Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Semantic Similarity for Zero-Shot Hate Speech Detection in Low-Resource Languages

[1]Ghadeer Al-Badani , [2]Muneer Alsurori, [3]Akram Alsubari

[1 2 3]Department of Computer Science and IT, Faculty of Science, University of Ibb, Yemen

Email: [1]ghadeeralbadani2023@gmail.com

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Multilingual Hate Speech Detection, Cross-Lingual Transfer Learning, Low-Resource Languages, Zero-Shot Learning, Language Similarity.

Abstract

This study investigates the role of semantic similarity in improving zero-shot and cross-lingual hate speech detection across several low-resource and typologically diverse languages, including Arabic, Hebrew, Persian, Russian, Chinese, Korean, and Amharic. A novel framework is proposed that clusters languages based on semantic and genealogical similarity using multilingual sentence embeddings derived from the XLM-R model. Each cluster was used to fine-tune the multilingual mDeBERTa-v3-base-mnli-xnli model, which was then evaluated in zero-shot settings on unseen languages within and across clusters. The results show that zero-shot transfer is highly influenced by linguistic proximity: models trained on Semitic languages (Arabic–Hebrew–Amharic) achieved strong zero-shot performance with F1 scores between 0.80 and 0.86 on unseen languages, while the Indo-European cluster (Russian–Persian) yielded competitive results (F1 ≈ 0.71–0.80). Training on typologically distant East Asian languages (Chinese–Korean) also demonstrated effective zero-shot generalization (F1 ≈ 0.75). Moreover, incorporating a newly developed Yemeni Arabic hate speech dataset enhanced Arabic performance and improved zero-shot transfer to related languages. These findings highlight the significance of semantic similarity in facilitating cross-lingual generalization and offer a scalable strategy for multilingual hate speech detection in low-resource settings.

1. Introduction

In recent years, social media platforms have transformed global communication by enabling users to rapidly share opinions, emotions, and personal experiences. This shift has provided valuable opportunities for researchers to analyze online discourse and public sentiment surrounding social and political issues. However, these same platforms have also facilitated the spread of harmful and offensive content, particularly hate speech [1].

Hate speech is defined as a toxic form of expression that targets individuals or groups based on intrinsic attributes such as race, religion, nationality, language, or ethnicity [2]. Its most alarming characteristic is the potential to incite real-world violence and discrimination, especially when directed at collective identities. Despite significant progress in automated detection, the majority of existing research has focused on high-resource languages, leaving low-resource languages largely unexplored [3]. Detecting hate speech in these languages remains crucial due to their wide presence on social media and the lack of sufficient annotated data for model training.

The increasing recognition of the societal harm caused by online hate speech [4] has encouraged the Natural Language Processing (NLP) community to propose diverse detection

approaches, ranging from traditional machine learning to advanced deep neural architectures. While early studies focused primarily on English or other high-resource monolingual settings, recent research has increasingly shifted toward multilingual and cross-lingual approaches to better capture the linguistic diversity of online communication [5]. Nevertheless, this expansion faces persistent challenges such as limited labeled corpora, cultural variation in offensive language, and differing linguistic structures across languages [6].

Several pioneering works have addressed monolingual hate speech detection in specific low-resource languages. For instance, research on Chinese [7] revealed challenges in annotation quality and cultural context sensitivity. In Russian, a dedicated dataset was developed to handle subtle negativity and ethnic bias issues [8]. In the Korean context, researchers fine-tuned several pre-trained models on a dataset of 20,000 political news comments, identifying KcELECTRA-base-v2022 as the most effective model for offensive language classification [9].

To address the limitations of scarce annotated data, zero-shot and cross-lingual transfer learning have emerged as promising paradigms. These approaches leverage multilingual pre-trained language models such as XLM-RoBERTa and mBERT to transfer knowledge from high-resource to low-resource languages [10]. Recent advancements further introduced complementary techniques, including data augmentation via machine translation and cross-lingual contrastive learning, to improve generalization and robustness [11].

However, selecting optimal source languages for transfer learning remains a critical open problem. Current practices often rely on intuition or linguistic family membership [12, 13], which does not always guarantee effective transfer due to intra-family structural variation [14]. While multi-source transfer methods have shown improved performance [15], they often overlook actual linguistic relationships and individual language contributions. Recent studies emphasize that linguistic and semantic similarity significantly influence cross-lingual transfer effectiveness, as morphosyntactic ally or genetically related languages tend to yield superior results [16].

This study contributes to multilingual hate speech detection by exploring under-researched low-resource languages from diverse linguistic families, with particular emphasis on Arabic through the inclusion of a Yemeni dialect dataset. It introduces a semantic clustering framework that links linguistic structures with semantic similarity, enabling systematic clustering of related languages. Furthermore, it analyzes transfer learning performance within and across these clusters, offering new insights into how linguistic and semantic proximity enhance zero-shot generalization across unseen languages.

The remainder of this paper is structured as follows. Section 2 reviews related work on cross-lingual and zero-shot hate speech detection. Section 3 describes the proposed methodology, including dataset construction, preprocessing, semantic similarity measurement, and experimental design, and discusses the experimental results, while Section 4 concludes the study and outlines future research directions.

2. Related Work

The detection of hate speech and offensive language in low-resource languages has witnessed significant advancements, driven primarily by cross-lingual machine learning techniques. Transferring knowledge from high-resource languages to their low-resource counterparts has become a dominant paradigm in this field, aiming to overcome the scarcity of annotated data. Previous research efforts within this framework can be categorized into several key methodological approaches, which are surveyed in this section.

Recent advances in hate speech detection have increasingly emphasized zero-shot and few-shot learning, aiming to extend model generalization to low-resource languages without requiring extensive labeled data. A large-scale study [17] trained models on English, Hindi, Urdu, and Bengali using more than twenty datasets, achieving a Macro-F1 of 79.62% on unseen Hindi and code-mixed Hindi–English data—demonstrating the effectiveness of multilingual pre-training in cross-lingual transfer. Similarly, [18] adapted the Stormfront English dataset for zero-shot evaluation on German (GermEval 2018), obtaining a Macro-F1 of 50.46% and further improvements through bootstrapping with unlabeled data, highlighting the importance of dataset alignment and label consistency in cross-lingual setups.

In another context, [19] investigated Turkish social media posts during the 2023 Türkiye–Syria earthquake, where zero-shot models like BERTurk achieved strong results (balanced accuracy = 0.814, F1 = 0.834) even without fine-tuning, indicating their robustness in real-world crisis scenarios. The potential of prompt-based zero-shot inference was further validated by [6] using large language models (e.g., FLAN-T5), which effectively distinguished between “respectful” and “toxic” content through verbalizer-driven templates.

To enhance cross-lingual generalization, several frameworks have been proposed. A teacher–student pseudo-labeling approach [10] improved performance by 7.6%, while HateMAML [20] employed meta-learning to outperform standard fine-tuning by up to 11%. Additionally, [21] leveraged contrastive learning and data augmentation to improve zero-shot transfer, and [22] integrated semi-supervised GANs with pre-trained models, achieving a 9.23% gain in F1-score with minimal labeled data

Beyond pure zero-shot settings, a substantial body of research has explored supervised and data-centric cross-lingual transfer approaches, emphasizing the creation, selection, and alignment of multilingual datasets to enhance performance across low-resource languages. For instance, the study NLPDove at SemEval-2020 Task 12 [23] demonstrated that careful data selection using Translation Embedding Distance (TED) can significantly improve transferability. By filtering highly transferable instances from English and Arabic datasets, the model achieved an impressive F1-score of 0.84 on Danish, showcasing the benefits of targeted source–target data alignment. Similarly, [24] proposed a Cross-Lingual Capsule Network (CCNL-Ex) that constructed parallel corpora through machine translation and processed both original and translated texts simultaneously. Their architecture attained state-of-the-art F1 scores (0.519–0.736) across English, Spanish, and Italian, outperforming strong baselines such as mBERT and XLM-R. Addressing the complexity of code-switched data, [25] introduced synthetic data generation techniques (SWAP and REWRITE) to augment multilingual datasets. Their XLM-RoBERTa model trained on synthetic code-switched samples achieved F1 = 0.67 on English–German mixed data, underscoring the effectiveness of data synthesis for handling intricate multilingual scenarios.

Parallel to data-centric strategies, recent research has delved into the linguistic and strategic foundations of transfer learning, examining how structural and typological properties of languages influence transfer success. A notable study by [3] systematically evaluated zero-shot transfer across seven languages from three families (Germanic, Slavic, Korean–Japonic), revealing that linguistic similarity strongly correlates with transfer performance—though it is not the sole determinant. Interestingly, Russian emerged as the most effective source language, even outperforming English when transferring to typologically distant languages such as Japanese and Korean. This observation aligns with findings from [3], who demonstrated that source language selection based on structural similarity metrics (e.g., linguistics, WALS-based indices) yields superior zero-shot performance compared to heuristic or family-based selection.

Further advances have integrated external linguistic knowledge into cross-lingual models. The Joint-Learning MUSE framework proposed by [26], which incorporates features from the HurtLex lexical resource, achieved stable results across six target languages, indicating that embedding external knowledge can mitigate model-only limitations. Complementary research in Indic languages has confirmed these trends: [27] found that MuRIL outperformed mBERT, and that zero-shot transfer was most successful between linguistically related languages, while emphasizing the necessity of limited gold target data for maximal effectiveness. Likewise, [28] observed that while multilingual models excelled for certain Indian languages (Hindi, Bangla, Bodo), monolingual models were superior for others (Marathi, Assamese). Their key insight—that script similarity (e.g., Devanagari) can drive transfer success more effectively than genetic family relations—offers a valuable linguistic perspective for advancing cross-script and zero-shot transfer in severely low-resource contexts.

Collectively, these studies demonstrate that successful cross-lingual transfer relies not only on data volume and model architecture but also on strategic data selection, synthetic augmentation, and linguistically informed pairing of source and target languages, providing a foundation for more adaptive and interpretable multilingual hate speech detection systems.

3. Methodology

This section follows the methodology shown in the figure1.

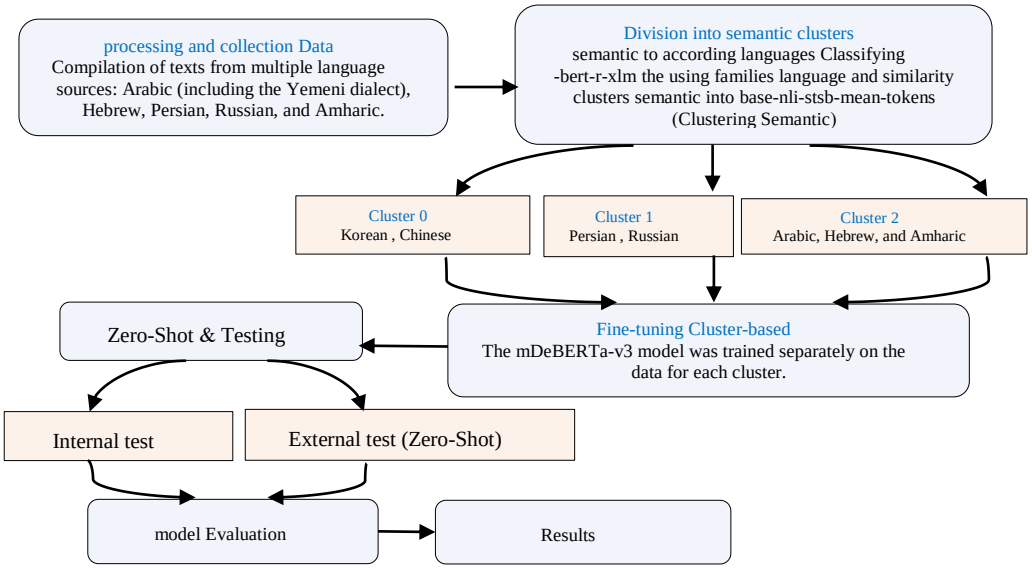

Fig. 1. Proposed Methodology for Semantic Cluster-Based Hate Speech Detection Across Languages

3.1 Data Collection and Description

In this section, the data sets used, their sources and characteristics are discussed.

Dataset 1: Researchers [29] constructed a multilingual dataset for detecting hate speech. They combined data from 24 different sources in 14 languages, including Arabic, English, Spanish, French, Turkish, German, Russian, Bengali, Chinese, Danish, Dutch, Portuguese, Korean, and Indonesian, they performed a binary classification: 1 = hate speech and 0 = normal. While some of the original groups were detailed (such as racism, religious discrimination, and misogyny), they were combined under the "hate speech" category to increase generalization and consistency. In this research, we used this dataset for the languages (Arabic, Russian, Chinese and Korean). The link to the dataset is here1.

Dataset 2: We then searched for other datasets for hate speech and for different languages (Hebrew, Persian, Hebrew, and African languages (Amharic), as shown in the following table1.

Table 1. The Linguistic Composition of the Dataset1: Training and Testing Set Sizes.

No	Language	Training data size	Test data size
1	Arabic	4155	463
2	Chinese	8065	897
3	Russian	14494	1611
4	Korean	7104	791
Total		33818	3762

Hebrew Dataset The Hebrew dataset was obtained from two sources. The first is a dataset compiled by the authors [30]. This dataset was compiled from Hebrew forums using a dictionary of offensive keywords. Each sample was manually identified by native speakers as abusive or non-offensive, with tags for severity and type of abuse (insults, threats, sexual profanity, hate speech). This dataset is one of the first large-scale resources for detecting offensive Hebrew. The link to the dataset2 , from which we obtained the training set (1750) and the validation set (250). The dataset consists of two classes (positive and negative). As for the second source, the following table shows the size of the dataset, its size is 5217. We took it from the link3 .

Persian Dataset The researchers presented a new dataset called PHATE [31], specifically designed to detect multi-category hate speech in Persian tweets. PHATE consists of over 7,000 manually annotated tweets, each of which

identifies the target category of hate speech and includes a rationale for assigning the category. By incorporating this additional information, PHATE facilitates the detection of targeted online harm and constitutes a valuable resource for research into the interpretability of hate speech detection models.

Researchers [32] created AfriHate, a multilingual dataset of hate speech and abusive language, covering 15 African languages. Each example in the dataset is a tweet with comments from native speakers with a sociocultural understanding of the context and language, addressing the crucial need for localized and community-driven moderation resources. The dataset consists of two categories: hate and non-hate. For this study, we selected only language Amharic.

Table 2. Dataset 2 Linguistic Structure: Class Sizes.

No	**Language**	**Training data size**	**Test data size**
1	Arabic	4155	463
2	Chinese	8065	897
3	Russian	14494	1611
4	Korean	7104	791
Total		33818	3762

Hebrew Dataset The Hebrew dataset was obtained from two sources. The first is a dataset compiled by the authors [30]. This dataset was compiled from Hebrew forums using a dictionary of offensive keywords. Each sample was manually identified by native speakers as abusive or non-offensive, with tags for severity and type of abuse (insults, threats, sexual profanity, hate speech). This dataset is one of the first large-scale resources for detecting offensive Hebrew. The link to the dataset4 , from which we obtained the training set (1750) and the validation set (250). The dataset consists of two classes (positive and negative). As for the second source, the following table shows the size of the dataset, its size is 5217. We took it from the link5 .

Persian Dataset The researchers presented a new dataset called PHATE [31], specifically designed to detect multi-category hate speech in Persian tweets. PHATE consists of over 7,000 manually annotated tweets, each of which identifies the target category of hate speech and includes a rationale for assigning the category. By incorporating this additional information, PHATE facilitates the detection of targeted online harm and constitutes a valuable resource for research into the interpretability of hate speech detection models.

Researchers [32] created AfriHate, a multilingual dataset of hate speech and abusive language, covering 15 African languages. Each example in the dataset is a tweet with comments from native speakers with a sociocultural understanding of the context and language, addressing the crucial need for localized and community-driven moderation resources. The dataset consists of two categories: hate and non-hate. For this study, we selected only language Amharic.

Table 2. Dataset 2 Linguistic Structure: Class Sizes.

Language	Class		Total
	non-hate speech	hate speech	
Amharic	1359	3599	4958
Persian	3860	3196	7056
Hebrew	4176	3030	7206
Arabic-Yemeni	4,466	4,203	8,669

Arabic-Yemeni dataset in order to increase data in Arabic and balance the categories This study introduces the first dedicated dataset for hate speech detection in the Yemeni Arabic dialect, a notably under-resourced variant in Natural Language Processing (NLP). To ensure representativeness and accuracy, a hybrid data collection methodology was employed, gathering

content from Twitter (X) and YouTube using specialized tools such as TwExtract and custom web-scraping scripts. To maintain contextual relevance, a keyword list derived from local Yemeni discourse was utilized, and non-Yemeni comments were filtered out with the assistance of linguistic experts. The collected data subsequently underwent a rigorous manual annotation process by three independent annotators, with a majority-voting mechanism adopted to ensure label reliability. The final dataset comprises 8,670 balanced instances between hate speech and non-hate speech categories (4,203 hate samples vs. 4,466 non-hate samples). This dataset holds significant importance as it captures the unique linguistic and cultural characteristics of the Yemeni dialect, including non-standard orthography and idiomatic expressions with offensive connotations specific to the Yemeni context. It thus serves as a valuable resource for developing more accurate and equitable hate speech detection models for marginalized Arabic dialects.

In this work, the Yemeni Arabic dataset and the Arabic dataset shown in Table 1 were combined.

3.2 Data Preprocessing

Data preprocessing is a critical step in the hate speech detection pipeline, significantly impacting model performance and reliability. In this study, we applied a comprehensive preprocessing framework to ensure data quality and consistency across the multilingual corpus. See Figure 2 for an overview of the complete preprocessing pipeline. These preprocessing steps are essential for reducing noise, minimizing overfitting, and enhancing the model's ability to generalize across diverse linguistic patterns, thereby improving the accuracy and robustness of the hate speech classification system.

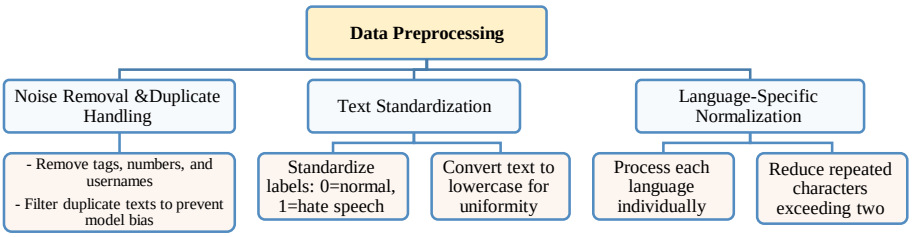

Fig. 2. Data preprocessing framework

3.3 Measuring Cross-Lingual Similarity Using Multilingual Embedding

This study implemented a methodology for calculating semantic similarity between languages using multilingual sentence embedding models to develop an advanced framework for cross-lingual hate speech detection. The research employed the xlm-r-bert-base-nli-stsb-mean-tokens6 model based on the XLM-RoBERTa architecture, which was trained on 100 languages through Natural Language Inference (NLI) and Semantic Textual Similarity (STS) tasks [33]. To enhance linguistic representation accuracy, rich informational textual descriptions were created for each language, including the official name, language family, and geographical region. For instance, Arabic was represented as "Arabic language Afro-Asiatic family spoken in Middle East." This representation is used as input to the model in order to extract a semantic vector that reflects the language properties in the semantic space of the model. The model then generated 768-dimensional embedding vectors for each language. Similarity matrices were calculated using cosine similarity between all language pairs according to the equation:

$$\text{Similarity}(A, B) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \quad (1)$$

where $\vec{A}$ and $\vec{B}$ represent the embedding vectors of languages A and B, respectively. The results revealed language clusters consistent with traditional linguistic classification, with the highest similarities observed between languages belonging to the same language family. For example, Semitic languages (Arabic, Hebrew, Amharic) showed similarities ranging between 0.65 and 0.85, while similarities between languages from different families were significantly lower (0.20-0.45) See Figure 3. For the application in multilingual hate speech detection, the proposed framework leverages linguistic similarity analysis to build a hate speech detection system operating through zero-shot transfer learning across languages, based on the principles of semantic proximity and cross-lingual model adaptation

Following the calculation of semantic similarity between languages, they were divided into clusters based on their linguistic families and

typological features. This resulted in three primary clusters:

Cluster 0: This cluster grouped together Chinese (a Sino-Tibetan language) and Korean (a Koreanic language), reflecting their typological and semantic proximity despite belonging to different language families.

Cluster 1: This cluster comprised languages from different families that showed a significant semantic affinity, grouping Persian (an Indo-European language) with Russian (also from the Indo-European family)

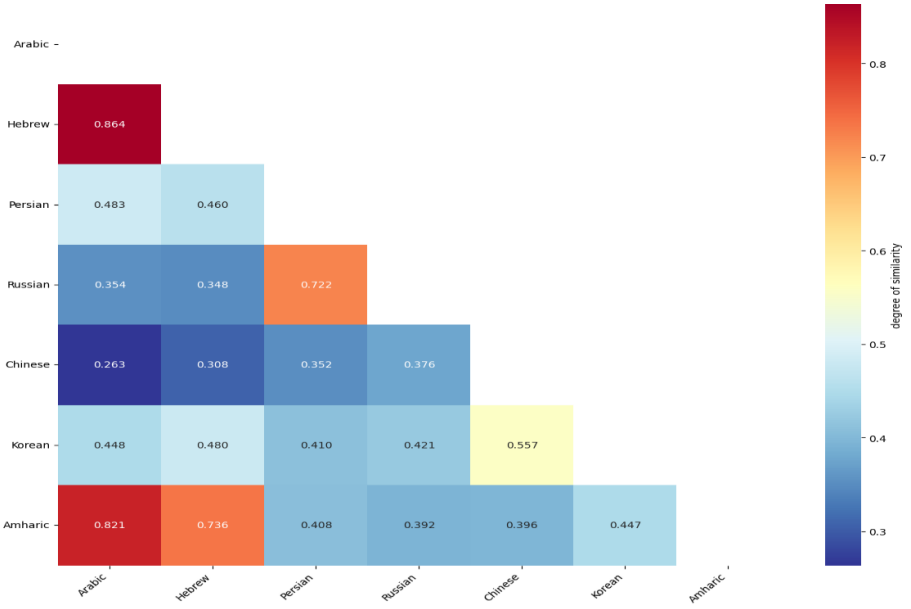

Fig. 3. Semantic similarity graph derived from XLM-R language embeddings. Edge weights represent cosine similarity. Stronger connectivity correlates with higher zero-shot transfer performance.

Cluster 2: This cluster included languages with strong historical and linguistic ties from the Afro-Asiatic family, namely Arabic, Hebrew, and Amharic.

This division demonstrates that semantic similarity does not always conform to the traditional family classification of languages, as languages from different families (such as Persian and Russian, or Chinese and Korean) can form coherent groups based on the criteria used in the analysis.

A review of previous work on cross-lingual transfer and hate speech detection (e.g., [3], [17], [18]) shows that most studies relied on evaluating the practical performance of models to estimate language similarity indirectly, using metrics such as F1-score or accuracy. In contrast, our methodology measures language similarity directly in the semantic space of multilingual models (XLM-RoBERTa) using the language embeddings themselves, providing a direct quantitative analysis of linguistic representations. Moreover, we introduced the integration of linguistic descriptions (language name, family, and geographic region) into the embedding generation process, which enhances language representation and reduces bias caused by varying textual data. Additionally, our approach combines quantitative and visual analyses through tools such as Cosine Similarity, Heatmaps, Network Graphs, and Hierarchical Clustering—a combination rarely employed in previous studies. Finally, this approach focuses on semantic similarity as perceived by the multilingual model, rather than traditional structural similarity, making it more relevant to cross-lingual transfer tasks and offering a novel and unprecedented method for measuring language similarity.

3.4 Models

In this section, the multilingual model used in the tuning process for transfer learning for languages unseen during training is discussed.

mDeBERTa-v3-mnli-xnli: The mDeBERTa-v3-mnli-xnli model is a multilingual variant of DeBERTa (Decoding-enhanced BERT 7 with Disentangled Attention), designed for advanced natural language understanding and particularly

effective in hate speech detection. Trained on text from 100 languages, it captures deep grammatical and semantic relationships across languages through its disentangled attention mechanism, which separates word content from positional information. The model is fine-tuned on MNLI and XNLI datasets to enhance its ability to reason about contextual relationships between sentences. For hate speech detection, it employs zero-shot or few-shot learning, interpreting text as a logical inference problem—assessing whether a given statement (hypothesis) like "This text contains hate speech" follows from the text (premise). This reasoning-based approach enables effective cross-lingual transfer, allowing the model to detect hate speech in languages it has never been explicitly trained on by leveraging its multilingual and inferential capabilities [34].

Previous work has relied heavily on models such as mBERT and XLM-RoBERTa, providing an important research opportunity to explore the capabilities of a specialized and improved model such as mDeBERTa-v3 in detecting hate speech across languages. Therefore, our study fills this gap by presenting the first use (to our knowledge) of the MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model as a baseline tool for evaluating hate speech detection across languages (Arabic, Hebrew, Persian, Russian, Chinese, and Korean). The transferability of zero-learning across languages was not demonstrated during fine-tuning the model.

3.5 Statistical Analysis

Our statistical evaluation employed multiple hypothesis testing approaches. We utilized Chi-square tests on confusion matrices to assess classification significance, one-way ANOVA with Tukey HSD post-hoc comparisons for between-group accuracy analysis, independent t-tests for in-language versus cross-language performance comparison, and Pearson correlation coefficients to examine relationships between evaluation metrics. Advanced performance metrics including Matthews Correlation Coefficient (MCC) and balanced accuracy were computed to ensure comprehensive assessment.

4 Experiments

To evaluate the proposed multilingual framework and examine the extent of cross-lingual transfer in hate speech detection, a series of controlled experiments were conducted. The experiments in this work were designed based on the clustering of languages according to their semantic and family similarity. For each cluster, the model was fine-tuned on the languages within that cluster and subsequently tested on languages outside the cluster to evaluate cross-cluster transferability. In addition, the model was fine-tuned on specific languages from the same cluster and tested on unseen languages belonging to that cluster. This experimental design aimed to analyze the portability of learned representations both within and across clusters, thereby assessing how linguistic similarity and shared semantic structures influence the model's ability to generalize across languages.

The dataset was divided into 70% for training and 30% for validation to monitor model performance and prevent overfitting. To evaluate the models' ability to generalize across languages, this experimental design tested the hypothesis that fine-tuning on training languages enables the learning of language-independent NLI representations that transfer effectively. During training, model parameters were optimized using backpropagation to minimize cross-entropy loss on the training set. For evaluation, the models were tested on unseen languages to assess their cross-lingual generalization ability. This experimental framework aimed to examine whether the hate speech patterns learned from the training data could effectively transfer across language boundaries, with performance measured using standard classification metrics, including accuracy, recall, and F1-score across all target languages.

To improve model performance in detecting hate speech in Arabic, the following experimental configuration was adopted. The model was trained for two epochs with a learning rate of 2e-5, epsilon of 1e-8, batch size of 8, and a maximum sequence length of 512 tokens. All experiments were conducted on GPU 0, providing efficient computational resources for transformer-based architectures. The Adam optimizer was used with these parameters to ensure stable gradient updates and effective convergence during the short training period. The limited number of epochs (2) was strategically chosen to prevent overfitting while maintaining computational efficiency, which is crucial under limited-resource conditions. The batch size of 8 represented an optimal balance between memory constraints and training stability, while the maximum sequence length of 512 tokens accommodated most textual inputs without excessive truncation. This experimental configuration optimized model efficiency under resource constraints while maintaining effective cross-lingual transfer to the seven target languages in our evaluation.

5 Results and Discussion

In this section, we present the results of our experiments on multilingual hate speech

detection, discussing the effectiveness of our proposed model in cross-lingual transferability and the implications of semantic similarity on performance across different languages.

5.1 Training on Cluster-0 Languages and Cross-Lingual Transfer Results

To evaluate the generalizability of our language clustering approach, we conducted a series of zero-shot cross-lingual experiments using models trained exclusively on languages from Cluster-0. This evaluation aims to assess how effectively linguistic knowledge transfers from this semantically coherent cluster to both related and distant languages across different language families. The following section presents the quantitative results of these experiments as shown in Tables 3 and 4, comparing performance across single-language and joint training paradigms.

Table 3. Zero-Shot Cross-Lingual Evaluation Results of the Multilingual Model Trained on Chinese and Korean

Training Languages	Testing Language	Accuracy	F1-Score
Chinese + Korean	Chinese	0.79	0.79
	Arabic	0.75	0.75
	Korean	0.77	0.77
	Russian	0.75	0.74
	Hebrew	0.69	0.68
	Persian	0.61	0.60
	Amharic	0.71	0.71

Table 4. Zero-Shot Cross-Lingual Performance of Single-Language Training Models (Chinese vs. Korean) on Multiple Target Languages

Training Languages	Test Language	Accuracy	F1-Score
Chinese	Chinese	0.81	0.81
	Korean	0.54	0.44
Korean	Korean	0.73	0.73
	Russian	0.73	0.72
	Amharic	0.72	0.72
	Arabic	0.69	0.68
	Hebrew	0.68	0.68
	Persian	0.65	0.64
	Chinese	0.60	0.56

Analysis and Cross-lingual Transfer Interpretation

Figure 3 illustrates the similarity network between the training languages (Korean and Chinese) and the unseen testing languages. The numerical values indicate a moderate similarity between Korean and Chinese (0.557), suggesting a reasonable potential for shared semantic representations during joint training. Their similarity to other languages varies depending on linguistic family affiliation: Korean–Arabic (0.372), Korean–Hebrew (0.380), Korean–Amharic (0.379), while Chinese–Arabic (0.440), Chinese–Hebrew (0.453), and Chinese–Amharic (0.437). Persian and Russian exhibit moderate similarity values ranging between 0.355 and 0.408.

These structural distances are reflected in the model's zero-shot performance, where the model achieved the highest results on the training languages themselves—Chinese (Accuracy = 0.79, F1 = 0.79) and Korean (Accuracy = 0.77, F1 = 0.77). When evaluated on unseen languages, performance gradually decreased as linguistic distance increased: Arabic (Accuracy = 0.75, F1 = 0.75), Russian (Accuracy = 0.75, F1 = 0.74), Amharic (Accuracy = 0.71, F1 = 0.71), Hebrew (Accuracy = 0.69, F1 = 0.68), and Persian (Accuracy = 0.61, F1 = 0.60).

These findings indicate that the model benefited from shared representations between Korean and Chinese—two typologically distinct languages (Korean and Sino-Tibetan)—allowing it to capture partially language-agnostic features transferable across linguistic boundaries. However, the effectiveness of transfer remained strongly correlated with linguistic proximity, as languages exhibiting higher structural or lexical

similarity to the training languages (e.g., Russian and Arabic) yielded better zero-shot results compared to more distant languages (e.g., Persian and Hebrew).

Overall, training on two typologically diverse East Asian languages enabled the model to acquire partially generalized multilingual embeddings, but the limited zero-shot performance on unrelated languages highlights the importance of incorporating typologically and morphologically diverse training languages to achieve robust cross-lingual transfer.

The results reveal a striking asymmetry in the effectiveness of cross-lingual transfer learning, even between languages grouped within the same semantic cluster. When Chinese was used as the source language, the model failed to generalize effectively to its cluster partner, Korean (F1: 0.81 vs. 0.44), demonstrating that shared cluster membership does not guarantee robust cross-lingual performance, particularly in the face of significant script differences and typological diversity.

In contrast, the model trained exclusively on Korean demonstrated exceptional generalization capability, achieving strong and consistent results across a wide range of linguistically diverse languages, including Russian, Amharic, and Hebrew (F1 scores ranging from 0.72 to 0.68). Notably, its performance when transferring to Chinese (F1: 0.56) was superior to the reverse scenario. This suggests that Korean serves as a highly effective source language, likely due to superior pre-trained language model representations or training data that captures more universal hate speech patterns, making it a optimal choice for building a broadly applicable multilingual hate speech detection model.

5.2 Training on Cluster-1 Languages and Cross-Lingual Transfer Results

In this section, we divide the generalizability of our approach to language grouping. We conducted a series of cross-border language experiments using models trained exclusively on languages from cluster-1, as shown in Table 5.

Table 5. Cross-linguistic performance of monolingual (Russian vs. Persian and Russian+ Persian) training models using the Zero-Shot Cross-Lingual technique on multiple target languages

Training Languages	Target Language	Accuracy	Macro F1-Score
Persian	Russian	0.72	0.71
	Arabic	0.69	0.68
	Amharic	0.64	0.60
	Hebrew	0.58	0.51
	Chinese	0.56	0.47
	Korean	0.54	0.44
Russian	Russian	0.88	0.88
	Amharic	0.68	0.66
	Persian	0.66	0.66
	Arabic	0.64	0.61
	Korean	0.64	0.64
	Chinese	0.59	0.59
	Hebrew	0.54	0.53
Persian+ Russian	Amharic	0.68	0.67
	Arabic	0.75	0.75
	Korean	0.62	0.58
	Chinese	0.60	0.57
	Hebrew	0.70	0.70

Analysis and Cross-lingual Transfer Interpretation: The experimental results demonstrate significant variations in cross-lingual transfer learning performance across different language pairs. When trained solely on Persian, the model achieved its highest performance on Russian (Accuracy: 0.72, Macro F1: 0.71), indicating substantial linguistic transfer between these languages. Conversely, training exclusively on Russian yielded superior overall results, particularly on itself (Accuracy: 0.88, Macro F1: 0.88) and reasonable transfer to Persian (Accuracy: 0.66, Macro F1: 0.66). The combined Persian-Russian training paradigm proved most effective, substantially improving performance on several target languages. Notably, Hebrew showed remarkable improvement from 0.54 to 0.70 accuracy, while Arabic increased from 0.64 to 0.75 accuracy. However, Asian languages (Korean and Chinese)

consistently presented the greatest challenge across all training configurations, suggesting greater linguistic distance from the source languages. These findings underscore the importance of strategic language selection for training and highlight the diminishing returns of cross-lingual transfer as linguistic divergence increases.

5.3 Training on Cluster-2 Languages and Cross-Lingual Transfer Results

In this section, we divide the generalizability of our approach to language grouping. We conducted a series of cross-border language experiments using models trained exclusively on languages from cluster-1, as shown in Table 6.

Table 6. Cross-linguistic performance of monolingual (Arabic,Hebrew,Amharic) training models using the Zero-Shot Cross-Lingual technique on multiple target languages

Training Languages	Testing Language	Accuracy	F1-Score (Macro)
Arabic, Hebrew, Amharic	Arabic	0.84	0.84
	Hebrew	0.62	0.62
	Amharic	0.86	0.86
	Chinese	0.61	0.58
	Korean	0.63	0.60
	Persian	0.71	0.71
	Russian	0.80	0.80
Arabic	Hebrew	0.72	0.72
	Amharic	0.70	0.69
	Chinese	0.61	0.60
	Korean	0.62	0.59
	Persian	0.71	0.71
	Russian	0.82	0.82

Analysis and Cross-lingual Transfer Interpretation : The experimental results highlight a strong correlation between linguistic similarity and the effectiveness of cross-lingual transfer learning for hate speech detection. When the model was trained on Arabic, Hebrew, and Amharic—languages with high mutual similarity (e.g., Arabic-Hebrew: 0.864, Arabic-Amharic: 0.821)—it achieved high accuracy on these languages (0.84–0.86). In contrast, languages with lower similarity to the training set, such as Korean and Chinese (similarity scores ≤ 0.48), yielded lower performance (accuracy: 0.61–0.63). Notably, Russian—despite its relatively low similarity to Arabic (0.394)—achieved strong results (0.80 accuracy). These findings underscore the importance of leveraging linguistically related language groups to enhance zero-shot and few-shot hate speech detection in low-resource settings.

The experimental results demonstrate a strong relationship between linguistic similarity and the success of cross-lingual transfer learning in hate speech detection, with Arabic serving as a key transfer bridge. Training on the Yemeni Arabic dataset, both alone and alongside Hebrew and Amharic, resulted in significant performance metrics (0.80 accuracy on Russian and 0.71 on Persian). The dataset effectively generalizes to various languages despite their low similarity to Arabic, revealing its value for multilingual applications. These findings underscore the necessity for high-quality regional resources to enhance cross-lingual hate speech detection, particularly in low-resource contexts.

5.4 Cross-Cluster Transfer Efficiency Analysis

The cross-cluster transfer efficiency matrix In Figure 4 reveals asymmetric patterns in zero-shot performance across language clusters. When models trained on Cluster 0 (East Asian languages) were evaluated on Cluster 1 (Indo-European languages), they achieved an average F1-score of 0.67, while the reverse direction (Cluster 1 → Cluster 0) yielded only 0.57, indicating unidirectional transfer patterns. Notably, Cluster 2 (Semitic languages) demonstrated the most effective cross-cluster transfer, particularly to Cluster 1 (F1 = 0.77), suggesting that Semitic languages capture generalizable hate speech patterns applicable to Indo-European languages.

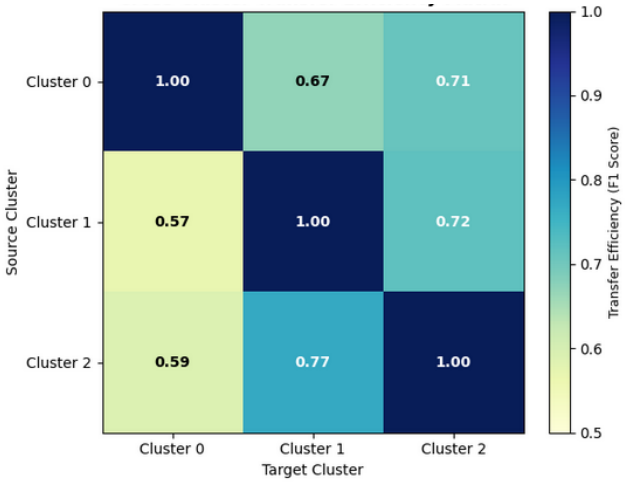

Fig. 4 Cross-Cluster Transfer Efficiency Matrix based on F1-Score

Statistical analysis showed significant differences in performance between in-language and cross-language tasks, with a t-value of 4.787 and a p-value of 0.0002. ANOVA results indicated no significant differences across various training configurations (F = 1.369, p = 0.259). All classification models assessed achieved statistical significance, evidenced by Chi-square tests with p-values below 0.001. A strong correlation (r = 0.992) was observed between accuracy and Matthews correlation coefficient (MCC), confirming the reliability of the metrics used. The lack of significant differences among multilingual training configurations suggests they can be strategically viewed as equivalent for cross-lingual transfer. Nonetheless, the notable performance gap between in-language and cross-language evaluations underscores the challenges associated with transfer learning. The strong relationship between conventional and advanced evaluation metrics further validates the evaluation framework employed.

5.5 Error Analysis

An error-oriented examination of the results presented in Tables 3–6 reveals that successful zero-shot transfer is primarily reflected in balanced class-wise detection performance rather than overall F1-scores alone. In semantically similar language pairs, positive transfer manifests through stable precision and recall for both hate and non-hate classes, suggesting that semantic proximity facilitates the transfer of robust semantic boundaries required for effective discrimination.

In contrast, performance degradation in linguistically distant languages is predominantly driven by asymmetric error patterns. Specifically, the confusion matrices[8] reveal a marked reduction in hate speech recall, while non-hate detection remains comparatively robust, leading to an increased rate of false negatives. These findings indicate that while semantic similarity supports cross-lingual generalization, the lexical and cultural realization of hate expressions plays a critical role in shaping error behavior and limiting effective knowledge transfer

5.6 Comparative Analysis with Previous Studies

To better contextualize our findings, the obtained zero-shot results were compared with previous cross-lingual hate speech detection studies that employed similar multilingual frameworks. This comparison highlights the relative improvements achieved by our semantically informed clustering approach over existing methods.

Recent zero-shot studies have shown the growing potential of multilingual models for hate speech detection across unseen languages. Kapil and Ekbal [17] demonstrated that multilingual pre-training enables effective transfer from high-resource to low-resource languages, achieving a Macro-F1 of 79.62% on unseen Hindi–English data. Similarly, Bigoulaeva et al. [18] obtained a Macro-F1 of 50.46% when transferring from English to German, highlighting the limitations of translation-based and bootstrapping methods. Eronen et al. [3] further emphasized that linguistic similarity plays a crucial role in cross-lingual transfer, showing that structurally related languages yield higher zero-shot accuracy. Building on these insights, our study extends the notion of similarity from structural to semantic

space by clustering languages using multilingual embeddings. This approach enabled the mDeBERTa-v3-mnli-xnli model to achieve superior zero-shot performance (F1 = 0.80–0.86) within semantically coherent clusters such as Arabic–Hebrew–Amharic, outperforming previous works [3,17,18] and confirming that semantic similarity provides a more predictive and interpretable foundation for cross-lingual hate speech detection.

6 Conclusion and Future Work

This study provides new insights into the role of linguistic and semantic similarity in cross-lingual transfer learning for hate speech detection. The results demonstrate that strategically clustering languages based on semantic similarity offers a reliable framework for predicting and interpreting transfer performance, particularly among low-resource and under-studied languages. Models fine-tuned on semantically related clusters—such as Arabic–Hebrew–Amharic (Afro-Asiatic) and Russian–Persian (Indo-European)—achieved strong mutual transfer performance, confirming that semantic proximity facilitates knowledge portability. However, the findings also reveal that transfer effectiveness is not solely determined by linguistic similarity but is mediated by the quality of source-language data and the representational power of the underlying pre-trained model. The robust performance observed across typologically diverse pairs, such as Chinese–Korean, further underscores the language-agnostic capabilities of the mDeBERTa-v3-base architecture, who's multilingual NLI pre-training enabled the extraction of deep, cross-lingual semantic representations.

Building upon these insights, several promising directions arise for future work. Expanding the framework to cover additional low-resource languages from diverse families would strengthen the empirical validation of semantic clustering. A systematic exploration of asymmetric transfer learning could optimize source-language selection for maximal generalization. Furthermore, integrating parameter-efficient adaptation techniques such as LoRA[9] or contrastive learning objectives may enhance cross-family transfer. Establishing robust data quality assessment metrics to detect and mitigate label noise—particularly in noisy datasets like Hebrew—would further improve model stability. Ultimately, developing a dynamic, adaptive multilingual system capable of automatically selecting optimal fine-tuning strategies based on each language's semantic position and resource availability remains an ambitious but essential step toward more equitable and scalable multilingual NLP.

References

1. A. Al-Hassan and H. Al-Dossari, Detection of hate speech in social networks: a survey on multilingual corpus, in 6th International Conference on Computer Science and Information Technology (ACM, New York, 2019), pp. 10-21.
2. A. Chhabra and D. K. Vishwakarma, A literature survey on multimodal and multilingual automatic hate speech identification, Multimedia Systems 29, 1203-1230 (2023).
3. J. Eronen, M. Ptaszynski, F. Masui, M. Arata, G. Leliwa, and M. Wroczynski, Transfer language selection for zero-shot cross-lingual abusive language detection, Information Processing & Management 59, 102981 (2022).
4. A. Jiang and A. Zubiaga, Cross-lingual offensive language detection: a systematic review of datasets, transfer approaches and challenges, arXiv preprint arXiv:2401.09244 (2024).
5. E. W. Pamungkas, V. Basile, and V. Patti, Towards multidomain and multilingual abusive language detection: a survey, Personal and Ubiquitous Computing 27, 17-43 (2023).
6. Gummadi, V. P. K. (2021). Streaming in Mule 4: High-volume processing. Journal of Information Systems Engineering and Management, 6(4), 1–9.
7. Y. Xiao, H. Bouamor, and W. Zaghouani, Chinese offensive language detection: current status and future directions, arXiv preprint arXiv:2403.18314 (2024).
8. E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, Detecting ethnicity-targeted hate speech in Russian social media texts, Information Processing & Management 58, 102674 (2021).
9. H. S. Ryu and J. K. Lee, Detection of political hate speech in Korean language, Language Resources and Evaluation 59, 1957-1988 (2025).
10. H. B. Zia, I. Castro, A. Zubiaga, and G. Tyson, Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models, in Proceedings of the International AAAI Conference on Web and Social Media (AAAI, 2022), Vol. 16, pp. 1435-1439.
11. L. Liu, D. Xu, P. Zhao, D. D. Zeng, P. J. H. Hu, Q. Zhang, and Z. Cao, A cross-lingual transfer

learning method for online COVID-19-related hate speech detection, Expert Systems with Applications 234, 121031 (2023).
12. E. P. Stabler and E. L. Keenan, Structural similarity within and among languages, Theoretical Computer Science 293, 345-363 (2003).
13. R. Cotterell and G. Heigold, Cross-lingual character-level neural morphological tagging, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, Copenhagen, 2017), pp. 748-759.
14. C. Gooskens, V. J. van Heuven, J. Golubović, A. Schüppert, F. Swarte, and S. Voigt, Mutual intelligibility between closely related languages in Europe, International Journal of Multilingualism 15, 169-193 (2018).
15. N. van der Heijden, H. Yannakoudakis, P. Mishra, and E. Shutova, Multilingual and cross-lingual document classification: a meta-learning approach, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics, 2021), pp. 1966-1976.
16. S. Gaikwad, T. Ranasinghe, M. Zampieri, and C. M. Homan, Cross-lingual offensive language identification for low resource languages: the case of Marathi, in Proceedings of the 7th Workshop on Noisy User-generated Text (Association for Computational Linguistics, 2021), pp. 63-67.
17. P. Kapil and A. Ekbal, Cross-Lingual Zero-Shot and Few-Shot Learning for Hate Speech Detection, SSRN 4902214 (2024).
18. I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, Label Modification and Bootstrapping for Zero-Shot Cross-Lingual Hate Speech Detection, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (Springer, Heidelberg, 2023), pp. 1-13.
19. U. Şahin, İ. E. Küçükkaya, O. Özçelik, and Ç. Toraman, Zero and Few-Shot Hate Speech Detection in Social Media Messages Related to Earthquake Disaster, in Proceedings of the IEEE International Conference on Signal Processing and Communications Applications (Springer, Heidelberg, 2023), pp. 1-4.
20. M. R. Awal, R. K. W. Lee, E. Tanwar, T. Garg, and T. Chakraborty, Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2023), pp. 1-10.
21. L. Liu, D. Xu, P. Zhao, D. D. Zeng, P. J. H. Hu, Q. Zhang, and Z. Cao, A Cross-Lingual Transfer Learning Method for Online COVID-19-Related Hate Speech Detection, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2023), pp. 1-15.
22. K. Mnassri, R. Farahbakhsh, and N. Crespi, Multilingual Hate Speech Detection: A Semi-Supervised Generative Adversarial Approach, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2024), pp. 1-13.
23. H. Ahn, J. Sun, C. Y. Park, and J. Seo, NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-Lingual Transfer, in Proceedings of the 14th International Workshop on Semantic Evaluation (Springer, Heidelberg, 2020), pp. 1-8.
24. A. Jiang and A. Zubiaga, Cross-Lingual Capsule Network for Hate Speech Detection in Social Media, in Proceedings of the 32nd ACM Conference on Hypertext and Social Media (Springer, Heidelberg, 2021), pp. 217-223.
25. C. Salaam, F. Dernoncourt, T. Bui, D. Rawat, and S. Yoon, Offensive Content Detection via Synthetic Code-Switched Text, in Proceedings of the 29th International Conference on Computational Linguistics (Springer, Heidelberg, 2022), pp. 6617-6624.
26. E. W. Pamungkas, V. Basile, and V. Patti, A Joint Learning Approach with Knowledge Injection for Zero-Shot Cross-Lingual Hate Speech Detection, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2021), pp. 1-13.
27. M. Das, S. Banerjee, and A. Mukherjee, Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages, in Proceedings of the 33rd ACM Conference on Hypertext and Social Media (Springer, Heidelberg, 2022), pp. 32-42.
28. K. Ghosh and A. Senapati, Hate Speech Detection in Low-Resource Indian Languages: An Analysis of Transformer-Based Monolingual and Multilingual Models with Cross-Lingual Experiments, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2025), pp. 1-22.
29. M. Akram, W. H. Moosa, and N. Zahra, Hate Speech Detection: A Social Network Story that Needs Serious Attention, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2024), pp. 97-108.
30. N. Hamad, M. Jarrar, M. Khalilia, and N. Nashif, Offensive Hebrew Corpus and Detection using BERT, in Proceedings of the IEEE/ACS International Conference on Computer

Systems and Applications (Springer, Heidelberg, 2023), pp. 1-12.
31. Z. Delbari, N. S. Moosavi, and M. T. Pilehvar, Spanning the Spectrum of Hatred Detection: A Persian Multi-Label Hate Speech Dataset with Annotator Rationales, in Proceedings of the AAAI Conference on Artificial Intelligence (Springer, Heidelberg, 2024), Vol. 38, pp. 17889-17897.
32. S. H. Muhammad, I. Abdulmumin, A. A. Ayele, D. I. Adelani, I. S. Ahmad, S. M. Aliyu, A. Yousif, and N. Ousidhoum, AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages, in Proceedings of the International Conference on Computational Linguistics (Springer, Heidelberg, 2025), pp. 1-15.
33. N. Reimers and I. Gurevych, Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (Springer, Heidelberg, 2019), pp. 1-15.
34. M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers, Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI, in Proceedings of the International Conference on Computational Linguistics (COLING 2024) (2024), pp. 1-17.