LinguaBERT-Arabic: Using AraBERT and Simple Style Features for Arabic Fake News Detection

Ajab Sanad Aldouseri , Yasir Alalawi

447540089 , 447540166|

Final NLP Research Project Paper - May 2026

Abstract

Arabic fake-news detection is a difficult NLP task because Arabic has rich morphology, dialectal variation, spelling differences, and fewer strong resources than English. In this project, we built and evaluated LinguaBERT-Arabic, a model that combines AraBERT with five simple stylometric features: sensationalism, type-token ratio, numerical density, elongation rate, and a negative-sentiment proxy. We used the official Arabic Fake News Dataset (AFND), removed duplicate article texts, excluded undecided sources, and used a source-disjoint split so that sources in the test set were not seen during training.

Our results show that this stricter evaluation is much harder than normal article-random splitting. The best overall model in our saved run is NB with TF-IDF, with Macro F1 = 0.4940. LinguaBERT-Arabic gets Macro F1 = 0.4332 and improves over vanilla AraBERT, which gets Macro F1 = 0.4055. Therefore, our main conclusion is careful: adding simple style features helped the transformer baseline, but it did not beat every classical model.

Keywords: Arabic NLP; Fake News Detection; AraBERT; Stylometric Features; AFND; Source-Disjoint Evaluation

1. Introduction

Arabic misinformation is an important problem because many people read and share Arabic news online every day. A false or misleading article can affect public health, politics, social trust, and crisis situations. Arabic text is also challenging for NLP because words can have many forms, dialects are common, and spelling is not always standardized.

In our project, we worked on binary Arabic fake-news/source-credibility classification. The model receives an Arabic news article and predicts whether it belongs to a credible or not-credible source. We used AFND because it is large and public. However, AFND labels are assigned at the source level, not by checking every article separately. This means that if the same source appears in training and testing, the model may learn the source's style instead of learning general fake-news patterns.

Because of this issue, we used a source-disjoint split. This means that all articles from one source go into only one split: train, validation, or test. This makes the task harder, but we think it gives a more honest test of generalization.

1.1 Objectives

- Use the official full AFND release for a binary credible/not-credible task.
- Apply a source-disjoint split to avoid source overlap between train, validation, and test.
- Build LinguaBERT-Arabic by adding five stylometric features to AraBERT.
- Compare our model with LR, SVM, NB, and vanilla AraBERT.
- Explain the results honestly, including where classical models perform better.

1.2 Contributions

- We prepared a full-AFND binary experiment with 373,369 deduplicated articles from 103 retained sources.
- We built a simple hybrid model that combines AraBERT with interpretable writing-style features.
- We showed that LinguaBERT-Arabic improves over vanilla AraBERT under source-disjoint evaluation.
- We also showed that NB with TF-IDF is still the strongest overall model in our saved run, which is important to report honestly.

2. Related Work

Khalil et al. introduced AFND, a large Arabic news credibility dataset. AFND is useful because of its size, but its labels come from source credibility. This creates a possible shortcut because a model may memorize source-specific words or templates.

AraBERT is a strong Arabic language model and has been used in many Arabic classification tasks. Previous AFND studies report strong results using deep learning and transformer models. Alkudah et al. report strong performance with hybrid AraBERT + BiLSTM. Al-Taie compares several machine-learning models and reports XGBoost as the best model in that study.

Other work studies Arabic manipulated news and AI-generated fake news. Nagoudi et al. work on Arabic manipulated/fake news, and Himdi et al. include human and AI-generated fake-news data. These studies helped us understand the field, but our project focuses on a stricter source-disjoint AFND evaluation.

2.1 Previous-Work Comparison

Study	Task / Domain	Dataset	Method	Performance	Advantages	Limitations
Khalil et al. (2022)	Arabic source credibility dataset	AFND; Arabic news	Dataset release	Not a model leaderboard	Large Arabic dataset	Source-level labels may cause source memorization
Nagoudi et	Arabic	AraFake	Arabic	Not	Focuses	Different task and

al. (2020)	manipulated/fake news	News	neural/transformer detection	directly comparable	on manipulated content	dataset
Alkudah et al. (2025)	Arabic fake-news detection	AFND	ML/DL including AraBERT + BiLSTM	Average F1 = 98.0%	Strong model comparison	Strict source-disjoint testing is not main focus
Al-Taie (2025)	Arabic fake-news detection	AFND	XGBoost, DNN, LSTM, classical ML	XGBoost Acc = 72.86%; F1 = 0.76	Compares several methods	Protocol differs from our setup
Himdi et al. (2025)	Human and AI-generated fake news	Arabic fake-news data	Dataset development and detection	Not directly comparable	Includes AI-generated fake news	Different data construction
Our project	Arabic source-credibility classification	Full binary AFND	TF-IDF, AraBERT, LinguaBERT-Arabic	NB F1 = 0.4940; LinguaBERT F1 = 0.4332	Strict source-disjoint split	Single-seed transformer training

2.2 Research Gap

The main gap we focused on is evaluation realism. If train and test data share the same news sources, a model may perform well for the wrong reason. We wanted to test a stricter setup where test sources are unseen. We also wanted to check if simple, interpretable style features could help AraBERT in this harder setting.

3. Methodology

Our workflow had five main steps: loading AFND, splitting sources, preprocessing Arabic text, extracting stylometric features, and training/evaluating models.

3.1 Dataset

The official AFND release has 606,912 articles from 134 source folders. Since our task is binary, we removed undecided sources and kept credible and not-credible sources. Before duplicate removal, the binary subset had 374,543 articles. After removing 1,174 exact duplicate article texts, we used 373,369 articles: 206,908 Real and 166,461 Fake.

3.2 Source-Disjoint Split

We split the data by source instead of by article. The final split used seed 1236. Training had 258,122 articles from 71 sources. Validation had 56,471 articles from 16 sources. Test had 58,776 articles from 16 sources. No source appears in more than one split.

3.3 Preprocessing

We cleaned the Arabic text by removing URLs and mentions, keeping Arabic characters, normalizing Alef, normalizing Teh marbuta, removing diacritics, reducing repeated characters, and removing very short texts. We applied the same preprocessing to all models so the comparison would be fair.

3.4 Stylometric Features

We extracted five simple features from each article. Sensationalism counts urgent or shocking words. TTR measures lexical diversity. Numerical density measures digit use. Elongation rate checks repeated-character patterns. Negative-sentiment proxy counts negation and negative words. These features are easy to interpret, which was one reason we selected them.

3.5 Model Architecture

Vanilla AraBERT uses the AraBERT encoder and a normal classification head. Our LinguaBERT-Arabic model uses the same encoder, but it also adds the five stylometric features to the [CLS] representation before classification. This gives the model direct access to global writing-style information.

3.6 Experimental Setup

We trained the transformer models with AdamW, learning rate 2e-5, batch size 16, maximum sequence length 256, up to 5 epochs, and early stopping on validation Macro F1. The saved run used an NVIDIA RTX 4070 SUPER GPU. We reported Accuracy, Precision, Recall, Macro F1, AUC, and McNemar's test.

4. Results

The results are lower than many AFND results in the literature because our split is stricter. Test sources are unseen, so the model cannot rely on source memorization as easily.

4.1 Feature and Corpus Results

The dataset split was fairly balanced. The train set had 55.2% Real articles, the validation set had 56.6% Real articles, and the test set had 55.4% Real articles. In the stylometric

features, TTR and the negative-sentiment proxy were the most useful. Sensationalism and numerical density were weak, and elongation was almost constant after preprocessing.

4.2 Main Results

NB (TF-IDF) was the strongest overall model, with Accuracy = 0.5057 and Macro F1 = 0.4940. LinguaBERT-Arabic got Accuracy = 0.4337 and Macro F1 = 0.4332. Vanilla AraBERT got Accuracy = 0.4120 and Macro F1 = 0.4055 in the final comparison table. So our model improved over vanilla AraBERT by 0.0277 Macro F1, but it did not beat NB.

Model	Accuracy	Precision	Recall	Macro F1	AUC
LR (TF-IDF)	0.4424	0.4439	0.4433	0.4415	0.4184
SVM (TF-IDF)	0.4208	0.4258	0.4259	0.4208	0.3907
NB (TF-IDF)	0.5057	0.4949	0.4951	0.4940	0.5055
Vanilla AraBERT	0.4120	0.4210	0.4281	0.4055	0.4166
LinguaBERT-Arabic	0.4337	0.4408	0.4417	0.4332	0.4219

4.3 Error Analysis

LinguaBERT-Arabic made 33,286 errors on the 58,776 test articles, which is a 56.63% error rate. Vanilla AraBERT made 34,563 errors, or 58.80%. We used rule-based categories to understand possible error patterns. Since these categories were based on keywords, we treat them as exploratory.

4.4 Ablation and Topic Results

For preprocessing ablation, we used LR + TF-IDF because it is faster than retraining transformers many times. Removing Alef normalization hurt performance the most. Topic-level results showed that LinguaBERT-Arabic improved over vanilla AraBERT in all five topic groups, but the gains were modest.

5. Discussion

The main result we can safely claim is that stylometric features helped AraBERT. The improvement over vanilla AraBERT is not huge, but it is consistent in the final results. This makes sense because AraBERT's [CLS] vector may not keep all article-level style information, especially when long articles are truncated.

At the same time, we should not overstate our model. NB with TF-IDF performed best overall. This shows that simpler models can still be strong, especially when the test distribution is different from the training distribution. For our project, the fair conclusion is that LinguaBERT-Arabic is better than vanilla AraBERT, not that it is the best model overall.

The low scores are also important. They show that source-disjoint AFND evaluation is hard. We think this is useful because it gives a more realistic view than only reporting high scores from easier splits.

6. Limitations and Future Work

Our project has several limitations. First, AFND labels are source-level labels, so they are not perfect article-level truth labels. Second, we trained the transformer models with one random seed only. Third, we tested only on AFND and did not check another dataset. Fourth, some stylometric features were weak. Fifth, our error analysis used rules, not full manual annotation.

In future work, we would run multiple seeds, compare random article splits directly with source-disjoint splits, try dialect-aware models such as CAMeL-BERT, improve the stylometric lexicons, and test on another Arabic fake-news dataset.

7. Conclusion

In this project, we built and evaluated LinguaBERT-Arabic for Arabic fake-news/source-credibility classification. We used the official AFND dataset and a strict source-disjoint split. Our results showed that NB with TF-IDF was the best overall model with Macro F1 = 0.4940. LinguaBERT-Arabic got Macro F1 = 0.4332 and improved over vanilla AraBERT, which got Macro F1 = 0.4055. Our final conclusion is that adding simple stylometric features helps AraBERT, but classical baselines must still be reported because they can perform better in this setting.

References

[1] Khalil, A., Jarrah, M., Aldwairi, M., & Jaradat, M. (2022). AFND: Arabic fake news dataset for the detection and classification of articles credibility. Data in Brief, 42, 108141. https://doi.org/10.1016/j.dib.2022.108141

[2] Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. OSACT4 Workshop, LREC 2020.

[3] Alkudah, N. M., Idris, N. B., & Abushariah, M. A. M. (2025). Fake news detection in Arabic media: comparative analysis of machine learning and deep learning algorithms using the Arabic Fake News Dataset. PeerJ Computer Science, 11, e3272. https://doi.org/10.7717/peerj-cs.3272

[4] 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, 3(1), 18-31. https://doi.org/10.3991/itdaf.v3i1.53575

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019.

[6] Nagoudi, E. M. B., Elmadany, A., Abdul-Mageed, M., Alhindi, T., & Cavusoglu, H. (2020). Machine generation and detection of Arabic manipulated and fake news. arXiv:2011.03092.

[7] Himdi, H. T., Zamzami, N., Najar, F., Alrehaili, M., & Bouguila, N. (2025). Arabic fake news dataset development: Humans and AI-generated contributions. IEEE Access, 13, 62234-62253. https://doi.org/10.1109/ACCESS.2025.3556376

[8] Abbas Yousef, M., ElKorany, A., & Bayomi, H. (2024). Fake-news detection: a survey of evaluation Arabic datasets. Social Network Analysis and Mining, 14, 225. https://doi.org/10.1007/s13278-024-01370-2

[9] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. ICLR 2019.

[10] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. WANLP 2021.

[11] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153-157.