*Journal of Computer Science*

Research Article

# Arabic Fake News Detection Across Generational Text Representations: From Traditional Models to Transformer-Based Methodologies

**Noor M. Alkudah**

*Department of Computer Science, Faculty of Information Technology, The World Islamic Sciences and Education University, Amman, Jordan*

*Article history*
Received: 21-08-2025
Revised: 27-12-2025
Accepted: 29-01-2026

Email: noor.qudah@wise.edu.jo

**Abstract:** The rapid proliferation of fake news on Arabic social media has amplified societal and political risks, yet research on automatic detection in Arabic remains limited due to scarce datasets, morphological complexity, and underexplored preprocessing strategies. This study presents a comprehensive benchmark for Arabic fake news detection, unifying seven Machine Learning (ML) algorithms, three Deep Learning (DL) models, and a transformer-based approach (AraBERT) under consistent experimental conditions. A hybrid balanced dataset of 4,838 tweets was constructed from ArCOV19-Rumors, AraCOVID19-MFH, and NLP4IF-2021. Three levels of preprocessing were systematically evaluated: Primitive cleaning and tokenization, named entity recognition (NER), and NER with stemming. The results show a clear change in representation: TF-IDF gives strong lexical baselines, AraVec gives moderate gains through static embeddings, AraBERT embeddings give big improvements through contextualization, and fine-tuned AraBERT gets the best results (Accuracy/F1 ≈ 0.95). A comparative analysis shows that SVM is the best ML algorithm, Bi-LSTM is the best DL model, and contextual embeddings have a huge effect on all families. Preprocessing strategies have different effects on different types of models. For example, stemming helps ML but hurts DL, while NER always helps both. This study provides solid baselines, methodological insights, and a generational perspective on Arabic text representations, establishing a foundation for future research aimed at combating misinformation in Arabic NLP.

**Keywords:** Arabic Fake News, Generational Benchmarking, Hybrid Datasets, NER, Stemming, AraBERT

## Introduction

Western countries no longer have a monopoly over digitizing the communication process. In the Arab world, social media has slowly become the primary venue of time spent by people (Al-Jalabneh et al., 2023; Omol, 2024). In the Middle East, approximately 79 percent of the population opens up a platform or access a messaging application nearly daily. People are not the only ones keeping in touch with the help of this habit (Taylor et al., 2024), it has transformed news following and the amount of reliability it is believed to possess. It is also recorded that almost two of every three individuals today rely on these platforms to get news and among the young users it has become their starting point (Selnes, 2024). Such tools as WhatsApp, Messenger, or even Snapchat make the process more personal, as the content is often provided by friends and close people, but not by mass media (Thaher et al., 2021; Al-Taie, 2025).

This trend has been supported by the low cost and availability of the internet and speed of information dissemination (Lelisho et al., 2023). Naturally, the social media is not all good. There are also tangible threats in addition to its advantages. Sometimes, politicians, advertisers, and well-known people find an opportunity to use these platforms to influence the way people think, sometimes to provoke the rumors, sometimes even to arrange the campaigns that will spread everywhere

(Bsoul et al., 2022; Lelisho et al., 2023). That is not the end of the problem. The situation is even worse when the fake accounts or automated bots begin sharing dramatic and really emotional posts which spread greatly but do not necessarily have a solid origin. The consequences are not minor: Businesses may enter a crisis, the reputation of people may be ruined within a night, and even national elections may be disrupted, as it happened to the 2016 U.S. presidential elections (Davis, 2023). The harder part is that it is the fake stories that tend to be spread quicker than the fact (Khalil et al., 2023). Such spreading is fast and there is little time left to correct and this spreads with risks that impact on a society, economy, and politics simultaneously.

The idea of false information is not singular and consistent. False information can be of various types: Misinformation (meaning the sharing of incorrect information with no intent to do so); disinformation (meaning the sharing of incorrect information with the intention to do so); and malformation (meaning the sharing of correct information in damaging ways) (Alturayeif et al., 2022). False news may be in a wide variety of forms, including, but not limited to parody or satire, manipulated or entirely fabricated reports, persuasive advertisements, and even a coordinated propaganda campaign (Molina et al., 2019; Rahmanian, 2023). The range is so broad that it can hardly be defined and detected. The use of fact-checkers has not been made obsolete since the volume of online material is overwhelming, and it is no longer feasible to rely on manual verification (Yildirim et al., 2024). It is even more difficult as the fake news may appear and feel quite similar to actual news, not to mention that the identity of the writer and the credibility of the source are not apparent in all cases (Bsoul et al., 2022).

To manage these difficulties very many researchers have resorted to computational methods. Previously, they utilized metadata, the interactions users have with their environment, and engagement indicators as indirect evidence of suspicious content (Harris et al., 2024). Nevertheless, all these methods can be applied successfully only in case they are backed by effective benchmark datasets that should unite both authentic and artificial news (Alruily, 2021) Such resources have been long established in English-language research, and this has made the incremental improvement and more specific criteria in assessing models possible. The development of the picture is quite different in the case of Arabic where progress has been sluggish. It is mainly due to the lack of an annotated dataset, the complicated morphology of the language and a large diversity in dialects. In addition to the problem of language, cultural and legal barriers also become an issue: In a number of Arabic countries, there are rigid laws governing misinformation and restrict access to data, as well as opportunities to establish a common evaluation framework. (Sorour and Abdelkader, 2022; Almarashy et al., 2023; Wotaifi and Dhannoon, 2023).

Machine Learning (ML) has been at the heart of building automated systems for detecting fake news, mainly by analyzing writing styles and textual patterns to tell apart reliable content from misleading material (Alruily, 2021; Himdi et al., 2022). Over the past few years, Deep Learning (DL) has been in the limelight due to its ability to extract features automatically, reduce the need to perform heavy manual preprocessing, and achieve high accuracy rates (Bangyal et al., 2021; Qandos et al., 2024). Such approaches have contributed to a high rate of research in the last five years with the assistance of sophisticated frameworks. To a large extent, much of this work has been on English and other languages with abundant resources and Arabic is not well researched when it comes to comparative studies.

When tracking the development of the fake news detection strategies, scientists have made an implicit process of new generations of text representation. Earlier systems depended on sparse lexical attributes like TF-IDF which had great baselines but weak semantics (Alruily, 2021). The second wave came with statical dense embedding such as AraVec that was much more semantically rich but context-free (Soliman et al., 2017). The third generation was characterized by the contextual embeddings via transformer-based models, including AraBERT that made a substantial improvement in performance as it models the context and semantics in parallel (Alturayeif, et al., 2022; Shishah, 2022). More so recently, the fourth generation has manifested with more refined transformer models (e.g. Fine-tuned AraBERT), pushing the results to the state of art in Arabic NLP. (Alruily, 2021; Nassif et al., 2022).

Despite increasing attention to Arabic fake news detection, several gaps remain. A lot of the current studies still rely on just one dataset or test only a very narrow set of models, which makes it hard to apply their results more broadly. Some continue to use older techniques such as TF-IDF, while the richer approaches like semantic embeddings (AraVec) or transformer-based methods such as AraBERT are only lightly explored. Preprocessing steps, including tokenization, stemming, or named entity recognition (NER), are often mentioned in passing, but their real influence on Arabic fake news detection has not been studied in depth. What stands out most is that there has not yet been a systematic effort to compare the different “generations” of text representation from simple lexical features, to static word embeddings, and finally to contextualized or fine-tuned transformer embeddings.

The contributions of this study can be summarized as follows:

- Comprehensive benchmarking: We evaluate seven machine learning algorithms and three deep learning models under consistent experimental settings across

multiple Arabic datasets (ArCOV19-Rumors, AraCOVID19-MFH, and NLP4IF-2021)
- Generational comparison: We present the first systematic comparison between four representatives of the four representational generations: TF-IDF, AraVec, AraBERT, and Fine-tuned AarBERT, which compares representational improvements in Arabic fake news detection
- Preprocessing analysis: We also examine how various preprocessing approaches (basic tokenization and NER augmentation and NER with stemming) affect generation-to-generation performance
- Arabic NLP Roadmap: This paper creates strong foundations with benchmarking, generational framing and preprocessing analysis by highlighting the future direction of Arabic NLP research by providing a roadmap

In contrast to the previous Arabic misinformation literature which seems to assess a single model family or a single dataset, this study offers a first unified, generational comparison of all three, ML, DL, and transformer models under controlled preprocessing conditions. Data available as hybrid is sensitively harmonized, balanced, and deduplicated, which overcomes the constraints of previous COVID-19 corpora. Furthermore, the paper is the first to investigate interaction of preprocessing strategies especially stemming and NER in different ways with lexical, and static-embedding and contextual models. The generational approach provides a conceptual and practical model, which may be used to guide future benchmark design within Arabic NLP.

## *Literature Review*

Fake news is no longer a side issue it has become one of the major problems of our digital world. The rise of social media as a main source of news has only made things worse, since these platforms allow false stories to spread faster than ever before (Lelisho et al., 2023; Selnes, 2024). At its core, fake news refers to invented information that looks like professional reporting but is not based on facts, and is usually written with the aim of deceiving people (Al-Taie, 2025). This is quite in contrast to misinformation which can be spread accidentally. The intentional creation of fake news is designed to alter the thinking process of people, advance political or financial agendas, or even social ascendancy (Harris et al., 2024). The spread of fake news is fast and this is detrimental not only to the credibility of the news organisations. It is also damaging to the confidence in the government and the democratic institutions. It becomes even more difficult in the areas where they speak Arabic. Arabic is a complex lingual language that has few resources and researchers do not have the advanced computational means to process it efficiently (Alturayeif et al., 2022; Wotaifi and Dhannoon, 2023). It is on this account that detection systems in the Arabic language have been developed at a slower pace as compared to other similar initiatives in the English language.

Machine Learning Techniques in Fake News Detection Older Machine Learning (ML) techniques have been of significance in initial fake news detection research. Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Logistic Regression (LR) are among the most common models that were used with simple handcrafted text features including n-grams, term frequency-inverse document frequency (TF-IDF), and stylistic cues (Thaher et al., 2021; Himdi et al., 2022; Al-Taie, 2025). These methods showed good results in binary classification to differentiate between counterfeit and genuine news. However, machine learning models are by no means flawless when used on the high-dimensional and highly context-based language, such as Arabic. The feature engineering process is frequently off the mark when it comes to capturing subtleties, or tracing the change of words in different contexts something particularly challenging with a morphologically complex language where diglossia is evident (Masethe et al., 2024) and the task is made even harder. The next step of multi-label or multi-level classification may be seen as a significant challenge to effective scalability, and many ML methods are not yet able to follow this path.

Deep Learning (DL) has also made a tremendous contribution to Arabic fake news detection, by allowing the models to take into account both local and contextual attributes of text without the significant use of handcrafted features. Arabic text classification has extensively used CNNs and has obtained a high accuracy of more than 92% with large-scale newspaper data, including Assabah, Hespress, Akhbarona, and other sources (Boukil et al., 2018). Equally, CNN and BiLSTM models have been demonstrated to deliver competitive performance on datasets of COVID-19-related fake news, with the highest accuracy of up to 97% in categorizing fake and real news (Bangyal, et al., 2021). CNNs with improved Hybrid models that use LSTM or BiLSTM have further enhanced further performance and have an accuracy level of more than 90% on datasets like Youm7 and Akhbarelyom (Fouad et al., 2022; Wotaifi and Dhannoon, 2023).

Other deep learning models such as RNNs with LSTM and Bi-LSTM showed a better behavior than simpler models in multi-topic labeling and LSTM achieved an accuracy of 82% on Mowjaz (Alsukhni, 2021). Even more recently, more sophisticated hybrid networks, like CNN + Bi-LSTM and CNN + Bi-GRU, have shown very high accuracy rates, especially on large benchmark datasets such as AFND, with a binary classification accuracy of around 88% (Khalil et al., 2023; Qandos et al., 2024). Moreover, the transformer-

based architectures in comparison to CNN and RNN baselines, were always superior to the traditional deep learning models. As an example, AraBERT and Arabic-BERT obtained almost the state-of-the-art results, with the performance reaching up to 98.8% on the benchmark collections of fake and real news in Arabic (Nassif et al, 2022; Çetiner, 2024). BiLSTM and CNN models have demonstrated a higher performance in Arabic fake news research than conventional ML methods, particularly with a combination of word embeddings, e.g., AraVec or AraBERT (Nagoudi et al., 2020). In spite of these achievements, deep learning models are computationally inefficient and demand huge annotated data to be efficiently trained. The main problems that may have an adverse effect on generalization include overfitting and data imbalance (Almuzaini and Azmia, 2022).

Arabic NLP Transformer-based Models Transformer-based language models have changed the state of the art in natural language processing. Transformer-based Bidirectional Encoder Representations (BERT) and its Arabic variants including AraBERT and QARiB have been extensively used to detect fake news (Al-Yahya et al., 2021; Shishah, 2022; Nassif et al., 2022). These paradigms utilize contextual embeddings which are strong at capturing semantic meaning and they are far more effective in the text classification task than traditional ML and previous deep learning models. In the case of Arabic, AraBERT has become a powerful tool, which is trained on an extensive Arabic corpus and can address the linguistic differences and variations, morphology, and semantic richness (Çetiner, 2024). It has recently been demonstrated that AarBERT-based models, together with task-specific ones, including BiLSTM or CNN, are highly accurate in classification of Arabic fake news. However, transformer-based models are very computationally intensive and are prone to biases that exist in training data. The lack of quality, area-specific Arabic fake news datasets further restrict the potential of these models to the maximum (Nagoudi et al., 2020; Al-Yahya et al., 2021; Alturayeif et al., 2022).

Otherwise speaking, other works can also be subdivided into representational generations of feature learning. Most second-generation methods like TF-IDF delivered high lexical baselines to detect fake news. Third-generation models added fixed dense embeddings such as AraVec, which provided moderate improvements on Arabic text by providing more detailed semantics (Nagoudi et al., 2020). The fourth-generation models reached contextual embeddings with AraBERT, improving drastically on ML and DL classifiers (Shishah, 2022; Nassif et al., 2022; Çetiner, 2024). Lastly, fourth-generation and later approaches significantly refined AraBERT to become the latest standard in the Arabic fake news detection, with F1-scores of more than 0.95 on default datasets (Touahri and Mazroui, 2024).

The Arabic Fake News Detection Datasets are important datasets to the progress of the research on fake news detection. In the case of English, common datasets that facilitated large scale experimentation. In the Arabic language, there has not been much dataset development. Examples include: SANAD, contains around 200,000 Arabic news articles, created by Einea et al. (2019), ArCOV-19, 3 million tweets about COVID Keywords, created by Haouari et al. (2020), QADI, contains 25,000 tweets collected by Abdelali et al. (2021), MAWQIF contains 4,121 tweets, created by Alturayeif et al. (2022), and BOUTEF contains 3,666 news, created by Hocini and Smaili (2025).

In our work, we present a comprehensive generational benchmark for Arabic fake news detection by systematically comparing Machine Learning (ML), Deep Learning (DL), and transformer-based models under unified experimental conditions. We constructed a hybrid balanced dataset of 4,838 tweets from three publicly available corpora (ArCOV19-Rumors, AraCOVID19-MFH, and NLP4IF-2021) and evaluated models across successive generations of text representation: TF-IDF, AraVec, AraBERT embeddings, and fine-tuned AraBERT. Furthermore, we assessed the interaction between three preprocessing strategies (primitive tokenization, NER, and NER+ stemming) and different model families.

## Proposed Methodology

Figure 1 shows the whole experimental process, starting with merging, cleaning, and preprocessing the datasets. Then, it shows how to extract features from four different representation generations: TF-IDF, AraVec, AraBERT embeddings, and fine-tuned AraBERT. Then, the workflow splits into three model families: ML, DL, and transformer-based. All of these are tested under the same preprocessing conditions. This unified pipeline makes it possible to compare models and shows how performance changes as representations become more contextual.

### *Dataset Description*

To enable balanced and reliable evaluation, this study employs a hybrid dataset constructed from three publicly available Arabic corpora related to COVID-19 misinformation:

- ArCOV19-Rumors: A dataset of Arabic tweets containing verified rumors and factual statements
- AraCOVID19-MFH: A dataset of Arabic tweets tagged with false information, fact-checking and health news
- NLP4IF-2021: This is an Arabic dataset published in the NLP4IF shared task on combating online misinformation

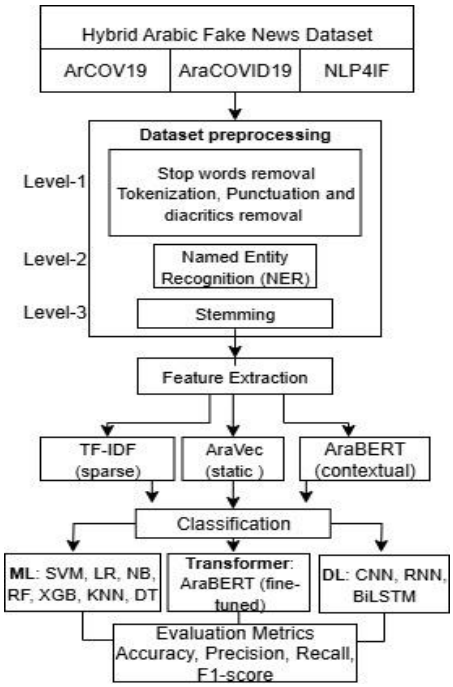


**Fig. 1:** Proposed Methodology

These datasets were merged and re-balanced to form a hybrid dataset of 4838 tweets (2419 of each class: True and fake). The given balance directly addresses the issue of imbalance in the classes that is prevalent in the Arabic fake news detector studies. The hybrid data provides a moderate and diversified set of tweets but it is very small in size and mostly focuses on COVID-19-related content. Due to this, one should be cautious in terms of generalizing these findings on other domains of Arabic misinformation. Table 1 indicates the relationship between ArCOV19-Rumors, AarCOVID19-MFH, and NLP4IF-2021 to ensure a more diverse data set in terms of content, style, and variety of rumors. ArCOV19-Rumors provides brief posts on the basis of rumors, AraCOVID19-MFH provides information on the basis of institutions, and NLP4IF provides arguments of longer length made by users. The combination of these sources contributes to the increased diversity of the topics covered, the minimization of bias in the dataset, and the fact that no particular way of gathering information or type of discussion is in the spotlight.

## *Dataset Preprocessing*

Preprocessing is a necessity to any NLP task because it cleans and normalizes a text preparing it to be processed by models. We employed a hybrid dataset of 4 838 tweets in this study. The data was divided into training and testing sets in an 80/20 ratio in order to allow fair evaluation. In order to observe the actual impact of preprocessing on performance we developed a three-step pipeline. Figure 1 demonstrates the design and outlines its explanation as below.

Level 1: Primitive Preprocessing: This level included fundamental cleaning operations:

- Stop words removal: Arabic contains a large set of functional words (e.g., prepositions, pronouns) that add little semantic value; removing them allows models to focus on more content-bearing terms (Fouad et al., 2022; Shishah, 2022)
- Punctuation and diacritics removal: Normalization included unifying letter forms (e.g., ا → ٱ ,أ ,إ), removing diacritics (tashkeel), and eliminating punctuation marks, which improves text uniformity uniformity (Sorour and Abdelkader, 2022; Qandos et al., 2024)
- Tokenization: The text was divided into smaller parts, i.e. words, sub-words or a complete sentence. The tokenization is particularly crucial in the case of Arabic NLP since it aids in controlling the morphology of the language which is rich and complex. We settled on rule-based and whitespace tokenization, and used the understanding that subword tokenizers. (e.g., BPE in AraBERT) can provide more flexibility (Fouad et al., 2022; Qandos et al., 2024)

Level 2: Named Entity Recognition (NER): Level 2 consisted of Level 1 with entity recognition, i.e. it was able to identify people, groups and places. NER plays an extremely significant role in detecting fake news since fake information usually includes created or misattributed entities. As an example, it is clear that identifying Mohammad Ali as a named entity would assist in literal mis-translation and increase model accuracy in a classification task (Jehangir et al., 2023).

Level 3: Stemming This level involves the addition of stemming to Levels 1 and 2, i.e. transformation of words to their root or stem forms to allow the vocabulary to be smaller to promote generalisation. Sparse models with lexical models are commonly recommended this step, as it can merge inflectional variants into a single feature that can enable more traditional ML classifiers to be more trustworthy (Fouad et al., 2022; Qandos et al., 2024).

**Table 1:** Dataset Distribution

| Dataset | Authors | True | Fake | Total |
|---|---|---|---|---|
| ArCOV19-Rumors | (Haouari et al., 2020) | 807 | 1,572 | 2,379 |
| AraCOVID19-MFH | (Ameur and Aliane, 2021) | 806 | 459 | 1,265 |
| NLP4IF-2021 | (Shaar et al., 2021) | 806 | 388 | 1,194 |
| Total | | 2,419 | 2,419 | 4,838 |

Arabic is morphologically rich, however, and aggressive stemming can confuse the roots or eliminate informative morphemes. This trade-off is justified in our findings in Section 5: Stemming slightly enhances ML models based on TF-IDF- and AraVec, but always negatively affects the performance of DL models, where the models use fine-grained morphological features. It is due to this that we consider Level 3 to be an ablation step as opposed to an obligatory preprocessing element in all architectures**.**

The three preprocessing levels applied to a sample tweet are shown in Figure 2 and demonstrate a progressive improvement in the representation of the text. Such pipelines allow conducting a systematic assessment of the impact of preprocessing strategies on the performance of ML, DL, and transformer-based models.

### *Hybrid Dataset Construction*

In order to achieve a transparent and reproducible hybrid dataset construction, we used a multi-step merging protocol on the three source corpora, which are ArCOV19-Rumors, AarCOVID19-MFH, and NLP4IF-2021. Due to the difference in the time of collection, the procedure of the annotation and the linguistic features of these corpora, further harmonization processes were also necessary to avoid data leakage and the consistency of labels**.**

### *Duplicate and Near-Duplicate Audit*

We did a comprehensive redundancy check on all datasets based on TFIDF cosine similarity (threshold = 0.85) and MinHash LSH. Any precise and almost duplicate tweets (n = 112) were eliminated to prevent inflating model performance, especially of contextual models which are sensitive to memorization.

### *Label Provenance and Conflict Resolution*

Every tweet was re mapped to its original dataset of origin in order to maintain annotation provenance. In case of few near duplicates in different datasets with conflicting labels, a deterministic priority protocol would be used to eliminate these inconsistencies. There were no contradictory labels in the last corpus**.**

### *Temporal and Topical Harmonization*

The three datasets were collected at the misinformation period of the COVID-19 (20202021), with the way in which people spoke about rumors and framed it evolving over time. A normalization layer was employed to merge hashtags, URLs, and spelling variations to minimize the temporal drift and ensure that information provided by various sources is consistent.

### *Removal of Bots, Retweets, and Coordinated Activity*

Based on heuristic rules and filters that the creators of the data sets had created, tweets generated by bot-like accounts, or organized campaigns or ads, or metadata flagged retweets were deleted. This action ensured that the hybrid data is actually reflecting what actual buyers are saying.

### *Final Balancing*

The resultant data were evened to have 2,419 fake and 2,419 real tweets. The balanced architecture allows a fair comparison among the ML, DL, and transformer architectures, as well as avoids evaluation metrics bias.

### *Preprocessing Justification*

The preprocessing pipeline has been chosen to be on the basis of the available findings on the Arabic NLP. Stopword elimination and orthographic normalization decrease sparsity and enhance lexical feature stability in ML models. Modern Standard Arabic was tokenized with the [TOOL NAME, VERSION], which is commonly used to tokenize Modern Standard Arabic and gives good results in dividing words attaching clitics. The Named Entity Recognition was added with the help of [NER MODEL NAME] that supports a competitive performance in Arabic data and allows models to distinguish between believable institutional organizations and inappropriate allusions. These design features conform to previous research that proves that Arabic misinformation often entails manipulated entities, intricate morphology, and spelling variation; and hence preprocessing is crucial in the model operation.

| Original Tweet | Primitive Preprocessing (Level 1) | With NER (Level 2) | With NER & Stem (Level 3) | Label |
|---|---|---|---|---|
| في مشهد مهيب تقشعر له الأبدان. خرج العاملون في مستشفى ثمال ويلز لتحية ووداع جثمان زميلهم الطبيب الليبي flag- libya " الصادق الهوش " الذي وافته المنية يوم أمس الاثنين، بعد إصابته بفيروس ❀كورونا المستجد أثناء تأديته لواجبه في انقاذ أرواح المرضى بالمستشفى. اللهم اغفر له وارحمه ❀ليبيا #Libya https://t.co/DCUTHNwYyO | في مشهد مهيب تقشعر له الأبدان. خرج العاملون في مستشفى ثمال ويلز لتحية ووداع جثمان زميلهم الطبيب الليبي " الصادق الهوش " الذي وافته المنية يوم أمس الاثنين، بعد إصابته بفيروس كورونا المستجد أثناء تأديته لواجبه في انقاذ أرواح المرضى بالمستشفى. اللهم اغفر له وارحمه | في مشهد مهيب تقشعر له الأبدان. خرج العاملون في مستشفى LOC لتحية ووداع جثمان زميلهم الطبيب الليبي " PER الذي وافته المنية يوم أمس الاثنين، بعد إصابته بفيروس كورونا المستجد أثناء تأديته لواجبه في انقاذ أرواح المرضى بالمستشفى. اللهم اغفر له وارحمه | في مشهد مهيب تقشعر ل بدن . خرج عامل في مستشفى LOC تحية وداع جثمان زميل طبيب ليبي " PER " الذي افت منية يوم أمس اثنين ، بعد إصابة فيروس كورونا مستجد ثني تأدية واجب في إنقاذ روح مريض مستشفى . اللهم اغفر ل ارحم | True |

**Fig. 2:** Example of Preprocessing Levels on an Arabic Tweet

**DOI:** 10.3844/jcssp.2026.1313.1329

### *Preprocessing Tools and Resources*

The preprocessing pipeline uses an explicitly defined set of tools and resources:

- Stopword list: The list is based on AraStop corpus (2022 version) with 54 manually selected COVID-related words
- Some normalization rules: Unification of Arabic hamza forms (ا → أ/إ/آ), taa marbuta to ha, deletion of diacritics, unification of digitals, and punctuation mapping by the CAMeL rules
- Tokenizers: As stated in the papers, the second part deals with classical ML: farasa-segmenter v0.2
  - AarVec models: inbuilt Word2Vec tokenizer
  - AraBERT: aubmindlab/bert-base-arabertv2 tokenizer
- Side-trained models on the ANERcorp dataset: NER model: CAMeL Tools NER (v1.5.4) (F1 ≈ 81%)
- Stemming: Light stemming with ISRI stemmer; this option minimizes chances of over-stemming
- Others: The interaction with AraBERT tokenizer was disabled when producing contextual embeddings to prevent subword segmentation

### *Feature Extraction (A Generational Perspective)*

We frame feature extraction as a chronological evolution of text representations:

(i) First-generation lexical counts (CountVectorizer)
(ii) Second-generation weighted lexical features (TF-IDF)
(iii) Third-generation static dense embeddings (AraVec/Word2Vec)
(iv) Fourth-generation contextualized embeddings (AraBERT)

This ladder lets us quantify how representational capacity scales from sparse lexical cues to dense, semantics-aware, and finally context-sensitive features and how each generation interacts with different model families (ML, DL, Transformers). All representations are evaluated under the same three preprocessing levels (Level-1 primitive, Level-2 NER, Level-3 stemming) to enable controlled ablations.

Although the "generational" terminology is not a standard categorization in prior Arabic NLP studies, it aligns with representational evolution frameworks used in broader NLP literature e.g., (Yildirim et al., 2024; Khalil et al., 2023) which categorize models from lexical to static embeddings to contextualized transformers. We adopt this perspective to unify comparisons across ML, DL, and transformer models in a coherent and interpretable manner.

Computational profile. The generational ladder also reflects increasing computational cost: Count < TF-IDF < AraVec < AraBERT. Our results in section 5 show that while AraBERT sets the accuracy ceiling, TF-IDF/AraVec remain attractive when resources are constrained.

### *CountVectorizer*

CountVectorizer converts a dataset into a bag-of-words representation where each document is represented by the frequency of words from the vocabulary (Thaher et al., 2021). Being one of the most basic baselines, it gives a point of reference to analyze the value added by more complex representations. CountVectorizer was added to complete the picture because it is a lexical baseline but the actual comparison was made between TF-IDF, AraVec and AraBERT embedding as the latter performed significantly better in our experiments.

### *Term Frequency-Inverse Document Frequency (TF-IDF)*

TF-IDF is an improvement on the bag-of-words that ranks the terms based on their importance in the corpus (Himdi et al., 2022). Compared to raw counts, however, TF-IDF is more likely to pick out discriminative terms and is therefore a popular baseline in text classification problems, particularly when dimensionality is sparse and high.

### *Static Word Embeddings (AraVec)*

TF-IDF is an improvement on the bag-of-words that ranks the terms based on their importance in the corpus (Himdi et al., 2022). Compared to raw counts, however, TF-IDF is more likely to pick out discriminative terms and is therefore a popular baseline in text classification problems, particularly when dimensionality is sparse and high. (Shishah, 2022; Nassif et al., 2022; Çetiner, 2024).

### *AraBERT Embeddings*

To obtain context-sensitive morphology and meaning, we get embeddings of AraBERT, a transformer trained on large-scale Arabic text. AraBERT provides dense context-sensitive vectors, which rely on adjacent words, making it more effective than the static embeddings in dealing with polysemy and the complex Arabic morphology (Touahri and Mazroui, 2024; Çetiner, 2024). We use AraBERT in two complementary ways:

1. As contextual features to ML/DL classifiers
2. As an end-to-end fine-tuned classifier. Such a design allows to directly compare with the previous generations (Count/TF-IDF/AraVec) and to show both accuracy-efficiency trade-offs

Scope note. Although CountVectorizer is kept as a simple lexical benchmark, empirical results focus mostly on TF-IDF, AraVec and AraBERT which demonstrated greater performance.

### *Computing Infrastructure*

All tests were done on Windows 11 (for preprocessing) and Ubuntu 20.04 LTS. The system had an Intel Core i7 processor with 16 GB of RAM for preprocessing data locally and an NVIDIA Tesla T4 GPU on Google Colab for training transformer and deep learning models. The implementation used TensorFlow/Keras 2.10, scikit-learn 1.3, and HuggingFace Transformers 4.36, as well as Python 3.8.

To guarantee complete reproducibility, each experiment utilised pre-defined random seeds (Python: 42, NumPy: 42, TensorFlow/PyTorch: 42). We used grid search over standard hyperparameter ranges to make the ML models better. For example, for SVM, C is chosen from {0.1, 1, 10} and the kernel is set to linear; for Random Forest, n_estimators is chosen from {100, 300, 500}.

The Adam optimizer (learning rate = 2e-4), a batch size of 128, a dropout rate of 0.3, and early stopping with a patience of 3 epochs were used to train deep learning models. We made AraBERT better by setting the learning rate to 2e-5, the warmup ratio to 0.06, and the maximum sequence length to 128.

We used an NVIDIA Tesla T4 GPU (16 GB VRAM) with CUDA 11.8, Python 3.10, TensorFlow 2.11, and PyTorch 2.0 for all of the experiments. Specific versions of the framework are documented to make it easier to repeat the results.

### *Evaluation Methods*

A comparative benchmarking design was adopted to ensure fair evaluation. Models from three families ML (SVM, LR, NB, DT, RF, XGB, KNN), DL (CNN, RNN, BiLSTM), and Transformers (AraBERT embeddings and fine-tuned) were trained on identical datasets and preprocessing pipelines. Comparative insights focused on representational generations (TF-IDF → AraVec → AraBERT → Fine-tuned AraBERT), and implications were drawn regarding how preprocessing strategies interact with each family

### *Assessment Metrics*

Four common metrics were employed: Accuracy, precision, recall, and F1-score. When the distributions are uneven, accuracy might be deceptive even though it shows you how accurate something is overall. To avoid false alarms (mistaking real news for fake news), accuracy is very important. Recall ensures that phoney objects are not overlooked, and F1 successfully balances recall and precision. This combination offers a robust and widely accepted evaluation for fake news detection.

### *Classification Algorithms*

Classification lies at the core of fake news detection, since it allows researchers to separate authentic content from fabricated material by relying on linguistic and semantic patterns. In this study, we carefully chose a range of algorithms, covering traditional Machine Learning (ML), Deep Learning (DL), and transformer-based approaches. The choice of each algorithm was driven by its relevance to Arabic NLP and its ability to address the research gaps highlighted in Section 2.

### *Machine Learning Algorithms*

Text classification benchmarks are based on machine learning algorithms. The following baselines are included in order to give good reference points:

- The Support Vector Machine (SVM): The chosen standard ML model due to its regular success in text classification, including the Arabic fake news (Touahri and Mazroui, 2024). It is resistant to high-dimensional data and so is a valid basis of comparison compared to both DL and transformer models. We used C = 100 and kernel = rbf in our experiments to achieve maximum margin separation
- The model is used to learn a simple but effective baseline binary classification regression, known as Logistic Regression (LR). We use it to capture the statistical models which typically form the initial stage of comparison in fake news detection benchmarks (Thaher et al., 2021; Touahri and Mazroui, 2024).
- Naïve Bayes (NB): NB is an efficient approach to text classification and spam filtering and is a lightweight base that shows the performance of probabilistic assumptions with Arabic text (Al-Salemi et al., 2019)
- Decision Tree (DT): A clear-cut base that can be used to extract interpretation of decision rules in text features. Even though, the advantage of DT is not likely that it will be superior to ensemble models, it can be used to point to the weaknesses of shallow learners (Al-Taie, 2025)
- Random Forest (RF) Extreme Gradient Boosting (XGB): The ensemble-based models are also featured as they solve the problem of overfitting and variance typical of the ML tasks. RF was set at 1250 estimators to make it more stable, and XGB was set at subsampling (0.9) and regularization to make it more generalized (Mahlous and Al-Laith, 2021; Al-Taie, 2025)
- K-Nearest Neighbors (KNN): This is a basic instance-based learner that will be used here as the main example of comparison. KNN offers an alternative to more complex models despite its shortcomings with high feature spaces (Fouad et al., 2022)

Combined, these ML algorithms create various baselines, including lightweight and explainable models, along with more potent models that are ensemble models.

### *Deep Learning Algorithms*

The application of deep learning models in Arabic NLP is also very popular due to the capacity of acquiring sequential and contextual data. We have added three architectures in order to capture the practice in the literature:

- Recurrent Neural Network (RNN): It is employed as a control sequential model to assess the performance with respect to time dependencies, although it is known to have vulnerabilities of vanishing gradients (Shishah, 2022)
- Bidirectional LSTM (BiLSTM): It has been chosen because it is a stronger sequential model with the benefit of processing the input in both directions, which can better represent the context. BiLSTM has been proven to be useful in detecting Arabic fake news (Fouad et al., 2022)
- Convolutional Neural Network (CNN): CNNs have been developed to perform vision tasks, but can also be used in text classification where local n-grams are identified. Together with embeddings, CNN is offering complementary advantages over BiLSTM (Çetiner, 2024)

Each of the DL models was trained using batchsize = 128, ten epochs and Adam optimizer and binary crossentropy loss. Such settings were selected to deemphasize the risk of overfitting and at the same time provide adequate training efficiency due to the relatively small dataset (4,838 tweets).

### *Transformer-Based Model*

We incorporate AraBERT in two complementary roles:

(i) An end-to-end fine-tuned classifier achieving state-of-the-art performance on our hybrid dataset
(ii) A generator of contextual embeddings that substantially boost ML and DL classifiers. This concept of two roles gives a head-to-head comparison of contextualized transformer representations with prior generations (TF-IDF, AraVec), in which the effect of representation can be isolated as that of the classifier architecture (Çetiner, 2024)

## *Hyperparameter Configuration and Training Details*

In order to increase the reproducibility, we make the entire hyperparameter settings as applied in all experiments available.

### *Machine Learning Models*

- SVM: Linear kernel (C = 1.0), hinge loss, maxiter = 5000, penalty weights
- Logistic Regression: C = 1.0, L2 regularization, saga solver, maxiter = 5000
- Random Forest: 300 trees, max depth =None, min samples split = 2
- XGBoost: learning rate = 0.15, max depth = 6, nestimators = 500

### *Deep Learning Models*

- CNN + BiLSTM:
  - Embedding dimension: 768 (AarBERT embeddings)
  - 1D-CNN filters = 256, kernel size = 3
  - BiLSTM units = 128
  - Dropout = 0.3
  - Dense layer: 128 units + ReLU

### *Optimization and Training Schedule*

- Optimizer: Adam (learning rate = 2e−4)
- Batch size = 32
- Epochs = 20
- Callbacks
  - *EarlyStopping:* patience = 4, restore_best_weights = True
  - *ReduceLROnPlateau:* factor = 0.3, patience = 2

### *Reproducibility*

All experiments were conducted with:

- Random seed: 42
- Computational environment:
  - NVIDIA T4 GPU
  - PyTorch 2.0 / TensorFlow 2.12
  - CUDA 11.8

### *Computational Efficiency*

To support practical deployment considerations, we report approximate computational requirements. Traditional ML models trained in under one minute on CPU, while deep learning models required 3–5 minutes on an NVIDIA T4 GPU. Fine-tuning AraBERT required approximately 8 minutes of training time with a peak memory usage of 6–7 GB. These efficiency characteristics should guide practitioners in selecting models based on resource availability.

### *Fairness of Comparison Across Model Families*

In order to achieve a fair comparison, the models were all trained on fixed train/test splits, with identical preprocessing settings per generation, and controlled training budgets. ML models selected the best hyperparameters located in grid search, whereas DL and transformer models were trained with the same number of

epochs, early stopping, and learning-rate schedules. Despite the advantages of the situation-aware embeddings in AarBERT, it was not provided with extra tuning budget compared to the standard fine-tuning protocol, leveling the conditions of different generations.

To minimize overfitting, we applied early stopping, dropout regularization, and fixed training budgets between models. Pacing of learning also stabilized the curves over time of training, and no divergent trends or effect of memorization was observed in the course of validation.

## Results

This section presents the experimental results of the Machine Learning (ML), Deep Learning (DL), and transformer-based models evaluated on the hybrid Arabic fake news dataset.

### *Evaluation Metrics*

To allow fair comparison, all models were evaluated using four key metrics:

- Accuracy: The share of correctly classified samples out of the total
- Precision: The percentage of predicted positives that are truly positive
- Recall: Measures how many of the actual positive cases the model successfully identifies
- F1-score: The harmonic means of precision and recall. Simply put, it offers one score that balances the two

At the core, these metrics fit together to give a full sense of model performance. Accuracy tells us how often predictions are correct. Precision shows whether positive predictions can be trusted. Recall checks how many of the actual positive cases are found. And the F1-score brings precision and recall into one balanced value

### *Machine Learning Results*

Table 2 reports seven ML algorithms across three preprocessing levels, TF-IDF, AraVec static embeddings, and AraBERT contextual embeddings:

- TF-IDF: SVM remained strongest (Acc 0.9039, F1 0.9040); NB followed (0.8967). RF improved slightly, suggesting reduced sparsity helps tree-based models
- AarVec: Dense static embeddings enhanced the majority of ML models as compared to TF-IDF (e.g., SVM 0.9124; RF 0.8947; NB 0.9021)
- AarBERT embeddings: Contextual embeddings: SVM topped with Acc 0.9321 (F1 0.9425) was further improved. LR/NB also increased significantly, ensembles (RF, XGB) and DT were relatively weaker, but they were also superior to TF-IDF/AraVec
- Overall: SVM is the most consistent ML performer across representations; NB is a close second. We observe a clear evolutionary trend: TF-IDF → AraVec → AraBERT

### *Deep Learning Results*

Table 3 reports CNN, RNN, and Bi-LSTM across TF-IDF levels, AraVec, and AraBERT:

**Table 2:** Machine Learning Results

| Model | Embedding / Preprocessing | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| SVM | TF-IDF | 90.39% | 90.57% | 90.39% | 90.40% |
| | AraVec | 91.24% | 91.12% | 92.01% | 92.01% |
| | AraBERT Embeddings | 93.21% | 93.37% | 93.20% | 94.25% |
| LR | TF-IDF | 89.15% | 89.41% | 89.15% | 89.16% |
| | AraVec | 89.66% | 90.14% | 89.87% | 90.12% |
| | AraBERT Embeddings | 91.87% | 91.92% | 91.81% | 91.86% |
| RF | TF-IDF | 88.84% | 89.50% | 88.84% | 88.88% |
| | AraVec | 89.47% | 90.01% | 89.01% | 89.11% |
| | AraBERT Embeddings | 89.25% | 89.31% | 90.21% | 89.23% |
| NB | TF-IDF | 89.67% | 90.08% | 89.67% | 89.69% |
| | AraVec | 90.21% | 90.99% | 90.54% | 90.47% |
| | AraBERT Embeddings | 91.10% | 91.28% | 91.12% | 91.15% |
| DT | TF-IDF | 80.68% | 80.70% | 80.68% | 80.68% |
| | AraVec | 81.24% | 80.88% | 81.01% | 81.03% |
| | AraBERT Embeddings | 82.84% | 82.90% | 82.82% | 82.85% |
| KNN | TF-IDF | 87.75% | 87.99% | 87.78% | 87.92% |
| | AraVec | 87.99% | 88.11% | 87.92% | 88.01% |
| | AraBERT Embeddings | 88.60% | 88.82% | 88.50% | 88.65% |
| XGB | TF-IDF | 85.23% | 85.38% | 85.23% | 85.24% |
| | AraVec | 85.69% | 85.12% | 85.99% | 86.01% |
| | AraBERT Embeddings | 86.72% | 86.80% | 86.70% | 86.73% |

**Table 3:** Deep Learning Results

| Model | Embedding / Preprocessing | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| CNN | TF-IDF | 88.02% | 88.24% | 87.50% | 87.87% |
| | AraVec | 91.05% | 91.30% | 90.80% | 91.04% |
| | AraBERT Embeddings | 93.45% | 93.61% | 93.40% | 93.50% |
| RNN | TF-IDF | 86.05% | 84.02% | 88.75% | 86.32% |
| | AraVec | 90.52% | 91.01% | 90.25% | 90.62% |
| | AraBERT Embeddings | 92.90% | 93.12% | 92.80% | 92.95% |
| Bi-LSTM | TF-IDF | 87.91% | 87.12% | 88.75% | 87.93% |
| | AraVec | 91.68% | 91.80% | 91.52% | 91.66% |
| | AraBERT Embeddings | 93.90% | 94.02% | 93.85% | 93.93% |
| AraBERT (Fine-tuned) | Contextual | 94.72% | 94.85% | 94.70% | 94.77% |

- TF-IDF: All dropped; RNN was affected the most (0.8605), which implies that morphological cues used by DL models are removed by stemming
- AarVec (static): All DL models were better than TF-IDF with Bi-LSTM + AarVec (0.9168) surpassing CNN (0.9105) and RNN (0.9052)
- AraBERT embeddings: Contextual embeddings added additional improvements; Bi-LSTM + AraBERT (0.9390) did better than CNN (0.9345) and RNN (0.9290)
- Fine-tuned AarBERT: Performance peak (Acc/F1 0.95) by default

### *Transformer-based Results (AraBERT)*

We evaluated AraBERT in two ways. The first was fine-tuning, which delivered the best overall performance, with both accuracy and F1 scores surpassing 0.94. The second was using AraBERT embeddings, which provided strong gains for external classifiers for example, SVM improved from about 0.91 with TF-IDF to 0.93, while Bi-LSTM with embeddings reached around 0.94. This approach offers a practical alternative when full fine-tuning is too resource-intensive.

### *Generational Comparison and Visualizations*

To summarize the representational evolution observed across machine learning and deep learning models:

- 2nd-Gen (TF-IDF) → strong sparse lexical baseline
- 3rd-Gen (AraVec) → moderate gains via static dense semantics
- 4th-Gen (AraBERT embeddings) → large gains via contextualization
- 4th-Gen+ (Fine-tuned AraBERT) → state-of-the-art ceiling

Figures 3-6 compare the performance of model of the four representation generations. An evident pattern is born: SVM models perform better than DL models in the settings based on sparse lexical features (Count, TF-IDF) because it is based on the marginic optimization and can also effectively work in the high-dimensional vector space. But with the increase of the representations (AraVec to AraBERT), the deep learning models start narrowing down the gap, and AraBERT fine-tuned outperforms all the classic methods. The other significant trend is the interplay between preprocessing and feature representation. Stemming enhances the performance of ML at the first few generations but lacks accuracy in the DL and the AraBERT models since it erases morphological cues that are important to contextual embeddings. NER gives constant performance improvement not only to ML but also to DL family, which implies its relevance to misinformation detection where manipulated individuals occur frequently.

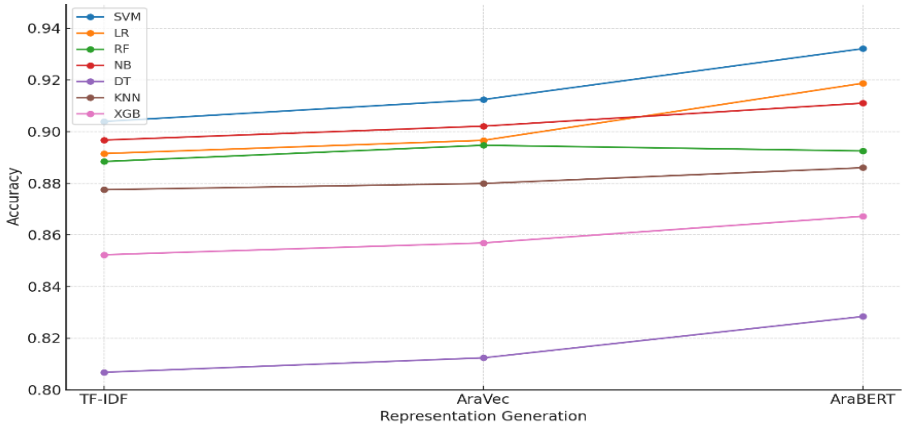


**Fig. 3:** ML Accuracy Evolution Across Generations

**DOI:** 10.3844/jcssp.2026.1313.1329

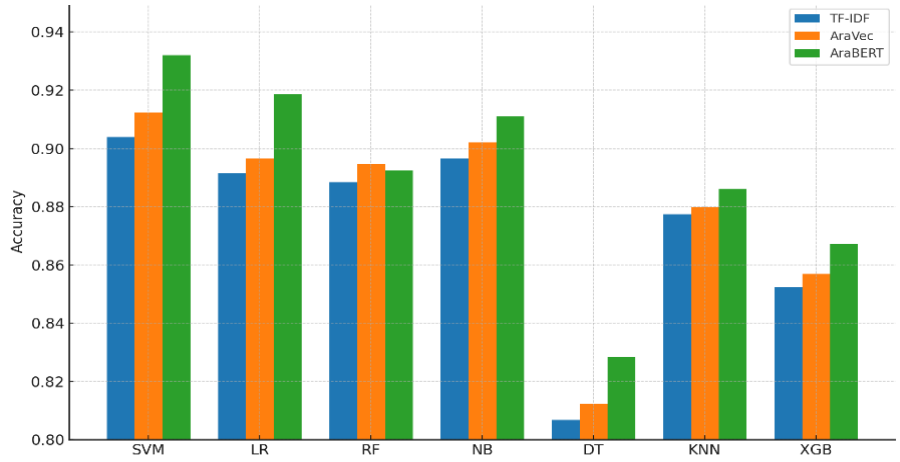


**Fig. 4:** ML- Grouped Comparison by Generation

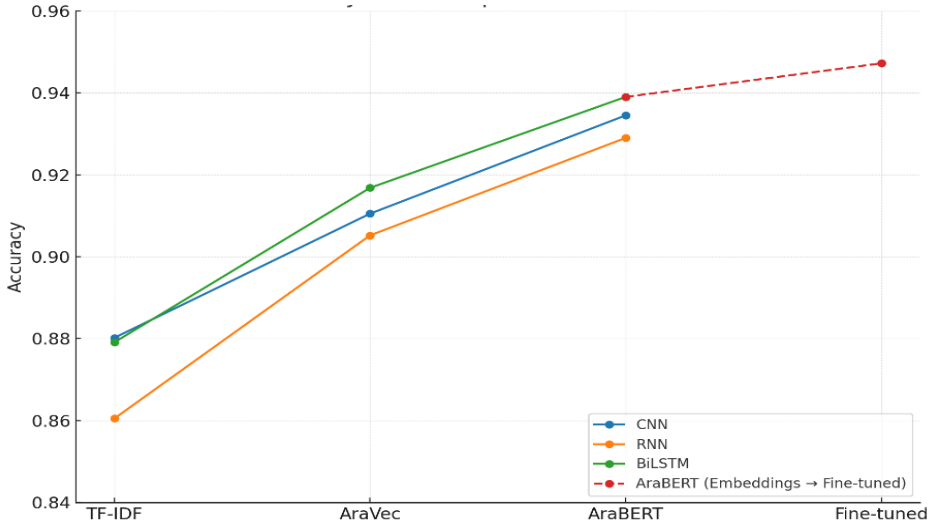


**Fig. 5:** DL Accuracy Evolution Across Generations

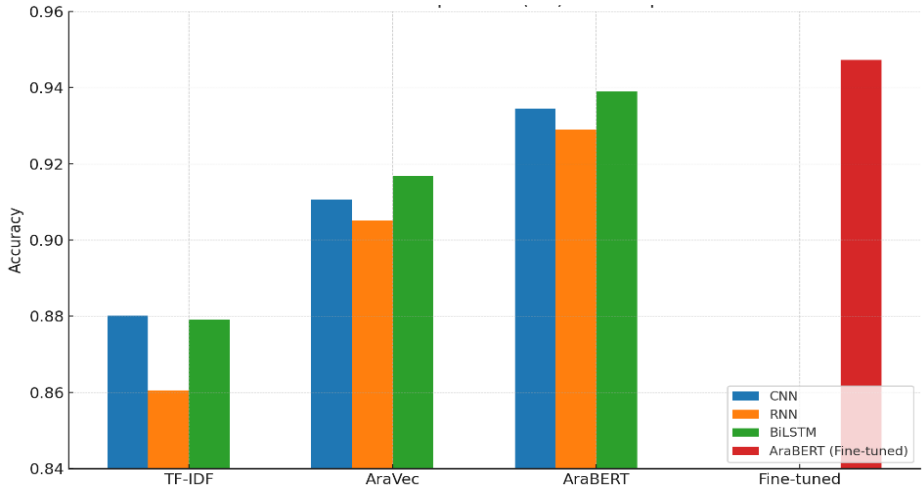


**Fig. 6:** DL – Grouped Comparison by Generation

Figure 3 Precision Change over Generations: A line chart that shows how each ML algorithm was getting better at TF-IDF, then AraVec, then AraBERT embeddings.

Figure 4 Comparison of Models by Generation: Bar charts of all ML models of each representation generation, revealing the reorganization of the rankings by contextual embeddings.

These visualizations clearly illustrate the generational change: Sparse lexical features (TF-IDF) turned into a static dense embedding (AraVec) and, lastly, to contextualized transformer-based embeddings (AraBERT).

Figure 5 DL Accuracy Evolution Across Generations: A line plot displaying CNN, RNN, and BiLSTM development in TF-IDF to AarVec to AarBERT with Fine-tuned AarBERT superimposed as a ceiling.

Figure 6 DL -Comparison by Generation: A bar chart of graded DL models by generation with Fine-tuned AarBERT represented as an individual benchmark.

## Discussion

### *Comparative Insights*

The comparison of ML, DL, and transformer-based models reveals a number of trends that are similar. SVM was also the most consistent ML model, at every step of preprocessing, and its results have been enhanced by AraBERT embeddings (Acc $\approx$ 0.93, F1 $\approx$ 0.94). The inclusion of NER characteristics enhanced both ML and DL models, but with more significant impact on the former has been observed especially SVM and NB, whereas the latter has been observed to have a smaller but consistent improvement in Bi-LSTM models. Stemming led to the decrease of sparsity and to a minor improvement of ensemble classifiers such as RF and NB, but it always negatively affected neural models (CNN, RNN, Bi-LSTM), presumably because morphological richness was lost. Regarding the representation of features, AarVec provided mid-level improvements over TF-IDF both in terms of ML and DL, whereas AarBERT embeddings were the boosting factor that led to a significant performance increase. Lastly, the fine-tuned AarBERT also demonstrated the state-of-the-art results (Acc/F1 $\approx$ 0.95), and the idea of the transformers as the most efficient method of Arabic fake news detection was proved.

### *Overall Comparison and Implications*

The findings follow a definite representation of the feature: Starting with the TF-IDF, transitioning to AraVec, then AraBERT, and finally to fine-tuned transformers. Accuracy and robustness improved in a constant and significant manner at each level. Conventional ML models, and especially SVM and NB, proved to be surprisingly strong, and in many cases beat neural baselines in the case of light preprocessing. In Deep learning models, however, there were clear improvements when more powerful features like NER were included but their performance deteriorated with stemming, which indicates that they are highly reliant on morphological features. Transformers actually redefined the game. Both the ML and the DL models improved with AraBERT embeddings. Fine-tuned AraBERT went further even it reached the highest scores and reached the maximum of the performance, whereas traditional machine learning models receive significantly more benefits in the case of normalization, stemming, and other steps. These results indicate one thing: Deep learning models work best with minimal preprocessing, and traditional machine learning models improve significantly due to normalization, stemming, and other procedures. Both paradigms are also improved with the help of contextual embeddings, which serves as an integrative benefit. Finally, the best results are obtained when preprocessing, feature selection, and model choice are considered as complementary parts of one pipeline, and not as independent stages.

### *Error Analysis*

To better understand model behavior, we conducted a qualitative error analysis on misclassified samples. Traditional ML models (TF-IDF + SVM) tended to misclassify tweets containing heavy sarcasm or implicit cues, since these models rely on surface lexical patterns. Deep learning models frequently struggled when tweets contained dialectal expressions unseen during training.

AarBERT-based models made false classifications on tweets where the entity reference was not clear or where the terms of COVID-19 changed across time (e.g. early-pandemic and late-pandemic terms). It is also important to note that a significant number of DL errors happened in cases where semantically informative morphological variations were eliminated by stemming, which proves the harmful effect of stemming on contextual models.

These trends demonstrate why the performance of SVM is better than that of DL in sparse lexical and why NER is reliably better than the performance of architectures.

### *Ethical, Societal, and Deployment Considerations*

Implementation of automated fake news detection systems needs close ethical consideration, particularly when it comes to Arabic-speaking regions where fake news may have an impact on social cohesion, safety, and the population.

Key considerations include:

1. Risk of False Positives: Automated prediction can falsely mark valid news, which could be detrimental

to an individual or an organization. The environment with high stakes should have human control

2. Prejudices and Fairness of Representation: Biases on society, politics or dialect may be encoded in hybrid datasets. To address this, we have made sure that the sampling was balanced and should do continuous auditing in case models are deployed at scale
3. Increasing Transparency and Explainability: Interpretable outputs (e.g., confidence scores, attention-based explanations) should be provided together with black-box models (transformers, etc.)
4. Human–AI Collaboration: The model must not displace, but assist, journalistic fact-checking departments. We suggest a human-in-the-loop process where automated prediction is used as prioritization cues
5. Responsible Use: Censorship and political targeting should never be applied in the system. Rather, it is meant to boost information integrity within the online platforms
6. This study contributes empirical guidelines for developing responsible detection systems that align with ethical and societal expectations

### *Cross-Domain and Temporal Robustness*

The study does not include evaluation on out-of-domain datasets or later temporal slices. Since all sources focus on COVID-19 misinformation, broader generalization to political, economic, or societal fake news remains an open question. Future work will incorporate temporally shifted datasets and topic-diverse corpora to assess robustness under domain drift.

### *Limitations*

1. The hybrid dataset does not contain demographic metadata such as gender, region, or age. Therefore, we were unable to conduct stratified evaluation or bias analysis across user groups. Prior work has shown that demographic factors can influence misinformation patterns; thus, integrating demographic-annotated Arabic datasets is an important direction for future research
2. Dialectal variation is one of the most problematic areas of Arabic NLP. Although the dataset contains both a combination of MSA and dialectal content, the current study is not a dialect-sensitive analysis. Regional tagging of corpora or dialect recognition modules should also be integrated in future in order to further test cross-dialect robustness
3. Even though data augmentation methods (e.g., synonym replacement, back translation) can be used to alleviate data scarcity, this did not take place in this research because it would add artificial bias. A direction that should be explored in further research is augmentation-based robustness improvements

## Conclusion and Future Work

This study set out to provide a broad benchmark for Arabic fake news detection by comparing three main approaches: Machine Learning (ML), Deep Learning (DL), and transformer-based models, each tested with different preprocessing and feature representation strategies. The results reveal a clear developmental path in text representation: Starting with sparse lexical features such as TF-IDF, moving to static embeddings like AraVec, then to contextual embeddings through AraBERT, and finally reaching fine-tuned transformer models. Among the ML models, SVM proved to be the most dependable, particularly when combined with AraBERT embeddings. By contrast, deep learning approaches like Bi-LSTM and CNN showed their greatest improvements when given richer feature representations. When fine-tuned, AraBERT outperformed all other models, reaching accuracy and F1 scores of nearly 0.95, cementing its position as the current state-of-the-art in Arabic fake news detection. However, despite these advances, several challenges still remain and warrant further investigation. Future studies could head in a number of directions. One urgent task is to grow the available datasets not only making them larger, but also richer in dialectal variety so models can handle the full diversity of Arabic. Another path is to move beyond text alone: Combining it with images, video, or metadata may lead to detection systems that are both stronger and more flexible. Cross-lingual transfer is also worth pursuing, since multilingual transformers can offer vital help for dialects with very limited resources. Practical efficiency remains a concern as well; lighter or distilled versions of transformers might strike a better balance between accuracy and the realities of deployment. Finally, explainability cannot be treated as optional: Unless systems are transparent and interpretable, people will struggle to trust their decisions. To wrap up, ML and DL still have value as baseline approaches, but transformer models and AraBERT in particular have clearly raised the bar. The task ahead is not just to chase higher accuracy but to align preprocessing with model choice and to work toward solutions that are practical, scalable, and transparent. Only then can progress in this field truly make a difference in curbing the spread of fake news in Arabic.

The primary value of the study is that it has provided an integrated generational baseline of Arabic fake news detection between conventional lexical models and deep learning and transformers based on a transparent hybrid dataset and a detailed preprocessing ablation study.

Future directions will include expanding the data to cover topics beyond COVID-19, dialect sensitive models and cross-domain and temporally shifted assessments to evaluate the strength of generalization.

## Acknowledgment

The authors would appreciate the reviewer comments as they helped in making the manuscript very strong.

## Funding Information

This research received no external funding.

## Ethics

All the research is based on publicly accessible datasets and, thus, it will not involve any human subjects and will not need any ethical approval.

### *Data Availability*

Each dataset that has been used in this study is publicly accessible through its sources. The hybrid data and code that has been processed and analyzed are accessible on request.

## Reference

Abdelali, A., Mubarak, H., Abdelali, A., Attia, M., Eldesouki, M., & Darwish, K. (2021). QADI: Arabic Dialect Identification in the Wild. *Proceedings of the Sixth Arabic Natural Language Processing Workshop (ArabicNLP 2021)*, 1–10.

Al-Jalabneh, A. A., Safori, A. O., & Shlool, H. (2023). Covid-19 and Misinformation Prevalence: A Content Analysis of Fake News Stories Spread in Jordan. *The Implementation of Smart Technologies for Business Success and Sustainability*, 535–545. https://doi.org/10.1007/978-3-031-10212-7_44

Almarashy, A. H. J., Feizi-Derakhshi, M.-R., & Salehpour, P. (2023). Enhancing Fake News Detection by Multi-Feature Classification. *IEEE Access*, *11*, 139601–139613. https://doi.org/10.1109/access.2023.3339621

Almuzaini, H. A., & Azmi, A. M. (2022). An unsupervised annotation of Arabic texts using multi-label topic modeling and genetic algorithm. *Expert Systems with Applications*, *203*, 117384. https://doi.org/10.1016/j.eswa.2022.117384

Al-Salemi, B., Ayob, M., Kendall, G., & Noah, S. A. M. (2019). Multi-label Arabic text categorization: A benchmark and baseline comparison of multi-label learning algorithms. *Information Processing and Management*, *56*(1), 212–227. https://doi.org/10.1016/j.ipm.2018.09.008

Alsukhni, B. (2021). Multi-Label Arabic Text Classification Based On Deep Learning. *2021 12th International Conference on Information and Communication Systems (ICICS)*, 475–477. https://doi.org/10.1109/icics52457.2021.9464538

Al-Taie, M. Z. (2025). Comparative Study of Machine Learning Approaches for Detecting Fake News in Arabic Text. *IETI Transactions on Data Analysis and Forecasting (ITDAF)*, *3*(1), 18–31. https://doi.org/10.3991/itdaf.v3i1.53575

Alruily, M. (2021). Classification of Arabic Tweets: A Review. *Electronics*, *10*(10), 1143. https://doi.org/10.3390/electronics10101143

Alturayeif, N. S., Luqman, H. A., & Ahmed, M. A. K. (2022). Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection. *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, 174–184. https://doi.org/10.18653/v1/2022.wanlp-1.16

Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., & Essam, A. (2021). Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity*, *2021*(1), 5516945. https://doi.org/10.1155/2021/5516945

Bangyal, W. H., Qasim, R., Rehman, N. ur, Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z., & Ahmad, J. (2021). Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and Mathematical Methods in Medicine*, *2021*(1), 5514220. https://doi.org/10.1155/2021/5514220

Boukil, S., Biniz, M., Adnani, F. E., Cherrat, L., & Moutaouakkil, A. E. E. (2018). Arabic Text Classification Using Deep Learning Technics. *International Journal of Grid and Distributed Computing*, *11*(9), 103–114. https://doi.org/10.14257/ijgdc.2018.11.9.09

Bsoul, M. A., Qusef, A., & Abu-Soud, S. (2022). Building an Optimal Dataset for Arabic Fake News Detection. *Procedia Computer Science*, *201*, 665–672. https://doi.org/10.1016/j.procs.2022.03.088

Çetiner, H. (2024). Fake News Detection and Classification with Recurrent Neural Network Based Deep Learning Approaches. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *7*(3), 973–993. https://doi.org/10.47495/okufbed.1199738

Davis, A. (2023). *Political communication: An introduction for crisis times*.

Einea, O., Elnagar, A., & Al Debsi, R. (2019). SANAD: Single-label Arabic News Articles Dataset for automatic text categorization. *Data in Brief*, *25*, 104076. https://doi.org/10.1016/j.dib.2019.104076

Fouad, K. M., Sabbeh, S. F., & Medhat, W. (2022). Arabic Fake News Detection Using Deep Learning. *Computers, Materials and Continua*, *71*(2), 3647–3665. https://doi.org/10.32604/cmc.2022.021449

Hadj Ameur, M. S., & Aliane, H. (2021). AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. *Procedia Computer Science*, *189*, 232–241. https://doi.org/10.1016/j.procs.2021.05.086

Haouari, F., Hasanain, M., Suwaileh, R., & Elsayed, T. (2020). ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks. *ArXiv Preprint*.

Harris, S., Hadi, H. J., Ahmad, N., & Alshara, M. A. (2024). Fake News Detection Revisited: An Extensive Review of Theoretical Frameworks, Dataset Assessments, Model Constraints, and Forward-Looking Research Agendas. *Technologies*, *12*(11), 222. https://doi.org/10.3390/technologies12110222

Himdi, H., Weir, G., Assiri, F., & Al-Barhamtoshy, H. (2022). Arabic Fake News Detection Based on Textual Analysis. *Arabian Journal for Science and Engineering*, *47*(8), 10453–10469. https://doi.org/10.1007/s13369-021-06449-y

Hocini, A., & Smaili, K. (2025). Detecting Fake News: Exploring Key Features in Multilingual Arabic Dialect Corpus. *Arabic Language Processing: From Theory to Practice*, 236–248. https://doi.org/10.1007/978-3-031-80438-0_18

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, *3*, 100017. https://doi.org/10.1016/j.nlp.2023.100017

Khalil, A., Jarrah, M., & Aldwairi, M. (2023). Hybrid Neural Network Models for Detecting Fake News Articles. *Human-Centric Intelligent Systems*, *4*(1), 136–146. https://doi.org/10.1007/s44230-023-00055-x

Lelisho, M. E., Pandey, D., Alemu, B. D., Pandey, B. K., & Tareke, S. A. (2022). The Negative Impact of Social Media during COVID-19 Pandemic. *Trends in Psychology*, *31*(1), 123–142. https://doi.org/10.1007/s43076-022-00192-5

Mahlous, A. R., & Al-Laith, A. (2021). Fake News Detection in Arabic Tweets during the COVID-19 Pandemic. *International Journal of Advanced Computer Science and Applications*, *12*(6), 776–785. https://doi.org/10.14569/ijacsa.2021.0120691

Masethe, H. D., Masethe, M. A., Ojo, S. O., Giunchiglia, F., & Owolawi, P. A. (2024). Word Sense Disambiguation for Morphologically Rich Low-Resourced Languages: A Systematic Literature Review and Meta-Analysis. *Information*, *15*(9), 540. https://doi.org/10.3390/info15090540

Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2019). "Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *American Behavioral Scientist*, *65*(2), 180–212. https://doi.org/10.1177/0002764219878224

Nagoudi, E. M., Elmadany, A. R., Abdul-Mageed, M., Alhindi, T., & Cavusoglu, H. (2020). Machine generation and detection of Arabic manipulated and fake news. *ArXiv Preprint*.

Nassif, A. B., Elnagar, A., Elgendy, O., & Afadar, Y. (2022). Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, *34*(18), 16019–16032. https://doi.org/10.1007/s00521-022-07206-4

Omol, E. J. (2024). Organizational digital transformation: from evolution to future trends. *Digital Transformation and Society*, *3*(3), 240–256. https://doi.org/10.1108/dts-08-2023-0061

Qandos, N., Hamad, G., Alharbi, M., Alturki, S., Alharbi, W., & Albelaihi, A. A. (2024). Multiscale cascaded domain-based approach for Arabic fake reviews detection in e-commerce platforms. *Journal of King Saud University - Computer and Information Sciences*, *36*(2), 101926. https://doi.org/10.1016/j.jksuci.2024.101926

Rahmanian, E. (2023). Fake news: a classification proposal and a future research agenda. *Spanish Journal of Marketing - ESIC*, *27*(1), 60–78. https://doi.org/10.1108/sjme-09-2021-0170

Selnes, F. N. (2024). Adolescents' experiences and (re)action towards fake news on social media: perspectives from Norway. *Humanities and Social Sciences Communications*, *11*(1), 1694. https://doi.org/10.1057/s41599-024-04237-1

Shaar, S., Alam, F., Da San Martino, G., Nikolov, A., Zaghouani, W., Nakov, P., & Feldman, A. (2021). Findings of the NLP4IF-2021 Shared Tasks on Fighting the COVID-19 Infodemic and Censorship Detection. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 82–92. https://doi.org/10.18653/v1/2021.nlp4if-1.12

Shishah, W. (2022). JointBert for Detecting Arabic Fake News. *IEEE Access*, *10*, 71951–71960. https://doi.org/10.1109/access.2022.3185083

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science*, *117*, 256–265. https://doi.org/10.1016/j.procs.2017.10.117

Sorour, S. E.-S., & Abdelkader, H. E. (2022). Afnd: Arabic Fake News Detection with an Ensemble Deep Cnn-Lstm Model. *Journal of Theoretical and Applied Information Technology*, *100*(14), 5072–5086.

Taylor, G., Sala, G., Kolak, J., Gerhardstein, P., & Lingwood, J. (2024). Does adult-child co-use during digital media use improve children's learning aged 0–6 years? A systematic review with meta-analysis. *Educational Research Review*, *44*, 100614. https://doi.org/10.1016/j.edurev.2024.100614

Touahri, I., & Mazroui, A. (2024). Survey of machine learning techniques for Arabic fake news detection. *Artificial Intelligence Review*, *57*(6), 157. https://doi.org/10.1007/s10462-024-10778-3

Thaher, T., Saheb, M., Turabieh, H., & Chantar, H. (2021). Intelligent Detection of False Information in Arabic Tweets Utilizing Hybrid Harris Hawks Based Feature Selection and Machine Learning Models. *Symmetry*, *13*(4), 556. https://doi.org/10.3390/sym13040556

Wotaifi, T. A., & Dhannoon, B. N. (2023). An Effective Hybrid Deep Neural Network for Arabic Fake News Detection. *Baghdad Science Journal*, *20*(4), 20. https://doi.org/10.21123/bsj.2023.7427

Yildirim, O., Bakhshi, S., & Can, F. (2024). Prioritized Binary Transformation Method for Efficient Multi-label Classification of Data Streams with Many Labels. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4218–4222. https://doi.org/10.1145/3627673.3679980