KFUPM
Journal of Undergraduate Research International

Diamond Open Access 67

Article Computer Science

Linguistic Features Embedded in Arabic Transformers: A Feature Ablation Study for E-Commerce Deepfake Detection

Rashed Jafry, Amal Sunba, and Tarek Helmy*

Cite https://doi.org/10.64589/juri/231620

Submitted: April 23, 2026 **Revised:** May 27, 2026 **Accepted:** July 09, 2026

ABSTRACT

Traditionally, Arabic natural language processing pipelines apply aggressive preprocessing, including removing punctuation, normalizing characters, and eliminating emojis, based on the assumption that cleaner inputs yield better model performance. This study investigates the linguistic features embedded in Arabic transformer models for detecting synthetic product listings in e-commerce environments. We conducted a systematic ablation study to evaluate the impact of removing individual linguistic feature categories on the classification performance of four Arabic transformer models. Our results show that preprocessing Arabic e-commerce text is not simply a matter of removing noise: the punctuation functions as a consistent structural signal across all architectures, numeric characters and English text behave as model-dependent specification-carrying signals, and emojis have no measurable effect across the models. These findings provide a practical framework for understanding the linguistic features embedded in Arabic transformers for e-commerce deepfake detection and suggest that preprocessing pipelines should preserve domain-relevant structures, rather than aggressively applying uniform cleaning.

Keywords: Arabic natural language processing, Arabic preprocessing, text preprocessing, Arabic transformers, E-commerce text classification

1. INTRODUCTION

Detecting artificial intelligence (AI)-generated content has become a critical challenge as large language models (LLMs) acquire greater capability to produce text similar to human writing. In the Arabic language ecosystem, the rapid development of Arabic-capable generative models, such as Falcon-Arabic [1], has raised concerns about the scale of AI-generated misinformation. Transformer-based detection approaches using AraBERT variants [2] and CAMeLBERT [3] have achieved significant results with news and social media text, but these models have limited utility with domain-specific content such as e-commerce product titles.

Arabic e-commerce titles are heavily code-switched, mixing Arabic script with English brand names, Latin technical specifications, numeric characters, and punctuation marks that function as structural delimiters between product components. This mix of languages and symbols is a known problem in Arabic natural language processing (NLP) tasks [4]. Transformer models can be sensitive to distributional mismatch and a non-standard input structure [5], which are the defining characteristics of product listings. Beyond code-switching, nearly all previous studies have focused on English [6], and only a few studies have examined the robustness of preprocessing choices [3,7]. Recent studies have shown that aggressive text cleaning can degrade the performance of transformer models [8,9], but this effect has not been fully explored for Arabic text. Moreover, in the e-commerce context, research has focused on review-based fraud [7] or translating product data [10], with product titles receiving limited attention in the context of deepfake detection, despite their direct influence on the purchasing decisions and search rankings.

One underexplored question is the impact of the preprocessing strategy on the detection performance in this setting. Practitioners routinely apply aggressive pipelines, such as removing punctuation and English characters, and eliminating emojis, assuming that cleaner inputs yield better results. Specifically, this assumption has not been tested for Arabic e-commerce deepfake detection, nor is it clear whether models actively discard these embedded signals.

We addressed this research gap by conducting a controlled ablation study using four Arabic transformer models, i.e., CAMeLBERT-sentiment, AraBERTv2, MarBERTv2, and ASAFAYAbert. Individual linguistic feature categories were removed and their downstream impact on the F1-score was measured. We aimed to establish the linguistic features that are embedded in these models for deepfake detection, and provide an empirically grounded framework for preprocessing decisions in this domain.

2. MATERIALS AND METHODS

2.1. Dataset Description. Our dataset contained 3,289 Arabic product titles sourced from Amazon.com, balancing authentic (human-written) and synthetic (code-modified) samples. The authentic samples were manually collected from marketplace listings, focusing on high-volume categories, such

Published under an exclusive license by KFUPM

Table 1. *Algorithmic strategies for generating synthetic titles based on the authentic product metadata*

Strategy	Description	Example
Attribute Inflation	Replaces a numeric specification (e.g., RAM or battery capacity) with the next higher value on a predefined ladder.	128 GB روم → **256 GB روم**
Variant Swap	Swaps a product tier, model variant, or network generation with an alternative baseline.	4G هاتف ذكي → **5G هاتف ذكي**
Brand Impersonation	Replaces the authentic brand token with an unrelated identifier or generic placeholder.	موتورولا → **براند**
Color Fabrication	Appends a non-existent, fabricated color variant token to the end of the title.	Title + Sampled_Color (Ø) (e.g., Appending: **الكريستالي باللون الأزرق**)
Feature Injection	Inserts a false or fabricated technical feature not present in the original product specification.	Title + Injected_Feature (False) (e.g., Appending: **الفائق يدعم الشحن اللاسلكي**)
Specification Mismatch	Appends a contradictory specification to create an internally inconsistent and logically flawed title.	Title + Contradictory_Spec (x != y) (e.g., Appending: **بشاشة 5 بوصة**)
Noise Styling	Appends low-quality promotional tokens or spam punctuation to mimic fraudulent e-commerce listing patterns.	Title + Heuristic_Noise (e.g., Appending: **عرض لفترة محدودة!!! أقوى!!!**)

as consumer electronics, and products listed before 2022, i.e., predating the generative AI boom.

The synthetic samples were generated by applying one of seven programmatic manipulation strategies to the collected authentic titles, which were distributed uniformly across the synthetic set. Table 1 lists each strategy with an illustrative example derived from a real product title in the dataset: "موتورولا جوال موتو جي 14 رو 128GB RAM 4GB هاتف ذكي 4G".

As shown in Table 2, the dataset was partitioned into training, validation, and testing splits. Each split maintained a near-perfect 49–51% class distribution between the authentic and synthetic samples, reducing the class imbalance as a confounding factor.

The dataset is not publicly available to prevent the development of adversarial detection-evasion models, but is available from the corresponding author on reasonable request.

2.2. Preprocessing Strategies. We evaluated four distinct preprocessing strategies, organized as ablation variants designed to test a specific feature dependency by removing one linguistic signal category. This allowed us to determine the features embedded in the models for accurate detection, and to assess the vulnerability of the model to targeted feature manipulation.

As summarized in Table 3, the four strategies involved the removal of (S1) all numeric characters, (S2) all Latin/English characters, (S3) all Unicode emojis, and (S4) all standard ASCII punctuation marks. In e-commerce, the punctuation marks act as structural delimiters between the product names and technical specifications. The English characters capture the brand names and technical terminology commonly used in code-switched Arabic e-commerce texts.

Table 2. *Dataset statistics and split configuration. The classes represent Authentic vs. Synthetic samples*

Split	Total Samples	Percentage (%)	Class Distribution (%)
Training	2,067	62.85	49–51
Validation	227	6.90	49–51
Testing	995	30.25	49–51
Total	3,289	100.00	Balanced

2.3. Model Architectures. We conducted the ablation study across a suite of transformer-based Arabic models, i.e., CAMeLBERT-sentiment[3], AraBERTv2[2], MarBERTv2[11], and ASAFAYAbert[12], to assess whether the preprocessing effects were model-specific. These architectures rely on learned contextual representations rather than explicit n-gram frequency features, enabling us to examine the response of the different pretraining settings to removing the signals, such as the punctuation, emojis, and brand names.

AraBERTv2 was pre-trained on 70 GB of formal Modern Standard Arabic (MSA) text sourced from news portals and Wikipedia, making it the primary formal-language benchmark. MarBERTv2, pre-trained on one billion Arabic tweets containing dialectal variations, was tested for its potential advantage in handling informal e-commerce text. CAMeLBERT-sentiment is a sentiment-tuned variant included in the recent AraGenEval shared task for AI-generated text detection[13,14]. We evaluated whether optimizing its affective and stylistic features improved deepfake detection. ArabicBERT is a bidirectional encoder representations from transformers (BERT)-based Arabic model with focused pre-training, which we refer to as ASAFAYAbert, to distinguish its specific pre-training characteristics.

Table 3. *Text preprocessing strategies for ablation analysis*

Strategy	Description
S1	Removing all numeric digits.
S2	Removing all Latin/ English characters.
S3	Removing all Unicode emojis.
S4	Removing all ASCII punctuation marks.



Figure 1. Confusion matrices for the four transformer models on the baseline (OriginalDataset) test set, aggregated across five random seeds

Figure 1 shows the confusion matrices for each model on the unmodified dataset aggregated across the five seeds. CAMeLBERT-sentiment achieved the highest baseline F1-score at 91.02%, with the fewest false negatives, whereas AraBERTv2, MarBERTv2, and ASAFAYAbert were clustered between 87–89%. All four models showed asymmetrical error distributions across both classes, with false negatives substantially outnumbering false positives across the board.

3. EXPERIMENTAL WORK

3.1. Experimental Protocol. We conducted a comprehensive cross-product experiment(4 transformer models ×4 preprocessing strategies = 16 configurations). Each configuration was trained across five random seeds (42, 1, 2, 3, and 4) to ensure statistical robustness, and yielded 80 transformer training runs.

All models were trained in a standardized environment using identical hyperparameters to isolate the preprocessing strategy as the sole experimental variable. All experiments were run with full-precision (FP32) on high-performance hardware (NVIDIA RTX 5090).

3.2. Test Setup. The statistical evaluation framework was designed to account for the paired nature of the ablation comparisons. For each model–feature pair, McNemar's test was applied to compare the predictions between the original dataset and each ablated variant under the same seed. Fisher's method was applied to aggregate the evidence across the five random seeds, and combine the resulting p-values into a single significance assessment per configuration. This approach ensured that the reported significance reflected consistent behavior across training runs, rather than a single random initialization.

4. RESULTS AND DISCUSSION

4.1. Feature Ablation Results. Table 4 presents the feature ablation results for the four transformer models, reporting the baseline F1-score on the original dataset, ablated F1-score after removing the specified feature category, change in performance expressed in percentage points (pp), change in the Matthews correlation coefficient (ΔMCC) for each configuration, and the combined p-value from Fisher's method across the five seeds. The ΔMCC results were consistent with the F1-score findings, confirming that the reported effects were not artifacts of the chosen evaluation metric.

Figure 2 summarizes the F1-score changes across all 16 model–strategy combinations. The pattern is consistent: punctuation removal (S4) produced the only significant decreases across all four models, whereas emoji removal (S3) produced no measurable effect in any model. Numeric (S1) and English characters (S2) had a strong impact on CAMeLBERT-sentiment but not on the other three architectures.

Figure 3 presents the same results as those of the absolute F1-scores, facilitating cross-model comparisons. CAMeLBERT-sentiment consistently outperformed the other three models at

Table 4. *Feature ablation results for the transformer models*

Model	Removed Feature	Baseline F1-score (%)	Ablated F1-score (%)	Delta F1-score (pp)	ΔMCC	p-value
CAMeLBERT-sentiment	Punctuation (S4)	91.02	88.10	-2.92	-0.0587	<0.001***
CAMeLBERT-sentiment	Numerics (S1)	91.02	87.14	-3.89	-0.0695	<0.001***
CAMeLBERT-sentiment	English (S2)	91.02	87.80	-3.23	-0.0549	<0.001***
CAMeLBERT-sentiment	Emojis (S3)	91.02	91.04	+0.02	+0.0006	0.725
AraBERTv2	Punctuation (S4)	87.73	86.06	-1.67	-0.0377	<0.001***
AraBERTv2	Numerics (S1)	87.73	87.68	-0.05	+0.0001	0.247
AraBERTv2	English (S2)	87.73	88.08	+0.35	+0.0097	0.612
AraBERTv2	Emojis (S3)	87.73	87.77	+0.04	+0.0035	0.968
MarBERTv2	Punctuation (S4)	87.77	85.24	-2.53	-0.0549	<0.001***
MarBERTv2	Numerics (S1)	87.77	87.32	-0.44	-0.0058	0.191
MarBERTv2	English (S2)	87.77	87.05	-0.72	-0.0121	0.070
MarBERTv2	Emojis (S3)	87.77	87.52	-0.25	-0.0042	0.852
ASAFAYAbert	Punctuation (S4)	88.93	86.39	-2.55	-0.0540	<0.001***
ASAFAYAbert	Numerics (S1)	88.93	88.90	-0.04	-0.0013	<0.001***
ASAFAYAbert	English (S2)	88.93	87.84	-1.10	-0.0177	<0.001***
ASAFAYAbert	Emojis (S3)	88.93	88.80	-0.14	-0.0035	0.941

*** denotes p <0.001.

the baseline but also showed the largest absolute degradation under S1 and S2, suggesting that its higher ceiling also presents greater feature dependency.

4.2. Feature Dependency Classification. The ablation results reveal a more meaningful pattern than a simple ranking of feature importance. Instead of showing that all non-Arabic or non-lexical elements behave as noise, the results indicate that Arabic e-commerce product titles contain task-relevant structural and specification signals that many preprocessing pipelines routinely remove. In this setting, the punctuation, numeric characters, and English text play different roles. Some of them function as organizing signals, others provide model-dependent specification information, whereas the remaining are insignificant. This distinction suggests that preprocessing for Arabic e-commerce deepfake detection should be understood as a selective signal-preservation problem, and not generic text cleaning.

4.2.1. Structural signals: punctuation. Removing the punctuation (S4) caused statistically significant F1-score decreases across all four transformer architectures: CAMeLBERT-sentiment (-2.92 pp, p <0.001), AraBERTv2 (-1.67 pp, p <0.001), MarBERTv2 (-2.53 pp, p <0.001), and ASAFAYAbert (-2.55 pp, p <0.001). This is the most consistent result of the study and suggests that the punctuation functions as

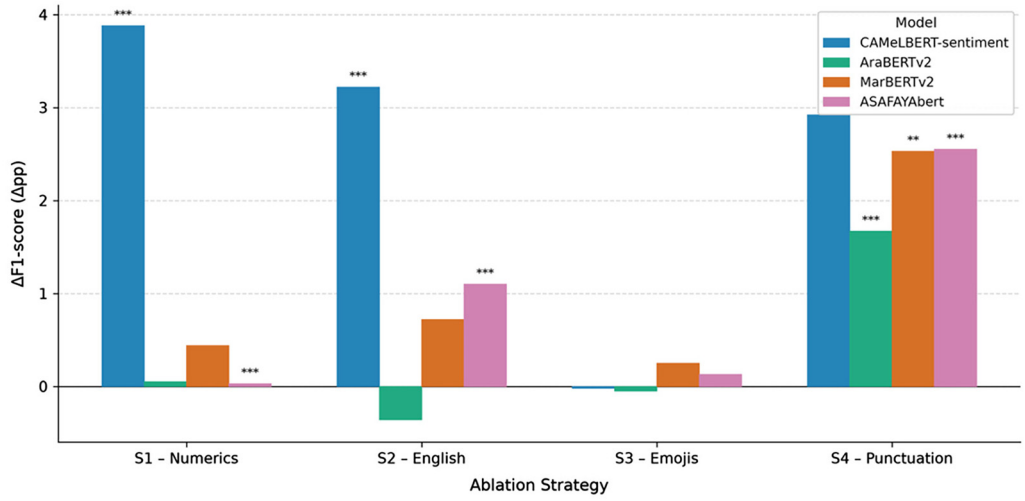


Figure 2. F1-score change (Δpp) per model across all four ablation strategies. ***denotes p <0.001, Fisher's method combined across the five seeds

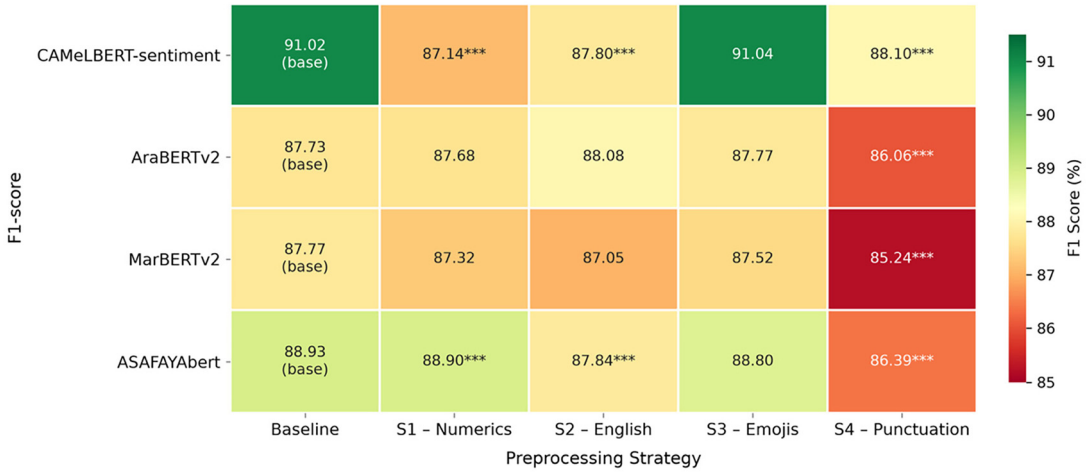


Figure 3. Absolute F1-scores (%) across all model–preprocessing combinations. *** denotes statistically significant decreases ($p < 0.001$)

a core structural signal in Arabic e-commerce titles. The product listings are not written as ordinary sentences; instead, they are compact, semi-structured strings that often use punctuation to separate the product names, model information, specifications, and variant details. This indicates that aggressively stripping the punctuation can discard information central to detection (Figure 4).

4.2.2. Specification-carrying signals: numeric characters and English text. The numeric characters and English text exhibited weaker and more model-dependent effects compared with the punctuation. CAMeLBERT-sentiment exhibited the strongest dependencies, with performance decreases of -3.89 pp for numeric characters and -3.23 pp for English text, both significant at $p < 0.001$. By contrast, AraBERTv2 and MarBERTv2 showed no statistically significant dependency on numeric characters, while the numeric character effect of ASAFAYAbert, although statistically significant, was practically negligible at -0.04 pp. A similar pattern was seen for the English text, with the impact substantial for CAMeLBERT-sentiment, moderate for ASAFAYAbert, and absent for AraBERTv2 and MarBERTv2. These findings suggest that numeric characters and English text behave less like universal structural signals and more like model-contingent carriers of product-specific information, such as brand names, technical terminology, and specifications.

The heavy reliance of CAMeLBERT-sentiment on these features may reflect its pre-training for sentiment analysis. Sentiment classification in code-switched Arabic contexts involves technical specifications and brand names; therefore, the model may learn to assign high weights to these representations. Other models, trained on different corpora with varying distributional properties, did not develop the same reliance.

4.2.3. Low-value decorative signals: emojis. Emoji removal (S3) had no statistically significant effect on any model. CAMeLBERT-sentiment improved marginally (+0.02 pp), whereas AraBERTv2, MarBERTv2, and ASAFAYAbert similarly showed null effects. Unlike the structural delimiters or specification-bearing tokens, emojis do not provide a stable discriminative value for deepfake detection. This result suggests that not all visually salient features in the product titles are equally meaningful; some encode the organization or product content, while others are insignificant.

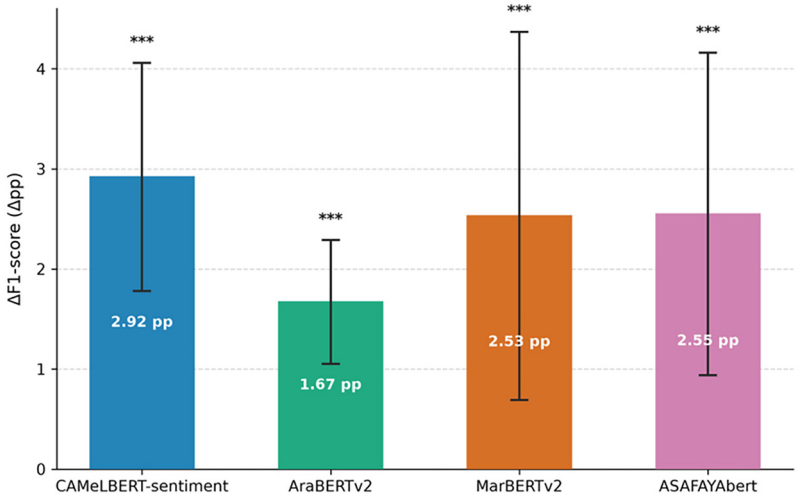


Figure 4. F1-score decreases after punctuation removal (S4) across all the models. The error bars represent ± SD across the five seeds. * * * denotes $p \leq 0.001$

4.3. The AraBERTv2 Anomaly. The marginal improvement of AraBERTv2 when the English characters were removed (+0.35 pp, $p = 0.612$) is counter-intuitive. All the other models degraded under the same conditions. We propose a tokenization mismatch hypothesis to explain this behavior, as detailed in the Supplementary Material.

4.4. Limitations of this Study. This study has several limitations that should be considered when interpreting the results. First, the dataset with 3,289 samples was relatively small by modern NLP standards, although sufficient for controlled ablation studies. The dataset also focused exclusively on consumer electronics, and its generalizability to other product categories, such as fashion and home goods, remains untested.

Second, all the statistical analyses were performed using a single held-out test set of 995 samples. Although we varied the random seeds during the training, the composition of the test set remained fixed. Cross-validation or multiple test sets drawn from different time periods or platforms would provide stronger evidence of generalization across the dynamic e-commerce landscape.

Third, the AraBERTv2 tokenization hypothesis was not analyzed or formally confirmed, and remains an open question for future research.

5. CONCLUSIONS

This study examined the linguistic features embedded in four Arabic transformer architectures for detecting synthetic product listings in Arabic e-commerce environments through a controlled feature ablation analysis.

The results show that preprocessing in Arabic e-commerce deepfake detection is not only about removing noisy or non-Arabic inputs. Instead, the ablation pattern suggests that the product titles contain different types of task-relevant signals. Punctuation emerged as the most consistent dependency across the four models, indicating that structural delimiters are central to transformer architectures interpreting compact product titles. Numeric characters and English text showed more variable effects across the models, suggesting that they function as specification-carrying signals whose values depend on the model architecture and pre-training background. By contrast, emoji removal produced no measurable degradation, indicating that visually salient features are not necessarily informative for detection. Together, these findings suggest that Arabic e-commerce titles should be treated as semi-structured, code-switched records and, effective preprocessing should preserve the domain-relevant structures rather than aggressively apply uniform cleaning.

An unexpected finding was the AraBERTv2 anomaly: removing the English characters marginally improved its performance, which we hypothesize may reflect a tokenization mismatch between its formal MSA pre-training and the informal code-switched nature of e-commerce titles. This warrants a dedicated investigation in future work (see the Supplementary Material).

Further research can also expand the dataset beyond consumer electronics, conduct formal tokenization analysis to validate the AraBERTv2 hypothesis, and test the adversarial robustness, given that the identified feature dependencies represent potential manipulation targets for evading detection.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Rashed Jafry – *Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0006-0652-8457
Email: rashed.jafry7670@alum.kfupm.edu.sa

Graduate Author

Amal Sunba – *Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia; Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia*; 0009-0002-4712-2104
Email: g202203780@kfupm.edu.sa

Corresponding Author

Tarek Helmy – *Research Mentor, Professor, Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0000-0001-9683-682X
Email: helmy@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the Department of Engineering and Interdisciplinary Research Center at the King Fahd University of Petroleum and Minerals.

REFERENCES

(1) Boussaha, B. E. A. et al. Adapting Falcon3-7B language model for Arabic: Methods, challenges, and outcomes. *Proc. 3rd Arab. Nat. Lang. Process. Conf.* 1–15 (2025). https://doi.org/10.18653/v1/2025.arabicnlp-main.1

(2) Antoun, W., Baly, F. & Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. *Proc. 4th Workshop Open-Source Arab. Corpora Process. Tools* 9–15 (2020). https://aclanthology.org/2020.osact-1.2

(3) Inoue, G. et al. The interplay of variant, size, and task type in Arabic pre-trained language models. *Proc. 6th Arab. Nat. Lang. Process. Workshop* 92–104 (2021). https://aclanthology.org/2021.wanlp-1.10

(4) Hamed, I. et al. A survey of code switched Arabic NLP: Progress, challenges, and future directions. Proc. Int. Conf. Comput. Linguist. (COLING 2025) (2025). https://doi.org/10.48550/arXiv.2501.13419

(5) Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. 2019 Conf. N. Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.* **1**, 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

(6) Gehrmann, S., Strobelt, H. & Rush, A. M. GLTR: Statistical detection and visualization of generated text. *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstrations* 111–116 (2019). https://doi.org/10.18653/v1/P19-3019

(7) Jindal, N. & Liu, B. Opinion spam and analysis. *Proc. 2008 Int. Conf. Web Search Data Min.* 219–230 (2008). https://doi.org/10.1145/1341531.1341560

(8) Siino, M., Tinnirello, I. & La Cascia, M. Is text preprocessing still worth the time? A comparative survey on the influence of popular

preprocessing methods on transformers and traditional classifiers. Inf. Syst. 121, 102342 (2024). https://doi.org/10.1016/j.is.2023.102342

(9) Matsko, V., Lyaskovska, S. & Starodub, O. Impact of text preprocessing on effectiveness of document classification by graph neural networks. Proc. SmartIndustry 2025, CEUR Workshop Proc. 3970 (2025). https://ceur-ws.org/Vol-3970/SHORT7.pdf

(10) Zhang, B., Danial, S. & Walter, S. Enhancing Arabic machine translation for e-commerce product information: Data quality challenges and innovative selection approaches. Proc. Arab. Nat. Lang. Process. Conf. 150–157 (2023). https://doi.org/10.18653/v1/2023.arabicnlp-1.13

(11) Abdul-Mageed, M., Elmadany, A. & Nagoudi, E. M. B. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. *Proc. 59th Annu. Meet. Assoc. Comput. Linguist.* 7088–7105 (2021). https://doi.org/10.18653/v1/2021.acl-long.551

(12) Safaya, A., Abdullatif, M. & Yuret, D. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. *Proc. 14th Workshop Semant. Eval.* 2054–2059 (2020). https://doi.org/10.18653/v1/2020.semeval-1.271

(13) Abudalfa, S. et al. The AraGenEval shared task on Arabic authorship style transfer and AI generated text detection. Proc. Third Arab. Nat. Lang. Process. Conf.: Shared Tasks 1–13 (2025). https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.1

(14) Zain, A., Farooqui, S. & Rafi, M. BUSTED at AraGenEval shared task: A comparative study of transformer based models for Arabic AI generated text detection. arXiv preprint arXiv:2510.20610 (2025). https://doi.org/10.48550/arXiv.2510.20610