Kashif-AI at AbjadGenEval Shared Task: A Transformer-based Approach for Arabic AI-Generated Text Detection

Fatimah Emad Eldin
Cairo University
12422024441586@pg.cu.edu.eg

Abstract

As Large Language Models (LLMs) become increasingly proficient at generating human-like text, distinguishing between human-written and machine-generated content has become a critical challenge for information integrity. This paper presents Kashif-AI, a system developed for the AbjadGenEval Task 1: AI-Generated Arabic Text Detection. The approach leverages fine-tuned Arabic Pre-trained Language Models (PLMs), specifically MARBERT and CAMeLBERT, to classify news articles. A rigorous ablation study was conducted to evaluate the impact of data augmentation, comparing models trained on the official shared task data against those trained on a combined corpus of over 47,000 samples. While near-perfect performance was observed during validation, the blind test set evaluation revealed a significant generalization gap. Contrary to expectations, data augmentation resulted in performance degradation due to domain shifts. The best-performing configuration, which utilized CAMeLBERT-Mix trained on the original dataset, achieved an F1-score of 66.29% and an Accuracy of 70.5% on the blind test set.

1 Introduction

The rapid evolution of Generative AI has lowered the barrier for creating high-quality, fluent text in various languages, including Arabic. While beneficial for productivity, this capability poses risks regarding misinformation, academic dishonesty, and the dilution of verified news sources (Solaiman et al., 2019; Alshammari and Elleithy, 2024). Consequently, the detection of AI-generated text has emerged as a vital research area for the Arabic NLP community, often categorized under broader efforts to maintain information integrity (Alshammari and Elleithy, 2024; Al-Shaibani and Ahmed, 2025; Jawahar et al., 2020). Comprehensive surveys have highlighted the evolving threat models and the necessity of robust detection mechanisms in the face of increasingly sophisticated LLMs (Crothers et al., 2022). This work addresses AbjadGenEval Task 1, which focuses on the binary classification (Human vs. Machine) of Arabic news articles (Ezzini et al., 2026).

The proposed system, Kashif-AI, treats this problem as a supervised sequence classification task. The efficacy of distinct Transformer architectures is explored, specifically MARBERT (Abdul-Mageed et al., 2021), which is optimized for both dialectal and MSA coverage, and CAMeLBERT-Mix (Abdul-Mageed et al., 2021), a model pre-trained on a diverse mix of Arabic texts. The contributions of this study include a comparative analysis of state-of-the-art Arabic PLMs under identical training conditions, an ablation study demonstrating the counter-intuitive impact of data augmentation where merging a large external corpus reduced generalization, and a detailed error analysis highlighting the discrepancy between validation results and blind test results. To ensure reproducibility and facilitate future research, open access is provided to all experimental code and fine-tuned models via GitHub[1] and Hugging Face[2].

2 Background and Related Work

The proliferation of Large Language Models (LLMs) has necessitated robust detection mechanisms to maintain information integrity. Early detection methodologies relied heavily on statistical artifacts, such as perplexity and burstiness features, or zero-shot methods utilizing the probability curvature of the generating model (Gehrmann et al., 2019; Mitchell et al., 2023). However, as generation quality has improved, these statistical signals have become less reliable.

[1] https://github.com/astral-fate/Kashif-AI/

[2] https://huggingface.co/collections/FatimahEmadEldin/kashif-ai

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 483–488

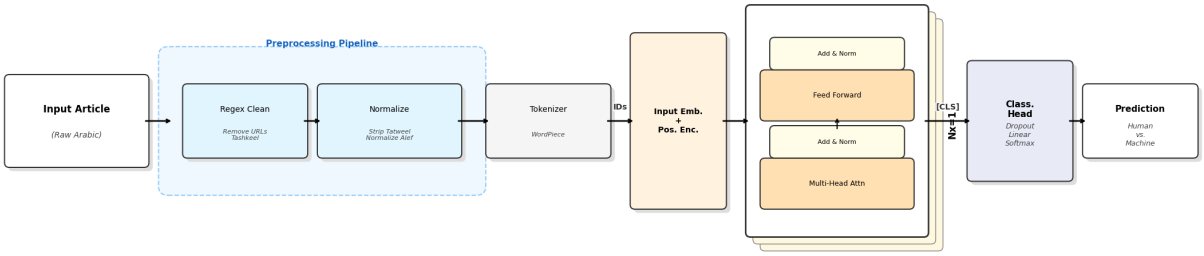

Figure 1: The system overview of Kashif-AI illustrating the data flow from input text to final classification.

Consequently, the paradigm has shifted towards supervised fine-tuning of Pre-trained Language Models (PLMs). Numerous studies have demonstrated that Transformer-based classifiers, such as BERT and RoBERTa, consistently achieve state-of-the-art performance in distinguishing human from machine text by learning high-dimensional semantic and stylistic representations (Devlin et al., 2019; Liu et al., 2020; Zellers et al., 2019).

In the context of Arabic NLP, this Transformer-based approach is increasingly critical due to the language's morphological complexity. Recent work has focused on domain-specific PLMs such as AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), and CAMeLBERT (Inoue et al., 2021). These models, pre-trained on massive Arabic corpora, have set new benchmarks for various classification tasks, including dialect identification and sentiment analysis. This study builds upon this established foundation by applying these specific Arabic PLMs (MARBERT and CAMeLBERT) to the novel task of AI-generated text detection.

2.1 Task Setup

The task is defined as a binary classification problem. Given an input text sequence S, the system must predict a label $y \in \{0, 1\}$, where 0 represents human-authored text and 1 represents machine-generated text (Ezzini et al., 2026; Abudalfa et al., 2025; Lamsiyah et al., 2025). The domain focuses on news articles, requiring the model to discern subtle stylistic artifacts inherent to neural generation rather than obvious semantic errors. This challenge is consistent with findings that human-like neural text can often be holistically confusing for both human judges and automated classifiers (Ippolito et al., 2020).

2.2 Dataset Details

Two primary sources of data were utilized for training: the Official AbjadGenEval Data and an External Dataset (Al-Shaibani and Ahmed, 2025). The official data consists of balanced human-written news articles and AI-generated content produced by models such as GPT-3.5, GPT-4, and Claude. To investigate the effects of scale, a large external dataset consisting of 41,940 samples was incorporated. The datasets were merged to create a "Combined" training corpus. The distribution of the utilized datasets is detailed in Table 1.

Dataset Source	Total Size	Human	Machine	Usage
Official Train	4,800	2,400	2,400	Base Training
External Data	41,940	8,637	33,303	Augmentation
Combined	**46,740**	**11,037**	**35,703**	Experiments 3 & 4
Official Test	2,000	-	-	Blind Evaluation

Table 1: Distribution of datasets used in the experiments.

3 Methodology

The proposed system, Kashif-AI, treats the detection of AI-generated text as a supervised binary sequence classification task. The architecture is designed as a modular pipeline comprising three distinct stages: preprocessing, feature extraction via pre-trained Transformer encoders, and classification. The high-level system architecture is illustrated in Figure 1.

3.1 System Architecture

The detection pipeline ingests raw Arabic news articles and processes them through a specialized regex-based cleaning module to remove noise such as URLs and tatweel (elongation). The normalized text is then tokenized and passed to a Transformer encoder (MARBERT or CAMeLBERT). The encoder outputs a sequence of contextualized vectors, where the vector corresponding to the special classification token (`[CLS]`) is extracted to serve as

Model	Training Data	Validation Split		Blind Test Set	
		F1-Score	Accuracy	F1-Score	Accuracy
MARBERT	Base (Original)	0.992	0.992	0.632	0.650
	Combined (Augmented)	0.998	0.998	0.636	0.570
CAMeLBERT	Base (Original)	0.995	0.995	**0.663**	**0.705**
	Combined (Augmented)	0.997	0.997	0.561	0.570

Table 2: Ablation Study: Comparison of MARBERT and CAMeLBERT performance across distinct data configurations. The table contrasts the internal **Validation** performance against the official **Blind Test** leaderboard results. The **Base** dataset refers to the official task data, while **Combined** includes external scraped data.

the aggregate semantic representation of the entire text (Devlin et al., 2019) .

3.2 Evaluation Metrics

Model performance was assessed using standard binary classification metrics. Given the potential for class imbalance in the augmented datasets, F1-Score was prioritized as the primary metric for model selection (Goutte and Gaussier, 2005). The metrics are defined as follows, where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The F1-Score represents the harmonic mean of Precision and Recall, providing a balanced assessment of the classifier's performance:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Accuracy was also recorded to measure the overall ratio of correctly predicted observations to total observations.

3.3 Experimental Setup

To ensure a rigorous comparative analysis, identical experimental conditions were maintained across all model configurations.

3.3.1 Training Configuration

A supervised fine-tuning approach was adopted. The models were trained using the **AdamW** optimizer (Loshchilov and Hutter, 2017) with a learning rate of $2e^{-5}$ and a weight decay of 0.01 to prevent overfitting. A linear learning rate scheduler was employed to stabilize convergence over 3 training epochs. To accommodate standard news article lengths, the maximum sequence length was fixed at 512 tokens, with a batch size of 16. The training procedure utilized a "save-best" strategy, where the model state achieving the highest F1-score on the validation set was preserved for final testing.

3.3.2 Dataset Splits

For internal evaluation, the training data was partitioned using stratified sampling to generate an 80/20 train-validation split. This stratification ensured that the class distribution (Human vs. Machine) remained consistent across subsets, preventing bias in the validation metrics.

4 Results

Four distinct experiments were conducted to isolate the effects of model architecture and data augmentation. Performance was evaluated on both the internal validation split (held-out from training) and the official Blind Test Set. Table 2 presents the comprehensive results of this ablation study. A stark contrast is observed between the validation metrics and the blind test metrics. During the training phase on the combined dataset, both models achieved exceptionally high performance on the held-out validation set, with F1-scores exceeding 0.99. However, the blind test results indicated a significant drop in performance.

4.1 Ablation Study

The impact of model architecture was evident, as CAMeLBERT-Mix consistently outperformed MARBERT on the blind test set when trained on the original data, achieving an accuracy of 70.5% compared to MARBERT's 65.0%. This suggests

that CAMeLBERT's pre-training on formal MSA texts aligns better with the news domain of the shared task compared to MARBERT's tweet-heavy pre-training.

Regarding the impact of data augmentation, contrary to the common assumption that increasing dataset size improves performance, the inclusion of the large external dataset caused a degradation in generalization on the blind test set. For CAMeLBERT, accuracy dropped significantly from 70.5% to 57.0%. This degradation is attributed to the heavy class imbalance introduced by the external data, which led the models to over-predict the "Machine" class. Additionally, the external data likely contained specific watermarks or stylistic artifacts not present in the blind test set, causing the models to learn dataset-specific shortcuts rather than generalized detection features. This phenomenon of "artifact learning" has been documented as a challenge in scenarios where models might pick up on subtle watermarking strategies or generation-specific biases (Kirchenbauer et al., 2023; Ippolito et al., 2020).

4.2 Error Analysis

To investigate the source of the generalization gap, a fine-grained error analysis was conducted on the held-out validation split ($N = 7,086$). As detailed in Appendix A, the model exhibited near-perfect classification capabilities within the training distribution, as illustrated by the confusion matrix in Figure 2. While OpenAI-generated abstracts were detected with 100% success, indicating the presence of strong learnable artifacts (Kirchenbauer et al., 2023; Ippolito et al., 2020), models such as Llama and JAIS (Sengupta et al., 2023) achieved slightly higher error rates (1.60% and 1.25% respectively). The complete breakdown of these rates is summarized in Table 3. The discrepancy between these results and the blind test performance reinforces the conclusion that the model overfitted to the specific generation patterns of the external training data.

5 Conclusion

In this paper, Kashif-AI was presented as a robust baseline for Arabic AI text detection. Through a systematic ablation study, it was demonstrated that fine-tuning CAMeLBERT-Mix on the balanced official dataset yielded the highest performance, achieving 70.5% accuracy. The study further revealed that naive data augmentation with imbalanced or out-of-domain external data can be detrimental to model generalization. Future work will focus on domain adaptation techniques to bridge the gap between training and testing distributions.

Acknowledgments

The organizers of the AbjadGenEval shared task are thanked for their valuable provision of the datasets and evaluation platform. Appreciation is also extended for their continuous support and guidance, which significantly facilitated this research.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah, and Hamzah Luqman. 2025. The AraGenEval shared task on Arabic authorship style transfer and AI generated text detection. In *Proceedings of The Third Arabic Natural Language Processing Conference at EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Maged S Al-Shaibani and Moataz Ahmed. 2025. The arabic AI fingerprint: Stylometric analysis and detection of large language models text.

Hamed Alshammari and Khaled Elleithy. 2024. Toward robust arabic ai-generated text detection: Tackling diacritics challenges. *Information*, 15(7).

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Evan Crothers, Nathalie Japkowicz, and Herna Viktor. 2022. Machine generated text: A comprehensive survey of threat models and detection methods.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Saad Ezzini, Irfan Ahmed, Salmane Chafik, Shadi Abudalfa, Mo El-Haj, Ahmed Abdelali, Mustafa Jarrar, Nadir Durrani, Hassan Sajjad, and Farah Adeeba. 2026. Abjadgeneval: Abjad ai generated text detection shared task for languages using arabic script at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. ECIR'05, page 345–359, Berlin, Heidelberg. Springer-Verlag.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Salima Lamsiyah, Saad Ezzini, Abdelkader El Mahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El amrany, Salmane Chafik, and Hicham Hammouchi. 2025. M-DAIGT: A Shared Task on Multi-Domain Detection of AI-Generated Text. In *Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text*, pages 1–9, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, and 3 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news.

A Appendix A: Detailed Error Analysis

This appendix provides a granular breakdown of the error analysis results obtained on the internal validation split.

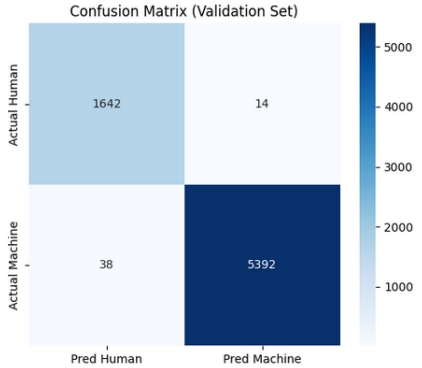

Figure 2: Confusion Matrix on the held-out Validation Set for CAMeLBERT-Mix, showing distinct separation between Human and Machine classes.

Table 3: Fooling Rate (False Negative Rate) by Source Model on the Validation Split.

Source Model	**Total Samples**	**Missed**	**Fooling Rate (%)**
Llama (Abstracts)	1,191	19	1.60%
JAIS (Abstracts)	1,284	16	1.25%
Official Task Data	392	2	0.51%
Allam (Abstracts)	1,282	1	0.08%
OpenAI (GPT)	1,281	0	0.00%