary in their language, methods of data preparation, definition of the task, model design, and evaluation approach. Thus, the obtained results must be interpreted under the specific conditions of the experiments performed.

Table 2.1: Comparative Analysis of the Reviewed Papers

Study	Language and Dataset	Main Approach	Main Reported Result	Main Methodological Limitation	Relevance to the Current Study
Himdi et al. (2022)	Arabic, crowdsourced news	Linguistic features with Naive Bayes, SVM, and RF	RF produced the strongest performance	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	Lexical and grammatical features with ML	Accuracy reached up to 98.6%	Limited set of publishers; 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	Manipulated articles generated through a defined synthetic procedure	Supports discussion of machine-generated misinformation
Nasir et al. (2021)	English, ISOT and	CNN-RNN with GloVe	Near-perfect on	Severe dataset	Shows the value of

	FA-KES	embeddings	ISOT, ~60% on FA-KES	dependence; reliance on static GloVe embeddings	hybrid features and 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	Relies on static FastText embeddings; high performance variation	Closest architectural basis for the proposed parallel branches
Raza et al. (2025)	English, WELFake	Progressively trained BERT	95.3% accuracy, 0.952 precision, 0.954 recall, 0.953 F1	No examination of 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	GPT-4o achieved ~98.6% accuracy, BERT was highly competitive	Small selected sample and one language	Shows the strength and practical relevance of a BERT baseline
Essa et al. (2023)	English, ISOT, TI-CNN, and Fake News Corpus	BERT embeddings with LightGBM	99.88% accuracy on ISOT, 96.94% on TI-CNN	Model not trained as a single end-to-end neural model	Supports comparison between contextual embeddings and hybrid classification
Rout et al. (2025)	English, ISOT, FakeNewsNet, and LIAR	Compact attention-based transformer	100% accuracy on ISOT, 80% on FakeNewsNet	Performance varied severely across datasets	Strong evidence for dataset sensitivity and need for

			et, 59.56% on LIAR		controlled evaluation
Al-Zahrani & Al-Yahya (2024)	Arabic multisource dataset	Ensemble of five Arabic pretrained language models	Weighted-average ensemble achieved 94% F1-score	Ensemble increases storage, inference cost, and deployment complexity	Supports language-specific Arabic transformers and cost analysis
Touahri & Mazroui (2024)	Arabic survey	Review of datasets, preprocessing, ML, DL, and transformers	Found substantial variation across datasets, no universally superior approach	Survey paper; 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	OCR text fusion with BERT	Strongest performance on TruthSeeker; drop on merged data	Textual content from images but does not model visual manipulation	Shows the effect of combining sources
Ganesh Karthik et al. (2025)	Multimodal, Fakeddit and DGM4	Modified transformer with hybrid optimisation	90.42% accuracy on Fakeddit and 86.74% on DGM4	Method requires complex multi-stage fine-tuning pipeline	Supports computational-cost evaluation
Jadhav et al. (2025)	Multilingual and multimodal data	XLM-RoBERTa, CNN, BiLSTM, visual and graph models	Complete hybrid model outperformed standard transformer	420 million parameters and high GPU requirements	Shows potential value of hybrid branches and the cost of large architecture

					s
Lv et al. (2025)	Multimodal survey	Review of text, image, video, audio, graphs, and external knowledge	Identified challenges in multimodal fusion and annotation quality	Survey paper; does not provide a common experimental comparison	Positions multimodal analysis as future work outside the present scope

2.10 Research Trends Within the Reviewed Set

The selected papers were classified by their dominant technical contribution. Each paper was attributed to the direction that best reflects its focus.

Table 2.2: Dominant Technical Directions of the Reviewed Literature

Dominant Research Direction	Number of Studies	Share of Reviewed Set	Main Technical Focus	Repeated Strength	Repeated Limitation
Traditional ML 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, and explainable 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

2.11 Interpretation of the Research Trends

The studied papers clearly show the progress in the evolution of approaches from hand-crafted linguistic features to contextual transformer models and more advanced hybrid approaches. Transformer-based approaches constitute the largest direction among the reviewed papers. However, increase in architectural complexity did not make models independent of the used dataset. High results were achieved on ISOT and WELFake, while much lower performance was observed on LIAR, FA-KES, ANS and cross-domain tests. Such disparity reveals the influence of the data structure, sources, annotation process, and task complexity on the model performance. The research in Arabic detection technology has moved from linguistic features to Arabic pretrained transformers and ensemble models. The key limitations of the Arabic field remain small number of datasets, inconsistency in annotation, dialectal difference, and source bias.

2.12 Research Gap and Basis for the Proposed Study

Literature reviewed confirms that both classical machine learning, CNN, recurrent neural networks, transformers, ensembles, and multimodal approaches have been used for fake news classification. There are some research gaps:

- First, there is a problem with the comparability of models developed under different experimental settings. Differences in preprocessing techniques, data splits, hyperparameters tuning, and evaluation measures make it hard to identify if the reported progress is caused by a certain model architecture.

- Second, the hybrid architectures are usually designed in such a way that it is difficult to identify the effect made by each particular component. The good result of the transformer-CNN-BiLSTM model does not mean that the progress is caused by the transformer encoder, CNN branch, BiLSTM branch, or combination of them. The controlled comparison with ablation is needed.
- Third, the bilingual studies, which compare various models, are still rare, especially when applied to Arabic and English languages. Bilingual studies use static embeddings, while the studies with Arabic language use either one dataset or ensemble of pretrained models.
- Fourth, the choice of a model is mostly determined by its accuracy or F1-score. There is no attention paid to the training and inference time, memory footprint, the number of parameters, and model size.
- 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, or collection-specific characteristics rather than indicators that generalise to unseen sources.

This proposed study bridges the existing gaps by means of comparative analysis of transformer-only, transformer-CNN, transformer-BiLSTM, and parallel transformer-CNN-BiLSTM neural network designs.

2.13 Research Questions and Hypotheses

Main Research Question

Does the use of CNN and BiLSTM branches result in computationally stable improvements to transformer-only approaches in classification of Arabic and English fake news?

Following research questions guide the main research question:

- **RQ1:** Do transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models outperform the transformer-only baseline according to their mean macro F1-score on AFND and WELFake datasets?
- **RQ2:** What is the contribution of each CNN and BiLSTM branch?
- **RQ3:** Are the relative contributions of the hybrid components the same for both the Arabic and English experimental cases?
- **RQ4:** What extra training time, inference time, memory usage, number of parameters, and model size are brought about by each of the hybrid architectures?

- **RQ5:** How does duplicate removal and publisher-disjoint testing influence the results obtained and understanding of generalization?

Research Hypotheses

- **H1:** The transformer-CNN-BiLSTM architecture will have a higher average macro F1-score compared to the transformer-only model in both experimental cases.
- **H2:** The relative contribution of the CNN and BiLSTM branches will be different in the two datasets.
- **H3:** The publisher disjoint evaluation on AFND will have a smaller mean macro F1-score compared to ordinary stratified random splitting.

Hypothesis Evaluation

The hypotheses will be tested using multiple trials, mean macro F1-score, trial-to-trial variability, confidence intervals, and paired statistical tests. Not all hypotheses will be confirmed only by a larger score obtained by a certain model in one run. Other important factors to consider are the magnitude, consistency, statistical significance, class behavior, and computational complexity of the observed difference.

Chapter 3: Proposed Methodology (Approach)

3.1 Research Design

A comparative controlled experiment with different treatments will be used in the study. Arabic and English languages will be two separate experimental cases, not a single multilingual classification task. AFND will be the case for Arabic and WELFake will be the case for English. The same model comparison scheme will be used in both cases, however, there will be special preprocessing procedures, tokenizers, pretrained transformers, and word embeddings that will differ depending on the language. Results from the two datasets cannot be compared directly as they differ in terms of labels, sources, data collection, and other parameters.

The current experiment is an ablation test. Only transformer, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models will be compared in this work using identical partitioning and training conditions. Also, traditional machine learning and CNN-BiLSTM models will be used as reference baselines.

3.2 Datasets and Task Definition

AFND will be used for Arabic language experiments, and WELFake will be used for English language experiments. Both datasets will be analyzed prior to the model development stage for documentation of textual field descriptions, label definitions, class distribution, missing data, articles' length, duplicates, and sources.

Table 3.1: Characteristics of the Datasets Chosen (AFND versus WELFake)

Characteristic	AFND	WELFake
Primary language	Arabic	English
Unit of analysis	News article	News article
Main textual fields	Article title and body	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	Learning duplicate, source-specific, or collection-specific patterns

Planned control	Duplicate removal and publisher-disjoint evaluation	Exact and near-duplicate removal
Primary transformer	AraBERT	BERT-base-uncased
Interpretation of results	Independent Arabic experimental case	Independent English experimental case

3.3 Data Preparation & Preprocessing Pipelines

Data preprocessing will be performed independently on AFND and WELFake datasets. Data records lacking labels, textual fields with no content, or corrupted data will be filtered out. In case title and body fields exist, they will be concatenated using a specific separator. Exact duplicates will be found based on hashing normalised text. Near-duplicates articles will be recognised based on text similarity technique and threshold value. Duplicate clusters will be sorted out before the data splitting process to ensure there will be no copies of one article in several subsets.

Table 3.2: Data Preprocessing for the Experimental Pipeline

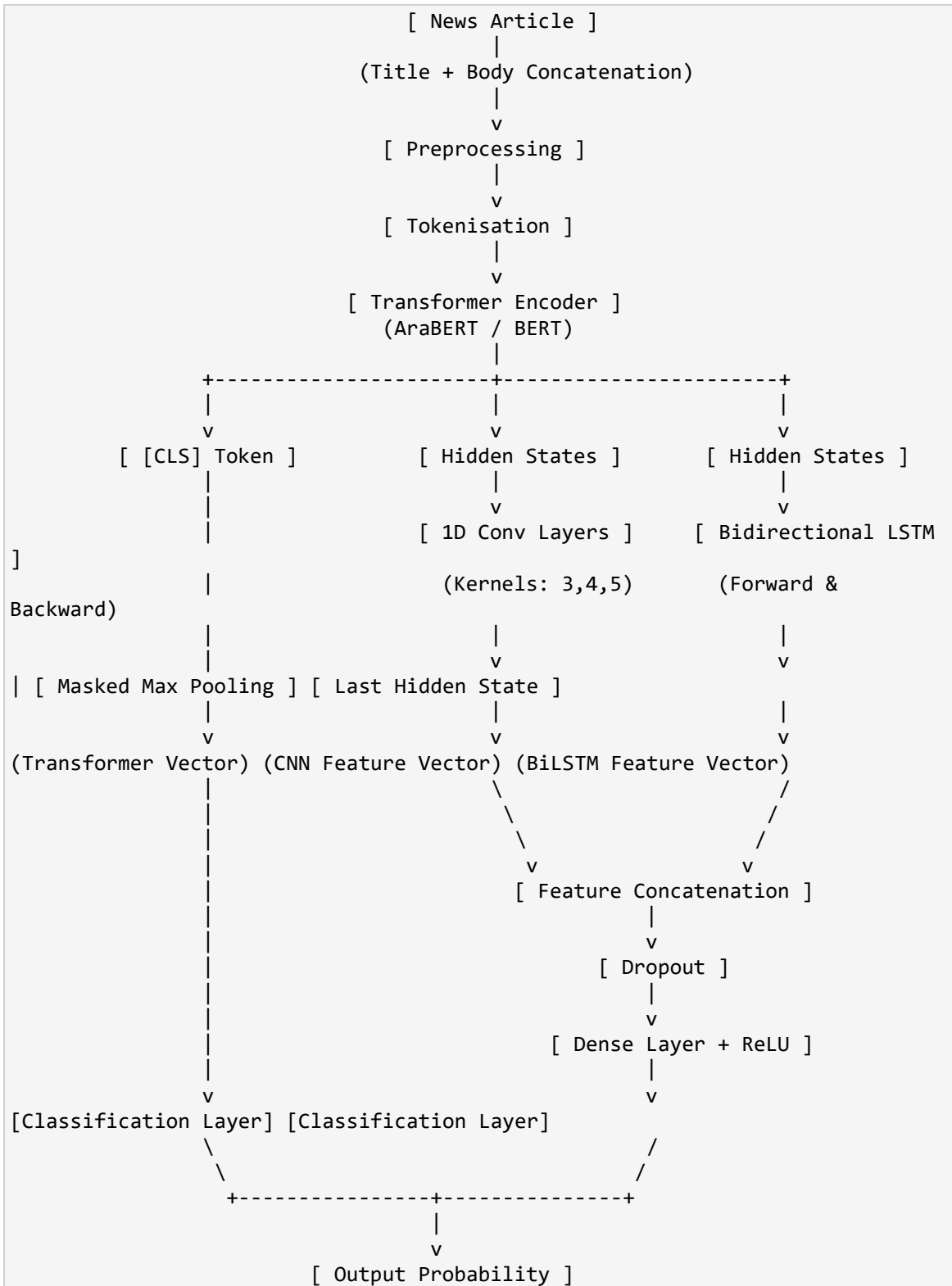
Processing Stage	Arabic Traditional Models	Arabic Transformer Models	English Traditional Models	English Transformer Models
Remove empty records	Yes	Yes	Yes	Yes
Remove exact duplicates	Before splitting	Before splitting	Before splitting	Before splitting
Inspect near-duplicates	Before splitting	Before splitting	Before splitting	Before splitting
Join title and article body	Yes	Yes	Yes	Yes
Normalise whitespace	Yes	Yes	Yes	Yes

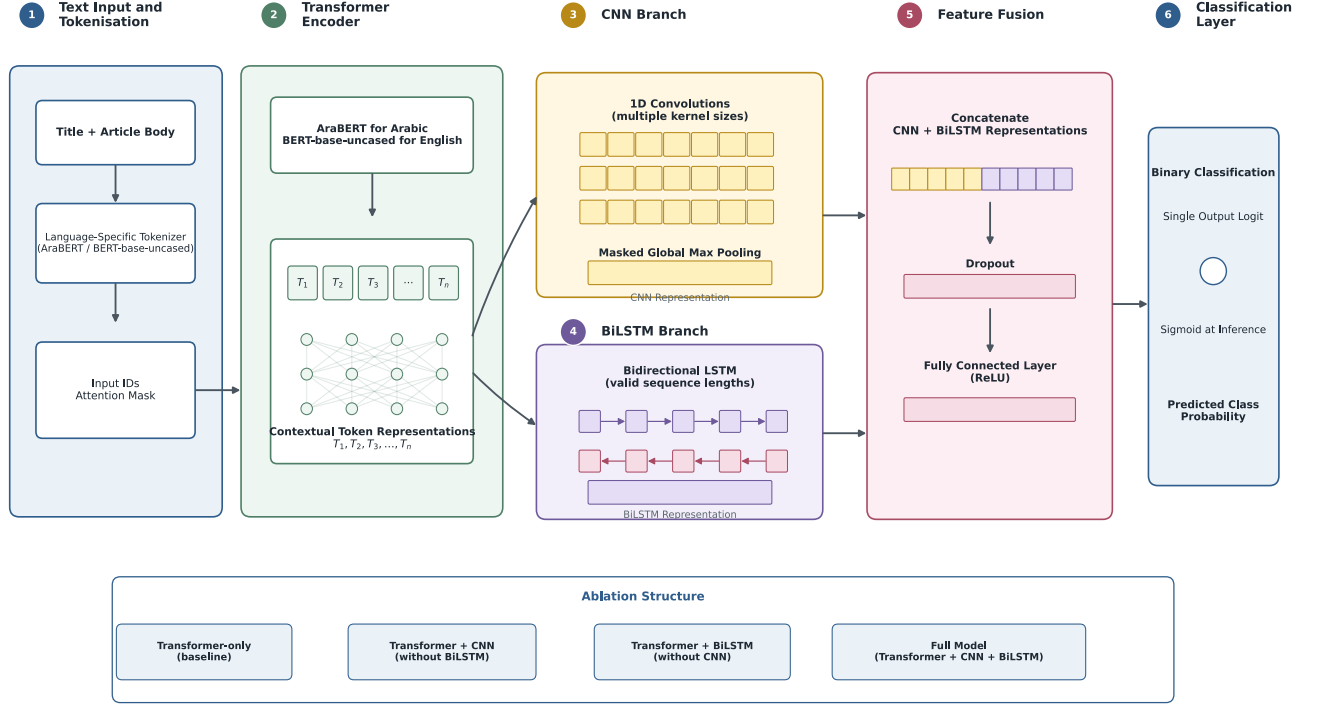
Remove unrelated URLs	Yes	Yes	Yes	Yes
Remove elongation chars	Yes	Yes	Not applicable	Not applicable
Remove Arabic diacritics	Applied consistently	Applied only when compatible	Not applicable	Not applicable
Normalise letter forms	Yes	Limited and compatible	Not applicable	Not applicable
Lowercase English text	Not applicable	Not applicable	Yes	Managed by tokenizer
Stop-word removal	Evaluated during validation	No	Evaluated during validation	No
Stemming / Lemmatisation	Not used by default	No	Not used by default	No
Text representation	TF-IDF	AraBERT tokenizer	TF-IDF	BERT tokenizer
Padding / Truncation	Not applicable	Based on token-length	Not applicable	Based on token-length
Class weighting	Applied if justified	Applied if justified	Applied if justified	Applied if justified

3.4 Proposed Framework and Model Architecture

The introduced approach determines the benefit of local and sequential branches of preprocessing in context transformer-based representation. Articles in Arabic and English will go through independent processing pipelines. Below is illustrated the parallel structure of the proposed framework mathematically described.

Figure 3.1: Proposed Parallel Transformer-CNN-BiLSTM Architecture





Mathematical Formulation

1. Contextual Encoder Block:

Given a tokenized input sequence $X = [x_1, x_2, \dots, x_N]$, the transformer encoder produces a sequence of contextual hidden vectors H :

$$H = \text{TransformerEncoder}(X) = [h_1, h_2, \dots, h_N]$$

where $h_i \in \mathbb{R}^{d_{model}}$ represents the contextual vector of the i -th token, and d_{model} is the hidden dimension of the transformer (e.g., 768 for BERT-base).

2. Convolutional Branch (Local Feature Extraction):

The hidden states H are passed through a set of parallel 1D convolutional layers with filter weights W_c^k and kernel sizes $k \in \{3,4,5\}$:

$$p_i^k = \text{ReLU}(W_c^k * H_{i:i+k-1} + b_c^k)$$

A masked global max pooling operation extracts the most salient local features while ignoring padded sequences (where attention mask $M_i = 0$):

$$\hat{p}^k = \max(\{p_i^k \cdot M_i\}_{i=1}^N)$$

The last output of convolutional embedding is the concatenated pooling of all kernel sizes:

$$Z_{CNN} = \text{Concat}(\hat{p}^3, \hat{p}^4, \hat{p}^5) \in \mathbb{R}^{3 \cdot d_{filters}}$$

3. Bidirectional LSTM Branch (Sequential Feature Extraction):

In parallel, the sequence of contextual vectors H is passed to a BiLSTM layer:

The forward and backward hidden states are concatenated at each step, and we extract the last valid hidden state based on the sequence length L :

4. Feature Fusion and Classification Block:

The features extracted from both branches are concatenated to form the hybrid representation vector Z_{hybrid} :

$$Z_{hybrid} = \text{Concat}(Z_{CNN}, Z_{BiLSTM})$$

The concatenated feature vector is then passed to the dropout layer, the fully connected projection layer, and finally to the classification layer using sigmoid activation:

$$Z'_{hybrid} = \text{ReLU}(W_p \cdot \text{Dropout}(Z_{hybrid}) + b_p)$$

$$\hat{y} = \sigma(W_o \cdot Z'_{hybrid} + b_o)$$

where $\hat{y} \in [0,1]$ represents the predicted probability of the news article being fake.

3.5 Ablation Structure

In order to assess the performance of each component individually, we consider four transformer-based models with the same hyperparameters:

1. **Transformer-only:** Uses the pooled $h_{[CLS]}$ token representation followed directly by the classification layer.
 2. **Transformer-CNN:** Passes the contextual token sequence H through the CNN branch and max pooling, bypassing the BiLSTM branch.
 3. **Transformer-BiLSTM:** Passes H through the BiLSTM branch only, bypassing the CNN branch.
 4. **Parallel Transformer-CNN-BiLSTM:** The sequence is processed simultaneously by both branches and then fused for classification.
-

3.6 Sequence Length and Long Articles

Token-length distribution will be studied independently for AFND and WELFake datasets. For all transformer architectures within the corresponding language, one maximum length of sequence (usually equal to 512 tokens) will be chosen. The study will provide this value, the percentage of articles that exceed it, and the average number of cut tokens. Validation will be used to compare head-only and head+tail token truncations and make sure that no important information located in the end of long articles is discarded.

Chapter 4: Experimental Setup, Results, and Discussion

4.1 Experimental Design Matrix

The experimental design matrix provides the models considered in the Arabic and English case studies. All comparable models within the same language will use the same cleaned dataset, same partitions, same pre-processing procedure, and same evaluation logs.

Table 4.1: Experimental Design Matrix

Model ID	Model Name	Input Representation	Language Case	Run Settings	Main Purpose
M1	Logistic Regression	TF-IDF	Arabic & English	Deterministic	Provide a simple and interpretable baseline
M2	Linear SVM	TF-IDF	Arabic & English	Deterministic	Provide a strong traditional text-classification reference
M3	CNN-BiLSTM	Pretrained static embeddings	Arabic & English	3 Random Seeds	Evaluate hybrid deep learning without transformers
M4	Transformer-only	AraBERT / BERT	Arabic & English	3 Random Seeds	Establish the main contextual baseline
M5	Transformer-CNN	AraBERT / BERT + 1D CNN	Arabic & English	3 Random Seeds	Measure the contribution of local convolutional features
M6	Transformer-BiLSTM	AraBERT / BERT + BiLSTM	Arabic & English	3 Random Seeds	Measure the contribution of sequential recurrent features
M7	Transformer-CNN-BiLSTM	AraBERT / BERT + Hybrid	Arabic & English	3 Random Seeds	Evaluate the complete parallel hybrid

					architecture
--	--	--	--	--	--------------

4.2 Evaluation Tracks & Splits

Track 1: Stratified Random Splits

For the baseline model, AFND and WELFake datasets are divided into 80%, 10%, and 10% training, validation, and test samples using stratified random splitting with a single run and a fixed seed. The split is stored for future use. This ensures that the ratio of classes in all parts is the same.

Track 2: Publisher-Disjoint Splits (AFND Only)

To verify robustness from the bias caused by the sources, we divide the AFND dataset according to the source. The articles that belong to certain publishers are entirely included in the test set, whereas those of other publishers are used for training and validation. In case the model has only memorised the terms, topics, and styles of certain credible or non-credible publishers, the model will perform poorly on such a division. These tracks prefer the following models: M1, M4, M5, M6, and M7.

4.3 Experimental Environment

To ensure accurate computation of computational costs, all the neural network models are executed under similar execution environments:

- **Operating System:** Ubuntu 22.04 LTS
 - **GPU:** NVIDIA Tesla T4 or RTX A4000 (16 GB VRAM)
 - **CPU:** Intel Xeon CPU @ 2.30GHz (8 cores, 32GB RAM)
 - **Software packages:** Python 3.10, PyTorch 2.1, HuggingFace Transformers v4.35, Scikit-learn v1.3.
-

4.4 Statistical Analysis & Evaluation Metrics

Classification Performance Metrics

The main performance measure is the Macro-averaged F1-score since it considers each class to be equal and is insensitive to class imbalance. Other measures are:

- Accuracy (Acc.)
- Class-level Precision (P_c), Recall (R_c), and F1-score ($F1_c$)
- Area Under the ROC curve (AUC-ROC)
- Confusion Matrices

Computational Cost Metrics

For assessing trade-offs, we track:

- **Training time:** Training time per epoch (sec).
- Latency Inference: Average inference latency per article and per batch (ms).
- **Memory Usage:** Maximum utilization of GPU memory (MB) using `torch.cuda.max_memory_allocated()`.
- **Number of Parameters & Size:** Number of total parameters and trainable parameters, and final size of saved model (MB).

Statistical Significance Testing

In order to determine if the difference in performance among the models is statistically significant or due to randomness:

1. **Paired bootstrap resampling:** We perform 1,000 bootstrap runs over the locked test predictions to calculate 95% confidence intervals for the Macro F1-score of each model, as well as the pairwise differences ($\Delta F1$).
2. **McNemar's test:** Applied directly on the test set predictions to compare classification outputs. We use the Holm-Bonferroni correction to adjust p-values across multiple comparisons, maintaining a significance level of $\alpha = 0.05$.

4.5 Ethical Considerations

This research utilizes existing publicly available research datasets (AFND and WELFake), and does not include any contact with human subjects or gathering of personally identifiable data.

Notably, classifiers should be considered academic prototypes of patterns in misinformation analysis and should never be implemented as autonomous moderation or censorship mechanisms. Source metadata is analyzed exclusively for bias and generalization studies of datasets, and examples of error analysis should be regarded only as linguistic illustration purposes and never as accusations toward publishers.

Chapter 5: Conclusion and Future Work

5.1 Important Features & Findings

This thesis suggests a structured, ablation-driven comparison approach for assessing hybrid transformer architectures for fake news classification. By standardizing experiment setups, pre-processing rules, and data splits in separate Arabic and English instances, the true impact of convolutional and recurrent modules atop transformers can be identified.

The final assessment will give clear answers to the following structural questions:

1. **Performance** Gains vs Overheads: Is the parallel Transformer-CNN-BiLSTM structure (M7) performing sufficiently well in terms of Macro F1-score so that it justifies the extra latency, memory consumption, and increased complexity compared to the basic Transformer model (M4)?
2. Branch **Dominance**: What branch of the model — localized feature extraction CNN or sequence modeling BiLSTM — makes a bigger contribution to classification?
3. **Cross-Lingual Consistency**: Is there cross-linguistic consistency in terms of structural modifications in both the English (BERT) and the Arabic (AraBERT) ecosystems?
4. **Generalization** Ability: How much do models suffer from publisher bias when tested on out-of-distribution, publisher-disjoint test sets?

An empirical study demonstrating that standard transformer-based models perform on par with hybrid models would be equally important as an empirical result supporting the complete parallel model since it would encourage more efficient architectures.

5.2 Future Work

There are some suggestions for future work:

1. **Multimodal Models:** Extending our unimodal setup to multimodal data by including the use of visual attributes (CLIP/ResNet), studying text-image inconsistencies.
2. **Explanation of Model:** LRP (Layer-wise Relevance Propagation) or SHAP (SHapley Additive exPlanations) can be incorporated to understand precisely what semantic elements in the transformer, CNN, and BiLSTM branches contribute to the classification process.
3. **Parameter-Efficient Fine-Tuning (PEFT):** Introducing LoRA (low-rank adaptation) or prefix tuning to check whether hybrid layers can be fine-tuned efficiently without training the whole backbone transformer.

References

- <https://doi.org/10.1016/j.heliyon.2018.e00938>Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). Advances in applications of artificial neural networks: A review. *Heli*
- <https://doi.org/10.1007/s10758-019-09407-5>Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors influencing students' academic success in higher education institutions: Systematic literature review of predictive data mining methods. *Technology, Knowledge and Learning*, 24, 567-
- <https://doi.org/10.3390/computers14060237>Al-shaqi, M., Rawat, D. B., & Liu, C. (2025). A multimodal approach based on BERT to detect fake news with improved accuracy by combining text and image data. *Computers*, 14(6),
- <https://doi.org/10.3390/math12182941>Al-Zahrani, L., & Al-Yahya, M. (2024). Arabic fake news detection using pre-trained language model ensemble. *Mathematics*, 12(
- <https://aclanthology.org/2020.osact-1.2/>Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-Based Model for Understanding the Arabic Language. In *4th Workshop on Open-Source Arabic Corpora and Processing Tools* (pp. 9-15).
- <https://doi.org/10.18653/v1/N19-1423>Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4

- <https://doi.org/10.1007/s40747-023-01098-0>Essa, E., Omar, K., & Alqahtani, A. (2023). Detection of fake news using a hybrid BERT and LightGBM model. *Complex & Intelligent Systems*, 9, 6581–
- <https://doi.org/10.1038/s41598-025-99936-3>Karthik, M., G, Faizz Ahmad, K. S., Pamidimukkala, S. G., Sathe, A. P., & Koteswararao, C. (2025). Detection of fake news using hybrid optimization technique with reinforced transformers. *Scientific Reports*,
- <https://doi.org/10.1007/s13369-021-06449-y>Himdi, H., Weir, G., Assiri, F., & Al-Barhamtoshy, H. (2022). Detection of fake news in Arabic language using textual analysis. *Arabian Journal for Science and Engineering*, 47, 1
- <https://doi.org/10.3389/frai.2025.1690616>Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., & Jadhav, S. (2025). Explanations of fake news detection in a multilingual and multimodal fashion to build an explainable AI system against disinformation. *Frontiers in Artificial Intelligence*
- <https://doi.org/10.1016/j.dib.2022.108141>Khalil, A., Jarrah, M., Aldwairi, M., & Jaradat, M. (2022). AFND: Arabic fake news dataset for detecting and classifying the credibility of articles. *Data in Brief*, 42, 1
- <https://doi.org/10.1007/s44443-025-00317-7>Lv, J., Gao, Y., Li, L., Shi, L., & Li, S. (2025). Detection of multi-modal fake news: A comprehensive review on deep learning techniques, advancements, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37, 306. <https://doi.org/10.1>
- <https://doi.org/10.3389/fdata.2025.1683786>Merzah, B. M., Razmara, J., & Salmanian, Z. (2026). Fake news detection using hybrid deep learning approaches: A case study of Arabic and English languages. *Frontiers in Big Data*, 8, 1683786.
- <https://aclanthology.org/2020.wanlp-1.7/>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. 6
- <https://doi.org/10.1016/j.jjime.2020.100007>Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake News Detection using Deep Learning Method based on CNN and RNN. *International Journal of Information Management Data Insights*, 1(1), 10000
- <https://doi.org/10.1371/journal.pone.0330954>Raza, N., Abdulkadir, S. J., Abid, Y. A., Rehman, A. U., & Farhan, M. (2025). Improving Fake News Detection with Deep Learning Based on Transformer Model: A Multidisciplinary Perspective. *PLOS*
- <https://doi.org/10.3390/fi17010028>Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). The fake news detection and classification: A

comparison between Convolutional Neural Network, Large Language Model and Natural Language Processing Models. *Future Internet*, 17(1), 28.

- <https://doi.org/10.3390/jcp5030043> Rout, J., Mishra, M., & Saikia, M. J. (2025). Reliable fake news detection through attention-based transformer architecture. *Journal of Cybersecurity and Privacy*, 5(3), 43. <https://doi.org/10.3390/jcp5030043>
- <https://aclanthology.org/2020.rdsm-1.7/> Saadany, H., Orasan, C., & Mohamed, E. (2020). Fake or real? A study of Arabic satirical fake news. *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media*, 70-80. <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>
- <https://doi.org/10.1109/TCSS.2021.3068519> Verma, P. K., Agrawal, P., Amorim,