

Android application for multilingual hate speech detection: using XLM-RoBERTa and ONNX optimized

Salam Thabet Doghmash^{*1}, İbrahim Alper Doğru², Kazım Kılıç³

¹Department of Computer Sciences, Institute of Informatics, Gazi University, Ankara, Türkiye

²Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Türkiye

³Department of Artificial Intelligence and Data Engineering, Faculty of Engineering and Natural Sciences,
Samsun University, Samsun, Türkiye

Cite this article as: Doghmash, S. T., Doğru, İ. A., & Kılıç, K. (2026). Android application for multilingual hate speech detection: using XLM-RoBERTa and ONNX optimized. *J Comp Electr Electron Eng Sci*, 4(2), 72-76. <https://doi.org/10.51271/JCEES-0039>

Received: 01.02.2026

Accepted: 07.05.2026

Published: 13.10.2026

ABSTRACT

Aims: Social media and digital platforms face increasing challenges in dealing with multilingual abusive content, which calls for the development of effective solutions for the automatic detection of hate speech. This research presents a practical system for monitoring hate speech in three languages (Arabic, Turkish, and English) based on the improved XLM-RoBERTa model, while exploring ensemble techniques as a baseline for comparison.

Methods: FrancophonIA data, consisting of 219,508 texts across the three languages, was used after processing. Baseline models were developed combining logistic regression, support vector machines, and random forest using soft voting with TF-IDF feature extraction. The main model was based on XLM-RoBERTa with graphics processing unit (GPU)-accelerated training and an early stopping strategy to improve efficiency. The final models were converted to ONNX for compatibility, and a mobile application was created for Android.

Results: The baseline model achieved a macro F1-score of 0.7796, outperforming the individual classifiers. In contrast, the optimized XLM-RoBERTa achieved a substantial improvement of 84.93% accuracy and a macro F1-score of 0.8493 across different languages. Optimization via open neural network exchange (ONNX) enabled efficient real-time inference in both the API and the mobile application.

Conclusion: The research proves the feasibility of employing transformer-based multilingual models in building practical content monitoring systems, with the potential for actual deployment on mobile devices to achieve real-time monitoring of hate speech in real-world environments.

Keywords: Multilingual hate speech detection, XLM-RoBERTa, ensemble baseline, ONNX, Android deployment

INTRODUCTION

Globally, the way people communicate and share information has changed dramatically as a result of the digital communication platforms' explosive rise. But this growth has also accelerated the dissemination of damaging content, such as hate and offensive speech, which has a detrimental impact on people's well-being and social cohesiveness (Vogels, 2021). A Pew Research Center survey found that 41% of American adults have personally been the victim of internet harassment, with 25% reporting more severe or protracted instances. Furthermore, recent reports from the EU highlight substantial challenges in content moderation, particularly the ineffectiveness of automated systems and human reviewers in detecting nuanced hateful speech across languages and dialects (Online Content Moderation-Current Challenges in Detecting Hate Speech | European Union Agency for Fundamental Rights, 2023).

Despite increasing attention, current hate speech detection approaches face several limitations. Studies show that approximately 85% of existing systems are designed

exclusively for English, with less than 15% addressing multilingual contexts, leaving a considerable gap for widely spoken languages such as Arabic and Turkish (Mnassri et al., 2024). Moreover, many models remain computationally expensive, resource-intensive, and difficult to deploy in real-world scenarios, particularly on mobile devices where content is frequently generated and consumed (Mahajan et al., 2024).

New developments in natural language processing (NLP) have shown promising cross-lingual performance, especially transformer-based architectures like XLM-RoBERTa. However, practical integration of these models into deployable, real-time applications remains limited due to infrastructure complexity and high resource demands (Usman et al., 2025). A recent systematic review by (Ahmad et al., 2024) compared over one hundred studies concluded that multilingual transformer models consistently outperform traditional methods, though they raise challenges in terms of computational cost and fairness across languages. (Conneau et al., 2020) expanded on multilingual capabilities by introducing XLM-R, showing

*Corresponding Author: Salam Thabet Doghmash, salam.th.d@gmail.com



that zero-shot cross-lingual transfer is much enhanced by scaling multilingual models with a variety of language data. For morphologically rich languages like Arabic and Turkish, (Rust et al., 2021) investigated language-specific fine-tuning techniques in more detail, addressing dialectal variances and code-switching that are common in social media contexts. For under-resourced languages, recent work has emphasized Arabic and Turkish (Ahmad et al., 2024) investigated Arabic hate speech detection, noting the difficulties posed by dialectal variation and informal writing styles, and highlighted the need for tailored preprocessing.

Addressing these challenges requires solutions that not only achieve high detection accuracy but also ensure accessibility, scalability, and efficiency for end-users. This paper presents a complete methodology for detecting hate speech in Arabic, Turkish, and English. The system combines an improved transformer model, XLM-RoBERTa, that has been fine-tuned on GPU with balanced sampling, with the traditional ensemble baselines of Random Forest, Support Vector Machines, and Logistic Regression coupled through soft-voting. In addition to performance, the models were deployed in a practical environment that included an enhanced open neural network exchange (ONNX) API for real-time inference and an Android mobile app for easy access by end users. This combination of improved accuracy and scalability demonstrates the feasibility of translating advanced multilingual NLP research into practical content management solutions.

METHODS

Ethics

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data. All procedures were carried out in accordance with the ethical rules and principles.

Study Design

This study was designed as an experimental analysis concentrating on building a system for detecting hate speech in multiple languages. Our methodology covered six stages: dataset preparation, preprocessing, baseline construction, transformer-based modeling, evaluation, and deployment as shown in **Figure 1**.

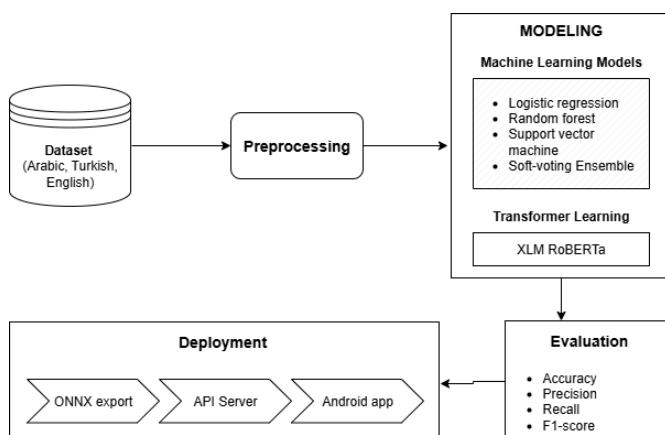


Figure 1. The methodology steps

Dataset: The FrancophonIA multilingual hate speech dataset was used, this dataset originally contained approximately 219,508 training samples in Arabic, English, and Turkish. The dataset shows a moderate imbalance in classes, with

hate speech comprising 38.5% (84,418 samples) and non-hate speech 61.5% (135,090 samples). To address this imbalance, under sampling was applied, randomly selecting 84,418 samples from the larger group. This resulted in a balanced dataset of 168,836 samples, with each group representing 50%. Subsequently, stratified sampling was used to divide the balanced dataset into three groups: a training set (138,122 samples), a validation set (15,349 samples), and a test set (15,365 samples). The 50% representation of each group was maintained across all groups. The final dataset distribution is summarized in **Table 1**. In addition, the dataset provides formal test sets dedicated to each language, used to assess performance separately for each language, including: Arabic (463 samples), Turkish (5,397 samples), and English (5,906 samples), as shown in **Table 2**.

Table 1. The dataset distribution

Subset	Non-hate samples	Hate samples	Total
Training	69,061	69,061	138,122
Validation	7,674	7,675	15,349
Test	7,683	7,682	15,365

Table 2. Official language test sets

Language	Total
Arabic	463
Turkish	5,397
English	5,906

Preprocessing: Preprocessing was applied consistently across the three languages using a five-stage pipeline. In the first stage, null records were identified and deleted, reducing the dataset from 219,981 to 219,508 samples (473 samples deleted). In the second stage, all text was converted to string format, removing extra spaces at the beginning and end of the text. In the third stage, URLs were removed using regular expressions. In the fourth stage, spacing was standardized by combining spaces and consecutive lines into a single space. In the fifth stage, character-level filtering was applied to retain only Unicode ranges associated with the target languages. These ranges include Arabic characters (U+0600-U+06FF), expanded Latin characters covering Turkish characters such as ç, ğ, ı, ö, ş, and ü (U+00C0-U+024F), and standard English letters and numbers, while all other special characters and symbols were removed. The average token length remained relatively constant before and after processing, at 18.66 tokens before processing and 18.97 tokens after processing per sample, with a maximum of 2,909 tokens and a minimum of 1 token. Finally, the labels were standardized into a binary format, so that the value 0 represents texts not classified as hate speech, and the value 1 represents hate speech, resulting in a data distribution of 135,090 non-hate samples (61.54%) and 84,418 hate samples (38.46%).

Baseline models: Classical machine learning models were implemented as a benchmark. TF-IDF vectorization was used to represent text, followed by classification with logistic regression, support vector machines (SVM), and random forest. Their outputs were combined through a soft-voting ensemble. The ensemble designed was primarily as a baseline for comparison with the transformer approach.

Transformer models: For the main structure, the XLM-RoBERTa transformer was fine-tuned on the balanced dataset. Training was conducted on an NVIDIA A100 GPU with optimized hyperparameters to prevent overfitting while maintaining generalization capability. **Table 3** summarizes the key training configurations and deployment specifications. The trained model was exported in ONNX format, ensuring portability and efficient inference across various deployment environments, including resource-constrained mobile devices.

Table 3. The key training configurations and deployment specifications

Parameter	Value/description
Model	XLM-RoBERTa-base
GPU	NVIDIA A100
Learning rate	1e-5
Max_length	128
Batch size	16
Epochs	8
Scheduler	linear_warmup_decay
Weight_decay	0.01
Early_stopping	patience=3 on val_loss
Random_seed	42
Num_labels	2
Dropout	0.1
Optimization	AdamW optimizer
Deployment target	Cross-platform mobile applications

XLM-RoBERTa: Cross-lingual language model-robustly optimized BERT approach, BERT: Bidirectional encoder representations from transformers, GPU: Graphics processing unit

Deployment: To ensure real-world applicability, deployment was emphasized as a central contribution. The ONNX-optimized transformer model enabled efficient inference with reduced latency, achieving real-time predictions in 22 ms on CPU. A lightweight API was developed for integration into applications, and an Android mobile app was built in Kotlin with a modern interface, confidence visualization, and network status indicators. This dual deployment API for scalability and Android for mobile accessibility demonstrates the feasibility of bridging state-of-the-art multilingual NLP research with practical solutions for multilingual content moderation.

RESULTS

The experimental evaluation was performed on the balanced internal test set of 15,365 samples. **Table 4** summarizes the results. The ensemble baseline achieved a macro F1-score of 0.7796, slightly outperforming individual classifiers. The multilingual BERT baseline (mBERT) achieved an F1-score of 0.8356 and accuracy of 83.57%. These results demonstrate the advantage of contextual embeddings in capturing multilingual hate speech semantics compared to TF-IDF-based approaches. The fine-tuned XLM-RoBERTa model achieved the highest performance with an F1-score of 0.8493 and accuracy of 84.93%, confirming its superiority for multilingual hate speech detection. Early stopping was triggered at epoch 6 for mBERT, while XLM-RoBERTa continued to improve through the full training cycle.

Table 4. The comparative performance of baseline models and the transformer model

Model	F1 macro	Accuracy	Precision	Recall
Ensemble baseline (LR+SVC+RF)	0.7796	0.7797	0.7801	0.7797
mBERT	0.8356	0.8357	0.8367	0.8357
XLM-RoBERTa (ours)	0.8493	0.8493	0.8493	0.8493

mBERT: Multilingual BERT, XLM-RoBERTa: Cross-lingual language model-robustly optimized BERT approach, BERT: Bidirectional encoder representations from transformers

To provide a more detailed evaluation, the transformer model was tested separately on the language-specific test sets provided by the FrancophonIA dataset, which includes Arabic (463 samples), Turkish (5,397 samples), and English (5,906 samples), totaling 11,766 samples. **Table 5** presents the results for each language individually. In English, the model achieved an F1 score of 0.9395 and an accuracy of 93.97%, reflecting strong performance in the language with the most readily available data. In Turkish, the F1 score was 0.8205 and the accuracy was 89.12%. Arabic performed relatively less well, with an F1 score of 0.7628 and an accuracy of 76.89%, attributed to the small size of the test group (463 samples) and the complexity associated with the diversity of Arabic dialects.

Table 5. Performance of the fine-tuned XLM-RoBERTa across languages

Language	Precision	Recall	F1 macro	Accuracy
Arabic	0.7612	0.7761	0.7628	0.7689
Turkish	0.8389	0.8055	0.8205	0.8912
English	0.9388	0.9405	0.9395	0.9397

XLM-RoBERTa: Cross-lingual language model-robustly optimized BERT approach, BERT: Bidirectional encoder representations from transformers

Confusion matrices for each language are shown in **Figure 2**, demonstrating that the model maintains a high capacity for recognizing non-hate speech texts across all languages, although its sensitivity to hate speech varies, particularly in Arabic and Turkish.

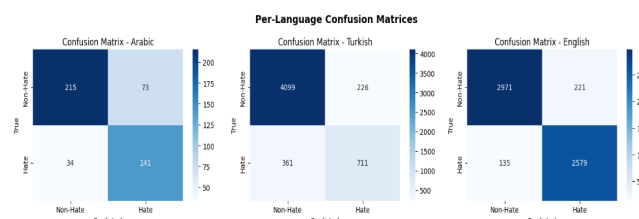


Figure 2. The confusion matrices for per-language using XLM-RoBERTa model

XLM-RoBERTa: Cross-lingual language model-robustly optimized BERT approach, BERT: Bidirectional encoder representations from transformers

In addition to the quantitative evaluation, the deployment aspect was validated through real-world testing. The fine-tuned XLM-RoBERTa model achieved an average inference latency of 9.26 ms per sample (± 7.14 ms) on GPU using PyTorch. Following ONNX export, the model was served through a lightweight API server, achieving real-time inference with response times below one second. The ONNX exported model has a file size of approximately 1.061 MB, comparable to the PyTorch model, while enabling CPU-only inference without requiring a GPU. Inference latency measurements, as shown in **Table 6**, revealed that the PyTorch model achieved a latency of 2.62 ms per sample when running on the A100 GPU, while the ONNX model achieved a latency of 22.41 ms per sample.

when running on the CPU without requiring GPU input. Despite the different operating environments, the ONNX model maintains real-time inference capability (less than 25 ms), highlighting its suitability for applications in resource-constrained environments such as mobile devices. To further demonstrate usability, an Android mobile application was developed with a simple interface for multilingual text input and real-time classification. **Figure 3** shows sample screenshots of the application, where the system successfully identifies hate speech across English, Turkish, and Arabic examples while displaying confidence scores.

Table 6. PyTorch vs ONNX: latency and model size comparison

Format	Hardware	Avg latency	Model size
PyTorch	GPU (A100)	2.62 ms	1,100 MB
ONNX	CPU	22.41 ms	1,061 MB

ONNX: Open neural network exchange, GPU: Graphics processing unit, CPU: Central processing unit, Avg: Average, ms: Milliseconds, MB: Megabytes

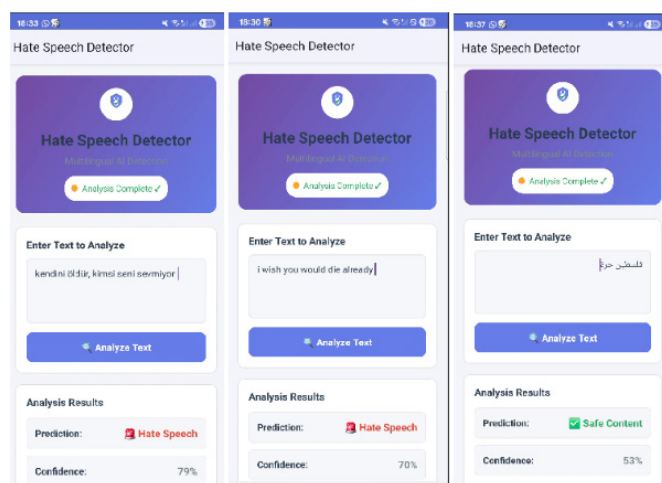


Figure 3. Sample screenshots from the application

Together, these results confirm that while ensemble baselines provide a reasonable benchmark, the transformer-based solution offers superior accuracy and practical deployability in both cloud and mobile environments.

DISCUSSION

The experimental results highlight the superiority of transformer-based models compared to classical machine learning approaches. While the ensemble baseline achieved moderate performance with a macro F1-score of 0.7796, the fine-tuned XLM-RoBERTa model substantially improved the results, achieving a macro F1-score of 0.8493 and an accuracy of 84.93%. This highlights how accurately contextual embeddings capture the complexity of multilingual hate speech. For comparison, the mBERT model achieved an overall F1 score of 0.8356, confirming that the broader multilingual pre-training of the XLM-RoBERTa model provides a significant advantage over the standard multilingual BERT model.

The evaluation for each language revealed a marked variation in performance among the three target languages. English achieved the best results with an F1 score of 0.9395 and an accuracy of 93.97%, while Turkish showed average performance F1=0.8205, accuracy 89.12%. Arabic showed relatively lower results F1=0.7628, accuracy=76.89%, which

can be attributed to two factors: the small size of the formal Arabic language test group (463 samples) and the inherent complexity of the diversity of Arabic dialects, for which the model was not specifically trained.

The strength of the study lies in its coverage of three languages (Arabic, Turkish, and English) within a unified model, with a significant reduction in false negatives thanks to data balancing and preprocessing. Furthermore, the successful ONNX export enabled real-time inference of 22.41 ms per sample on the CPU. This confirms its practical deployment on resource constrained devices without relying on the GPU. The Android application also underscores the system's overall feasibility for managing multilingual content in the real world.

Limitations

However, some limitations remain. The dataset does not adequately reflect the diversity of Arabic dialects, as Egyptian, Levantine, and Gulf dialects differ significantly from the patterns observed in the training data. Furthermore, the system is limited to text only and does not support multimedia speech that combines images or audio with text, which is increasingly common on social media platforms. Social media language also evolves over time, meaning the model's performance may decline without periodic retraining.

Overall, the results demonstrate that transformer-based architectures constitute a robust and deployable solution for multilingual hate speech detection, providing a solid foundation for future developments in practical content moderation systems.

CONCLUSION

This study presents a multilingual Android-based system for detecting hate speech, covering Arabic, Turkish, and English. Based on experimental results, the modified XLM-RoBERTa transformer performed significantly better than traditional baselines, achieving an overall F1-score of 0.8493 and an accuracy of 84.93%, while the baseline achieved an overall F1-score of 0.7796. These results confirm the effectiveness of contextual embeddings in dealing with multilingual hate speech detection compared to traditional feature-based methods. In addition to the model's performance, the system was successfully deployed with ONNX export and Android app, ensuring instant inference with lightweight deployment. This highlights the feasibility of linking advanced natural language processing research with deployable solutions for online content management.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: STD, İAD, KK; Design: STD, İAD, KK; Control: STD, İAD, KK; Resources: STD, İAD, KK; Materials: STD, İAD, KK; Data Collection and/or Processing: STD, İAD, KK; Analysis and/or Interpretation: STD, İAD, KK; Literature Review: STD, İAD, KK; Writing the Article: STD, İAD, KK; Critical Review: STD, İAD, KK.

REFERENCES

- Ahmad, A., Azzeh, M., Alnagi, E., Abu Al-Haija, Q., Halabi, D., ..., & AbuHour, Y. (2024). Hate speech detection in the Arabic language: corpus design, construction, and evaluation. *Frontiers in Artificial Intelligence*, 7, 1345445. <https://doi.org/10.3389/frai.2024.1345445>
- Barkhordar, E., Topçu, I. S., & Hürriyetoglu, A. (2024). Team Curie at HSD-2Lang 2024: hate speech detection in Turkish and Arabic tweets using BERT-based models. In A. Hürriyetoglu, H. Tanev, S. Thapa, & G. Uludoğan (Eds.), *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)* (pp. 215-220). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.case-1.30>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ..., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schuster, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440-8451). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.747>
- European Union Agency for Fundamental Rights. (2023). Online content moderation: current challenges in detecting hate speech. Publications Office of the European Union. <https://doi.org/10.2811/923316>
- Mahajan, E., Mahajan, H., & Kumar, S. (2024). EnsMulHateCyb: multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications*, 236, 121228. <https://doi.org/10.1016/j.eswa.2023.121228>
- Mnassri, K., Farahbakhsh, R., Chalehchaleh, R., Rajapaksha, P., Jafari, A. R., ..., & Crespi, N. (2024). A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10, e1934. <https://doi.org/10.7717/peerj-cs.1934>
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 3118-3135). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.243>
- Usman, M., Ahmad, M., Tash, M. S., Gelbukh, I., Tellez, R. Q., & Sidorov, G. (2025). Multilingual hate speech detection in social media using translation-based approaches with large language models. arXiv. <https://doi.org/10.48550/arXiv.2506.08147>
- Vogels, E. A. (2021, January 13). The state of online harassment. Pew Research Center. <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/>