Credibility Assessment for Arabic News on the Gaza War: A Hybrid Neural-Symbolic Pipeline

Sanaa Abril, Sihame Mouanid, El Habib Benlahmar, Omar Zahour
Faculty of Science Ben M'Sick, Laboratory of Information Processing and Modeling (LTIM),
Hassan II University, Casablanca, Morocco
{sanaa.abril-etu, sihame.mouanid-etu, elhabib.benlahmar}@etu.univh2c.ma

Abstract

While misinformation has long circulated online, the Gaza conflict has intensified its visibility and spread across news websites, online portals, and social media, complicating the credibility and long-term curation of conflict-related Arabic records, including historical accounts and written testimonies. This work proposes a hybrid framework for Arabic fake news detection that combines interpretable linguistic cues with contextual semantic representations. The approach integrates fuzzy logic-based handcrafted features capturing exaggerated and sensational linguistic patterns, AraBERT contextual embeddings for semantic understanding, and a CNN-based text feature extractor for local textual patterns. These complementary features are combined into a unified representation for downstream classification. Multiple machine learning and deep learning classifiers are evaluated to identify the most effective detection model. The resulting system is deployed as a real-time web browser plugin, enabling users to automatically assess the credibility of Arabic news content during browsing.

Keywords: Arabic fake news detection, News classification, hybrid models, Machine learning, Deep learning, fuzzy logic, browser plugin, misinformation

1. Introduction

Conflict-related Arabic content is produced and consumed at scale across news websites and social media, and it increasingly becomes part of the digital record that researchers, journalists, and archivists consult when documenting events and public discourse. However, contemporary conflict reporting takes place in a high-velocity information environment where emotionally charged framing and coordinated disinformation can blur the line between verified and fabricated claims, affecting public opinion and trust (Institute for Strategic Dialogue, 2025). This makes it difficult to curate reliable collections and to support downstream tasks such as search, retrieval, and analysis of conflict-related content.

This challenge is especially acute in Arabic, where automated credibility assessment is constrained by domain variation and the relative scarcity of standardized evaluation resources (Yousef et al., 2024). Building on commonly used Arabic benchmarks such as AFND (Khalil et al., 2022), we focus the war-related Arabic news collected from online sources and study an interpretable detection pipeline that can be deployed as a practical user tool during browsing.

We design a pipeline tailored to Arabic news that combines semantic and stylistic signals relevant to misinformation detection. The proposed system integrates fuzzy linguistic cues, a CNN-based text feature extractor, and BERT-based contextual embeddings within a gated fusion framework. The fuzzy component uses an extravagant-words membership function to assign a graded exaggeration-oriented score based on the presence of highly emotional or exaggerated language. In addition, a CNN is used to extract local n-gram patterns from the text.

The objective of this work is to combine interpretable linguistic cues and n-gram patterns with deep semantic features to improve the accuracy, robustness, and explainability of Arabic fake news detection. We evaluate multiple machine learning models, deep learning architectures, and large language models to compare performance and select the most effective classifier.

Finally, we deploy the selected model as a real-time browser plugin that provides credibility predictions with a confidence score during browsing, enabling users to assess Arabic news content directly in a practical setting.

2. Related Work

During the Gaza and Israel war, social media became a major vector for real and fake news dissemination. Framing strategies, emotionally charged content, and coordinated disinformation campaigns (Alsharairi et al., 2024; Institute for Strategic Dialogue, 2025) significantly influenced public opinion, amplified polarization, and blurred the distinction between verified information and misleading or fabricated narratives.

Recent work has quantified how conflict narratives and emotions evolve across platforms. At

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 50–59

the public opinion level, digital activism during the Gaza conflict, reflected through social media signals (posts, comments, and hashtags), highlights measurable shifts in sentiment and engagement that shape perceptions and consumer attitudes (Mohammed et al., 2025). Event-driven discourse shifts further reveal clear platform-specific pathways for narrative diffusion during the Gaza war (Antonakaki and Ioannidis, 2025).

In NLP-based fake news detection, several supervised classifiers have been explored, including Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, Passive Aggressive Classifiers, and XGBoost. Jouhar et al. (Jouhar et al., 2024) evaluated these methods on the ISOT dataset, showing that ensemble and boosting techniques achieved superior performance. Deep learning models further improved classification performance by learning contextual representations, with reported accuracy levels of up to 90% (Fouad et al., 2022).

More recently, BERT-based embeddings have enabled models to capture both Modern Standard Arabic and dialectal variations effectively (Awane et al., 2021; Al-yahya et al., 2021). When combined with hyperparameter optimization strategies across architectures such as RNNs, LSTMs, and transformers, these approaches can significantly enhance detection performance and overall model efficiency (Ettaik and Ben Lahmar, 2021).

Bidirectional recurrent models are especially suitable for deceptive-language detection because cues such as exaggeration, negation, and stance depend on both preceding and following context, making BiLSTM a strong candidate for Arabic news classification (Merzah et al., 2025; Turki et al., 2025). Recent hybrid deep learning designs combine CNN and BiLSTM to jointly capture local patterns and sequential context; Merzah et al. (Merzah et al., 2025) report that a multi-channel CNN with dual BiLSTM (FastText embeddings) achieves strong performance on Arabic benchmarks (AFND/ANS) and an English dataset (WELFake).

Beyond single-architecture models, hybrid frameworks integrate complementary techniques to enrich representations. Ennaouri et al. (Ennaouri and Zellou, 2025) proposed a model combining fuzzy logic-based membership functions to capture linguistic exaggeration with BERT embeddings for contextual semantics, achieving significant improvements over standard classifiers such as SVM and Logistic Regression. The fuzzy layer enables the assignment of degrees of truth rather than strict binary decisions (Mirza Saad et al., 2025).

Current literature shows that fake news detection performance depends not only on the classifier itself, but also on the choice of features and the way complementary signals are combined (Hamed et al., 2023). Our work builds on these findings by combining contextual Arabic embeddings, local neural pattern extraction, and fuzzy exaggeration-aware cues within a unified pipeline. The proposed framework is intended to capture complementary aspects of Arabic conflict-related misinformation that prior studies have typically addressed separately or only through partial combinations.

In parallel, mitigation efforts increasingly focus on user tools; for example, *Aletheia* introduces a browser extension that leverages retrieval-augmented generation and large language models to support fake news detection with evidence based explanations and user engagement features (Sallami et al., 2025). Inspired by this direction, we operationalize our best-performing model as a browser plugin that reports real/fake predictions with a confidence score during browsing.

3. Data

We assembled a multi-source Arabic news corpus for binary fake news classification. First, we incorporated the AFND benchmark dataset (Khalil et al., 2022), retaining its original labels. In addition, we collected public posts from *X* (formerly Twitter) using keyword-based queries related to the Gaza war and the broader Palestine–Israel conflict. Within this subset, posts were assigned to the fake-news class when they originated from bots or non-authentic accounts, or when they were explicitly associated with hashtags indicating fake news. To balance the corpus with verifiable information, we also automatically collected recent articles from established Arabic news outlets, including *Al Jazeera* Arabic and *Al Arab* newspaper, which were assigned to the real news class.

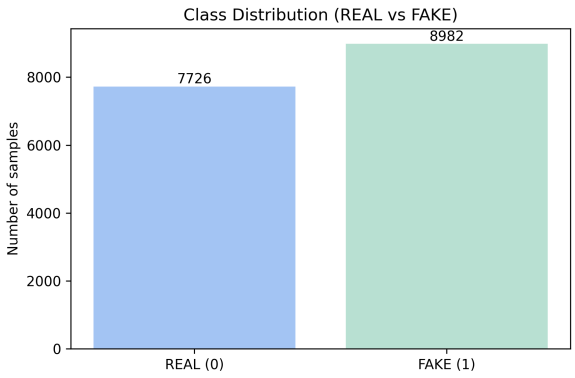

Figure 1: Distribution of fake and real news instances in the collected corpus.

No manual annotation was performed; instead, the

labeling process relied on inherited benchmark labels and source-based assignment rules. This initial data collection process resulted in a total of **16,708** labeled instances. We then applied keyword filtering using conflict terms (such as: غزة، القدس، فلسطين، الضفة، الاحتلال، إسرائيل، حماس، رفح، خان يونس، جباليا، النازحين) in order to retain topically relevant content. The final dataset was restricted to Modern Standard Arabic; Dialectal Arabic was outside the scope of the present study. The distribution between fake and real news classes is illustrated in Figure 1.

4. Proposed Method

The proposed method introduces a hybrid framework for Arabic fake news detection that processes news content through multiple stages to improve classification performance.

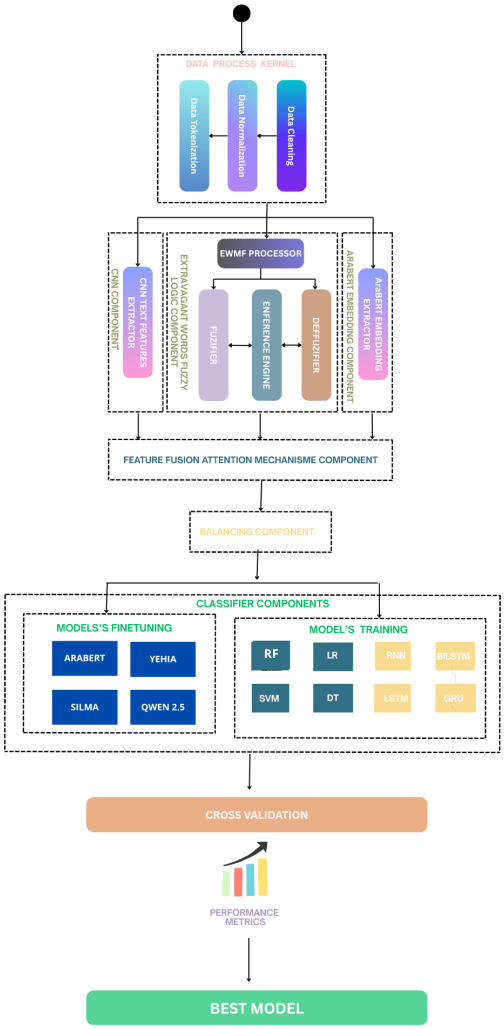

Figure 2: Overview of the proposed hybrid pipeline for Arabic news classification.

As illustrated in Figure 2, the architecture consists of several key components, including a preprocessing module, a fuzzy logic-based linguistic feature extractor, an AraBERT embedding module, a CNN-based text feature extractor, a gated feature fusion component, a data balancing component, and a classification module. Each component contributes to transforming raw textual data into enriched representations that support accurate and robust classification.

4.1. Preprocessing Component

Preprocessing is a necessary step in our pipeline because Arabic news text collected from the web often contains noise and inconsistencies that can degrade downstream classification. We first perform data cleaning by removing duplicates, handling missing fields, and filtering non-Arabic content. We then eliminate web artifacts such as URLs, user mentions, and hashtags, and normalize the text by removing special characters, stop words, and redundant whitespace. Next, Arabic normalization is applied to standardize common character variants (e.g., different forms of *alef* and *yaa*). Spelling, grammatical, and orthographic errors were not explicitly modeled as standalone predictive features; instead, normalization was applied to reduce superficial variation and support more stable downstream representations. Finally, the cleaned text is transformed into normalized input suitable for downstream feature extraction and modeling.

4.2. AraBERT Embedding Component

We represent each article using contextual embeddings from AraBERT. Texts are tokenized with the AraBERT tokenizer and encoded in the standard input format `[CLS] x [SEP]`, then converted to token IDs according to the pre-trained vocabulary. We freeze all AraBERT parameters and extract the final-layer hidden states (768 dimensions per token). A fixed-length sentence embedding is obtained by mean pooling over non-padding tokens using the attention mask, producing one 768-dimensional vector per text.

4.3. Fuzzy Logic Component

While contextual embeddings capture semantics effectively, deception in news is often expressed through stylistic cues such as sensationalism, urgency, and exaggerated framing. To complement AraBERT representations with an interpretable signal, we introduce a fuzzy-logic feature engineering component. We design an Arabic *Extravagant Word Membership Function* (EWMF) that maps the observed frequency of exaggerated expressions in a text to a continuous membership score,

enabling a smooth transition from low to high levels of sensational language. The underlying lexicon consists of commonly used sensational markers in Arabic misinformation (e.g., فاضح، مدمر، مهلك، قاتل، حصري، عاجل، سري، خطير، حساس، مكشوف، مفضوح), and the resulting membership value is used as an additional input feature alongside learned embeddings.

The process of Arabic Extravagant Word Membership Function (EWMF) described in Algorithm 1.

Algorithm 1: Extravagant Word Membership Function (EWMF)

Input: Arabic texts $\mathcal{T} = \{t_1, \ldots, t_n\}$, threshold T_0, scaling factor σ
Output: Extravagant word counts $\mathcal{C}$, fuzzy membership values $\mathcal{M}$
Initialization:
Define extravagant word set $\mathcal{W}_{ext}$;
Define intensifier set $\mathcal{W}_{int}$;
Define clickbait pattern set $\mathcal{P}_{click}$;
$\mathcal{V}_{sens} \leftarrow \mathcal{W}_{ext} \cup \mathcal{W}_{int} \cup \mathcal{P}_{click}$;
Initialize $\mathcal{C} \leftarrow \emptyset$, $\mathcal{M} \leftarrow \emptyset$;
foreach $t_i \in \mathcal{T}$ **do**
  $count_i \leftarrow 0$;
  $\mathcal{W}_i \leftarrow$ ExtractArabicWords(t_i);
  foreach $w \in \mathcal{W}_i$ **do**
    if $w \in \mathcal{V}_{sens}$ **then**
      $count_i \leftarrow count_i + 1$;
  if $count_i = 0$ **then**
    $m_i \leftarrow 0$;
  else if $count_i \geq T_0$ **then**
    $m_i \leftarrow 1$;
  else
    $m_i \leftarrow 1 - e^{-\frac{\sigma \cdot count_i}{T_0}}$;
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{count_i\}$;
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{m_i\}$;
return $(\mathcal{C}, \mathcal{M})$

The membership function outputs 0 when no sensational cues are found ($x = 0$). As the cue count x increases, the membership value increases according to an exponential mapping. In our implementation, the mapping is defined piecewise: $m(x) = 0$ for $x = 0$, $m(x) = 1$ for $x \geq T_0$, and otherwise $m(x) = 1 - e^{-\frac{\sigma x}{T_0}}$. We apply this function to each news item by parsing the text, counting matches against the sensational cue sets, and converting the resulting count into a fuzzy score.

4.4. CNN Extractor Component

To capture local n-gram patterns, we employ a 1D CNN feature extractor. The input sequence is first mapped through a Keras embedding layer initialized with a pre-trained embedding matrix. We then apply three parallel `Conv1D` branches with 100 filters each and kernel sizes $\{3, 4, 5\}$, using ReLU activation. Each branch is followed by Global Max Pooling to obtain a fixed-length vector. The pooled outputs are concatenated into a 300 dimensional representation (100×3), which is used as the CNN feature vector. This vector is later fused with AraBERT embeddings and fuzzy scores in feature fusion component.

4.5. Feature Fusion Component

This component combines three complementary feature types into a single representation: (i) the fuzzy membership score produced by the Extravagant Word Membership Function (EWMF), (ii) AraBERT embeddings capturing contextual semantics, and (iii) CNN features capturing local n-gram patterns. The goal is to integrate interpretable stylistic cues with learned semantic and surface-level representations.

In our implementation, we use the EWMF score as a gating signal that scales the AraBERT and CNN feature vectors before concatenation. For each sample i, let $f_i \in [0, 1]$ denote the fuzzy score, $\mathbf{b}_i \in R^{d_B}$ the AraBERT embedding, and $\mathbf{c}_i \in R^{d_C}$ the CNN feature vector. The gated representations are computed as:

$$\tilde{\mathbf{b}}_i = f_i \cdot \mathbf{b}_i, \qquad \tilde{\mathbf{c}}_i = f_i \cdot \mathbf{c}_i.$$

The final fused representation is then formed by concatenation:

$$\mathbf{z}_i = [\, f_i;\ \tilde{\mathbf{b}}_i;\ \tilde{\mathbf{c}}_i \,].$$

This fused vector is used as input to the downstream classifiers.

4.6. Classifier Component

We evaluate a range of machine learning and deep learning classifiers, as well as selected large language models, for distinguishing real versus fake news in the Palestine/Israel war setting. The tested models include traditional classifiers (e.g., Random Forest, SVM, Logistic Regression, and Decision Tree), deep neural models (RNN, LSTM, BiLSTM, and GRU), and LLM-based baselines (Qwen, AraBERT, Silma, and Yehia).

For a robust performance estimate, we apply 5-fold cross-validation, where the dataset is split into five folds and models are trained and validated iteratively. For deep learning models and LLM-based runs, we additionally use early stopping based on validation performance. Model performance is reported using Accuracy, Precision, Recall, and F1-score.

5. Results and Discussion

We report model performance using standard classification metrics derived from the confusion matrix.

Accuracy measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Recall (Sensitivity) measures the proportion of actual positive instances correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Precision measures the proportion of predicted positive instances that are correct:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

F1-score is the mean of Precision and Recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Figures 3-4-5 report an ablation comparison of models trained with and without the proposed EWMF+CNN components. Across machine learning and deep learning families, incorporating these components yields consistent improvements in Accuracy, Precision, Recall, and F1, indicating that exaggerated language cues and local n-gram patterns provide complementary signals to contextual embeddings. The strongest deep learning results are obtained with BiLSTM (Fig. 3), which is in line with prior evidence that bidirectional recurrent modeling benefits deception detection by leveraging both preceding and following context (Merzah et al., 2025; Turki et al., 2025). For classical machine learning models (Fig. 4), the gains are more pronounced for weaker baselines, suggesting that feature enrichment is particularly beneficial when the classifier cannot implicitly learn such stylistic cues. This observation is consistent with hybrid approaches that combine fuzzy linguistic indicators with contextual representations (Ennaouri and Zellou, 2025). Finally, Fig. 5 shows that adapting large language models to the proposed pipeline substantially improves performance, while the Arabic-specialized AraBERT model remains the most reliable among the evaluated LLMs, supporting findings that domain and language aligned encoders are critical for Arabic credibility assessment (Fouad et al., 2022; Al-yahya et al., 2021; Papageorgiou et al., 2024).

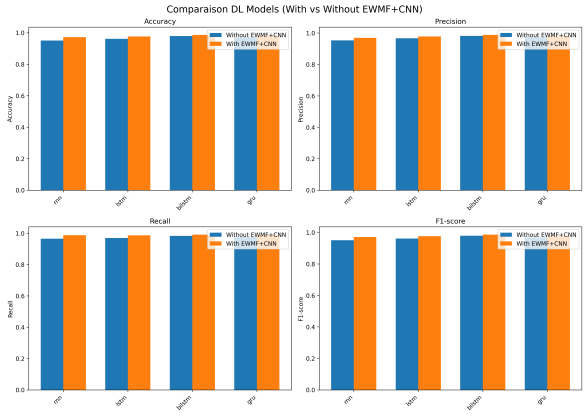

Figure 3: Performance comparison of deep learning models with and without the EWMF and CNN components.

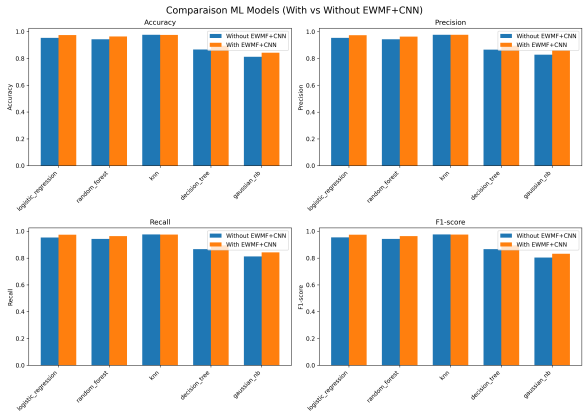

Figure 4: Performance comparison of machine learning models with and without the EWMF and CNN components.

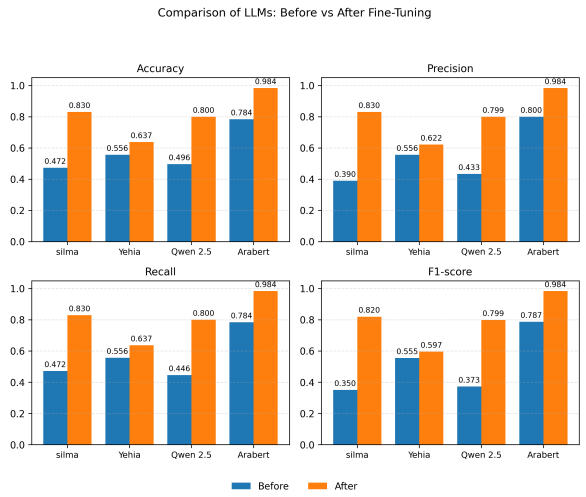

Figure 5: Performance comparison of large language models with and without the EWMF and CNN components.

Model	Acc.	Prec.	Rec.	F1
Logistic Regression	0.9742	0.9742	0.9742	0.9742
Random Forest	0.9634	0.9637	0.9634	0.9632
KNN	0.9756	0.9765	0.9756	0.9754
Decision Tree	0.8948	0.8950	0.8948	0.8949
Gaussian NB	0.8429	0.8620	0.8429	0.8319
RNN	0.9716	0.9693	0.9872	0.9715
LSTM	0.9767	0.9777	0.9863	0.9767
BiLSTM	**0.9867**	0.9881	0.9912	0.9867
GRU	0.9807	0.9816	0.9885	0.9807
Silma	0.8300	0.8300	0.8300	0.8200
Yehia	0.6375	0.6220	0.6375	0.5965
Qwen 2.5	0.8000	0.7991	0.8000	0.7995
AraBERT	0.9838	0.9839	0.9838	0.9838

Table 1: Performance comparison of models using the proposed pipeline.

To enable a fair comparison of Palestinian fake news detection approaches, we evaluate all models on the same dataset under a unified preprocessing and evaluation protocol. This is important because prior work often relies on different datasets tied to specific events or platforms, and reported metrics are not always consistent across studies, which limits direct comparability. Therefore, Table 1 reports our baseline results across three families of methods: (i) conventional machine learning classifiers, (ii) deep learning sequence models, and (iii) LLM baselines.

Overall, the results indicate that deep learning models achieve the highest performance on this dataset, with BiLSTM obtaining the best scores among the evaluated architectures. The observed hierarchy can be interpreted in terms of (1) how each model family represents context in Arabic news, and (2) whether it benefits from complementary cues beyond semantics, such as exaggeration-oriented fuzzy scores and local n-gram patterns.

Conventional ML baselines. Logistic Regression and KNN provide strong results (both around 97% accuracy/F1), suggesting that the representation used in our pipeline yields a feature space where the classes are largely separable with relatively simple decision boundaries. However, ML models do not explicitly model sequential dependencies; consequently, they may miss long-range context signals (e.g., stance, negation, or discourse-level cues) that are relevant to deception in news narratives.

Sequence DL baselines. Recurrent architectures consistently outperform conventional ML models, and the BiLSTM achieves the best overall performance. This trend is aligned with recent Arabic fake news studies that report gains from BiLSTM-based designs when contextual representations are combined with additional textual cues (e.g., hybrid contextual features coupled with BiLSTM classifiers) (Turki et al., 2025). Similarly, hybrid CNN–BiLSTM designs have been shown to be effective because CNN captures local patterns while BiLSTM models sequential context, leading to stronger performance than using either family alone (Merzah et al., 2025). Our results support this line of evidence: once local pattern cues and exaggeration-oriented fuzzy scores are integrated alongside contextual semantics, BiLSTM is the most effective downstream classifier on this dataset.

LLM baselines. General-purpose LLM baselines (e.g., Silma, Yehia, Qwen 2.5) underperform supervised Arabic models in our controlled setting, whereas AraBERT remains competitive with the best recurrent architectures. This gap is consistent with recent evidence that LLM outputs may be less reliable for politically sensitive or high-stakes judgments due to susceptibility to bias and inconsistent reasoning, especially when the task requires stable label decisions rather than open-ended generation (Gubelmann and Karray, 2025; Barman et al., 2024). Moreover, comparative studies in fake-news detection frequently report that encoder-style transformers (BERT-like) remain strong for supervised classification pipelines, while GPT-like models often require careful adaptation to match them (Papageorgiou et al., 2024; Raza et al., 2025). Overall, the strong AraBERT baseline and weaker general LLM baselines suggest that performance here depends primarily on domain-aligned representations and supervised training, rather than generative capacity alone.

The strong scores observed across model families can be partly explained by the focused nature of the dataset and by the fused representation used in our pipeline. In this setting, fake and real news differ through a combination of semantic context, local lexical patterns, and exaggeration-oriented stylistic cues, which increases class separability once these signals are jointly encoded. This also helps explain why conventional ML baselines remain highly competitive on the learned feature space. Taken together, these results indicate that (i) contextual modeling is critical for conflict-related Arabic news classification, and (ii) integrating complementary signals, including local n-gram cues and exaggeration-oriented fuzzy scores, provides additional discriminative power that is most effectively exploited by bidirectional sequence models.

5.1. Deployment as a Browser Plugin

To demonstrate practical usability beyond offline evaluation, we deployed the best-performing model (BiLSTM) as a real-time browser plugin for Arabic news credibility assessment. During browsing, the extension extracts the article text and sends it to a backend analysis endpoint,

which applies the same preprocessing and inference pipeline used in our experiments and returns a predicted label with an associated confidence score. To improve transparency, we additionally provide token-level rationales by highlighting influential spans in the article, computed using Integrated Gradients and refined with a deletion test that retains only spans whose removal produces a measurable confidence drop.

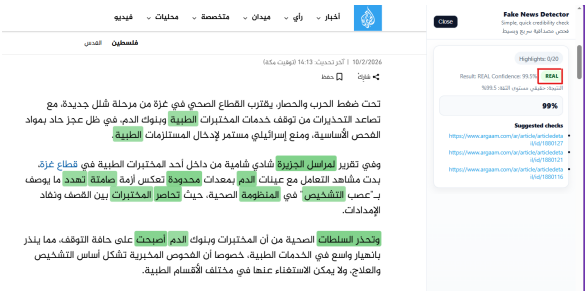

Figure 6: Screenshot of the deployed browser plugin. The BiLSTM model predicts the credibility label and returns a confidence score for the visited Arabic news content.

6. Limitations and Future Work

The strong scores reported in Table 1 should be interpreted with caution. Because the dataset is focused on conflict-related Arabic news, class separability may be partly supported by strong lexical, semantic, and stylistic regularities within this domain. Although exact duplicate entries were removed during preprocessing, near-duplicate or lightly rewritten content may still remain, especially when similar narratives circulate across multiple channels. In addition, our source-based labeling strategy may introduce bias, since real news instances were primarily collected from established news outlets, whereas part of the fake news class originated from benchmark misinformation data and selected social media posts. As a result, some of the predictive signal may reflect source-related or topic-specific regularities in addition to misinformation-specific cues. These factors do not invalidate the reported results, but they suggest that future work should consider stricter deduplication, source-balanced sampling, and broader cross-domain evaluation.

A second limitation concerns the scope of the proposed pipeline, which is developed and evaluated primarily for Arabic news text. Extending the approach to multilingual settings would require additional language-specific preprocessing, resources, and evaluation across more diverse benchmarks.

Third, although our data target conflict-related news, the deployed browser plugin is currently optimized for relatively clean article pages. Its performance may degrade on social media content, where text is shorter, more conversational, noisier, and often mixed with screenshots, images, or embedded videos.

In addition, the system produces a model-based credibility label and confidence score, which should not be interpreted as definitive factual verification. This is particularly important in cases where misinformation is subtle, partially true, or dependent on external context. A useful extension would therefore be to integrate retrieval-augmented components that gather related coverage from trusted sources and present supporting evidence alongside predictions.

Finally, although the framework incorporates interpretable linguistic cues such as exaggerated or sensational expressions, these cues are not exhaustive and may evolve over time, especially in rapidly changing conflict discourse. Continuous lexicon updates, monitoring of domain shift, and periodic re-evaluation on newly collected data would be necessary to preserve reliability.

7. Conclusion

This paper addresses credibility assessment for Arabic conflict-related news, a setting where manipulated discourse and emotionally charged framing can distort public understanding and the preservation of Palestine/Israel narratives as language resources. We proposed an interpretable hybrid pipeline that combines contextual semantics from AraBERT embeddings with complementary signals: (i) an Extravagant Word Membership Function (EWMF) that quantifies sensational and clickbait oriented cues via fuzzy scoring, and (ii) a CNN-based extractor that captures local n-gram patterns. Across a unified evaluation protocol on Palestine/Israel war related Arabic news, deep sequence models achieved the strongest results, with BiLSTM performing best (Accuracy/F1 = 0.9867, Recall = 0.9912), while AraBERT remained highly competitive (Accuracy/F1 = 0.9838). Conventional machine learning baselines were strong but consistently below the top recurrent models, and general purpose LLM baselines lagged substantially, suggesting that domain aligned representations and supervised training are crucial for reliable classification in this sensitive context. To demonstrate practical utility beyond offline experiments, we operationalized the best performing model as a browser plugin that outputs a real/fake prediction with a confidence score during browsing. Limitations include restricted language and domain coverage, limited validation on social media discourse, and the absence of evidence verification.

Future work should study multilingual generaliza-

tion and robustness on social media discourse, incorporate retrieval-augmented evidence to support model predictions, and address temporal domain shift via continuous lexicon maintenance and regular re-training/evaluation on newly collected data.

7.1. Ethics and Compliance:

We limit collection to publicly available content, honor platform and site policies, and remove items upon request. For X/Twitter, we follow the developer policy on redistribution. For news sites, we use feeds, public datasets, or permissions and respect Terms of Service.

References

Maha Al-yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and A. Essam. 2021. Arabic fake news detection: Comparative study of neural networks and transformer-based approaches. *Complexity*, 2021.

Ahmad Alsharairi, Hana Abdul-Rahman Al-Souob, Mohammad Farouq AlQadi, and Sora Mohammad Shatnawi. 2024. Social media communication and framing of the gaza conflict: Impact on public opinion. ResearchGate.

Despoina Antonakaki and Sotiris Ioannidis. 2025. Cross-platform digital discourse analysis of the israel-hamas conflict: Sentiment, topics, and event dynamics.

Widad Awane, El Habib Ben Lahmar, and Ayoub El Falaki. 2021. Hate speech in the arab electronic press and social networks. *Revue d'Intelligence Artificielle*, 35(6):457.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545.

Mohammed Ennaouri and Ahmed Zellou. 2025. Enhancing fake review detection using linguistic exaggeration, bert embeddings, and fuzzy logic. *IEEE Access*, 13:135957–135968.

Noureddine Ettaik and El Habib Ben Lahmar. 2021. Hyperparameter optimisation in nlp architectures. In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, pages 55–62, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Khaled M. Fouad, Sahar F. Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2):3647–3665.

Reto Gubelmann and Ghassen Karray. 2025. Assessing reliability and political bias in LLMs' judgements of formal and material inferences with partisan conclusions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30005–30031, Vienna, Austria. Association for Computational Linguistics.

Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. 2023. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, 9(10):e20382.

Institute for Strategic Dialogue. 2025. *Conflict Amplified: Disinformation and Hate in the Israel-Hamas War*. Institute for Strategic Dialogue.

Jumana Jouhar, Anju Pratap, Neharin Tijo, and Meenakshi Mony. 2024. Fake news detection using python and machine learning. *Procedia Computer Science*, 233:763–771. 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024).

Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, and Manar Jaradat. 2022. Afnd: Arabic fake news dataset for the detection and classification of articles credibility. *Data in Brief*, 42:108141.

Baqer M. Merzah, Jafar Razmara, and Zolfaghar Salmanian. 2025. Hybrid deep learning models for fake news detection: Case study on arabic and english languages. *Frontiers in Big Data*.

Mahmud Mirza Saad, Shahzad Muhammad Khuram, Ali Syed Imran, Jabeen Farzana, and Moetesum Momina. 2025. Fond: Fuzzy-based optimized fake-news detection using big-bird and longformer. *SSRN Electronic Journal*.

Ashraf Bany Mohammed, Manaf Al-Okaily, Dhia Qasim, Shafique Ur Rehman, and Latifa Abdalla. 2025. Digital activism and public opinion: understanding the role of social media during the gaza conflict. *Journal of Islamic Marketing*, 16(11):3366–3393.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data.

H. Sallami, E. Aïmeur, et al. 2025. Aletheia: Detect, discuss, and stay informed on fake news. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Demonstrations Track*.

Hussain Mohammed Turki, Essam Al Daoud, Ghassan Samara, Raed Alazaidah, Mais Haj Qasem, Mohammad Aljaidi, Suhaila Abuowaida, and Nawaf Alshdaifat. 2025. Arabic fake news detection using hybrid contextual features. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(1):836–845.

Mohammed Abbas Yousef, Abeer ElKorany, and Hanaa Bayomi. 2024. Fake-news detection: a survey of evaluation arabic datasets. *Social Network Analysis and Mining*, 14:225.

A. Data

A.1. Detailed Preprocessing Steps

The preprocessing pipeline applied to the collected corpus consisted of the following steps:

1. Removal of duplicate and incomplete entries.
2. Retention of Arabic texts only through language and script filtering.
3. Removal of web-specific artifacts, including URLs, user mentions, and hashtags.
4. Removal of non-Arabic and non-linguistic characters.
5. Diacritic removal and whitespace normalization.
6. Arabic character normalization to reduce orthographic variation.
7. Removal of very short texts with limited linguistic content.

A.2. Data availability

The dataset is anonymously available at: `https://huggingface.co/datasets/Anonymous-123-de/credibility-assessment`

B. Experimental Configuration

B.1. General Settings

Setting	Value
Train/test split	80/20, stratified
Cross-validation	5-fold StratifiedKFold
Random state	42
Feature scaling	StandardScaler

Table 2: General experimental settings.

B.2. AraBERT Settings

Parameter	Value
Pretrained model	`aubmindlab/bert-base-arabertv02`
Maximum sequence length	128
Batch size	8

Table 3: AraBERT feature extraction settings.

B.3. CNN Settings

Parameter	Value
Embedding dimension	128
Filter sizes	3, 4, 5
Number of filters	100 per filter size
Dropout	0.3
Maximum sequence length	256
Batch size	16

Table 4: CNN feature extractor settings.

B.4. Classical Machine Learning Settings

Model	Configuration
Logistic Regression	max_iter=2000
Random Forest	n_estimators=100
KNN	n_neighbors=5
Decision Tree	default settings
Gaussian Naive Bayes	default settings

Table 5: Classical machine learning model settings.

B.5. Neural Model Settings

Parameter	Value
Models	RNN, LSTM, BiLSTM, GRU
Number of layers	2
Dropout	0.3
Optimizer	Adam
Learning rate	0.001
Weight decay	10^{-5}
Loss function	CrossEntropyLoss
Maximum epochs	50
Early stopping patience	5

Table 6: Neural model training settings.

B.6. Notebook availability

The implementation is anonymously available at: https://github.com/Annonymous975-del/Credibility-Assessment

C. Plugin Code availibitlity

https://github.com/Sanaa99-ab/-fake-news-detection-plugin