Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRANSFORMER-BASED CYBERBULLYING DETECTION IN MOROCCAN DARIJA: A WILLARD TAXONOMY APPROACH

ABDELKARIM TAHIRI[1], BOUBKER SBIHI[2], RACHID GOUARTI[3], FATIMA MOURCHID[4], ALI EL KSIMI[5]

Lyrica Laboratory, School of Information Sciences (ESI), Rabat, Morocco

E-mail: [1]abdelkarim.tahiri@esi.ac.ma, [2]bsbihi@esi.ac.ma, [3]rgouarti@esi.ac.ma, [4]fmourchid@esi.ac.ma, [5]ael-ksimi@esi.ac.ma

ABSTRACT

Cyberbullying remains a serious threat, yet detection tools for Arabic dialects are scarce. Moroccan Darija poses particular challenges for automated systems. This paper evaluates three transformer models on the Offensive Moroccan Comments Dataset. Results show MARBERT achieves 85.07% F1-score, outperforming AraBERT (83.84%) and the multilingual baseline (80.10%), confirming that dialectal pre-training matters for low-resource varieties. We also propose a severity framework based on Willard's (2007) cyberbullying taxonomy, which distinguishes eight behavioral types. Our system maps these into three levels: CRITICAL (cyberstalking, harassment, outing) for immediate action, MODERATE (flaming, denigration, exclusion, trickery, impersonation) for standard review, and NONE for safe content. Dataset analysis shows 9.6% of comments fall into the CRITICAL category. On this high-risk class, MARBERT reaches 86.05% F1-score with 84.80% recall, ensuring most dangerous content gets flagged. These results offer practical guidance for deploying content moderation systems for Arabic communities.

Keywords: *Cyberbullying Detection, Arabic NLP, Willard Taxonomy, Transformer Models, Severity Classification*

1. INTRODUCTION

Social media platforms have transformed communication worldwide. While offering opportunities for connection and self expression, these platforms have simultaneously enabled cyberbullying via digital technologies to harass, threaten or embarrass others. Research consistently associates cyberbullying victimization to adverse mental health outcomes, including depression, anxiety, and suicidal ideation among youth [1], [2]. The anonymous nature of online harassment intensifies the psychological effect of these attacks.

Representing more than 420 million individuals globally, Arabic speaking communities face acute challenges in this domain. While considerable research has addressed cyberbullying detection in English, Arabic social media platforms remain under-served by automated content moderation systems [3]. This disparity becomes particularly obvious for Arabic dialects, which differ markedly from Modern Standard Arabic (MSA) in vocabulary and syntax. Moroccan Darija presents more challenges due to extensive borrowing from French and Spanish, alongside the switching between Arabic script and Latin-based transliteration.

Recent advances in natural language processing, particularly transformer-based architectures, have demonstrated good capabilities across text classification tasks. Arabic transformers including AraBERT [4] and MARBERT [5] offer promising results addressing the dialectal challenge. However, their application to cyberbullying detection in Moroccan Darija remains unexplored, and existing systems built for English rely on binary classification without accounting for the varying severity of online aggression

Beyond binary classification, effective cyberbullying intervention requires nuanced severity assessment. Willard [6] established a foundational taxonomy identifying eight distinct forms of cyberbullying: cyberstalking, harassment, outing, trickery, denigration, impersonation, flaming, and exclusion. These categories differ substantially in their psychological impact and intervention urgency. Traditional binary systems

fail to capture these varying degrees of severity, treating a death threat equivalently to mild inappropriate language.

This paper addresses the aforementioned gaps by presenting a comprehensive framework for cyberbullying detection in Moroccan Darija with integrated severity assessment grounded in Willard's taxonomy. Our contributions include: (1) systematic comparison of three transformer architectures on the OMCD dataset; (2) a theoretically-grounded severity framework mapping eight cyberbullying types into three operational levels; (3) detailed evaluation demonstrating MARBERT's superior performance for both binary classification and CRITICAL content detection.

Our results demonstrate that MARBERT achieves superior performance with an F1-score of 85.07%. This represents improvements over both multilingual BERT (80.10%) and AraBERT (83.84%). Our three-level system based on Willard Severity Assessment categorizes detected content as CRITICAL (cyberstalking, harassment, outing), MODERATE (flaming, denigration, exclusion, trickery, impersonation), or NONE (non-offensive). Analysis exposes that 9.6% of the dataset needs immediate intervention (CRITICAL), while 44.1% warrants standard review within 48 hours (MODERATE). For CRITICAL content detection, MARBERT achieves 86.05% F1-score with 84.80% recall, ensuring that the majority of dangerous content is recognized for urgent action.

The remainder of this paper is organized as follows. Section 2 reviews related work across three domains: cyberbullying detection using machine learning, natural language processing for Arabic and its dialects, and severity assessment systems for content moderation. Section 3 describes our methodology in detail, including the dataset characteristics, preprocessing pipeline, model architectures, and our proposed Willard-based severity stratification framework. Section 4 presents results including comparative model performance and severity classification metrics. Section 5 discusses key findings, and limitations. And we conclude with a summary of our contributions and future research.

2. RELATED WORK

Cyberbullying detection has attracted considerable research attention over the past decade. [7] pioneered work using SVMs trained on bag-of-words and sentiment features. The advent of deep learning transformed detection capabilities. CNNs and LSTMs demonstrated superior performance by learning hierarchical representations without extensive feature engineering [8], [9].

Transformer-based models have set new benchmarks. [10] developed a heterogeneous ensemble deep learning model for Arabic sentiment analysis, demonstrating the effectiveness of combining multiple architectures for Arabic text classification. [11] combined BERT with deep learning for Twitter cyberbullying detection, reporting 84.7% F1-score. [12] explored attention mechanisms showing that explicit attention weights help identify offensive portions. Despite these advances, research specifically addressing dialectal Arabic cyberbullying remains practically unexplored.

Arabic presents unique NLP challenges due to morphological complexity and substantial dialectal variation. [13] addressed Arabic cyberbullying using deep learning, reporting 81.2% F1-score on MSA datasets but acknowledging limited generalization to dialectal content. [4] introduced AraBERT, pre-trained on 77GB Arabic text. [5] developed MARBERT, pre-trained on 1 billion Arabic tweets spanning multiple dialects. [14] created the OMCD dataset for Moroccan Darija, [15] evaluated several Arabic language models on this dataset, with MARBERTv2 achieving the best performance at 86.04% F1-score.

Willard [6] established a foundational taxonomy identifying eight distinct cyberbullying types. Table 1 describes the specific behavioral patterns of each type and their psychological impacts. Most computational cyberbullying detection research treats offensive language as a binary classification problem, ignoring this variation in severity. [16] proposed multi-level severity classification based on linguistic features. [17] proposed an ensemble method based on feature fusion for suicidal ideation, demonstrating that combining multiple feature types improves classification. However, existing severity work has not operationalized established typologies like Willard's taxonomy for automated classification, nor addressed dialectal Arabic contexts.

Table 1 : Willard's Cyberbullying Types

cyberbullying types	Description
Cyberstalking	Repeated harassment with threats causing fear for safety. This includes death threats, rape threats, and stalking behavior.

Harassment	Repeatedly sending offensive, rude, and insulting messages to a target.
Outing	Publicly sharing private information, or embarrassing images without consent.
Trickery	Deceiving someone into revealing secrets, then sharing that information publicly.
Denigration	Posting cruel gossip, rumors, or false statements to damage someone's reputation.
Exclusion	Intentionally excluding someone from an online group or community.
Flaming	Online arguments using vulgar, angry, or provocative language.
Impersonation	Pretending to be someone else online to damage their reputation, relationships, or social standing

3. METHOD

We utilize the Offensive Moroccan Comments Dataset (OMCD) [14], comprising 8,023 YouTube comments in Moroccan Darija. Comments are written in mixed scripts including Arabic and Arabizi, reflecting authentic social media discourse. Text undergoes minimal preprocessing: URLs and user mentions are removed, emojis are stripped, and repeated characters are reduced to a maximum of two, while no stemming or lemmatization is applied in order to preserve the morphological cues exploited by transformer encoders [5]. The dataset is drawn from publicly available YouTube comments, all identifiable information was removed by the original authors [14], and no additional data collection was conducted for this study. Table 2 summarizes dataset statistics.

Table 2: Dataset Statistics

Split	Offensive	Non-offensive	Total
Train	3,415 (53.2%)	3,003 (46.8%)	6,418
Test	888 (55.3%)	717 (44.7%)	1,605
Total	4,303 (53.6%)	3,720 (46.4%)	8,023

3.1 Model and Training Configuration

We evaluate three transformer models representing different levels of Arabic linguistic specialization:

BERT-Base-Multilingual (mBERT): Pre-trained on Wikipedia from 104 languages, serving as our multilingual baseline. Architecture comprises 12 transformer layers, 768 hidden dimensions, 12 attention heads, and approximately 110M parameters [18].

AraBERT: Arabic-specific BERT pre-trained on 77GB Arabic text including news, Wikipedia, and books. The AraBERTv2 variant incorporated dialectal data during pre-training, with 135M parameters [4].

MARBERT: Pre-trained specifically on 1 billion Arabic tweets spanning MSA and multiple regional dialects including Maghrebi varieties. With 163M parameters, MARBERT was explicitly designed for informal social media Arabic [5].

All models undergo fine-tuning with identical hyperparameters to ensure fair comparison: AdamW optimizer, learning rate 2×10^{-5}, batch size 16, 3 epochs, max sequence length 128 tokens, weight decay 0.01. A classification head consisting of dropout ($p=0.1$) followed by linear transformation maps the [CLS] token to two output logits. Experiments were conducted on Google Colab with Tesla T4 GPU. Random seeds were fixed to 42 for reproducibility.

3.2 Willard-Based Severity Framework

Our severity framework extends binary classification by categorizing offensive content according to Willard's [6] taxonomy. We map eight cyberbullying types into three operational levels based on psychological harm potential and intervention urgency. Table 3 presents the mapping with justification.

Table 3: Willard Taxonomy To Severity Level Mapping

Severity	Willard types	Action	Score
CRITICAL	Cyberstalking, Harassment, Outing	Immediate	2
MODERATE	Flaming, Denigration, Exclusion, Trickery, Impersonation	48h review	1
NONE	Non-offensive	No action	0

The CRITICAL classification encompasses cyberstalking, harassment, and outing because these three types share distinguishing characteristics: (1) high psychological impact associated with depression and suicidal ideation; (2) potential for real-world harm as cyberstalking often escalates to physical threats; (3) time-sensitive nature requiring immediate removal; (4) legal implications in many jurisdictions.

To operationalize the taxonomy, we developed keyword lexicons for each Willard type adapted to Moroccan Darija linguistic patterns. The lexicon comprises 152 keywords covering all eight types in both Arabic script and Arabizi transliteration. Classification proceeds as follows: if binary prediction is non-offensive → NONE; if offensive, scan text for keyword matches; if CRITICAL keywords detected → CRITICAL; otherwise → MODERATE (default as FLAMING).

It is important to acknowledge that our mapping from Willard's eight cyberbullying types to three severity levels represents an operational approximation rather than a clinically validated classification. the assignment of specific Willard types to CRITICAL versus MODERATE categories reflects our interpretation of harm potential based on existing literature. Despite this limitation, we argue that an approximate theoretically-grounded framework provides more principled guidance for intervention prioritization than purely data-driven approaches lacking conceptual foundations.

4. RESULTS

4.1 Classification Results

Table 4 presents comparative performance across the three transformer models on the OMCD test set (1,605 samples). MARBERT achieves the strongest performance with 85.07% F1-score, representing an improvement of 4.97 percentage points over mBERT and 1.23 points over AraBERT. [19] test confirms all pairwise differences are statistically significant ($p < 0.001$). This validates that dialectal pre-training on social media text provides substantial advantages for processing informal Moroccan Darija.

Table 4: Binary Classification Results

Model	Acc	Pre	Recall	F1
mBERT	0.7925	0.8535	0.7545	0.8010
AraBERT	0.8299	0.8839	0.7973	0.8384
MARBERT	0.8386	0.8713	0.8311	0.8507

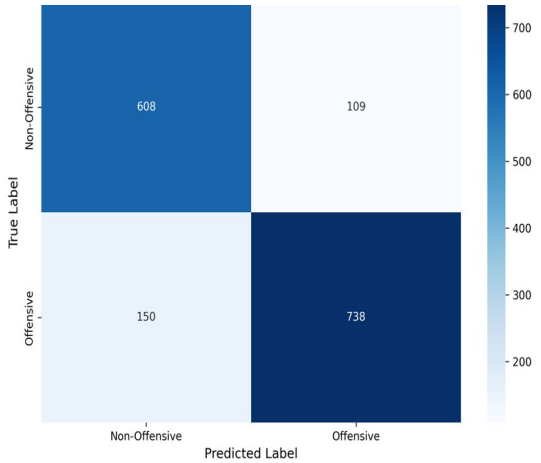

Figure 1 : Confusion Matrix for MARBERT Binary Classification on Test Set

AraBERT exhibits the highest precision (88.39%) but lower recall (79.73%), suggesting conservative predictions. MARBERT achieves better balance between precision (87.13%) and recall (83.11%), indicating superior calibration essential for detecting the majority of offensive content without excessive false positives. Figure 1 shows Confusion Matrix For MARBERT. As illustrated, the model correctly identifies 738 offensive comments (TP) and 608 non-offensive comments (TN). The 150 false negatives represent missed offensive content, while 109 false positives are benign comments incorrectly flagged.

4.2 Severity Analysis

Table 5 presents the severity distribution across the full dataset using Willard-based classification. As shown in Table 5, 9.6% of content falls into the CRITICAL category requiring immediate intervention, while 44.1% is MODERATE and 46.4% is non-offensive. Figure 2 shows that critical content represents 17.8% among offensive class. This distribution enables efficient resource allocation for content moderation teams.

Table 5: Severity Distribution (N=8,023)

Level	Count	%	Action
CRITICAL	768	9.6%	Immediate intervention
MODERATE	3,535	44.1%	Review within 48h
NONE	3,720	46.4%	No action required

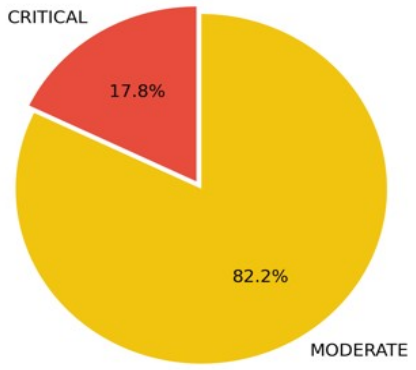

Figure 2: Distribution of Severity Levels among offensive content

As shown in Figure 3, among CRITICAL content, harassment (440 cases, 5.5%) is most prevalent, followed by outing (240 cases, 3.0%) and cyberstalking (88 cases, 1.1%). Flaming dominates the MODERATE category (2,784 cases, 34.7%), consistent with the argumentative nature of comments on political content. The 9.6% CRITICAL prevalence underscores the importance of severity-based prioritization for efficient resource allocation.

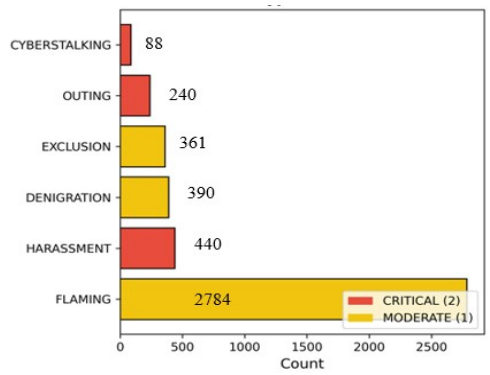

Figure 3: Willard types distribution

4.3 Multi-Class Evaluation

We evaluated all models on their ability to correctly classify content into the three severity levels. Table 6 presents CRITICAL class performance, the focus for safety applications.

Table 6: Critical Class Performance

Model	Precision	Recall	F1-Score
mBERT	87.84%	76.02%	81.50%
AraBERT	93.88%	80.70%	86.79%
MARBERT	87.35%	84.80%	86.05%

MARBERT achieves the best recall on CRITICAL content (84.80%), detecting more dangerous cases than other models. AraBERT shows higher precision (93.88%) but lower recall (80.70%), missing more threats. For safety-focused deployment prioritizing victim protection, MARBERT's balanced performance is preferred—the 84.80% recall, as shown in Figure 4, ensures that the vast majority of dangerous content requiring immediate intervention is successfully identified.

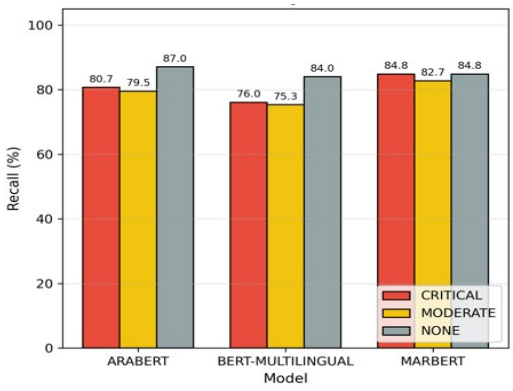

Figure 4 : Recall by class for the three models

Analysis of missed CRITICAL cases reveals all 26 errors (15.2%) were misclassified as NONE rather than MODERATE. This indicates errors occur at the binary classification stage rather than severity assignment—the Willard-based keyword detection performs reliably once content is correctly identified as offensive.

4.4 Performance Comparison

To contextualize our results within the broader research landscape, we conduct a detailed comparison with existing cyberbullying and offensive language detection systems across multiple dimensions: classification performance, language coverage, and severity assessment capabilities. Table 7 presents a comprehensive comparison of our work with related studies in cyberbullying detection.

As shown in Table 7, our MARBERT model achieves an F1-score of 85.07%, which is lower than English-based systems. [12] achieved 94.49% using deep learning with attention mechanisms, [20] obtained 92.80% with traditional machine learning, and [16] reported 92.90% with Random Forest. This performance gap (about 7 to 10 percentage points) can be attributed to three factors: first, the limited availability of pre-training data for Moroccan Darija compared to English. Second, the inherent challenges of dialectal Arabic including orthographic variation and code-switching, and finally the significantly smaller dataset size (8,023 compared to 24,000-37,000 samples).

When compared to [15], who evaluated MARBERTv2 on the same dataset (OMCD) achieving 86.04% F1-score, our model performs comparably with only a 0.97 percentage point difference. This marginal gap demonstrates that our approach maintains competitive performance while adding severity classification capability, a feature absent from the baseline evaluation.

Among all compared studies, only [16] and our work incorporate severity assessment. However, their approach relies on confidence-based thresholds derived from classifier output probabilities, whereas our method employs the theoretically-grounded Willard taxonomy. This provides more interpretable and consistent severity categorization based on established cyberbullying research rather than arbitrary confidence intervals.

Table 7 : Comparison with related studies

Study	Language	Dataset size	Method	Metrics	Severity
Fati et al. (2023) [12]	English	37,373	Conv1DLSTM	F1-score: 94.49%	No
Muneer & Fati (2020) [20]	English	37,373	Logistic Regression	F1-score: 92.80%	No
Talpur & O'Sullivan (2020) [16]	English	24,158	Random Forest	F1-score: 92.90%	Yes (Confidence)
Qarah & Alsanoosy (2025) [15]	Darija	8,023	MARBERTv2	F1-score: 86.04%	No
Our work	**Darija**	**8,023**	**MARBERT**	**F1-score: 85.07%**	**Yes (Willard)**

Our work represents the first severity-aware cyberbullying detection system for Moroccan Darija, bridging the gap between high-performing English systems and the underexplored domain of dialectal Arabic social media moderation.

5. DISCUSSION

Our results demonstrate that dialectal pre-training proves critical for low-resource Arabic varieties. MARBERT's improvement over mBERT stems from its pre-training on dialectal social media text resembling YouTube comments. The Willard-based severity framework offers several advantages over confidence-based approaches used in prior work [16]: (1) theoretically-grounded severity levels based on decades of cyberbullying research rather than ad-hoc engineering decisions; (2) interpretable classifications where moderators understand why content is CRITICAL; (3) reproducibility across models as severity depends on behavioral type rather than model-specific probability distributions.

The framework enables efficient resource allocation: 46.4% requires no review, 44.1% receives standard processing, and 9.6% flagged as critical demands immediate attention. For organizations processing thousands of comments daily, this differentiation translates to operational efficiency while maintaining comprehensive coverage of unsafe content.

5.1 Practical Implications

Our findings carry several important implications for deployment of automated cyberbullying detection systems in Arabic speaking contexts. For Mental Health Support Platforms, the system's reliable CRITICAL detection (86.05% F1, 84.80% recall) makes it suitable for integration into mental health crisis intervention platforms. When CRITICAL content is detected, the platform can automatically initiate appropriate responses.

Schools and universities can deploy the system to monitor student communications. The Willard-based framework provides educational value beyond mere detection: administrators can understand not just that cyberbullying occurred, but what type, informing appropriate disciplinary and support responses aligned with the behavioral patterns exhibited.

Concerning Social Media Moderation for Arabic-language platforms, the three-level framework enables tiered moderation workflows. CRITICAL content receives immediate review by specialized teams. MODERATE content enters standard review queues. This differentiation improves response time for cases threatening user safety.

In Research Applications, the Willard-based annotation enables epidemiological research on cyberbullying patterns. Researchers can study not just prevalence but type distribution. Such insights inform prevention program design and resource allocation.

Finally, for transferability to Other Dialects. While our work focuses on Moroccan Darija, the methodology transfers to other Arabic dialects. The Willard taxonomy is language-agnostic; only the keyword lexicons require adaptation.

5.2 Limitations

Several limitations should be acknowledged. First, OMCD contains only YouTube comments on Moroccan political topics; cross platform generalization to Facebook, Instagram or TikTok remains unvalidated.

Second, keyword-based severity detection may miss novel expressions or coded language not in the lexicon (the 0% detection of TRICKERY and IMPERSONATION types may reflect lexicon gaps).

Third, the 15.2% CRITICAL miss rate indicates room for improvement in detecting subtle threatening content. Also, social media language evolves rapidly; periodic lexicon updates and

model retraining may be necessary for sustained performance.

About the severity classification, our manual annotation of the OMCD dataset according to Willard's taxonomy, was performed by a single annotator following the keyword-based approach described in our methodology. While this approach ensures consistency, it lacks the inter-annotator agreement validation typically required for gold-standard datasets.

The absence of multiple independent annotators means we cannot report Cohen's kappa or similar reliability metrics. Future work should involve multiple annotators with cyberbullying expertise to validate the severity labels and establish annotation guidelines that could be applied to other Arabic dialect datasets.

5.3 Ethical Consideration

Deploying automated cyberbullying detection systems raises important ethical considerations that must be prudently addressed for responsible implementation.

Monitoring user generated content, even for protecting purposes, constitutes surveillance that users may find offensive. Organizations implementing such systems must keep transparency about monitoring practices and obtain appropriate consent where legally required. The balance between protecting vulnerable people and respecting privacy requires ongoing ethical deliberation.

Despite achieving 87.35% precision on CRITICAL content, false positives remain inevitable. Users wrongly flagged may experience frustration or unfair restrictions. To mitigate these risks, robust appeals processes and human review for CRITICAL classifications should be implemented before disciplinary actions. Automated systems should augment rather than replace human judgment.

The 15.2% CRITICAL miss rate means some risky content escapes detection. Victims of undetected attacks receive no initiated intervention. Organizations must supplement automated detection with accessible manual reporting mechanisms, ensuring victims can flag incidents the system misses.

Machine learning models can perpetuate biases present in training data. If annotations reflect systematic biases, the model may exhibit discriminatory behavior. Regular bias audits and diverse annotator teams help identify and mitigate such issues, though complete elimination remains challenging.

Transformer models work as "black boxes," making explanation difficult. Our Willard based framework partially addresses this by providing interpretable categories, moderators can explain that content was CRITICAL due to harassment indicators. Future work should explore additional explainability techniques.

While designed for protection, this technology could be misused for censorship or surveillance of legitimate speech. Responsible deployment policies and regulatory frameworks should prevent abuse of content detection technologies.

6. CONCLUSION

This paper addresses cyberbullying detection for Moroccan Darija, establishing performance benchmarks on the OMCD dataset (8,023 YouTube comments). Through systematic comparison of three transformer architectures, we demonstrate that MARBERT achieves 85.07% F1-score for binary classification, significantly outperforming AraBERT (83.84%) and multilingual BERT (80.10%). Statistical significance testing confirms these differences ($p < 0.001$), validating that dialectal pre-training matters for low-resource varieties.

We introduce a severity framework grounded in Willard's (2007) cyberbullying taxonomy, mapping eight behavioral types into three operational levels: CRITICAL (cyberstalking, harassment, outing), MODERATE (flaming, denigration, exclusion, trickery, impersonation), and NONE. Dataset analysis reveals 9.6% of content requires immediate intervention. On CRITICAL detection, MARBERT achieves 86.05% F1-score with 84.80% recall, ensuring most dangerous content gets flagged for urgent review.

These contributions establish foundations for deploying theoretically grounded content moderation systems protecting Arabic speaking youth from online harassment. Future work should address enhanced binary recall for implicit threats, expanded Darija lexicons, cross platform evaluation, and longitudinal validation of clinical effectiveness in mental health intervention settings.

REFERENCES:

[1] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Archives of Suicide Research, Vol. 14, No. 3, 2010, pp. 206-221.

[2] S. M. B. Bottino et al., "Cyberbullying and adolescent mental health: Systematic review," Cadernos de Saúde Pública, Vol. 31, No. 3, 2015, pp. 463-475.

[3] E. A. Vogels, "The state of online harassment," Pew Research Center, 2021. [Online]. Available: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[4] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," Proceedings of OSACT4, 2020, pp. 9-15.

[5] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," Proceedings of ACL-IJCNLP, 2021, pp. 7088-7105.

[6] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Champaign, IL: Research Press, 2007.

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," Proceedings of ICWSM, Vol. 5, No. 3, 2011, pp. 11-17.

[8] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," Proceedings of IEEE DCOSS, 2016, pp. 43-48.

[9] X. Zhang et al., "Cyberbullying detection with a pronunciation based convolutional neural network," Proceedings of IEEE ICMLA, 2016, pp. 740-745.

[10] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, and T. Alkhalifah, "Heterogeneous Ensemble Deep Learning Model for Enhanced Arabic Sentiment Analysis," Sensors, Vol. 22, No. 10, 2022, Art. no. 3707. DOI: 10.3390/s22103707.

[11] Ç. O. Aliyeva and M. Yağanoğlu, "Deep learning approach to detect cyberbullying on twitter," Multimedia Tools and Applications, Vol. 84, 2025, pp. 20497-20520. DOI: 10.1007/s11042-024-19869-3.

[12] S.M. Fati, A. Muneer, A. Alwadain, and A.O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," Mathematics, vol. 11, no. 16, p. 3567, Aug. 2023. DOI: 10.3390/math11163567

[13] B. Haidar, M. Chamoun, and A. Serhrouchni, "Arabic cyberbullying detection: Using deep learning," Proceedings of ICTA, 2019, pp. 1-6.

[14] K. Essefar, H. Ait Baha, A. El Mahdaouy, A. El Mekki, and I. Berrada, "OMCD: Offensive Moroccan Comments Dataset," Language Resources and Evaluation, Vol. 57, No. 4, 2023, pp. 1745-1765. DOI: 10.1007/s10579-023-09663-2.

[15] F. Qarah and T. Alsanoosy, "Evaluation of Arabic Large Language Models on Moroccan Dialect," Engineering, Technology & Applied Science Research, vol. 15, no. 3, pp. 22478–22485, Jun. 2025. DOI: https://doi.org/10.48084/etasr.10331

[16] M. S. Talpur and D. O'Sullivan, "Multi-level severity classification for cyberbullying detection using deep learning," PLOS ONE, Vol. 15, No. 10, 2020, Art. no. e0240924.

[17] J. Liu, M. Shi, and H. Jiang, "Detecting Suicidal Ideation in Social Media: An Ensemble Method Based on Feature Fusion," International Journal of Environmental Research and Public Health, vol. 19, no. 13, p. 8197, 2022. DOI: 10.3390/ijerph19138197

[18] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of NAACL-HLT, 2019, pp. 4171-4186.

[19] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, Vol. 12, No. 2, 1947, pp. 153-157.

[20] A. Muneer and S.M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," Future Internet, vol. 12, no. 11, p. 187, Oct. 2020. DOI: 10.3390/fi12110187