



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

Prepared By

Mohammed Aljaaidi ID No.: 232000843

Mohammed Abusultan ID No.: 253002925

Khalid Abbas Elzein ID No.: 261001961

**SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
MASTER OF SCIENCE IN ARTIFICIAL INTELLIGENCE**

At

College of Informatics, Midocean University, Comoros

Supervisor Name: Dr. Yasser Abdelhamid

August 2026

Acknowledgements

We would like to express our sincere gratitude to our supervisor, Dr. Yasser Abdelhamid, for his valuable guidance, constructive feedback, and continuous support throughout the development of this research.

We also extend our appreciation to the College of Informatics at Midocean University for providing the academic environment and resources that supported this work.

Finally, we would like to thank our families, colleagues, and everyone who contributed directly or indirectly to the completion of this study.

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 powerful 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. Arabic Fake News Dataset (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 Bidirectional Encoder Representations from Transformers (BERT-base-uncased). The comparison will include TF-IDF-based Logistic Regression and Linear Support Vector Machine, Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (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, Receiver Operating Characteristic Area Under the Curve (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.

List of Publications

No publications arising from this research have been published or accepted for publication to date.

Table of Contents

Acknowledgements.....	i
Abstract.....	ii
List of Publications.....	iv
Table of Contents	v
List of Abbreviations	viii
List of Tables.....	ix
List of Figures	ix
:Chapter (1).....	Introduction
.....	1
1.1 Overview.....	1
1.2 Problem Definition	3
1.2.1 Research Questions.....	4
1.2.2 Research Hypotheses	5
1.2.3 Hypothesis Evaluation.....	5
1.3 Thesis Objectives.....	5
1.4 Thesis Contribution.....	6
1.5 Thesis Organization	8
Chapter (2): Background and Related Work.....	9
2.1 Background.....	9
2.1.1 Fake News as a Text Classification Problem	9
2.1.2 Traditional Machine Learning Approaches	9
2.1.3 Deep Learning and Hybrid Models.....	10
2.1.4 Transformer-Based Language Models.....	10
2.1.5 Arabic Language Challenges.....	11

2.1.6	Dataset Bias and Generalization.....	11
2.2	Related Work	12
2.2.1	Review Scope and Study Selection	12
2.2.2	Traditional Machine Learning and Linguistic Features	14
2.2.3	CNN and Recurrent Neural Networks.....	14
2.2.4	Transformer-Based Classification	15
2.2.5	Arabic Pretrained Language Models and Ensembles	17
2.2.6	Multimodal, Multilingual, and Explainable Approaches	17
2.2.7	Dataset Dependence and Generalization.....	19
2.2.8	Comparative Synthesis of the Reviewed Studies	19
2.2.9	Research Trends Within the Reviewed Set.....	20
2.2.10	Interpretation of Research Trends.....	28
2.2.11	Research Gap and Basis for the Proposed Study.....	28
2.3	Chapter Summary	30
Chapter (3): Proposed Approach (Methodology)		32
3.1	Research Design.....	32
3.2	Datasets and Task Definition	32
3.3	Data Preparation	33
3.4	Experimental Splits	34
3.5	Models.....	35
3.6	Training Procedure	36
3.7	Evaluation.....	37
3.8	Robustness and Error Analysis	38
3.9	Dataset Profile and Labelling Structure	38
3.10	Preprocessing Pipelines.....	41
3.11	Proposed Framework and Model Architecture	43

3.11.1	Text Input and Tokenization	43
3.11.2	Transformer Encoder	43
3.11.3	CNN Branch	43
3.11.4	BiLSTM Branch	44
3.11.5	Feature Fusion	44
3.11.6	Classification Layer	44
3.11.7	Ablation Structure	46
3.11.8	Sequence Length and Long Articles	46
3.12	Experimental Design Matrix	47
3.12.1	Evaluation Tracks	49
3.13	Experimental Environment	49
3.14	Statistical Analysis	50
3.15	Ethical Considerations	51
	References	52

List of Abbreviations

Abbreviation	Full Term
AFND	Arabic Fake News Dataset
AI	Artificial Intelligence
AraBERT	AraBERT
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
F1	F1-Score
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
OCR	Optical Character Recognition
RNN	Recurrent Neural Network
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
WELFake	Word Embedding over Linguistic Features for Fake News Detection
XLM-RoBERTa	Cross-lingual Language Model RoBERTa

List of Tables

Table 1-1: Thesis Organization Explanation	8
Table 2-1: Comparative Summary of the Reviewed Studies	21
Table 2-2: Dominant Research Trends Within the Reviewed Studies.....	27
Table 3-1: Profile of the Selected Datasets	39
Table 3-2: Preprocessing Procedures for the Experimental Pipelines	41
Table 3-3: Experimental Design Matrix	48

List of Figures

Figure 3-1: Proposed Parallel Transformer-CNN-BiLSTM Architecture	45
---	----

Chapter (1): Introduction

1.1 Overview

Digital news channels and social media have sped up the rate of publishing and distributing the news articles. While this has resulted in easier access to news, it has also made faster the delivery of false news content to a large number of people. Large volume of data available online makes manual verification very difficult and has driven the researchers to devise automated techniques for fake news detection (Touahri & Mazroui 2024; Lv et al. 2025).

Previous approaches to news classification were mainly dependent on traditional machine learning approaches like Logistic Regression, Support Vector Machine (SVM), Naive Bayes and Random Forest. These methods usually represented the text using bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF), and hand-crafted linguistic features. However, deep learning models used CNN, LSTM, and BiLSTM to learn local and sequential text patterns without the need of handcrafted features (Touahri & Mazroui, 2024).

Pretrained transformer models helped in creating contextual representation of the language and demonstrated outstanding results in several text classification tasks (Devlin et al., 2019). BERT and Arabic pretrained models like Arabic Bidirectional Encoder Representations from Transformers (AraBERT) (Antoun et al., 2020) can generate representations of the texts based on their contexts.

Recently, research on fake news detection has employed transformers together with CNN, recurrent neural networks, ensemble techniques, and multimodal architectures. The added complexity of the mentioned approaches does not necessarily lead to consistent improvements for different datasets, languages, and evaluations (Al-Zahrani & Al-Yahya, 2024; Jadhav et al., 2025; Touahri & Mazroui, 2024).

Classification of fake news in Arabic involves several problems related to morphology, spelling variation, dialectical variety, and inconsistent labelling of the datasets (Touahri & Mazroui, 2024). The construction of dataset affects evaluation of the proposed approaches. When labels refer to the credibility of the publishing entity, or duplicated articles and source-related content occur in the train and test partitions, then the classifier can recognize the source or its editorial style, instead of common features of misleading information (Khalil et al., 2022; Verma et al., 2021).

The current study aims to review recently published research on fake news classification in Arabic and English and suggests comparing traditional, deep learning, transformer-only, and hybrid transformer models. In addition, AFND and WELFake will be considered as two distinct experimental cases related to fake news in Arabic and English. The study will check whether CNN and BiLSTM branches are consistently able to improve the performance of the transformer-only baseline models and whether such improvement is worth additional costs of computations.

The study considers fake news classification as the supervised problem using labels assigned in the selected datasets. This study does not conduct independent verification of the claims in order to assess the truthfulness. Contributions of the study include controlled comparison of components, computational resources, bias related to the source, and generalization under consistent experimental settings.

1.2 Problem Definition

Transformers performed well in fake news classification, however, the reported performances varied considerably depending on the dataset, language, and evaluation setting. A direct comparison between studies was not feasible as they used different data preprocessing methods, data splitting, model setup, and evaluation metrics (Touahri & Mazroui, 2024; Merzah et al., 2026; Rout et al., 2025).

Several studies include CNN and BiLSTM components in addition to transformers in order to capture local and sequential textual features (Jadhav et al., 2025). High performance of a hybrid architecture does not imply that each added component adds something valuable. It should be estimated relative to a transformer baseline under the same conditions and taking into account the cost of computation.

The existing literature also does not cover Arabic and English. There are morphological differences between these languages, including spelling variants and differences in language resources and dataset creation process (Antoun et al., 2020; Touahri & Mazroui, 2024). A hybrid architecture, which performs better with English data, does not necessarily lead to performance increase for Arabic data using a language-specific transformer.

There is another issue related to the structure of the dataset. AFND assigns publisher's credibility to articles, so there is a possibility that models do not use credibility but learn source style (Khalil et al., 2022). WELFake consists of articles from several previous datasets and could contain duplications or specific features of particular dataset (Verma et al., 2021). So, random article-level splitting may lead to overly optimistic results due to overlapping content and repeated source characteristics in training and testing data (Khalil et al., 2022; Verma et al., 2021).

The research problem is the absence of evidence about whether the inclusion of CNN and BiLSTM branches is supported by stable, statistically reliable, and

computationally reasonable improvements over transformer-only baselines for Arabic and English fake news detection.

The research question is addressed by the ablation-based comparison of model performances, computational costs, presence of duplicate data, and source biases without a prior assumption about the superiority of a full hybrid architecture.

1.2.1 Research Questions

The research is focused on answering the following primary research question:

Main Research Question

Is there evidence of consistent and computationally sensible performance gains from CNN and BiLSTM branches against transformer-based baselines for Arabic and English fake news detection?

The following research questions will be used to complement the primary research question:

RQ1: Can transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM architectures exhibit higher macro F1-scores than a transformer baseline architecture in AFND and WELFake datasets?

RQ2: What contributions do CNN and BiLSTM branches make independently and together?

RQ3: Do the relative impacts of the hybrid elements remain the same between the Arabic and English experimental setups?

RQ4: How much extra training time, inference time, memory usage, number of parameters, and model size are incurred by each hybrid structure?

RQ5: How does duplication removal and publisher-disjoint testing impact the results presented?

1.2.2 Research Hypotheses

The following hypotheses are to be tested:

H1: The transformer-CNN-BiLSTM model will have a greater mean macro F1-score compared to the transformer only model within each experimental setup.

H2: The CNN and BiLSTM branches' relative impacts on classification will vary between AFND and WELFake.

H3: Evaluation that separates instances by publisher on AFND will result in a smaller mean macro F1-score than evaluation using stratified random splitting.

1.2.3 Hypothesis Evaluation

The hypotheses will be tested using repeated experiments, mean macro F1-score, consistency across runs, confidence intervals, and paired tests.

Hypotheses will not be accepted because of the relative scores of one model over another in one experiment. Differences in size, stability, statistical significance, class performance, and computational complexity will also factor into the decision about hypotheses acceptance.

1.3 Thesis Objectives

The general goal of this research project is the controlled assessment of transformer-based hybrid models for fake news detection in Arabic and English, and whether the extra benefits justify the extra costs.

The specific objectives are:

1. For the exploration of the dataset's structure, labelling, class distribution, missing values, duplicates, article length, and source availability in AFND and WELFake.
2. For the creation of independent Arabic and English data processing pipelines for traditional machine learning, deep learning, and transformers.
3. For implementing baselines such as traditional baselines based on TF-IDF, CNN-BiLSTM baselines, transformers, and transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models.
4. For performing an ablation study under a controlled environment with the same data splits, training methods, evaluation, and fixed random seed values.
5. For assessing classification accuracy, statistical significance, and computational efficiency in terms of macro F1-score, class-level scores, confidence interval, training time, inference time, memory usage, number of parameters, and model size.
6. In order to investigate the impact of duplicated content and biases due to sources via deduplication, publisher-independent evaluation on AFND, and error analysis.
7. In order to find out which architectural design offers the best justification for balancing predictive performance, computational expense, and generalization capacity.

1.4 Thesis Contribution

This research has four primary contributions. First, it performs a systematic synthesis of recently conducted research on Arabic and English fake news classification. Second, it conducts an ablation-based comparison separately for Arabic and English experiments with language-specific transformers. Third, it considers predictive performance in conjunction with computational expense

and repeated runs' stability. Finally, it explores whether duplicated content and biases due to publishers give rise to overly optimistic generalization estimates.

The importance of this work is in the controlled comparison of model components instead of simply comparing architectures. Fake news detection studies performed using traditional machine learning, deep learning, transformers, and hybrid architectures have reported impressive results; however, due to the difference in the data used and experiment design, it is impossible to compare the results (Nasir et al., 2021; Touahri & Mazroui, 2024; Merzah et al., 2026).

This study will be able to compare transformer-only and hybrid models under identical data splits, preprocessing rules, training protocols, evaluation metrics, and computational resources. This will allow the researcher to determine whether the differences in performance are caused by CNN and BiLSTM branches rather than other factors.

Moreover, the research relates to Arabic Natural Language Processing. There are fewer Arabic fake news datasets compared to English ones, and they differ in terms of the language variation, annotation process, and source diversity (Touahri & Mazroui, 2024). In the Arabic experiment, a language-specific pretrained transformer will be applied, instead of the English model without linguistic adjustments.

On the other hand, the English experiment will give another example of a similar architectural problem. Neither raw scores of AFND and WELFake will be analyzed for choosing the easier language since the main objective is to explore whether the same architectural changes provide similar performance within each specific dataset.

Another novelty of the proposed research is a consideration of computational cost and robustness to the source. For instance, a small increment of macro F1-score may not be significant if it demands more time and memory for training and inference processes. Evaluation of publisher-disjoint performance on

AFND will give even better information about the generalization capacity of each model outside familiar sources.

In any case, regardless of the model's performance, the study will prove to be relevant. Evidence of the equal performance of a simpler transformer model at lower cost will be just as useful as the results that support the whole hybrid architecture.

1.5 Thesis Organization

Table 1-1: Thesis Organization Explanation

Chapter	Description
1	Presents research problems, questions, objectives, and contributions.
2	Provides the background and reviews the related literature.
3	Presents the proposed methodology and experimental framework.
4	Reports the experimental results and discussion.
5	Presents the conclusions and future work

Chapter (2): Background and Related Work

2.1 Background

2.1.1 Fake News as a Text Classification Problem

In general, fake news is an umbrella concept that includes false or misleading pieces of information delivered in a way that imitates a news article. In computational datasets, it is understood as a set of different concepts, such as a fabrication, misinformation, satire, rumor, manipulation, or news from unreliable sources.

Most of the automated techniques solve the problem through supervised text classification (Touahri & Mazroui, 2024). The classifier learns patterns from labeled examples and then classifies unseen text accordingly. Automatic fact verification, on the contrary, requires evidence from outside and evaluation of the factual claims of an article.

In the current research, performance of the classifier is assessed based on the labels assigned by AFND and WELFake. Hence, the results of this paper are defined by the definitions used by these datasets and do not imply independent verification of the content of articles.

2.1.2 Traditional Machine Learning Approaches

Traditionally, algorithms for fake news classification were based on Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest. The textual representation included bag-of-words, TF-IDF or manually chosen linguistic features (Himdi et al., 2022; Touahri & Mazroui, 2024).

Such techniques provide computationally efficient baselines for comparison. Yet, their success largely depends on the choice of features and nature of the

training dataset (Himdi et al., 2022; Touahri & Mazroui, 2024). Patterns found in linguistic features associated with a certain publisher, topic or period do not have to be universal (Saadany et al., 2020; Touahri & Mazroui, 2024).

2.1.3 Deep Learning and Hybrid Models

With deep learning, the need to design features manually was eliminated. CNN models detect patterns in local expressions and short tokens while LSTM and BiLSTM models capture sequential patterns in longer texts (Nasir et al., 2021; Merzah et al., 2026).

The hybrid CNN and recurrent networks integrate local and sequential information (Nasir et al., 2021; Merzah et al., 2026).

Past research demonstrated good results on certain benchmarking datasets; however, the performance was highly inconsistent across the sources and domains (Nasir et al., 2021; Merzah et al., 2026). Such inconsistency shows that complex architecture does not ensure good generalization.

2.1.4 Transformer-Based Language Models

Transformer models produce contextual representations based on self-attention where the token can be described with regard to the sequence it belongs to (Devlin et al., 2019). BERT and other transformer models have already become strong baselines for fake news detection, whereas Arabic pretrained models produce contextually-aware representations tailored for Arabic vocabulary and morphology (Raza et al., 2025; Antoun et al., 2020). Adding layers of CNN or BiLSTM after the encoder of the transformer model makes it possible to process data additionally using local and sequential information. In addition, those components increase the number of parameters, the duration of training, memory usage, and inference costs (Jadhav et al., 2025; Ganesh Karthik et al., 2025). It is necessary to prove the importance of those

components directly comparing the hybrid model with the transformer-only model.

2.1.5 Arabic Language Challenges

The classification of Arabic texts is influenced by the richness of morphology, prefixes, and suffixes, derived word forms, different spellings, and difference between Modern Standard Arabic and dialects. Thus, the preprocessing for traditional and transformer-based models should be different since in case of the former, normalizing and feature selection might be helpful, whereas aggressive stemming and stop-word removal might delete information which will be taken into account in case of the pretrained tokenizer (Antoun et al., 2020; Touahri & Mazroui, 2024).

The Arabic fake news datasets differ in size, composition of the sources, languages used, annotation approach, and definition of labels. Thus, their comparison becomes problematic (Touahri & Mazroui, 2024).

2.1.6 Dataset Bias and Generalization

The performance of a model is affected by the manner of construction and partitioning of the dataset used to evaluate its performance. For instance, in case the split of a random article-level is done, it is possible that articles from the same publisher, topic or original dataset appear in both train and test datasets. In such conditions, a classifier may learn patterns typical for the same source and not transferable patterns (Kaufman et al., 2012).

Another issue to be considered is exact and near duplicates of articles. This way, in case similar versions of an article exist in different partitions of the dataset, a reported test result will be an overestimate of the performance on unseen data (Lee et al., 2022).

Significant differences between datasets were noted by previous studies in ISOT, WELFake, LIAR, FA-KES and ANS datasets (Nasir et al., 2021; Merzah et al., 2026; Rout et al., 2025). As a consequence, in order to consider separately performance inside a dataset and its generalization, duplicate elimination and evaluation with disjoint publishers are included into the suggested design of the experiments.

2.2 Related Work

In accordance with their main technical contributions, the selected studies are divided into categories. Thus, at first, there are presented studies using traditional machine learning and linguistic features. After that, there are considered CNN and recurrent networks, transformer-based classification, Arabic pre-trained models and multimodal or multilingual systems. The last category deals with dataset dependence and generalization.

Such thematic approach allows comparing results obtained by different studies in spite of their use of various languages, datasets, labels and procedures of evaluation.

2.2.1 Review Scope and Study Selection

A narrative analysis is utilized to conduct an evaluation of current advances in classification of Arabic and English fake news. The analysis will take into consideration articles published from 2020 until 2026, which investigate approaches such as traditional machine learning, deep learning, transformers, hybrids, multilingual, or multimodal methods.

The first set of articles for analysis was collected via extensive searches on academic websites and publisher websites such as ACL Anthology, IEEE Xplore, ScienceDirect, SpringerLink, Frontiers, MDPI, and PLOS.

A literature search was carried out for papers published from 2020 to 2026 on issues including classification of fake news, Arabic natural language processing, transformer models, convolutional neural network (CNN), recurrent neural networks (RNNs), hybrid models, biases in datasets, and generalization. It involved an analysis of their titles, abstracts, complete articles, publication information, and DOIs.

The inclusion criteria for studies included the following:

1. They dealt with either fake news, manipulation news, satire identification, news verification, or other similar misinformation tasks.
2. They proposed a machine learning, deep learning, or transformer-based approach.
3. They had identifiable data sets, metrics, or methodological results.
4. They contributed directly to either Arabic classification, English benchmarking, architecture hybridization, generalization of data sets, or computational cost consideration.
5. The entire research could be accessed for examination.

Researches that lacked adequate methodological information, did not assess a classification approach, were duplicate of selected research, or concentrated on irrelevant task were excluded from the review.

A total of fifteen papers were selected for comparative analysis after the screening phase.

The selected studies were categorized based on their main technical contribution. Information was gathered about the purpose of the research, the language used, the data, preprocessing, architecture of the model, evaluation metrics, findings, limitations, and relevance to this particular research.

2.2.2 Traditional Machine Learning and Linguistic Features

Conventional machine learning approaches are still valuable because they offer computationally inexpensive baselines. Himdi et al. (2022) conducted a study on identifying Arabic fake news. The features used included emotion, polarity, part of speech, negation, certainty, intensification, and reporting style. The most effective classifier found in the experiment was Random Forest. The paper indicated that deceiving Arabic news contains valuable linguistic cues; however, the usage of handcrafted features limits its applicability to other dialects and topics.

Saadany et al. (2020) analyzed satirical news in Arabic. Their work was focused on lexico-grammatical features and showed high effectiveness in separating articles between selected satirical and non-satirical publishers. Thus, satire is identifiable through specific linguistic characteristics. However, it raised concerns about the methodology since the classifier might have learned the unique styles of some publishers instead of detecting factual and deceptive reporting.

Nagoudi et al. (2020) researched the creation and identification of fabricated Arabic news. Their research introduced an Arabic resource and tested both human readers and automatic classifiers. The most effective automatic classifier found in the experiments showed lower performance than some English benchmark data sets.

However, these research papers also demonstrated that fixed feature selection has its limitations. For instance, a particular characteristic of one source or misinformation category will lose its meaning when the same model will come into contact with the other topic, dialect, or publication.

2.2.3 CNN and Recurrent Neural Networks

Moreover, deep learning techniques significantly decreased the need for manual features selection. CNN models were used to capture local patterns and

short token sequences while LSTM and BiLSTM architectures were applied to capture long-range sequential relations (Nasir et al., 2021; Merzah et al., 2026).

Thus, Nasir et al. (2021) employed a combination of CNN and recurrent layers with pretrained GloVe embeddings. It reached almost perfect accuracy in ISOT while significantly worse accuracy was obtained for FA-KES. This contrast proves that only architecture does not affect performance. Dataset composition, writing style, number and variety of sources, and task complexity play the key role too.

Merzah et al. (2026) examined the parallel CNN-BiLSTM model with FastText embeddings trained on Arabic and English datasets. 98.85% of accuracy was reached for WELFake, 94.43% – for AFND, and 71.63% – for ANS. Ablation studies showed that CNN and BiLSTM layers contributed complementary features. At the same time, a difference between AFND and ANS proved that outstanding accuracy in one Arabic dataset does not guarantee the same result in another.

Therefore, the research papers prove the effectiveness of a combination of local and sequential features. However, they do not prove generalizability of CNN-BiLSTM architectures over the sources and languages. Static embeddings were also used in both studies.

2.2.4 Transformer-Based Classification

Transformers introduced context embeddings where a representation of a token was a function of the entire sequence. BERT and other transformer-based architectures became standard baselines in the literature of detecting fake news (Devlin et al., 2019; Raza et al., 2025).

Raza et al. (2025) examined the performance of progressively trained BERT on WELFake. The model showed an accuracy of 95.3%, with equal balance in precision, recall, and F1-score. The research indicates how a well-thought-out approach can enhance the performance of an algorithm without making any

changes to the architecture. However, as the experiment was conducted on only one English dataset, the results cannot be generalized across the languages.

Roumeliotis et al. (2025) compared the performance of CNN, BERT, GPT-4o, and GPT-4o-mini on a balanced subsample of WELFake. The fine-tuned language models proved to be the most effective, while the basic CNN lagged behind them. The BERT performed comparably to the large models but with the easier setup. The use of a limited English sample limits the possibilities for transferability of the result to other datasets.

Essa et al. (2023) used LightGBM with BERT features and obtained good performance on ISOT, TI-CNN, and the Fake News Corpus datasets. The research demonstrated how contextualized transformer features can serve as a foundation for the efficient external classifier. The presented architecture was not trained as one end-to-end neural model, and it should be noted how close to perfection its ISOT result is related to the characteristics of the dataset.

Rout et al. (2025) proposed a compact transformer architecture and analyzed its performance on ISOT, FakeNewsNet, and LIAR. The model achieved 100% accuracy on ISOT, 80% on FakeNewsNet, and 59.56% on LIAR. The discrepancy between the metrics provides additional information on the problem: a model that performed perfectly on one benchmark failed when it comes to another.

Collectively, these papers prove transformers as efficient baselines for contextual modeling. Moreover, they demonstrate the fact that reported performance is determined by both the abilities of models and the qualities of used datasets. Thus, a good result obtained with the help of transformer should be proved first in order to consider CNN and BiLSTM addition as valuable.

2.2.5 Arabic Pretrained Language Models and Ensembles

Pretrained Arabic models have been proposed to produce contextual embeddings adapted to Arabic language (Antoun et al., 2020).

Al-Zahrani and Al-Yahya (2024) analyzed an ensemble of AraBERT, MARBERT, AraELECTRA, AraGPT2, and ARBERT. The best ensemble gave an F1-score of 94% using weighted averaging. The result indicates that different decision boundaries were learned by the participating models. Besides, the paper revealed that larger ensembles do not necessarily lead to better results. The use of ensembles increases storage, computational time, and difficulty of deployment.

Touahri and Mazroui (2024) analyzed Arabic fake news classification methods, datasets, preprocessing techniques, linguistic features, deep learning architectures, and Arabic transformers. As a result of their analysis, they revealed that no architecture was the best in all experiments conducted. The authors highlighted the lack of available datasets, annotation technique inconsistency, dialect variability, and domain dependence as current problems.

Thus, Arabic transformers provide the evidence of the necessity of language-specific pretraining. This approach does not eliminate the effect of data quality, annotation method, source diversity, or computational complexity.

2.2.6 Multimodal, Multilingual, and Explainable Approaches

In recent research, the scope of the task is expanded beyond simple article text. Images, text in images, propagation networks, social interactions, external knowledge, and explanation methods have been incorporated into more complex systems (Jadhav et al., 2025; Lv et al., 2025).

Al-alshaqi et al. (2025) used Optical Character Recognition to extract written information from images and incorporate it with the article text. The best performance was shown for TRUTHSEEKER, whereas performance

deteriorated for the merged dataset. The discrepancy shows that merging different sources increased difficulty of classification. The model worked with the information extracted from images in text format but did not consider any visual manipulation or relationship misrepresentation.

Ganesh Karthik et al. (2025) used hybrid optimization of the modified transformer on Fakeddit and DGM4. The model achieved better performance compared to some baselines but had a multistage optimization and fine-tuning procedure during training. The results were different for the two datasets and again prove that optimization does not eliminate dependency on the dataset.

Jadhav et al. (2025) developed a multilingual, multimodal, and explainable framework, using such models as XLM-RoBERTa, CNN, BiLSTM, ResNet50, GraphSAGE, and attention-based fusion. The model showed better performance than a transformer and produced explanations. It has an architecture with about 420 million parameters and requires significant amounts of Graphics Processing Unit (GPU) capacity. Arabic is not among the evaluated languages.

Lv et al. (2025) conducted a survey on multimodal fake news detection on text, images, video, audio, social context, propagation structures, graphs, external knowledge, and large language models. The authors noted such issues as annotation quality, heterogeneous fusion, interpretability, and computational cost.

These studies illustrate the advantages of going beyond the article text and provide datasets and required infrastructure beyond the current project. However, their results can be useful in order to understand the future direction of the area but cannot be taken as baselines for the proposed study with a text-only input.

2.2.7 Dataset Dependence and Generalization

The common pattern of these studies is the distinction between the within-dataset performance and generalization. In many cases, high performance was achieved on ISOT and WELFake and the same architecture showed lower performance on LIAR, FA-KES, ANS, and merged datasets, as well as in cross-domain settings (Nasir et al., 2021; Merzah et al., 2026; Rout et al., 2025; Alalshaqi et al., 2025).

It happens because of the nature of the task. Some datasets include whole articles, others – short claims; some use binary labels, others – multilevel labels; labels of some datasets are based on the fact verification of the content, labels of others – on the credibility of the publishing source (Touahri & Mazroui, 2024; Khalil et al., 2022).

In AFND, article labels are assigned based on the credibility category of their publisher (Khalil et al., 2022). Thus, a great Arabic dataset is possible but at the same time, there is a possibility that the machine learning algorithm will get used to the publishing language and style. WELFake is the merging of articles from various previous datasets and it poses another challenge of source-dependent patterns and duplicate content (Verma et al., 2021).

Random article-level split of the dataset could mean that there would be content from one source in both train and test splits. In that case, the achieved score would indicate the performance on the unseen articles but not on the unseen publishers. Thus, the publisher-disjoint evaluation and duplication analysis are crucial for interpretation of the obtained results (Kaufman et al., 2012; Lee et al., 2022).

2.2.8 Comparative Synthesis of the Reviewed Studies

These studies varied in language, dataset construction, problem definition, model architecture, and evaluation procedure. The results presented in them should thus be considered in the context of every individual experiment rather

than compared as if they all were achieved using one common benchmark. Table 2-1: Comparative Summary of the Reviewed Studies provides an overview of the datasets, methods used, results obtained, methodological limitations, and relevance of the studied papers to the proposed research.

2.2.9 Research Trends Within the Reviewed Set

The reviewed studies were classified according to the prevailing technical trend demonstrated in them. Although some of the papers featured methods of more than one type, each study was categorized by the trend which seemed to define its primary focus. Table 2-2: Dominant Research Trends Within the Reviewed Studies describes the trends within the reviewed papers, their advantages, and recurring limitations.

Table 2-1: Comparative Summary of the Reviewed Studies

Study	Language Dataset	and	Main Approach	Main Result	Reported	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 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 Dataset	and	Main Approach	Main Result	Reported	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 BiLSTM FastText	CNN- with 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 generalization.	Supports the transformer-only baseline

Study	Language Dataset	and	Main Approach	Main Result	Reported	Main Methodological Limitation	Relevance to the Current Study
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
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

Study	Language Dataset	and	Main Approach	Main Result	Reported	Main Methodological Limitation	Relevance to the Current Study
Rout et al. (2025)	English, FakeNewsNet, and LIAR	ISOT,	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.	Convincing 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
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

Study	Language Dataset	and	Main Approach	Main Result	Reported	Main Methodological Limitation	Relevance to the Current Study
Al-alshaqi et al. (2025)	English, WELFake, TRUTHSEEKER, and merged data	ISOT,	OCR text fusion with BERT	The strongest performance was reported on TRUTHSEEKER, while performance decreased 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 transformers with hybrid optimization	The model achieved 90.42% accuracy on Fakeddit and 86.74% on DGM4.		The method requires a complex optimization and multi-stage fine-tuning pipeline.	Supports computational-cost evaluation
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

Study	Language Dataset	and	Main Approach	Main Result	Reported	Main Methodological Limitation	Relevance to the Current Study
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 2-2: Dominant Research Trends Within the Reviewed Studies

Dominant Research Direction	Number of Studies	Share of Reviewed Set	Main Focus	Technical	Repeated Strength	Repeated Limitation
Traditional machine learning and linguistic features	3	20.0%	Manual grammatical, and stylistic features	lexical, and	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 datasets 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 optimization-based systems	4	26.7%	Text-image fusion, graph learning, explainability, and optimization		Richer evidence and broader task coverage	High computational demand and complex data requirements

2.2.10 Interpretation of Research Trends

It is obvious from the reviewed literature that there is a tendency of moving from hand-designed linguistic features to transformer-based representations and even more sophisticated hybrid models. When the standard transformer classification approach, Arabic pretrained transformers, and transformer-based multimodal systems are taken into account, this trend turns out to be the biggest one in the set of reviewed papers.

Even with the increase in architectural complexity, the dataset dependency was not eliminated. Some studies yielded extremely good results on ISOT and WELFake, but much worse scores on LIAR, FA-KES, ANS, combined datasets, and cross-domain experiments. This suggests that the dataset structure, source variability, the process of labeling, and task complexity affect the performance of the models (Nasir et al., 2021; Merzah et al., 2026; Rout et al., 2025).

The trend in the Arab world went from linguistic features and classical classifiers to Arabic pretrained transformers and ensembles. It still suffers from a limited number of public datasets, unstable labeling, dialectal difference, and source bias (Touahri & Mazroui, 2024). The reviewed literature also does not present enough information on the performance of the same hybrid components when they are used independently for Arabic and English languages.

Accordingly, the main research gap is not in combining BERT, CNN, and BiLSTM in the mentioned manner but in conducting the controlled experiments for evaluating each of the component within the same experimental conditions, including language-appropriate transformers, runs, computational measurements, and analysis without overlapping publications.

2.2.11 Research Gap and Basis for the Proposed Study

The examined literature indicates that traditional machine learning, CNN, recurrent networks, transformers, ensembles, and multimodal architectures have all been applied to fake news classification. Additionally, it indicates that

the performance results depend significantly on the choice of the language, datasets, annotation approach, preprocessing, and experiment design (Touahri & Mazroui, 2024; Lv et al., 2025).

Several research gaps remain.

Firstly, many works conduct comparisons between models developed in different experimental conditions, which makes it hard to understand whether the reported difference is caused by the model's architecture itself (Touahri & Mazroui, 2024).

Secondly, the majority of works with hybrid architectures combine multiple components without comparing the contribution of each of the component separately. The high result of the transformer-CNN-BiLSTM model cannot be used to identify what component (transformer encoder, CNN, BiLSTM, or all together) brings an improvement. The controlled ablation study is needed in such cases.

Finally, there are very few works that apply the similar comparative structure for Arabic and English languages with appropriate language transformers. Most bilingual studies use static word embedding, whereas Arabic transformers are mostly based on a single dataset or ensembles (Merzah et al., 2026; Al-Zahrani & Al-Yahya, 2024).

Finally, model choice is primarily guided by accuracy or F1-score. Other criteria such as training time, inference time, memory usage, number of parameters, and model size are relatively disregarded, whereas complex architectures can require considerable computation resources (Jadhav et al., 2025; Lv et al., 2025).

Additionally, data set construction can impact the interpretation of reported results. Specifically, AFND assigns article labels based on credibility of the publishers, and WELFake is built upon several earlier data sets (Khalil et al., 2022; Verma et al., 2021).

The suggested study will fill the identified gaps by means of controlled experiment with four types of models: transformer-only, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM architecture. AFND and

WELFake will be considered separate Arabic and English cases. All models will follow the same data preprocessing, division of data into parts, training rules, performance metrics, and repeated experiments within each language.

The novelty of the work will consist in the evidence-based evaluation of whether the CNN and BiLSTM branches offer stable and significant advantage compared to transformer-only models. A conclusion about advantage of a simpler transformer model is no less important than a conclusion about the entire hybrid model.

2.3 Chapter Summary

From the literature review, one can conclude that classification of fake news evolved from manually designed linguistic features and machine learning algorithms to deep learning, pretrained transformers, hybrid models, ensembles, and multimodal systems. Despite all the progress, however, reported results are strongly affected by data set construction, data source, language, labeling approach, and evaluation process.

There is no proof from the literature that adding CNN and BiLSTM components to transformer representations results in a meaningful improvement. These models become more complex and computationally expensive while their advantages tend to be inconsistent for different data sets.

This research suggests carrying out the comparison of the traditional machine learning approach, CNN-BiLSTM architecture, the transformer representation, the transformer representation enhanced with CNN, the transformer representation enhanced with BiLSTM, and the transformer representation enhanced with both CNN and BiLSTM. AFND and WELFake will be considered as distinct cases for Arabic and English languages respectively using appropriate transformer encoders.

The aspects of evaluation will include the performance, consistency of runs, statistical significance, computational cost, duplicated content, and generalization based on the source. The macro F1-score will be used as the

primary metric of evaluation supplemented by per-class statistics, confidence intervals, pairwise comparisons, and publisher-independent evaluation for AFND.

The value of this research study will not be determined by the highest performance of the complete hybrid architecture. It can be equally interesting to obtain the evidence that a more simple transformer model offers similar performance with lower cost. The final conclusion will reveal the most defensible trade-off between the performance, cost, and generalization under the specified conditions of experiments.

Chapter (3): Proposed Approach (Methodology)

3.1 Research Design

A controlled comparative experimental design will be employed in this research. Arabic and English will be considered as two separate cases of experiments rather than one multilingual classification task.

In both of these experiments, however, the model comparison approach will be the same. However, language-dependent preprocessing, tokenizers, transformers, and word embeddings will be utilized. Scores on the two data sets will not be compared directly because they differ in labels, sources, gathering processes, and tasks.

It is an ablation-based comparison that will test Transformer only, Transformer + CNN, Transformer + BiLSTM, and Transformer + CNN + BiLSTM. Traditional machine learning and CNN + BiLSTM models will act as additional baseline models for reference.

The experiment is not about proving superiority of the most complicated model. The goal is measuring contributions of each component and evaluating whether any gain obtained is robust, statistically significant, and proportional to the increased computational load.

3.2 Datasets and Task Definition

AFND will be the dataset used for the Arabic experiment and WELFake for the English experiment. Each of the data sets will be examined before the development of the models. They should be analyzed for textual fields, label descriptions, class distribution, missing instances, article length, duplicated texts, and available sources.

A single news article is the primary unit of analysis. If there are title and body fields in the record, they will be concatenated with a certain separator. Instances with inadequate text or invalid labels will be filtered out.

The main experiment will use binary text classification. WELFake provides binary labels – real and fake (Verma et al., 2021). For AFND, there are three initial labels: credible, not credible, and undecided credibility (Khalil et al., 2022); credible and not credible articles will be used for the binary task, whereas undecided articles will be eliminated during splitting.

This task will assess agreement with the provided by the dataset labels. The task will not prove or disprove the factual truth of each claim using external sources. The results of experiments on AFND will be interpreted as classification according to publisher credibility labels (Khalil et al., 2022), while the results of WELFake will reflect binary labels from the original source datasets (Verma et al., 2021).

All cleaning, label assignments, and duplicates elimination will be done prior to the splitting of datasets into the train, validate, and test sets.

3.3 Data Preparation

Data preparation will be done independently for AFND and WELFake. The original datasets will be kept intact, and cleaning and transforming operations will be done on documented working copies of the datasets.

Articles without any label, without textual information, having corrupted data, or containing an insufficient number of words in the body will be discarded according to certain criteria. When the title and body columns are present, these columns will be joined by the same separator.

The exact copies will be detected by hashing of normalized text. The near duplicates will be detected by means of a documented similarity algorithm and threshold. The confirmed duplicate sets will be reconciled prior to data

partitioning so that no related duplicates would be included into different partitions.

Preprocessing of Arabic texts will be performed through unicode normalization, de-elaboration of characters, fixing inconsistent whitespace, and normalization of a limited number of letter forms. No aggressive stemming, root extraction and stopword removal will be applied to transformer inputs.

Preprocessing of English texts will consist of removal of invalid markup, duplicate whitespaces, and irrelevant markup. Lowercasing and tokenization for transformer models will be done following the chosen BERT tokenizer.

Traditional machine learning models will employ separately trained TF-IDF vectors. Tokenization of transformers, TF-IDF vocabulary, class weights and any feature selection will be based on the training subset only.

3.4 Experimental Splits

After data cleaning and duplicate removal, every dataset will be split into training, validation and test subsets in the 80:10:10 proportion. The main evaluation procedure will rely on stratified random splitting.

The splits will be generated once with fixed random seeds and saved for all comparable models to use. The test split will remain sealed throughout model training. Hyperparameters will be tuned, checkpoints chosen, and preprocessing and model comparison done using only the training and validation datasets.

Near-duplicate samples from the same identified family will stay in the same partition. Thus, articles which are similar to each other will not get separated into the training and test datasets.

The second track will consist of splitting for AFND in a publisher-disjoint manner if metadata allows for a reliable separation. Test publishers will not show up in training and validation splits.

Publisher-disjoint results will be reported independently from stratified random results because the two types of splitting represent generalization under different conditions. In cases where metadata in WELFake does not allow for a reliable separation of sources, there will be no publisher-disjoint results reported.

3.5 Models

Seven different model configurations will be considered in the study:

1. Logistic Regression with TF-IDF.
2. Linear SVM with TF-IDF.
3. CNN-BiLSTM with pre-trained static embeddings specific for each language.
4. Transformer-only classification.
5. Transformer-CNN.
6. Transformer-BiLSTM.
7. Parallel transformer-CNN-BiLSTM.

Traditional machine learning algorithms will serve as efficient baselines. CNN-BiLSTM model will serve as a deep learning hybrid approach without the use of contextual transformers.

Arabic transformer settings will be based on AraBERT checkpoint selection (Antoun et al., 2020), whereas English transformer settings will be based on BERT-base-uncased (Devlin et al., 2019). Exact version of AraBERT will be provided together with experimental settings.

All configurations in the same language based on transformers will share the same pretrained encoder checkpoint, tokenizer, data splits, and max sequence length. In such way, differences between transformer-only and hybrid architectures will be mainly due to the CNN and BiLSTM branches added.

In the full hybrid network, token embeddings obtained from the transformer encoder will be fed to CNN and BiLSTM branches alongside. Their outputs will

be pooled and concatenated together with dropout regularization followed by passing them to the classification layer.

The transformer-CNN and transformer-BiLSTM will allow for direct ablation of the complete architecture.

3.6 Training Procedure

Development of models will be done following the same training and validation approach in each experiment setup. Comparable models will use the same data splits and fixed random seed.

Hyperparameters will be chosen based on the validation set alone. The search space will consist of learning rate, batch size, dropout rate, max sequence length, number of epochs, CNN kernels, and BiLSTM hidden size where applicable. Comparing transformer models will have the same budget allocated to them.

Early stopping will rely on macro F1-score on validation. The best performing checkpoint on validation will be kept for the final evaluation step. Test set will not be used for hyperparameter selection, for setting the number of training epochs, for preprocessing steps, and for model checkpoint choice.

Dropout and weight decay techniques will be used for overfitting reduction. Gradient clipping will be utilized for recurrent layers. Training/validation losses, macro F1-scores, and learning rate evolution will be monitored per epoch.

Each neural network architecture will be trained with at least three fixed random seeds. Mean and standard deviation of results from repeated runs will be provided, not the best run result.

All directly compared models will be trained and evaluated on the same hardware setup. In case of any hardware issues during the experiment run, it will be repeated with the initial parameters.

3.7 Evaluation

Macro F1-score will be used as the primary evaluation measure because macro averages treat classes equally, which makes such measures useful when class level performance should be evaluated regardless of class frequency (Sokolova & Lapalme, 2009).

Some of the methods to support classification measures would be:

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 research project will measure the computational needs of the neural architectures:

1. Training time.
2. Time required for inference per document and per batch.
3. Number of trainable parameters.
4. Peak GPU memory consumption.
5. Saved model size.

The timing and memory measures will be measured with the same machine and environment. The initialization time and loading of the data once will be separate from the inference times.

The primary measure will be through averaging out of repeat measurements. Just numerical improvement will not confirm the superiority of one architecture. Stability, class-level results, confidence intervals, statistics tests, robustness to the publisher-disjoint test, and computation will also factor into the assessment.

3.8 Robustness and Error Analysis

Robustness study will look into the dependence of model performance on duplicates, source diversity, class imbalance, and document length.

Publisher-disjoint results will be compared to the stratified ones on AFND. A noticeable drop in performance under publisher-disjoint evaluation will suggest that some part of performance achieved via random split depended on certain source-related aspects.

The effect of eliminating the duplicates will be analyzed through counting the amount of exact and near-duplicates found before splitting. Model evaluation will be done using clean data.

False positives and false negatives will be looked into in error analysis of the transformer-only and hybrid models. Error analysis will include:

1. Article length and truncation.
2. Publisher or source.
3. Topic and frequent named entities.
4. Repetitive phrases and stylistic elements.
5. Arabic orthography and morphological differences.
6. Prediction confidence.
7. Differences in the evaluated architectures.

Instead of making factual statements on the basis of particular publishers or articles, examples will be provided to illustrate error types.

3.9 Dataset Profile and Labelling Structure

The architectures will be demonstrated in two cases: AFND for Arabic, and WELFake for English. These have differences in language, fields of text, label generation, source metadata, and collection process.

AFND has a set of credible, not credible, and undecided labels according to the credibility rating of the source of publication (Khalil et al., 2022). The main AFND experiment will be conducted using credible and not credible records as a binary classification problem. Undecided records will be discarded before the division into classes, and the number of discarded records will be provided.

WELFake has a binary set of real and fake labels and was created on the basis of records taken from different sources (Verma et al., 2021). Thus, its collecting process implies that the dataset should be checked for possible duplicates and source-specific patterns before starting training the models.

Table 3-1: Profile of the Selected Datasets shows the main features, methodological problems, and control measures of the two datasets under consideration. The study will report the initial number of records, number of records excluded, number of removed duplicates, and the final number of classes in the dataset after cleaning the data.

Table 3-1: 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 the experiment will be determined after removing empty records, irrelevant labels, exact duplicates, and confirmed near-duplicates. The study will provide information about the record count in each cleaning step instead of reporting only the final size of the dataset.

Moreover, the class distribution in the original dataset will be analyzed before any kind of balancing procedure. Weighting the classes will be preferable to removing many valid records. Oversampling and undersampling procedures will be implemented only if validation suggests it clearly.

Scores obtained from AFND and WELFake will not be compared directly in order to find out which of the languages is easier. Distinguishing factors for the two data sets include label construction, publisher composition, article length, topic distribution, and history of data collection. Instead, each of the data sets will be used to test if the same architectural modifications lead to the same patterns of performance increase within each individual experiment separately.

3.10 Preprocessing Pipelines

splitting while vocabulary fitting, tokenization settings, class weight, and feature selection will be done based on the training set only.

The pre-processing of Arabic articles will involve controlled Unicode and letter normalization to preserve input specifications of the selected Arabic transformer (Antoun et al., 2020; Touahri & Mazroui, 2024), while inputs for English transformers will follow the tokenizing and input specifications of BERT-base-uncased (Devlin et al., 2019).

Table 3-2: Preprocessing Procedures for the Experimental Pipelines lists the processes used in each of the experimental pipelines.

Table 3-2: 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
Normalize 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
Normalize 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 lemmatization	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

3.11 Proposed Framework and Model Architecture

The proposed framework tests if local and sequential processing branches provide any value to contextual representation of the data produced by transformers. Arabic and English articles will be processed by separate language-specific pipelines. In case of AFND, it will be an Arabic transformer while WELFake will be processed using BERT-base-uncased.

The architecture for parallel runs includes the following steps:

3.11.1 Text Input and Tokenization

The title and content of the article will be concatenated together in case both fields are present, using an appropriate delimiter. The article will then undergo tokenization using the appropriate tokenizer corresponding to the language-specific transformer.

3.11.2 Transformer Encoder

The tokenized input will be passed through the chosen pretrained transformer. The output from the encoder will be the contextualized representation of every token.

For the transformer-only baseline model, the pooled representation corresponding to the first classification token will be used along with dropout and a classification layer. For the hybrid models, the contextualized representation of every token will be preserved to be processed in CNN and BiLSTM branches.

3.11.3 CNN Branch

The contextualized token sequence will be passed through the one-dimensional convolutional filters having particular kernel sizes.

Masked global max pooling will help create a fixed-sized CNN representation, without the padded positions having any effect on the pooling operation.

3.11.4 BiLSTM Branch

The exact same contextualized tokens will go through a bidirectional LSTM, where the valid lengths based on attention masks will ensure that the padded positions do not affect the recurrent representation.

The forward and backward hidden states will be concatenated together to form the representation with information in both directions.

3.11.5 Feature Fusion

The pooled CNN representation and the BiLSTM representation will be concatenated together to form one single representation, which will then be passed through a fully connected layer after applying dropout.

3.11.6 Classification Layer

The main task is binary classification. There will be one output logit for each article, and the classification loss function will be binary cross entropy with logits.

Class balancing will only be done where it is needed based on the training set distribution. The same class balancing criteria will be used in comparable models for each dataset.

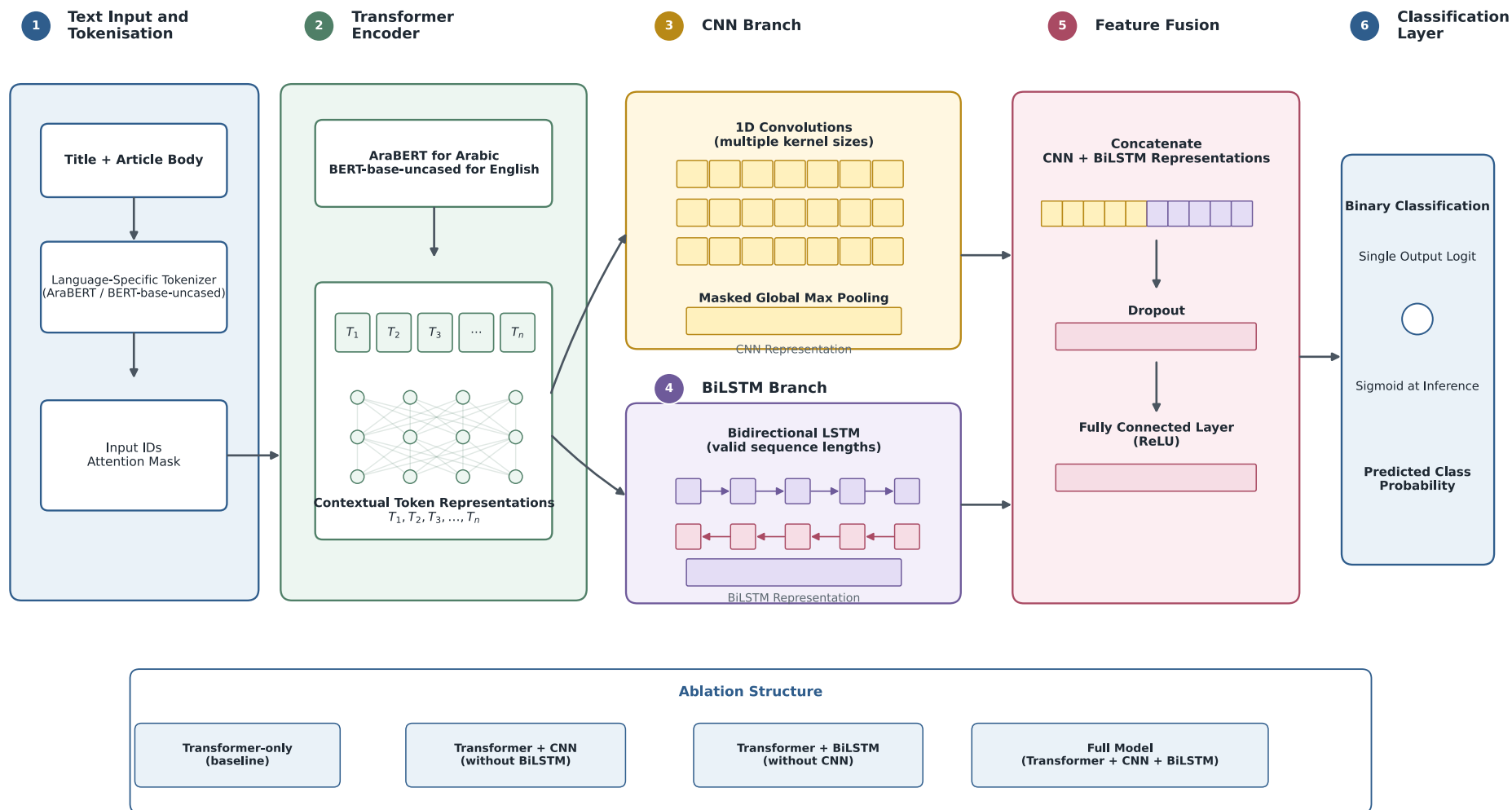


Figure 3-1: Proposed Parallel Transformer-CNN-BiLSTM Architecture

3.11.7 Ablation Structure

The following four transformer configurations will be considered based on the same encoder checkpoint, tokenizer, data split, maximum sequence length, and training schedule for each language:

1. **Transformer-only:** Uses the transformer representation only followed by the classification head.
2. **Transformer-CNN:** The contextualized token sequence is passed to the CNN head before classification.
3. **Transformer-BiLSTM:** The contextualized token sequence is passed to the BiLSTM head before classification.
4. **Transformer-CNN-BiLSTM:** Token sequences are processed using the context through parallel CNN and BiLSTM branches and then the two representations are fused before classification.

This model helps identify the individual effect of the added branch. The entire hybrid model cannot be regarded as better until its increase is consistent with the added complexity of computation.

3.11.8 Sequence Length and Long Articles

Distributions of the token lengths will be analyzed individually for AFND and WELFake. A single maximum length will be chosen for all transformer models for a particular language.

The paper will present the choice length, the percentage of articles surpassing it, and the average number of discarded tokens. Head and head-tail truncations will be analyzed with validation sets if resources allow. The test set will have no impact on choosing the truncation strategy.

3.12 Experimental Design Matrix

Matrix of the experimental design represents the models tested in Arabic and English cases. All models, which are comparable within one language, will utilize the same preprocessed data and partitioning.

Deterministic baselines from traditional machine learning methods will be used. At least three different random seeds will be used to measure the performance of neural architectures.

Table 3-3: 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 three seeds	Evaluate hybrid deep learning without contextual transformers
Transformer-only		AraBERT	BERT-base-uncased	At least three seeds	Establish the main contextual baseline
Transformer-CNN		AraBERT-CNN	BERT-CNN	At least three seeds	Measure the contribution of local convolutional features
Transformer-BiLSTM		AraBERT-BiLSTM	BERT-BiLSTM	At least three seeds	Measure the contribution of sequential recurrent features
Transformer-CNN-BiLSTM		AraBERT-CNN-BiLSTM	BERT-CNN-BiLSTM	At least three seeds	Evaluate the complete parallel hybrid architecture

3.12.1 Evaluation Tracks

The first track of evaluations will use cleaned and deduplicated data with stratified train, validation, and test splits. All models for the same language case will use the same splits.

The second track of evaluations will use publisher-disjoint splits for AFND, assuming that metadata allows for source separation. Test sets will not include any publishers present in the training or validation splits.

Publisher-disjoint track will focus on Logistic Regression, transformer-only, transformer-CNN, transformer-BiLSTM, and transformer-CNN-BiLSTM models. The evaluation will check if models maintain their performance when seeing the publishers which were not presented during training.

Two different evaluation tracks will be used independently. Stratified evaluation tests the models on unseen articles from the same structured dataset, whereas publisher-disjoint gives more strict estimates of generalization.

3.13 Experimental Environment

The experiments will be implemented using Python with PyTorch and Hugging Face Transformers for training neural networks. Scikit-learn will be used for baselines such as TF-IDF, classification metrics and auxiliary statistics. For dataset preprocessing, Pandas and NumPy libraries will be used.

Training will take place in the GPU-enabled Compute Unified Device Architecture (CUDA) environment. Compared directly neural models will be trained and timed in the same environment to ensure that any discrepancy in training time, inference time and memory usage is due to differences between the architecture itself and not the execution environment.

In the final report, information about the GPU, Central Processing Unit (CPU), system memory, operating system, Python version, PyTorch version, Transformers version and CUDA version will be reported. If applicable, automatic mixed-precision training will be utilized but only consistent among comparable transformer setups.

Sequence length, batch size and number of runs will be chosen based on available memory while keeping the comparison as fair as possible between evaluated architectures.

3.14 Statistical Analysis

The main analysis will consist of comparing macro F1-score across repetitions. Accuracy, weighted F1-score, per-class precision, recall and F1-score will complement the main comparison.

For every neural architecture, the experiment will present the mean and the standard deviation with respect to pre-defined random seeds. The best individual run will not be chosen as the primary measure for the comparative analysis of the models.

Confidence intervals for macro F1-score will be estimated via bootstrapping the locked test set predictions. The confidence intervals for the differences between the corresponding hybrid architecture and the transformer model (without hybrids) will be estimated via paired bootstrap comparisons (Efron & Tibshirani, 1993).

McNemar's test will be used to compare the results of paired classifiers on the same test data (McNemar, 1947). Given that multiple hybrid models will be compared against the same baseline transformer model, Holm correction will be used (Holm, 1979).

A level of statistical significance equal to 0.05 will be chosen. Statistical significance will be evaluated taking into account the magnitude of the performance increase, the confidence interval, the behavior per class, stability of the performance gain across runs, results on publisher-disjoint dataset, and the computational cost.

The hybrid architecture will be regarded as preferable only if the improvement is consistent across runs and commensurate with the increased computational costs.

Negative or inconclusive results will still be part of the analysis. Inability to detect any significant improvements will show that CNNs and/or BiLSTMs were not valuable enough for the given conditions.

3.15 Ethical Considerations

The study will be using previously existing research data sets and will not have any interaction with human participants or acquisition of personal information. AFND and WELFake will be used in accordance with their access and redistribution rules.

The labeling of articles will not be considered as independent verification of their truthfulness. AFND uses article labeling as part of publisher credibility ratings (Khalil et al., 2022), whereas WELFake uses the labels inherited from the data sets used in its creation (Verma et al., 2021). Therefore, all results will be discussed as classifications by the given data set, but not final assessments of the publishers or their journalists.

Classification mistakes may cause reputational or informational damage. Thus, the proposed models will be considered as research prototypes and not autonomous fact-checkers or moderators. Their predictions will need additional review by humans before any practical application.

Metadata of sources will be used solely for the purpose of analyzing the bias and generalizability of the data set. Error analysis will be used just to show patterns of methodology and not to accuse individual publishers.

Details on cleaning, discarded entries, data splits, models used, and failed outcomes will be kept. The raw content from articles will not be shared unless allowed by the data set licensing agreement.

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/>
- <https://doi.org/10.18653/v1/N19-1423> Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*
- Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall
- <https://doi.org/10.1007/s40747-023-01098-0> Essa, E., Omar, K., & Alqahtani, A. (2023). Fake News Detection Using a Hybrid Approach of BERT and LightGBM Models. *Complex & Intelligent Systems*, 9(1), 65
- <https://doi.org/10.1038/s41598-025-99936-3> Karthik, Ganesh, M., Ahmad, Faizz, K. S., Pamidimukkala, S. G., Sathe, A. P., Sirisha, G. N. V. G., Ram, Sitha, M., & Koteswararao, C. (2025). Fake news detection using hybrid reinforcement-based transformer models. *Scientific*
- <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 via textual analysis. *Arabian Journal for Science and Engineering*, 47, 1
- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2),
- <https://doi.org/10.3389/frai.2025.1690616> Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., & Jadhav, S. (2025). Explainable multilingual and multimodal fake news detection: Towards robust and

- trustworthy AI to fight against disinformation. *Frontiers in Artificial Intelligence*, 8,
<https://doi.org/10.1145/2382577.2382579>Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and prevention. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 1
<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 credibility of articles. *Data in Brief*, 42, 10
<https://doi.org/10.18653/v1/2022.acl-long.577>Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Removing duplicates from training data improves language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8424–8445). Association
<https://doi.org/10.1007/s44443-025-00317-7>Lv, J., Gao, Y., Li, L., Shi, L., & Li, S. (2025). Fake news detection using multiple modes: An extensive survey of deep learning techniques. *Journal of King Saud University Computer and Information Sciences*, 37(3), 306. <https://doi.org/10.1007>
<https://doi.org/10.1007/BF02295996>McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(3), 15
<https://doi.org/10.3389/fdata.2025.1683786>Merzah, B. M., Razmara, J., & Salmanian, Z. (2026). Deep hybrid learning models for detection of fake news: Study case on 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). Automatic generation and detection of manipulated and fake news in Arabic. *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (pp. 69-84). Association for Computational Linguistics. <https://doi.org/10.1016/j.jjimei.2020.100007>
<https://doi.org/10.1016/j.jjimei.2020.100007>Nasir, J. A., Khan, O. S., and Varlamis, I. (2021). Detection of fake news: A hybrid deep learning approach using CNN and RNN. *International Journal of Information Management Data Insights*, 1(1), 100
<https://doi.org/10.1371/journal.pone.0330954>Raza, N., Abdulkadir, S. J., Abid, Y. A., Albouq, S. S., Alwadain, A., Rehman, A. U., Sumiea, E. H., &

- Farhan, M. (2025). Improving detection of fake news with deep learning using transformers: A multidisc
<https://doi.org/10.3390/fi17010028>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Detecting and classifying fake news: A comparative study on convolutional neural network, large language model, and natural language processing models. *Future Internet*, 17(1), 2
<https://doi.org/10.3390/jcp5030043>
- Rout, J., Mishra, M., & Saikia, M. J. (2025). Towards a reliable detection of fake news: Attention based transformer model. *Journal of Cybersecurity and Privacy*, 5(3), 4
<https://aclanthology.org/2020.rdsm-1.7/>
- Saadany, H., Orasan, C., & Mohamed, E. (2020). Fake or real? An investigation into 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/20>
- Sokolova, M., & Lapalme, G. (2009). Systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 42
<https://doi.org/10.1007/s10462-024-10778-3>
- Touahri, I., & Mazroui, A. (2024). Machine learning-based approaches to detecting fake news in Arabic language. *Artificial Intelligence Review*, 57, 157
<https://doi.org/10.1109/TCSS.2021.3068519>
- 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), 88