A Hybrid Confidence-Aware Framework for Arabic Toxicity Detection in Social Media

Fawzia Alanazi [1], Asma Alamri[1], Arwa bin Saleh[1], Abdullah Alharbi[1,2]

[1]College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University, Riyadh 11432, Saudi Arabia
[2]King Salman Global Academy for Arabic Language (KSAA), Riyadh, Saudi Arabia
{446012765, 446012594, 446012875}@sm.imamu.edu.sa
aialharbii@imamu.edu.sa, aialharbi@ksaa.gov.sa

Abstract

Automatic detection of toxic and offensive content in Arabic social media is a challenging task due to rich morphology, dialectal variation, and noisy writing styles. While transformer-based language models have achieved strong performance, they often produce uncertain predictions in borderline cases. This paper presents a hybrid framework for Arabic toxicity detection that combines a pretrained Arabic-specific transformer model with a confidence-aware rule-based mechanism. The proposed approach activates automatically induced lexical rules only when the model prediction falls within a predefined gray zone of uncertainty, preserving neural dominance while improving robustness and interpretability. Experiments conducted on a manually annotated dataset of 35,000 Arabic posts demonstrate that the hybrid approach achieves consistent improvements over the baseline model, particularly in reducing false negatives for toxic content. The results indicate that selective rule activation is an effective strategy for enhancing reliability in real-world Arabic social media moderation systems.

1 Introduction

The rapid growth of social media platforms has led to a significant increase in user-generated content, enabling open communication while also facilitating the spread of toxic and offensive language such as insults, harassment, and hate speech. This type of content can negatively affect individuals and online communities, making automatic toxicity detection a crucial task for maintaining safe and healthy digital environments Fortuna and Nunes (2018) .

Toxicity detection is commonly formulated as a text classification problem. Early approaches relied on traditional machine learning techniques with hand-crafted features, while recent advances in deep learning particularly transformer-based models have substantially improved performance by capturing contextual and semantic information more effectively Devlin et al. (2019). Despite these advances, toxicity detection in Arabic remains challenging due to rich morphology, dialectal variation, and the informal nature of social media text Darwish et al. (2021).

Recent studies have shown that Arabic-specific transformer models achieve strong performance in toxicity and offensive language detection across multiple dialects and social media domains Asiri and Saleh (2024); Magnossão et al. (2022). Despite these advances, purely neural approaches often struggle in borderline or ambiguous cases and typically provide limited interpretability, particularly in culturally nuanced or implicitly offensive content Alharbi and Lee (2020). In contrast, lexicon-based methods offer linguistic transparency and explainable decisions, but suffer from limited coverage and poor generalization when used in isolation, especially in the presence of dialectal variation and creative language use.

Moreover, in dialectal and culturally grounded Arabic contexts, toxic meaning is frequently conveyed through sparse lexical cues, indirect references, sarcasm, or locally salient expressions that are underrepresented in training data and difficult to capture even with large-scale pretraining Asiri and Saleh (2024). As a result, model confidence naturally degrades on such inputs, yet standard inference pipelines typically treat all predictions uniformly, regardless of confidence or uncertainty. This mismatch between prediction confidence and decision strategy highlights the need for uncertainty-aware mechanisms that adapt the decision process itself, rather than solely increasing model capacity.

Motivated by these complementary strengths and limitations, this work proposes a confidence-aware hybrid decision framework for Arabic toxicity detection that augments transformer-based language models with an automatically induced, data-driven lexicon activated exclusively under predictive un-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 364–370
March 28, 2026. ©2026 Association for Computational Linguistics

certainty. This selective intervention strategy preserves neural dominance while enhancing robustness, reducing false negatives, and maintaining interpretability in a principled and controlled manner.

2 Related Work

2.1 Arabic toxicity detection

Arabic toxicity detection has gained increasing attention due to the rapid growth of social media content and the linguistic complexity of Arabic, particularly its dialectal varieties. Early approaches mainly focused on sentence-level classification, while recent studies have explored fine-grained detection, dialect-aware modeling, and linguistic resources tailored to Arabic social media text.

One research direction extended toxicity detection beyond sentence-level classification to identify character-level toxic spans within Arabic tweets. This approach modeled toxicity detection as a sequence labeling task and combined Arabic word embeddings with transformer-based contextual representations. The results demonstrated effective localization of toxic expressions and improved interpretability, although the focus remained on span extraction rather than sentence-level robustness Radman et al. (2022). Another line of work emphasized dialect-specific toxicity detection, particularly for Moroccan Arabic. Transformer-based models were fine-tuned on a large, dialect-specific dataset collected from social media, showing that dialect-adapted BERT models substantially outperform generic models. These findings highlight the importance of dialect-aware pretraining and large-scale annotated datasets for Arabic toxicity detection Rachidi et al. (2025).

Research on Gulf Arabic further contributed by introducing a large-scale Saudi dialect dataset annotated using a hierarchical offensive language scheme. Experiments with machine learning, deep learning, and transformer-based models demonstrated strong performance, with additional improvements achieved through data augmentation techniques to address class imbalance. This work underscores the value of dialect-specific datasets for improving Arabic offensive language detection Asiri and Saleh (2024). Lexicon-based approaches have shown that manually constructed sentiment lexicons tailored to Saudi dialect tweets can outperform larger automatically generated lexicons when neutral content is considered. Although these studies focus on sentiment analysis rather than toxicity detection, they demonstrate the effectiveness of dialect-specific lexical resources and motivate their integration into hybrid detection frameworks Al-Thubaity et al. (2018) .

2.2 Transformer Model

Recent studies in toxicity detection have mainly addressed binary classification (toxic vs. non-toxic). Early machine learning methods such as Bag-of-Words and SVM were limited in capturing contextual meaning, while LSTM-based models improved sequential modeling but struggled with long-range dependencies. Transformer-based models significantly advanced this task through self-attention and bidirectional context modeling, with encoder-based models like BERT and RoBERTa consistently outperforming traditional and recurrent approaches.

Recent findings show that language-specific Transformer models outperform multilingual ones when applied to a single language. Barrón-Cedeño and García-Silva (2025) demonstrated that BETO, a Spanish-specific Transformer, achieved superior results in binary toxicity classification. Their BETO-MP model reached an accuracy of 0.9649 and an F1 score of 0.9645, outperforming multilingual models such as mBERT and XLM-RoBERTa. These results highlight the importance of language-specific Transformers for effective toxicity detection, especially in linguistically rich contexts.

Earlier studies on Arabic abusive content detection mainly relied on traditional machine learning and later CNNs and LSTMs, which showed limited ability to capture semantic and dialectal complexity. With the introduction of Transformer-based models, pretrained Arabic models such as AraBERT and MARBERT achieved superior performance. However, most studies treated the task as single-dimensional. To overcome this limitation, Alrashidi and AlGhamdi (2023) proposed a multi-aspect annotation framework with Multi-Task Learning, showing that dialect-aware models like MARBERT significantly outperform traditional and single-task approaches in fine-grained Arabic abusive content detection.

Magnossão et al. (2022) evaluated several Arabic-specific and multilingual Transformer models on offensive language detection, hate speech detection, and fine-grained hate speech classification. Their experiments showed strong improvements over baselines, with AraBERT achieving an F1-macro of 0.827 for offensive language detection, and an ensemble approach reaching 0.792 for hate speech detection. However, fine-grained classification remained challenging, achieving lower F1-macro scores due to severe class imbalance, indicating that while Transformers perform well in binary tasks, detailed hate speech classification remains difficult.

While large Transformer models such as BERT and RoBERTa achieve state-of-the-art results, their high computational cost limits practical deployment. To address this issue, Kamphuis and van der Meer (2024) proposed *Tiny-toxic-detector*, a compact Transformer model with only 2.1 million parameters. Despite its small size and the absence of large-scale pretraining, the model achieved an accuracy of 90.97% on the ToxiGen dataset and 86.98% on the Jigsaw dataset. These results demonstrate that efficient, task-specific Transformer architectures can deliver competitive toxicity detection performance while significantly reducing computational requirements.

3 System Architecture

Figure 1 presents the overall architecture of the proposed hybrid toxicity detection system. The framework is designed to integrate deep contextual representations learned by AraBERT with lightweight symbolic reasoning modules to enhance robustness in ambiguous cases. Specifically, the model first produces a probabilistic prediction, which is then evaluated by a confidence-aware controller to determine whether rule-based lexical verification should be applied. This design preserves the dominance of neural inference while enabling targeted correction in low-confidence scenarios.

3.1 Dataset

The dataset used in this study is the “X Posts Hate Speech Dataset for the Saudi Dialect”, published on Mendeley Data by A. Alhazmi in 2024.[1] It contains 35,000 posts manually annotated by native speakers. Each post is labeled as either Toxic (Offensive) or Non-toxic (Non-offensive).

For clarity and consistency, we adopt an operational definition of toxic (offensive) language consistent with the annotation guidelines of the Saudi X Posts Hate Speech Dataset. A post is labeled as toxic if it contains direct or indirect personal attacks, insults, harassment, tribal or religious slurs, demeaning language, or expressions intended to offend or degrade individuals or groups. Non-toxic posts include neutral, polite, critical, or emotionally expressive content that does not target individuals or groups offensively. This definition aligns with prior Arabic offensive language detection studies while explicitly accounting for dialectal and culturally grounded expressions common in Saudi social media.

The dataset is balanced, with approximately 52.6% non-toxic and 47.4% toxic posts (Alhazmi, 2024).

Unlike many Arabic corpora, this dataset focuses on the Saudi dialect and includes tribal slurs, religiously offensive expressions, and implicit sarcasm. For instance, toxic samples contain dialectal expressions such as اول مره اشوف حيوان بيتكلم and اسكت يا حمار, while non-toxic posts reflect neutral or polite interactions commonly used in daily online communication. These unique linguistic traits challenge traditional MSA-trained models. By fine-tuning AraBERT on this dataset, we aim to adapt the model to better capture dialectal toxicity patterns in Saudi social media.

3.2 Data Preprocessing

To prepare the raw posts for fine-tuning, a comprehensive preprocessing pipeline was implemented. The objective was to clean the data, preserve dialectal features, and optimize the input for Arabic language modeling.

First, a cleaning phase was applied to remove irrelevant or noisy artifacts, including URLs, user mentions (@user), hashtags, punctuation marks, numeric tokens, and special characters. Diacritics, elongated characters, and repeated letters were normalized to reduce lexical sparsity. Emojis were retained due to their semantic contribution in informal communication, particularly in conveying sentiment and sarcasm.

Second, text normalization was performed. Common orthographic variants specific to the Saudi dialect were preserved to enable the model to learn from authentic linguistic variations. No mapping to Modern Standard Arabic was conducted, ensuring dialectal richness was maintained.

Third, tokenization was performed using the AraBERT tokenizer, with a maximum sequence length of 128 tokens. The tokenizer was configured with `do_lower_case=False` to retain case-sensitive features where applicable.

Finally, the dataset was stratified and split into

[1] https://data.mendeley.com/datasets/c2jpnv9yk6/4

training, validation, and test sets using a 70%–15%–15% ratio. This ensured balanced class distributions across all subsets, maintaining the integrity of both toxic and non-toxic samples throughout the modeling pipeline, as shown in Table 1.

Split	Non-Toxic (Non-Offensive)	Toxic (Offensive)	Total
Train (70%)	12,893	11,607	24,500
Validation (15%)	2,763	2,487	5,250
Test (15%)	2,763	2,487	5,250
Total	**18,419 (52.6%)**	**16,581 (47.4%)**	**35,000**

Table 1: Dataset distribution across training, validation, and test splits.

3.3 Model Training

To build a robust classifier for dialectal toxicity, we fine-tuned the AraBERT-base-v2 model on the Saudi X posts dataset.AraBERT-base-v2 was selected as the backbone model due to its strong performance in Arabic natural language processing tasks. Previous studies have shown that AraBERT consistently outperforms multilingual transformer models and traditional machine learning approaches in Arabic text classification Antoun et al. (2020); Abu Kwaik et al. (2020). This makes it a suitable choice for modeling dialectal arabic social media content. The training pipeline integrated deep contextualized representations with confidence-aware control mechanisms and symbolic rule activation. Tokenized input was passed through the AraBERT encoder, and the resulting [CLS] embedding was fed into a binary classification head. Training was conducted using binary cross-entropy loss, with the AdamW optimizer (learning rate = 2×10^{-5}) and early stopping (patience = 5 epochs). Dropout and gradient accumulation were applied to mitigate overfitting and optimize performance under computational constraints. The best-performing model was selected based on validation F1-score.

3.4 Auto-Lexian Module

A toxicity lexicon was automatically induced from the training data to capture words that are strongly indicative of toxic (offensive) content. Tokens were scored based on their relative frequency in toxic versus non-toxic samples. Only tokens satisfying specific filtering criteria were retained, including minimum token length, minimum occurrence in toxic texts, class ratio thresholds, and exclusion of tokens frequently appearing in non-toxic contexts.

As a result, a compact lexicon of 60 toxic-indicative tokens was generated, representing highly discriminative terms learned directly from the data. This lexicon was used to enhance the model's sensitivity to dialectal toxicity patterns and reduce false negatives during classification.

3.5 Gray-Zone only Rule Activation

To enhance the reliability and interpretability of toxicity detection in ambiguous inputs, a rule-based activation module was integrated into the model inference pipeline. This component leverages the model's confidence scores, computed via softmax probabilities over the final classification logits, to selectively trigger linguistic rules when the model exhibits uncertainty. Two thresholds are defined: a lower bound (`gray_low` = 0.2) and an upper bound (`gray_high` = 0.8), delineating a gray zone of confidence.

These gray-zone thresholds (`gray_low` = 0.2, `gray_high` = 0.8) were selected empirically based on preliminary experiments on the validation set. These values were chosen to isolate predictions with high uncertainty while preserving neural dominance for confident decisions. Specifically, confidence scores above 0.8 consistently corresponded to correct predictions, whereas scores below 0.2 reliably indicated non-toxic content. Intermediate scores exhibited higher error rates, making them suitable candidates for rule-based verification. Sensitivity analysis showed that small variations around these thresholds did not significantly affect performance.

The system operates as follows.

If the confidence score is greater than or equal to `gray_high`, the model prediction is accepted directly without intervention.

If the confidence lies within the gray zone, the input is checked against a previously induced toxicity lexicon (see Section 3.4). If a match is found, for example the presence of tribal slurs, sarcasm markers, or sentiment-laden emojis, the prediction is overridden to the toxic (offensive) class.

If no lexicon hit is detected within the gray zone, the original model prediction is retained.

If the confidence score is below `gray_low`, the model prediction is also accepted as is, assuming high certainty in a non-toxic classification.

This hybrid mechanism ensures that the model remains primarily data-driven while still benefiting from symbolic reasoning in edge cases. It also offers a practical compromise between deep learning flexibility and rule-based interpretability, which is particularly important in real-world applications such as moderation of dialectal content with strong sociolinguistic nuance.

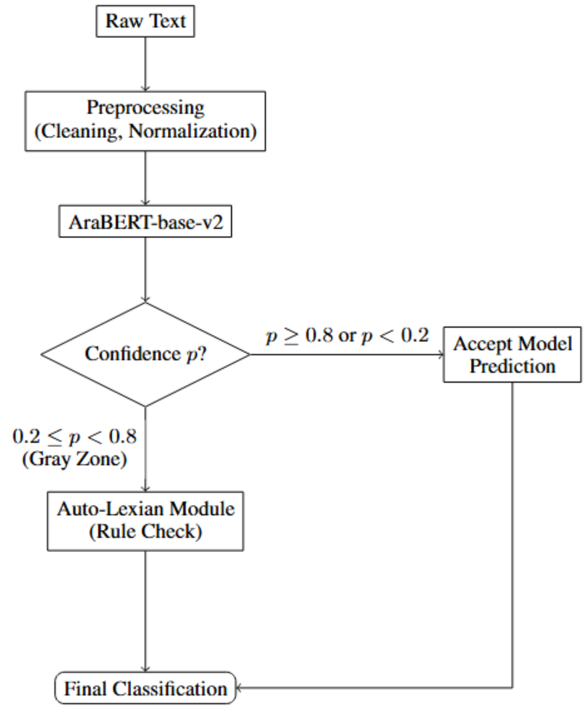

Figure 1: Hybrid System Architecture: AraBERT with Gray-Zone Rule Activation.

4 Experimental Results and Discussion

To evaluate the effectiveness of the proposed hybrid system, we compare its performance against a baseline AraBERT model fine-tuned on the Saudi X posts Hate Speech dataset. All models were evaluated on a held-out test set using precision, recall, and F1-score metrics.

4.1 Baseline AraBERT Model

The baseline system consists of the pre-trained AraBERT-base-v2 model fine-tuned on the raw annotated posts. It achieved an overall accuracy of 88.22% and a macro-averaged F1-score of 88.10%. As shown in Table 2, the model exhibits high recall for the *toxic* class (0.8870), suggesting good sensitivity, but slightly lower precision (0.8640), indicating occasional over-flagging.

Class	Precision	Recall	F1-score
Non-Toxic (Non-Offensive)	0.8953	0.8692	0.8821
Toxic (Offensive)	0.8640	0.8870	0.8754
Macro avg	**0.8796**	**0.8781**	**0.8810**

Table 2: Baseline Model Performance

4.2 Proposed Hybrid Model (AraBERT + Lexicon + Rule Activation)

The proposed model extends the baseline by integrating two symbolic reasoning components:

- A lexicon-based feature enhancer that checks for toxic tokens learned from training data.
- A gray-zone rule activation module triggered when the model confidence lies between thresholds $[0.2, 0.8]$.

This hybrid design led to noticeable improvements across all metrics. As presented in Table 3, the model achieved an accuracy of 88.87% and a macro-averaged F1-score of 88.85%. Compared to the baseline, there was a +0.75% increase in F1-score and a reduction in false negatives for the toxic (offensive) class due to rule-based correction in low-confidence scenarios.

To assess whether the observed improvement is statistically significant, we apply McNemar's test on paired predictions of the baseline and hybrid models. The difference does not reach statistical significance ($p = 0.125$).

Class	Precision	Recall	F1-score
Non-Toxic (Non-Offensive)	0.9011	0.8856	0.8933
Toxic (Offensive)	0.8755	0.8922	0.8838
Macro avg	**0.8883**	**0.8889**	**0.8885**

Table 3: Hybrid Model Performance

5 Error Analysis

To better understand the limitations of the proposed hybrid framework, we conduct a qualitative analysis of misclassified instances in the test set. Out of 3,541 samples, the model produces 369 errors, including 171 false positives and 198 false negatives, achieving a macro-averaged F1-score of 0.895. These errors reveal several recurring linguistic patterns that remain challenging in dialectal Arabic toxicity detection.

A substantial portion of the misclassifications arises from emotionally charged but non-offensive expressions, where strong negative sentiment is used to criticize opinions or situations rather than to directly attack individuals. In such cases, the model occasionally overestimates toxicity due to the presence of harsh wording or intensifiers commonly used in Saudi dialect discourse. Although this behavior leads to some false alarms, it reflects a conservative tendency that is generally preferable in moderation systems.

Conversely, many misclassified toxic instances correspond to implicit insults, sarcasm, or pragmatically offensive constructions that do not contain explicit profanity. Examples include dialectal expressions such as مدرب غبي جدا and sarcastic remarks like حمار لو صدقتك مره ثانيه, where offensive intent is conveyed indirectly through tone and cultural context rather than lexical cues alone. These cases illustrate the difficulty of capturing pragmatic meaning using sentence-level representations, even when supported by lexical verification.

Additional errors occur in context-dependent utterances, where toxicity becomes evident only when considering the conversational target or prior discourse. Without explicit contextual grounding, such posts may appear neutral or ambiguous in isolation, leading to incorrect predictions.

Despite these challenges, the hybrid model exhibits improved robustness in ambiguous cases where the baseline shows uncertainty. The rule activation module successfully corrects a subset of misclassifications by leveraging lexicon matches (e.g., tribal slurs and sentiment-bearing emojis), particularly when the neural model's confidence is low. Importantly, these symbolic augmentations enhance recall without harming precision, resulting in a more balanced trade-off between false positives and false negatives.

Beyond these cases, a particularly challenging class of errors involves pragmatically offensive content, where toxicity is conveyed indirectly through sarcasm, irony, or culturally grounded implications rather than explicit profanity.

Overall, these observations highlight that toxicity in Saudi social media is often expressed through subtle pragmatic strategies, including sarcasm, indirect disparagement, and culturally grounded expressions. While the confidence-aware rule activation mechanism improves robustness in uncertain cases, fully resolving such errors will likely require incorporating broader conversational context and more advanced pragmatic modeling. Expanding dialect-specific lexical resources and integrating context-aware architectures remain promising directions for future work. Table 4 summarizes the distribution of classification errors across the test set.

Error Type	**Count**
False Positives (Non-toxic → Toxic)	171
False Negatives (Toxic → Non-toxic)	198
Total Errors	369

Table 4: Distribution of classification errors on the test set.

6 Conclusion and Future Work

In this study, we presented a hybrid framework for Arabic toxicity detection in Saudi dialect posts, combining the contextual depth of AraBERT with symbolic reasoning through a confidence-aware rule activation module and an automatically induced lexicon. Our approach effectively mitigates misclassifications in gray-zone predictions by leveraging shallow linguistic cues such as emojis, slurs, and dialectal indicators. Experimental results demonstrate that this hybrid mechanism yields substantial gains in F1-score and recall, particularly in detecting toxic content that often escapes purely neural models.

For future work, we aim to improve the model's morphological generalization enabling it to robustly recognize multiple surface forms of the same word without requiring explicit dictionary entries. Furthermore, we will explore incorporating lightweight morphological analyzers and subword-level rule augmentation to reduce the impact of sparsity in dialectal variants. While the automatically induced lexicon has proven effective, we also plan to enhance its coverage to ensure it performs reliably even when encountering unconventional or naive expressions of toxicity.

Limitations

Despite the strong performance of the proposed hybrid framework, several limitations should be acknowledged. First, the experiments were conducted solely on Saudi dialect Arabic posts, which may limit the generalizability of the results to other Arabic dialects with distinct linguistic and sociocultural characteristics. Second, the confidence-aware rule activation mechanism relies on predefined uncertainty thresholds to identify gray-zone predictions. While effective in this setting, these thresholds were empirically determined and may require recalibration when applied to different datasets or deployment environments.

Furthermore, while the automatically induced lexicon contributes to reducing false negatives, its coverage remains inherently limited and may fail to capture implicit, sarcastic, or highly context-dependent expressions of toxicity. In particular, the lexicon-based component is sensitive to surface-form variation and may not fully capture rich morphological variation and rare or creatively altered word forms commonly observed in dialectal Arabic. Finally, while symbolic rule activation is selectively applied, integrating rule-based components with neural models increases system complexity and may pose scalability challenges in large-scale, real-world content moderation systems.

References

Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad, and Richard Johansson. 2020. An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self training. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 1–8, Marseille, France. European Language Resource Association.

Abdulmohsen Al-Thubaity, Qubayl Alqahtani, and Abdulaziz Aljandal. 2018. Sentiment lexicon for sentiment analysis of saudi dialect tweets. *Procedia Computer Science*, 142:301–307.

Abdullah I. Alharbi and Mark Lee. 2020. Combining character and word embeddings for the detection of offensive language in Arabic. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 91–96, Marseille, France. European Language Resource Association.

Mohammed Alrashidi and Mashael AlGhamdi. 2023. Multi-task learning for fine-grained abusive language detection in arabic. *Arabic Language Processing*, 10(2):143–157.

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 (OSACT)*.

Abdulaziz Asiri and Mohammed Saleh. 2024. Sod: A corpus for saudi offensive language detection. *Computers*, 13(1).

Alberto Barrón-Cedeño and Andrés García-Silva. 2025. Nlp4good: Beto-mp model for spanish toxicity classification. *Journal of Computational Linguistics*, 51(1):85–101.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Bert Kamphuis and Sjoerd van der Meer. 2024. Tiny-toxic-detector: A compact transformer for toxicity detection. *Transactions of the ACL*, 12(3):215–229.

Lucas Magnossão, Gabriela Cardenas, and José L. Martín. 2022. Upv at arabic hate speech 2022: Multi-dialect transformer ensembles. In *Proceedings of the WANLP Shared Task on Hate Speech and Offensive Language in Arabic*, pages 66–75.

Rabia Rachidi, Mohamed Amine Ouassil, Mouaad Errami, Mounir Omari, Bouchaib Cherradi, and Hassan Silkan. 2025. Leveraging BERT models for toxicity detection in moroccan dialect. In *Proceedings of the 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Fez, Morocco.

Azzam Radman, Mohammed Atros, and Rehab Duwairi. 2022. Spans detection of toxic phrases in arabic tweets. In *Proceedings of the 13th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan.