computers

MDPI

Article

UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers

Muhammad Ahmad [1], Muhammad Waqas [2], Ameer Hamza [2], Sardar Usman [3], Ildar Batyrshin [1] and Grigori Sidorov [1,*]

[1] Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC-PN), Mexico City 07738, Mexico; mahmad2024@cic.ipn.mx (M.A.); batyr1@cic.ipn.mx (I.B.)
[2] Department of Software Engineering, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; muhmmadwaqas449@gmail.com (M.W.); hamzariaz1142@gmail.com (A.H.)
[3] School of Informatics and Robotics, Institute of Arts and Culture, Lahore 54000, Pakistan; sardar.usman@iac.edu.pk
* Correspondence: sidorov@cic.ipn.mx

Abstract: The rise in social media has improved communication but also amplified the spread of hate speech, creating serious societal risks. Automated detection remains difficult due to subjectivity, linguistic diversity, and implicit language. While prior research focuses on high-resource languages, this study addresses the underexplored multilingual challenges of Arabic and Urdu hate speech through a comprehensive approach. To achieve this objective, this study makes four different key contributions. First, we have created a unique multi-lingual, manually annotated binary and multi-class dataset (UA-HSD-2025) sourced from X, which contains the five most important multi-class categories of hate speech. Secondly, we created detailed annotation guidelines to make a robust and perfect hate speech dataset. Third, we explore two strategies to address the challenges of multilingual data: a joint multilingual and translation-based approach. The translation-based approach involves converting all input text into a single target language before applying a classifier. In contrast, the joint multilingual approach employs a unified model trained to handle multiple languages simultaneously, enabling it to classify text across different languages without translation. Finally, we have employed state-of-the-art 54 different experiments using different machine learning using TF-IDF, deep learning using advanced pre-trained word embeddings such as FastText and Glove, and pre-trained language-based models using advanced contextual embeddings. Based on the analysis of the results, our language-based model (XLM-R) outperformed traditional supervised learning approaches, achieving 0.99 accuracy in binary classification for Arabic, Urdu, and joint-multilingual datasets, and 0.95, 0.94, and 0.94 accuracy in multi-class classification for joint-multilingual, Arabic, and Urdu datasets, respectively.

Keywords: social media; deep learning; machine learning; transfer learning; data mining; SVM; BERT; RoBERTa; hate speech; Urdu hate speech; Arabic hate speech

check for updates

Academic Editor: Hovhannes Harutyunyan

Received: 27 May 2025
Revised: 11 June 2025
Accepted: 13 June 2025
Published: 18 June 2025

Citation: Ahmad, M.; Waqas, M.; Hamza, A.; Usman, S.; Batyrshin, I.; Sidorov, G. UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers. *Computers* **2025**, *14*, 239. https://doi.org/10.3390/computers14060239

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the era of digital communication, online social media platforms such as Facebook, Instagram, X (formerly Twitter), and YouTube have become vital channels for interaction, idea sharing, and self-expression in ways that were previously unimaginable. The rapid increase in user-generated textual, audio, and video content has fundamentally transformed how people connect, communicate, and engage with global communities. These platforms

also serve as important sources of data for various natural language processing (NLP) tasks, including hope speech detection [1–3], and other text classification problems [4,5]. However, this exponential growth has also facilitated the spread of harmful content, including hate speech [6–8], cyberbullying [9–15], misinformation [16–21], and offensive language [22–26], which pose significant challenges for maintaining safe and supportive online environments.

Hate speech is a form of communication that expresses hatred, discrimination, or incites violence against individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, or disability. Due to the rise in textual data, hate speech has become a serious issue, spreading negativity [27] and promoting division [28]. Many victims of hate speech experience deep emotional distress, and in some cases, the online hostility translates into real-world violence, social unrest, and even hate crimes. As a result, hate speech detection and mitigation have become a growing concern for researchers, policymakers, and technology companies to create safer digital spaces.

Hate speech detection is a crucial task, due to the subjective nature of what constitutes "hate speech" and the linguistic diversity in online discussions. Many different cultures, communities, and legal frameworks have changing definitions of hate speech, making it exciting to establish universal criteria for identification and regulation. Furthermore, hate speech is often discussed in subtle, implicit, or coded forms, including sarcasm [29], humor [30], or euphemisms, which traditional keyword-based detection methods struggle to identify. Additionally, many social media users use some slang [31,32], abbreviations, misspellings, or word play to avoid automated moderation systems, which makes further complications in the detection of hate speech. Due to these challenges, an automated hate speech detection model must incorporate sophisticated natural language processing (NLP) techniques to effectively analyze and classify harmful content on social media platforms.

The rapid growth of artificial intelligence (AI) has transformed hate speech detection, enabling researchers to develop more accurate and scalable solutions. Traditional approaches often misclassified social media content, but transfer learning models with advanced pre-trained embeddings now offer better contextual understanding. These embedding-based models analyze vast amounts of textual data to identify subtle linguistic patterns. However, challenges persist as social media data evolves with new slang and coded expressions. Ethical concerns such as censorship, bias, and freedom of speech also complicate moderation. AI models can be biased if they learn from a biased dataset. It is essential to remove harmful data while still allowing free speech. To make the model unbiased and perfect, we need more ideas from language experts, psychologists, and legal experts. By working together, we can create better and more responsible hate speech detection systems.

Detection of hateful content in both Arabic and Urdu is crucial due to the rapid growth of these languages on social media platforms. Arabic, one of the top five most spoken languages globally, has around 450 million speakers [33]. As of April 2024, Twitter has about 16.28 million active Arabic-speaking users [34]. However, hate speech detection in Arabic is challenging due to its diverse dialects, complex morphology, and the lack of annotated datasets. Similarly, Urdu, spoken by around 230 million people worldwide [35], is among the top 20 most spoken languages [36]. With about 3.5 million active Urdu-speaking users on Twitter in April 2024, hate speech detection in Urdu faces its own set of challenges, including its complex syntax, regional variations, and a shortage of annotated datasets.

Traditional approaches for the detection of hate speech content using social media data have failed owing to their inability to tackle the difficulty in grammar, various dialects, and regional differences in both languages. Additionally, the lack of annotated datasets for Arabic and Urdu further complicates manual detection efforts. As social media data grows rapidly, there is an urgent need for automated, real-time detection systems [37] based on

deep learning and transformer models that can effectively handle linguistic diversity and complexity. As a result, these traditional methods cannot effectively find and stop harmful content. Given the rapid increase in Arabic and Urdu data on social media platforms, there is an urgent need for automated detection of hate speech content, ensuring a safer and more inclusive digital space for speakers of these languages.

To achieve this objective, we constructed two novel, manually annotated hate speech datasets in Arabic and Urdu, ensuring high-quality labels for both binary and multi-class classification tasks. We collected a diverse range of tweets from Twitter, applied preprocessing techniques, and developed comprehensive data annotation guidelines. Following annotation, we employed various machine learning models using TF-IDF features, deep learning models with pre-trained word embeddings such as FastText and GloVe, and advanced language models utilizing contextual embeddings to enhance classification performance. Our goal was to identify the most effective model for detecting harmful content on social media platforms. The resulting datasets serve as a valuable resource for advancing hate speech detection in Arabic and Urdu.

This study makes the following key contribution:

- ✓ To the best of our knowledge, joint multi-lingual and joint-translated approaches were not explored earlier for multi-lingual hate speech detection in Urdu and Arabic Languages.
- ✓ We created a comprehensive Arabic and Urdu hate speech dataset named UA-HSD-2025, manually annotated for both binary (hate, not hate) and multi-class such as direct hate speech, disguised hate speech, sarcastic hate speech, and exclusionary hate speech, to enhance annotation and classification.
- ✓ We developed annotation guidelines to ensure consistent and accurate labeling for binary and multi-class classification. These guidelines reduce ambiguity, enhance dataset reliability, and improve inter-annotator agreement, making the data more suitable for robust machine learning models.
- ✓ We presented a comprehensive and reproducible pseudocode framework that systematically integrates translation, manual refinement, dataset merging, preprocessing, and fine-tuned multilingual classification, facilitating transparent replication and extension of hate speech detection across Arabic and Urdu languages.
- ✓ We conducted 54 experiments to evaluate and compare the performance of machine learning, deep learning, and transfer learning models on both binary and multi-class classification tasks, with the aim of identifying the most suitable model for our hate speech detection task.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the dataset construction process in detail. Section 4 presents the methodology. Section 5 discusses the results and analysis. Section discusses the limitations of our proposed solution, and finally, Section 6 concludes the paper and suggests future directions.

2. Literature Review

Recent research on hate speech detection spans various methodologies, datasets, and languages, reflecting the field's rapid evolution and persistent challenges. To better understand these developments, this section reviews key studies focusing on different aspects such as dataset creation, modeling techniques, and multilingual approaches.

A broad overview of the landscape is provided by Alkomah et al. [38], who reviewed 138 studies focusing on datasets, features, and machine learning models. They observed significant variation in approaches, with many studies combining multiple deep learning models to enhance performance. However, they also pointed out that the small size and

low reliability of many datasets limit model effectiveness, offering guidance for future research directions.

The integration of large language models (LLMs) into hate speech detection has received increasing attention. Albladi et al. [39] explored how models like GPT-3 and BERT are applied in this domain, analyzing their strengths and limitations. Their review emphasizes the ongoing challenges of maintaining fairness and accuracy in LLM-based systems while offering insights into potential future improvements to enhance their reliability and equity.

Addressing the growing need for inclusive solutions, several studies have focused on low-resource languages. Hashmi et al. [40] tackled hate speech detection in Norwegian using the Barlow Twins Method (BTM). By combining text augmentation and self-training, they developed Nor-BERT—a semi-supervised model that outperformed existing approaches, despite limited annotated data.

Similarly, Chavinda et al. [41] addressed hate speech detection in under-resourced languages like Sinhala and Tamil. Their approach combined Multilingual Large Language Models (MLLMs) with Dual Contrastive Learning (DCL), using both self-supervised and supervised learning techniques. Fine-tuning the Twitter/twhin-bert-base model on domain-specific data led to superior performance over traditional models like CNN and LSTM.

Work on Devanagari-script languages such as Hindi and Nepali has also gained traction. Thapa et al. [42] organized a shared task targeting language identification, hate speech detection, and target classification. Using curated datasets and transformer models, 113 participants developed multilingual approaches that advanced NLP solutions for these languages. In a complementary effort, Khadka et al. [43] evaluated both machine learning and transformer-based models. Their use of a multilingual RoBERTa model, pre-trained on social media data, yielded impressive results: 99.5% accuracy in language identification, 88.3% F1-score in hate speech detection, and 68.6% in target classification.

Unlike previous studies that explore hate speech detection using techniques such as Dual Contrastive Learning [41], the Barlow Twins Method [40], or transformer-based models like RoBERTa [41], our work introduces a novel multilingual approach focused specifically on the underrepresented Arabic and Urdu scripts. While earlier research addresses low-resource languages generally, none offer a comprehensive framework tailored to Arabic and Urdu. In contrast, our study contributes a new manually annotated dataset (UA-HSD-2025) with binary and multi-class labels, applies both joint multilingual and translation-based classification strategies, and evaluates performance using 54 experiments across traditional machine learning, deep learning, and state-of-the-art transfer learning models.

3. Methodology

3.1. Construction of the Dataset

For the data collection, we compiled a dataset related to hate speech in the Arabic and Urdu languages, sourced from Twitter using the Tweepy API, a Python (https://www.python.org/) library that enables developers to extract and analyze Twitter data, including collecting tweets. Our objective was to gather a diverse set of tweets containing hateful content. To do this, we first created a dictionary of commonly used hate speech keywords in Arabic, including "كراهية" (hatred), "عنصرية" (racism), "إهانة" (insult), "تحريض" (incitement), "تمييز" (discrimination), "إبادة" (genocide), "كفار" (infidels), "مجرمون" (criminals), "اذبحهم" (slaughter them), and "تطرف" (extremism). To enhance diversity, we used logical combinations such as "عنصرية AND كراهية", "إهانة OR سب" (insult OR curse), and "تحريض NOT سياسة" (incitement NOT politics). We further extended our approach

by incorporating hate speech keywords in the Urdu language, including "نفرت" (hatred), "گالی" (abuse), "دشمن" (enemy), "ظالم" (oppressor), "لعنت" and "جہنم میں جاؤ" (curse, go to hell), "کافر" (infidel), "بلوچ دہشتگرد" (Baloch terrorist), "دو ٹکي کی عورت" (a derogatory term for women), and "غدار" (traitor). This multilingual keyword-based search enabled us to capture a more comprehensive set of hate speech instances across different linguistic communities. After filtering out irrelevant tweets, we finalized a dataset of 30,000 tweets, each manually reviewed for relevance to hate speech detection. The dataset was saved in a CSV (Comma Separated Values) format for structured storage and further analysis. Consisting entirely of Arabic and Urdu text, this dataset serves as a valuable resource for studying hate speech in both Arabic-speaking and Urdu-speaking communities. Figure 1 shows the proposed architecture of our methodology.

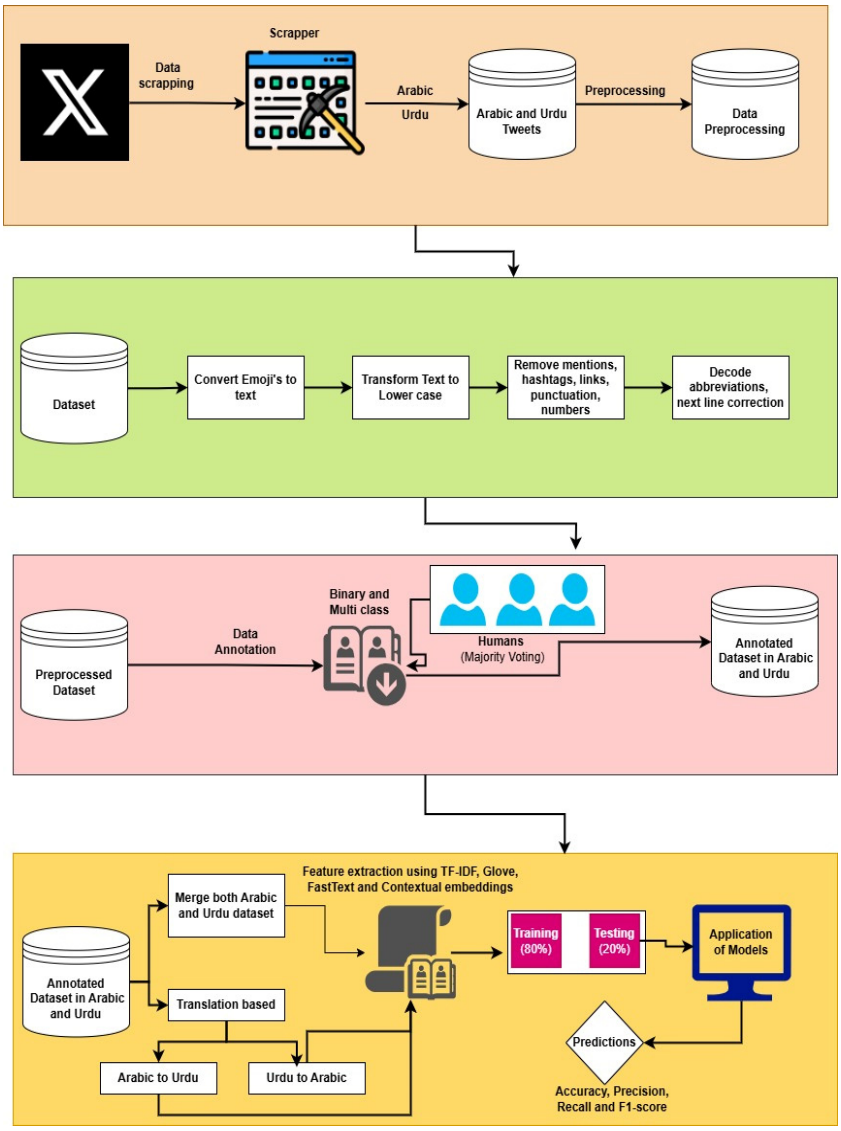


Figure 1. Proposed methodology and design.

3.2. Pre-Processing

Social media posts often contain noisy and unstructured content, including spelling errors, emojis, and special characters. Such elements introduce challenges for automated processing and make the data difficult for machine learning models to interpret. Therefore, it is essential to clean and preprocess the data to enhance its quality and suitability for

downstream machine learning tasks. This process is called preprocessing. We conducted multiple pre-processing steps to clean our dataset:

(1) Uppercase Text is transformed to lowercase.
(2) Removal of irrelevant tweets which does not show the hateful content.
(3) Removal of special characters and punctuation from the text.
(4) Convergence of digits to text.
(5) Removal of stop words from the text for better understanding of machine learning.
(6) Normalize elongated words from the tweets.
(7) Removal of hashtags and mentions from the tweets.
(8) Remove duplicates and tweets with less than 20-character tweets.
(9) Decodes all the short text, such as thnx to Thanks, plz to please, etc.

After the pre-processing of 30,000 tweets, only 5240 original tweets remained to create a multi-lingual hate speech detection dataset.

3.3. Annotation Process

3.3.1. Annotation Guidelines

Hate speech detection in our dataset is complex because it adds another layer of complexity to distinguishing between hate and non-hate before classifying the specific type of hate speech. Some posts comprise explicit hate speech with direct insults or threats, making them easy to classify. However, many posts use disguised hate speech, where the hateful intent is hidden through indirect or coded language. Sarcastic hate speech is another challenge, as it may appear harmless at first but carries an offensive meaning when read in context. Exclusionary hate speech is subtler, as it suggests that a group should be removed or denied rights without using direct insults. Additionally, tweets may contain ambiguous or mixed sentiments, making it harder to decide if they should be classified as hate speech. Given these complexities, a structured approach is necessary in the first level, where we categorize posts into binary classes such as hate speech and non-hate speech. In the second level, we additionally classify hateful posts into four fine-grained categories of hate speech based on their characteristics, as shown in Figure 2, while Table 1 shows some sample tweets from the dataset. These categories include:

1. Direct Hate Speech: In this type of hate speech, someone uses openly offensive or abusive language targeting an individual or group of people or communities.
2. Disguised Hate Speech: In this type of hate speech, someone uses indirect hateful language that may use sarcasm, coded words, or humor to mask their intent.
3. Sarcastic Hate Speech: In this type of hate speech, someone uses Hate speech that appears neutral or positive but, in context, conveys hostility or mockery.
4. Exclusionary Hate Speech: In this type of hate speech, someone uses Speech that suggests a group should be excluded, dehumanized, or denied rights.
5. Not Hate: The tweet does not indicate any offensive, discriminatory, or exclusionary language or a post that will indicate respect, unity, and inclusivity, without targeting or discriminating against any group.

Since hate speech can be highly context-dependent, a structured annotation process is needed to differentiate between these categories accurately.

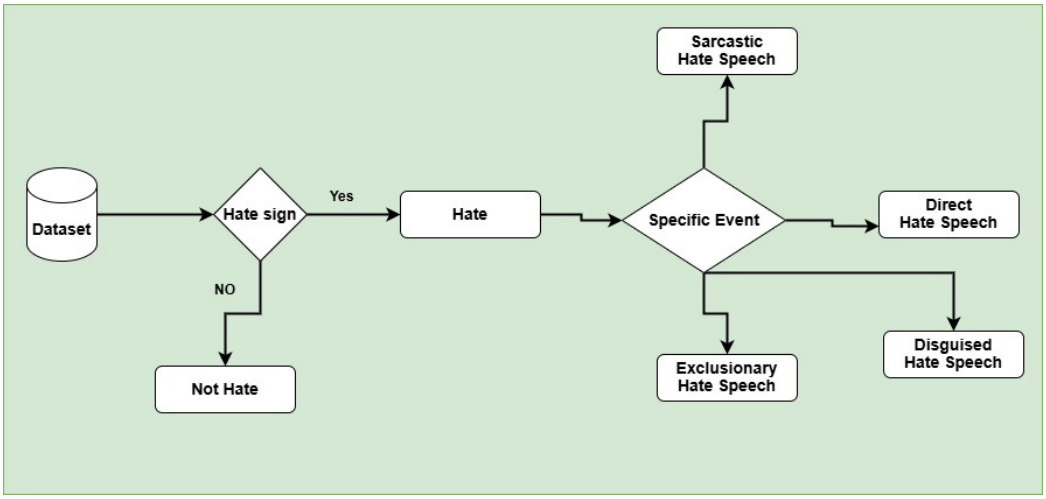


Figure 2. Annotation Procedure of Hate Speech Detection.

Table 1. Sample Tweets from the Dataset.

Language	Tweet	English Translation	Class
Arabic	#USER هؤلاء الحشرات يجب التخلص منهم فورًا.	These insects should be eliminated immediately.	Direct Hate Speech
	الغريب أن بعض الناس يطالبون بحقوق في بلاد ليست لهم.	Strange how some people demand rights in a country that isn't theirs.	Disguised Hate Speech
	في الفشل، كعادتهم "خبراء"نعم، هم فعلاً😂	Yes, they're real "experts" in failure, as usual 😂	Sarcastic Hate Speech
	إن لم تكن من ثقافتنا أو ديننا، فمكانك ليس هنا.	If you're not from our culture or religion, you don't belong here.	Exclusionary Hate Speech
Urdu	#USER ان کچرے کے ڈھیر لوگوں کو سیدھا جہنم بھیجنا چاہیے	These piles of trash should be sent straight to hell.	Direct Hate Speech
	#USERکچھ لوگ ایسے ہوتے ہیں جو ہر جگہ مسائل کھڑے کرتے ہیں، لیکن پھر بھی معصوم بن جاتے ہیں #USER	Some people create problems everywhere but still play innocent.	Disguised Hate Speech
	#USER #USERجی ہاں، ان کی عقل تو ناسا میں بھی نہیں ملے گی 😂	Oh yes, even NASA won't find their level of genius 😂	Sarcastic Hate Speech
	#USERجو ہماری تہذیب کے مطابق نہیں چل سکتے، ان کو یہاں رہنے کی اجازت نہیں ہونی چاہیے #USER	Those who can't adapt to our values shouldn't be allowed to stay here.	Exclusionary Hate Speech
	#USER #USERہمیں ہر انسان سے برابری اور عزت کے ساتھ پیش آنا چاہیے	We should treat every person with equality and respect.	Not Hate

3.3.2. Annotation Selection

Since distinguishing between different types of hate speech was a major challenge during dataset annotation, a rigorous process was implemented to carefully select qualified annotators for this task. Initially, we created an account on Freelancer.com, which is one of the most well-known freelancing platforms, with over 60 million active users (accessed on 10 January 2025). We posted a project with detailed annotation guidelines to find qualified annotators. The project cost was 30 USD to 500 USD. After posting the project, there were several data annotators who reached out to us and bid on the project. After reviewing applications, we carefully selected eight data annotators from the UAE and Saudi Arabia for the Arabic annotation task and six data annotators from Pakistan for the Urdu annotation task. All annotators held at least a Master's degree in Computer Science. To evaluate their annotation skills, we provided each group with an initial set of 300 Twitter posts in their

respective languages. Upon reviewing their annotations, we found that three annotators from the Arabic group and two from the Urdu group had incorrectly labeled several samples. The remaining five Arabic annotators and four Urdu annotators demonstrated a satisfactory level of consistency in their labeling.

To further validate their annotation quality, we assigned a second set of 300 posts to these selected annotators. After reviewing their work in the second round, we observed that two of the five remaining Arabic annotators and one of the four remaining Urdu annotators produced inconsistent or irrelevant annotations. As a result, we finalized three annotators from each group—Arabic and Urdu—who consistently demonstrated high-quality annotations across both rounds. These six annotators were selected to label the entire dataset and finalize the annotations. We created individual Google Forms for each annotator to classify the sample independently. Each annotator was paid 0.03 USD per sample to ensure fair compensation for their efforts. If any confusion or disagreement arose during the annotation process, we arranged a meeting with the annotators to discuss and resolve the issue. This collaborative approach helped us reach a final agreed-upon label for ambiguous cases, ensuring consistency and accuracy in the dataset.

3.4. Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) evaluates the consistency among annotators by accounting for the likelihood of agreement by chance. For the binary hate speech dataset, a Cohen's Kappa Coefficient [44] of 85% was achieved for Arabic, while the annotators working on the Urdu dataset achieved a Cohen's Kappa of 83% for binary classification. For the multiclass hate speech dataset, a Fleiss' Kappa Score [45] of 82% was obtained in Arabic, while the Urdu annotators achieved an even higher Fleiss' Kappa of 85%. These high agreement scores reflect the robustness of the datasets, which were developed through a rigorous and meticulous annotation process.

To ensure consistent interpretation of guidelines over time, we conducted periodic calibration sessions where annotators discussed ambiguous and difficult cases. When disagreements arose during annotation, they were resolved through group discussions using a majority voting approach. This collaborative approach helped maintain annotation quality and consistency throughout the project.

3.5. Dataset Statistics

Figure 3 presents detailed statistics for an Arabic and Urdu hate speech dataset, showcasing key metrics for each language. Starting with the number of tweets, Arabic has a slightly higher count (3072) than Urdu (2167), suggesting a broader collection for Arabic. Interestingly, despite Urdu having fewer tweets, it has more total words (23,479) than Arabic (22,031), resulting in a significantly higher average word count per tweet (10.35 for Urdu vs. 6.35 for Arabic). This implies that Urdu tweets are generally more verbose. However, when looking at the average number of sentences per tweet, Urdu tweets tend to be shorter and more concise in structure, with only one sentence per tweet, while Arabic tweets average about 1.12 sentences. In terms of vocabulary richness, Arabic has a more diverse vocabulary (2269 unique words) compared to Urdu (1772). Finally, the total sentence count is also higher in Arabic (3466) than in Urdu (2267), reinforcing that Arabic tweets tend to be more segmented and possibly more expressive. Overall, this comparison provides valuable insight into the structural and linguistic differences in hate speech expression between the two languages. Figure 4 presents the word cloud of hate speech-related keywords in the dataset, highlighting the most frequent words and revealing common patterns across different hate speech categories. Figure 5 shows the Label distribution in the binary and multi-class hate speech detection in both Arabic and

Urdu datasets. We ensured a balanced multi-class dataset to enhance model performance, improve generalization, and prevent bias toward any specific category. This balance helps achieve better accuracy, F1-score, and robustness in classification.

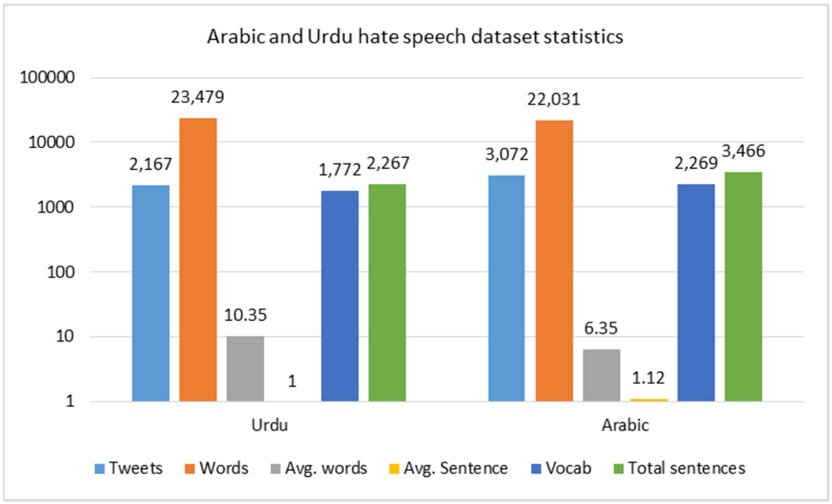


Figure 3. Statistics of the dataset.



Figure 4. Word cloud of hate speech-related keywords in the dataset (**a**) Urdu and (**b**) Arabic.

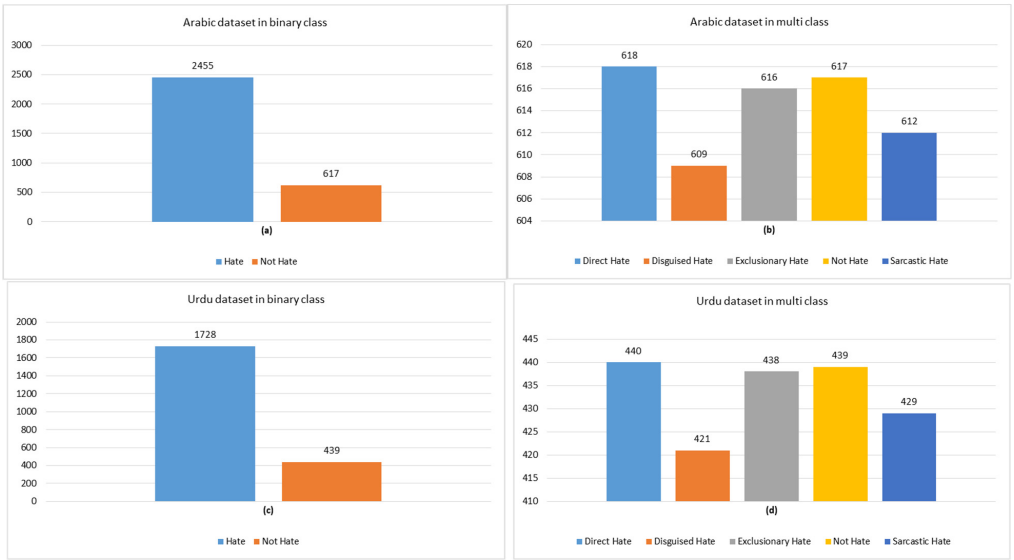


Figure 5. Label distribution for binary and multi-class hate speech detection in Arabic and Urdu: (**a**) Arabic binary, (**b**) Arabic multi-class, (**c**) Urdu binary, and (**d**) Urdu multi-class.

3.6. Translation and Joint Multi-Lingual Steps

The objective of the translation and joint multi-lingual process is to transform multi-lingual Twitter posts into a single universal language. Three datasets, such as Arabic,

Urdu, and joint multi-lingual, are constructed as detailed below. The pseudo-code of both approaches is presented in Table 2.

1. Initially, we employed Google Translate API to convert Arabic Twitter posts into Urdu. Following the translation, we manually refined the text to correct any inconsistencies or translation errors. The finalized Urdu translations were then merged with the original Urdu dataset for subsequent processing.
2. For the second dataset, Urdu comments were translated into Arabic using the same process. After manually refining the translated text to address any inaccuracies, the cleaned translations were integrated with the original Arabic dataset for further analysis.
3. For the third dataset, we combined the original Urdu and original Arabic datasets to create a unified joint multi-lingual dataset.

Table 2. Pseudo-code outlining the multilingual hate speech classification process for Arabic and Urdu datasets.

Step	Procedure/Operation	Description
1	Multilingual-hate-speech-classification ()	Procedure for multilingual hate speech classification (Arabic–Urdu dataset)
2	D1 ← Arabic-Dataset (3073)	Load the original Arabic dataset
3	D2 ← Urdu-Dataset (2167)	Load the original Urdu dataset
4	Urdu-Translated ← Translate-To-Arabic(D2)	Translate the Urdu dataset into Arabic using the Google Translate API
5	Arabic-Translated ← Translate-To-Urdu(D1)	Translate the Arabic dataset into Urdu using Google Translate API
6	Review-Translations (Urdu-Translated, Arabic-Translated)	Manually review and correct translations (especially for slang, idioms)
7	Combined-Arabic ← Merge (D1, Urdu-Translated)	Merge the Arabic original with the translated Urdu (in Arabic)
8	Combined-Urdu ← Merge (D2, Arabic-Translated)	Merge the Urdu original with the translated Arabic (in Urdu)
9	Combined-Multilingual ← Merge (D1, D2)	Merge the original Arabic and Urdu into a single multilingual dataset
10	Pre-Processed-Arabic ← Pre-Processing(Combined-Arabic)	Clean and normalize the Arabic dataset
11	Pre-Processed-Urdu ← Pre-Processing(Combined-Urdu)	Clean and normalize the Urdu dataset
12	Pre-Processed-Multi ← Pre-Processing(Combined-Multilingual)	Clean and normalize the original multilingual dataset
13	Tokenized-Arabic ← XLMR-Tokenizer(Pre-Processed-Arabic)	Tokenize the Arabic dataset using xlm-roberta-base
14	Tokenized-Urdu ← XLMR-Tokenizer(Pre-Processed-Urdu)	Tokenize Urdu dataset using xlm-roberta-base

Table 2. *Cont.*

Step	Procedure/Operation	Description
15	Tokenized-Multi ← XLMR-Tokenizer(Pre-Processed-Multi)	Tokenize the joint Arabic–Urdu dataset using xlm-roberta-base
16	Classification(Tokenized-Arabic, 1)	Classification using fine-tuned xlm-roberta-base on Arabic
17	Classification(Tokenized-Urdu, 2)	Classification using fine-tuned xlm-roberta-base on Urdu
18	Classification(Tokenized-Multi, 3)	Classification using fine-tuned xlm-roberta-base on joint Arabic–Urdu
19	Pre-Processing(D)	Procedure to clean and normalize text
20	D1 ← Cleaning(D)	Remove HTML, hashtags, URLs, mentions, punctuation, etc.
21	D2 ← Lower-Case(D1)	Convert text to lowercase
22	D3 ← Replace-Emoji(D2)	Replace emojis with corresponding words
23	Classification(D, mode)	Fine-tune and classify using xlm-roberta-base
24	Model ← fine-tuning(D, 'xlm-roberta-base', 80-20)	Train/test split (80/20) and model training
25	Confusion-matrix ← generate-results(Model)	Generate a confusion matrix
26	Accuracy ← compute-accuracy(confusion-matrix)	Compute accuracy
27	Precision ← compute-precision(confusion-matrix)	Compute precision
28	Recall ← compute-recall(confusion-matrix)	Compute recall
29	F1-score ← compute-F1(confusion-matrix)	Compute F1-score

3.7. Application of Models

In this section, we will explore the structured approach to pre-processing and applying various machine learning techniques such as Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), and Multinomial Naive Bayes (MNB) using TF-IDF, deep learning such as BiLSTM and CNN using advanced pre-trained word embeddings such as FastText, and three advanced language-based models such as bert-base-multilingual-cased, Robert-base, xlm-roberta-base to identify hate speech detection in more accurate way and find the best model for hate speech detection. We chose these models based on their higher performance. The process begins by splitting the dataset into training (20%) and testing sets (80%). The training data is used to train the model, while the testing data is used to evaluate the performance of the trained model, as shown in Figure 6. After the training and testing cycle, the model's accuracy is calculated using four different metrics, such as precision, recall, F1-score, and Accuracy. These metrics provide a detailed understanding of how well the model is performing, not just in terms of accuracy but also in terms of its ability to correctly identify positive cases and avoid false positives and negatives.

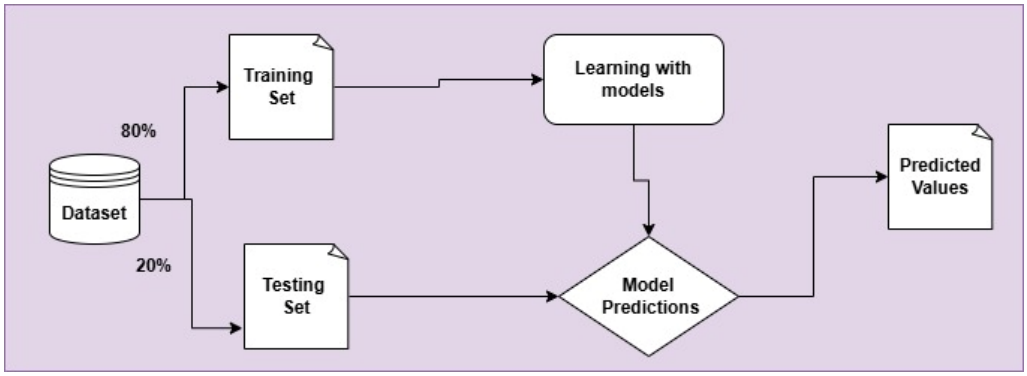


Figure 6. Application of machine learning and deep learning.

4. Results and Analysis

This study employs a range of machine learning, deep learning, and transformer-based models to evaluate their effectiveness in hate speech detection across Arabic, Urdu, and multilingual datasets. Traditional ML models such as SVM, KNN, DT, and MNB are used with TF-IDF feature representations. These models are known for their simplicity, interpretability, and strong performance on structured text features. For DL approaches, we utilize neural network architectures such as CNN and BiLSTM networks, which are enhanced by pre-trained word embeddings like FastText and GloVe. These models are capable of capturing local patterns (CNN) and long-range dependencies (BiLSTM) in text. The core focus of this study lies in transformer-based models, which leverage self-attention mechanisms to process input sequences more effectively. Specifically, we use Multilingual BERT, XLM-RoBERTa, and Multilingual RoBERTa—pre-trained models that support cross-lingual understanding. Multilingual BERT is trained on 104 languages and provides a shared representation space. XLM-RoBERTa enhances this with improved training on a larger multilingual corpus, leading to better contextual understanding. Multilingual RoBERTa adapts the RoBERTa architecture for multiple languages, offering competitive performance in multilingual tasks. All transformer models utilize their respective tokenizers and process input sequences with a maximum length of 512 tokens. By comparing these three modeling paradigms, we highlight the advantages and limitations of each in detecting hate speech in low-resource, multilingual contexts.

4.1. Results for Machine Learning

Table 3 outlines the key hyperparameters used for tuning traditional machine learning models in the hate speech detection experiments. For K-Nearest Neighbors (KNN), the number of neighbors was set to 2, allowing the model to make predictions based on the closest two data points. Multinomial Naive Bayes (MNB) used a smoothing parameter (alpha) of 1.0, which helps manage zero probabilities in text classification. The Decision Tree (DT) was configured with a maximum depth of 20, and a fixed random state of 42 to ensure reproducibility. Lastly, the Support Vector Machine (SVM) model was trained using a polynomial kernel of degree 2, a regularization parameter (C) of 0.1, and the same random state of 42, enabling consistent model behavior across runs. These hyperparameter choices were tailored to balance complexity, overfitting, and generalization across the multilingual hate speech dataset.

Table 4 shows the results on hate speech detection across multilingual contexts, specifically evaluating the performance of various traditional machine learning models—KNN, Multinomial Naive Bayes (MNB), Decision Tree (DT), and Support Vector Machine (SVM) on binary classification tasks in Arabic, Urdu, and combined multilingual settings. The results show that the Decision Tree consistently performs well, particularly in Urdu, where

it achieves the highest precision, recall, F1-score, and accuracy (0.95). In the multilingual and Arabic settings, performance is slightly more varied, with all models performing competitively. Notably, translations between Arabic and Urdu did not significantly impact model effectiveness, suggesting that hate speech patterns are robustly captured across both languages. Overall, these findings highlight the importance of considering both linguistic diversity and model selection in building effective hate speech detection systems.

Table 3. Best hyperparameter values for machine learning models.

Model	Hyperparameter	Value
K-Nearest Neighbors (KNN)	n_neighbors	2
Multinomial Naive Bayes (MNB)	alpha	1.0
Decision Tree (DT)	max_depth	20
	random_state	42
Support Vector Machine (SVM)	kernel	poly
	degree	2
	C	0.1
	random_state	42

In the multi-class hate speech detection task, this study extends its evaluation of machine learning models—KNN, MNB, DT, and SVM—across multilingual, Arabic, and Urdu datasets. Performance trends reveal that classification is more challenging in multilingual and Arabic contexts, where precision, recall, and F1-scores hover around the mid-50s to low 60s. Among the models, Decision Tree and SVM tend to perform more consistently across languages, with DT slightly outperforming others in the multilingual setting (F1-score: 0.62) and SVM leading in Arabic (precision: 0.65). However, in the Urdu dataset, all models perform significantly better, especially KNN and DT, both achieving high F1-scores of 0.77, indicating that hate speech in Urdu is more distinctly classifiable in multi-class scenarios. These results emphasize the increased complexity of multi-class hate speech detection in diverse linguistic settings and highlight Urdu as a language with clearer hate speech patterns for machine learning models to detect.

Table 4. Results for machine learning models.

Class	Language	Models	Precision	Recall	F1-Score	Accuracy
Binary	Multi-lingual	KNN	0.86	0.86	0.84	0.86
		MNB	0.81	0.83	0.81	0.83
		DT	0.86	0.87	0.85	0.87
		SVM	0.85	0.86	0.84	0.86
	Arabic	KNN	0.87	0.87	0.83	0.87
		MNB	0.86	0.87	0.86	0.87
		DT	0.88	0.86	0.87	0.86
		SVM	0.87	0.87	0.84	0.87
	Urdu	KNN	0.93	0.93	0.93	0.93
		MNB	0.88	0.88	0.86	0.88
		DT	0.95	0.95	0.95	0.95
		SVM	0.93	0.93	0.93	0.93

Table 4. *Cont.*

Class	Language	Models	Precision	Recall	F1-Score	Accuracy
Multi class	Multi-lingual	KNN	0.63	0.57	0.58	0.57
		MNB	0.51	0.5	0.48	0.5
		DT	0.62	0.62	0.62	0.62
		SVM	0.62	0.58	0.59	0.58
	Arabic	KNN	0.56	0.51	0.52	0.51
		MNB	0.57	0.56	0.55	0.56
		DT	0.63	0.6	0.6	0.6
		SVM	0.65	0.59	0.59	0.59
	Urdu	KNN	0.78	0.77	0.77	0.77
		MNB	0.63	0.63	0.63	0.63
		DT	0.78	0.77	0.77	0.77
		SVM	0.75	0.74	0.74	0.74

4.2. Results for Deep Learning

Two deep learning models, such as Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), are individually trained using FastText embeddings on Arabic, Urdu, and joint multi-lingual hate speech detection datasets.

Table 5 outlines the parameters used for training deep learning models for our hate speech detection task. Both models were trained for five epochs, using the Adam optimizer and categorical cross-entropy as the loss function, which is suitable for multi-class classification tasks. The embedding size for both models is set to 300, and they share a learning rate (lr) of 0.001 and a dropout rate of 0.1 to prevent overfitting. The activation function for the output layer is softmax, which is ideal for multi-class classification as its output probability distributions across classes. The CNN model has additional parameters, such as filters = 128, which are used to extract features from the input text at different granularities.

Table 5. Parameters for deep learning models.

Parameter	CNN	BiLSTM
Epochs	5 per fold	5 per fold
Optimizer	Adam	Adam
Loss	Categorical Crossentropy	Categorical Crossentropy
Number of Filters	128	–
Embedding Size	300	300
Learning Rate (lr)	0.001	0.001
Dropout	0.1	0.1
Activation	Softmax	Softmax

Table 6 shows the performance of two different deep learning models using pretrained word embeddings, such as FastText, for binary and multi-classification tasks. In the binary hate speech detection task using FastText word embeddings, both CNN and BiLSTM models demonstrate strong and consistent performance across multilingual, Arabic, and Urdu datasets. Notably, BiLSTM slightly outperforms CNN overall, achieving the highest scores in the Arabic dataset with near-perfect precision, recall, F1-score, and accuracy at

0.99. In both multilingual and Urdu contexts, the models maintain high effectiveness with F1-scores ranging from 0.96 to 0.98, underscoring their capability to capture contextual meaning even across diverse linguistic structures. These results highlight the strength of combining FastText's subword-level embeddings with deep learning models—especially BiLSTM—for accurately identifying hate speech in binary classification settings.

Table 6. Results for deep learning models.

Class	Language	Models	Precision	Recall	F1-Score	Accuracy
Binary	Multi-lingual	CNN	0.96	0.96	0.96	0.96
		BiLSTM	0.98	0.98	0.98	0.98
	Arabic	CNN	0.98	0.98	0.98	0.98
		BiLSTM	0.99	0.99	0.99	0.99
	Urdu	CNN	0.98	0.98	0.98	0.98
		BiLSTM	0.98	0.98	0.98	0.98
Multi class	Multi-lingual	CNN	0.81	0.8	0.8	0.8
		BiLSTM	0.81	0.78	0.78	0.78
	Arabic	CNN	0.86	0.86	0.86	0.86
		BiLSTM	0.87	0.87	0.87	0.87
	Urdu	CNN	0.86	0.86	0.86	0.86
		BiLSTM	0.86	0.86	0.86	0.86

In the multi-class hate speech detection task using FastText word embeddings, both CNN and BiLSTM models perform effectively across multilingual, Arabic, and Urdu datasets, though with slightly lower scores than in binary classification—reflecting the added complexity of distinguishing between multiple hate speech categories. BiLSTM edges out CNN in the Arabic dataset with an F1-score of 0.87, while both models perform equally well in Urdu with consistent metrics of 0.86. In the multilingual setting, CNN shows a slight advantage with an F1-score of 0.80 compared to BiLSTM's 0.78. These results suggest that while FastText-enhanced models remain strong in multilingual and morphologically rich environments, their performance slightly varies depending on the language and task complexity. Overall, the models show promising capability for multi-class hate speech detection, with Arabic and Urdu yielding the most balanced and high-performing outcomes.

4.3. Transformer Results

Table 7 outlines the training setup used for multilingual transformer models—specifically Multilingual BERT, XLM-RoBERTa, and multilingual RoBERTa—for hate speech detection. These models utilize their respective tokenizers (e.g., bert-base-multilingual-uncased, XLM-RoBERTa and roberta-base) to process input text, with a maximum sequence length capped at 512 tokens. The training is performed in small batches of 16 samples, using the AdamW optimizer with a learning rate of 3×10^{-5}, which helps balance training speed and stability. The model is trained over 3 epochs, with 80% of the data used for training and 20% for testing. To ensure fair evaluation and avoid class imbalance issues, CrossEntropyLoss is employed as the loss function, and the data is shuffled during training. The model runs on either CUDA. For evaluation, accuracy, F1-score, and the confusion matrix are used to assess performance. Additionally, padding and truncation are enabled during tokenization to ensure all input sequences are uniform in length, which is critical for transformer models

to function properly. Overall, the setup reflects a standard yet effective configuration for fine-tuning multilingual language models on classification tasks.

Table 7. Fine-tuning parameters for transformer-based models.

Model Name	Hyperparameter	Value
Multilingual BERT, XLM-RoBERTa, multi-lingual RoBERTa	Tokenizer	BERT Tokenizer (bert-base-multilingual-uncased), RoBERTa Tokenizer (roberta-base),
	Max Sequence Length	512
	Batch Size	16
	Learning Rate	3×10^{-5}
	Optimizer	AdamW
	Epochs	3
	Train/Test Split	80%/20%
	Loss Function	CrossEntropyLoss
	Shuffle	True
	Device	CUDA
	Evaluation Metrics	Accuracy, F1-score, Confusion Matrix
	Tokenization: Padding	True
	Tokenization: Truncation	True

Table 8 shows the performance of language-based transformer models using advanced contextual embeddings in both binary and multi-class settings. In the binary hate speech detection task, the results demonstrate exceptional performance across all language settings, such as joint multi-lingual, Arabic translation, and Urdu translation. Among the models evaluated, XLM-RoBERTa-base consistently achieves near-perfect scores with a precision, recall, F1-score, and accuracy of 0.99 across all languages, highlighting its robustness and generalizability. BERT-base-multilingual-cased also performs comparably well, particularly in multilingual and Arabic contexts, while RoBERTa-base slightly trails with metrics in the 0.97–0.98 range. These findings underscore the effectiveness of multilingual transformer models in capturing nuanced hate speech patterns and suggest that deep learning approaches significantly outperform traditional machine learning models, especially when handling complex and diverse linguistic data.

In the multi-class hate speech detection task, transformer-based models again show strong performance, with XLM-RoBERTa-base leading across all three language settings—multilingual, Arabic, and Urdu—achieving F1-scores of 0.95. BERT-base-multilingual-cased also delivers high accuracy and balanced scores, particularly in Urdu (0.94 F1-score), indicating its effectiveness in capturing nuanced hate speech categories. However, RoBERTa-base shows slightly lower performance, especially in Urdu (F1-score: 0.72), suggesting a limitation in adapting to low-resource or morphologically rich languages. Overall, these results reaffirm the superior capability of multilingual transformer models—especially XLM-RoBERTa—in handling complex, multi-class hate speech detection tasks across diverse languages, reinforcing their value in cross-lingual and translation-based moderation systems.

Table 8. Results for language-based transformer models.

Class	Language	Models	Precision	Recall	F1-Score	Accuracy
Binary	Multi-lingual	bert-base-multilingual-cased	0.99	0.99	0.99	0.99
		Robert-base	0.97	0.97	0.97	0.97
		xlm-roberta-base	0.99	0.99	0.99	0.99
	Arabic	bert-base-multilingual-cased	0.99	0.99	0.99	0.99
		Robert-base	0.98	0.98	0.98	0.98
		xlm-roberta-base	0.99	0.99	0.99	0.99
	Urdu	bert-base-multilingual-cased	0.98	0.98	0.98	0.98
		Robert-base	0.97	0.97	0.97	0.97
		xlm-roberta-base	0.99	0.99	0.99	0.99
Multi class	Multi-lingual	bert-base-multilingual-cased	0.91	0.9	0.9	0.9
		Robert-base	0.86	0.86	0.86	0.86
		xlm-roberta-base	0.95	0.95	0.95	0.95
	Arabic	bert-base-multilingual-cased	0.91	0.91	0.91	0.91
		Robert-base	0.87	0.86	0.86	0.86
		xlm-roberta-base	0.95	0.94	0.94	0.94
	Urdu	bert-base-multilingual-cased	0.94	0.94	0.94	0.94
		Robert-base	0.74	0.72	0.72	0.72
		xlm-roberta-base	0.94	0.93	0.93	0.93

4.4. Error Analysis

Table 9 presents a comprehensive comparison of the top-performing models for hate speech detection across three different learning approaches—machine learning, deep learning, and transfer learning—applied to datasets in three languages: multi-lingual (combined), Arabic, and Urdu. Each model is evaluated on both binary (hate vs. non-hate) and multi-class classification tasks using key metrics: precision, recall, F1-score, and accuracy. While Figures 7–9 show the confusion matrix of our proposed methodology, xlm-roberta-base in multi-lingual, Urdu translation, and Arabic translations.

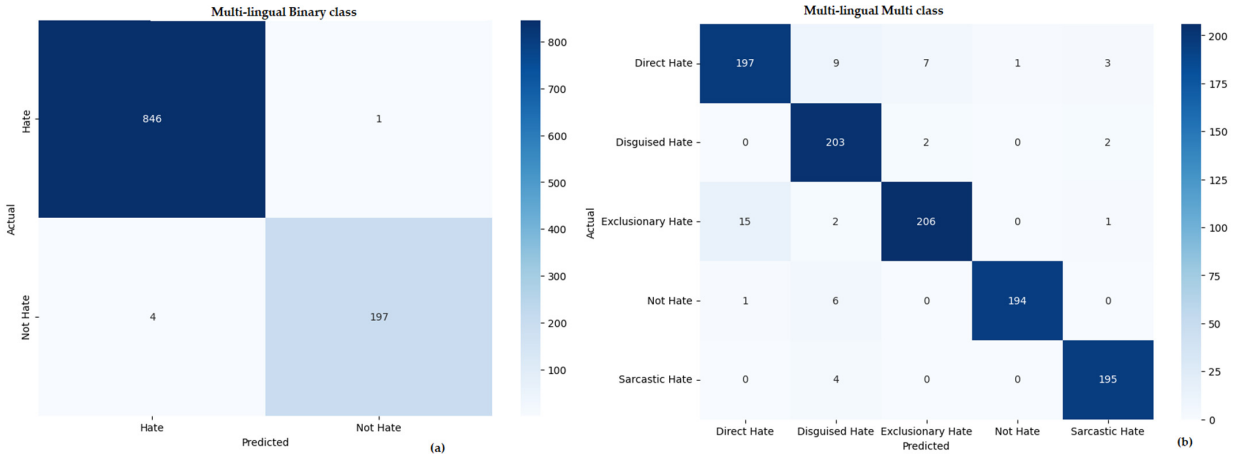


Figure 7. Confusion matrix for binary and multi-class hate speech detection in a multilingual setting: (**a**) joint multilingual binary classification, and (**b**) joint multilingual multi-class classification.

Table 9. Top-performing models in each learning approach employed in this study.

Class	Model	Language	Precision	Recall	F1-Score	Accuracy
Machine learning						
Binary	DT	Multi-lingual	0.86	0.87	0.85	0.87
	SVM	Arabic	0.87	0.87	0.84	0.87
	DT	Urdu	0.95	0.95	0.95	0.95
Multi-class	DT	Multi-lingual	0.62	0.62	0.62	0.62
	SVM	Arabic	0.63	0.6	0.6	0.6
	KNN	Urdu	0.78	0.77	0.77	0.77
Deep Learning						
Binary	BiLSTM	Multi-lingual	0.98	0.98	0.98	0.98
	BiLSTM	Arabic	0.99	0.99	0.99	0.99
	BiLSTM	Urdu	0.98	0.98	0.98	0.98
Multi-class	CNN	Multi-lingual	0.81	0.8	0.8	0.8
	BiLSTM	Arabic	0.87	0.87	0.87	0.87
	CNN	Urdu	0.86	0.86	0.86	0.86
Transfer Learning						
Binary	XLM-R	Multi-lingual	0.99	0.99	0.99	0.99
		Arabic	0.99	0.99	0.99	0.99
		Urdu	0.99	0.99	0.99	0.99
Multi-class	XLM-R	Multi-lingual	0.95	0.95	0.95	0.95
		Arabic	0.95	0.94	0.94	0.94
		Urdu	0.94	0.93	0.93	0.93

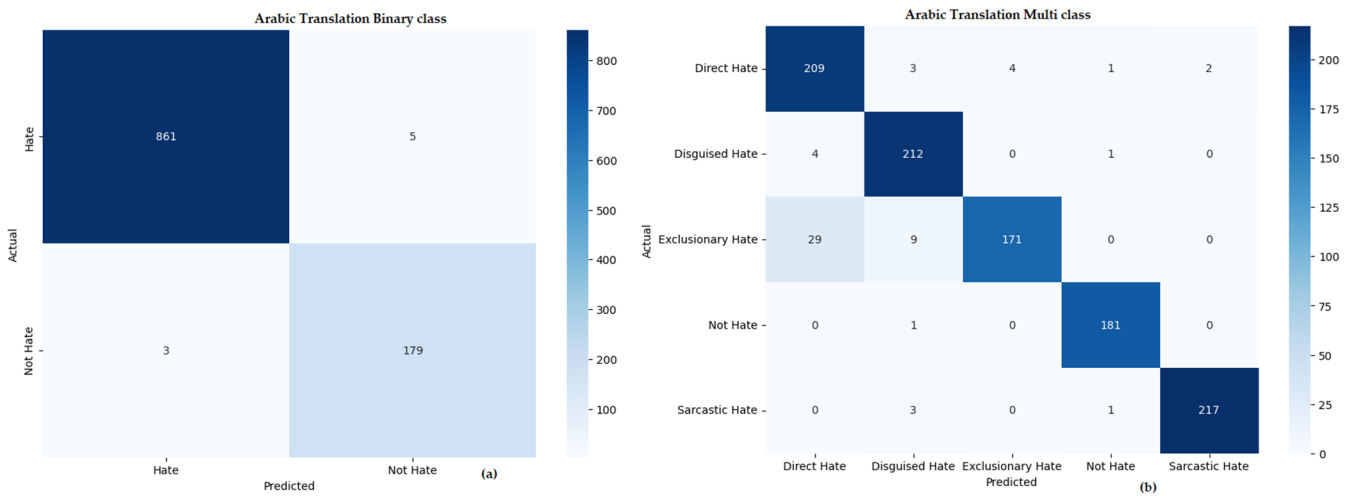


Figure 8. Confusion matrix for binary and multi-class hate speech detection in the Arabic translation setting: (**a**) Arabic translation (binary classification), and (**b**) Arabic translation (multi-class classification).

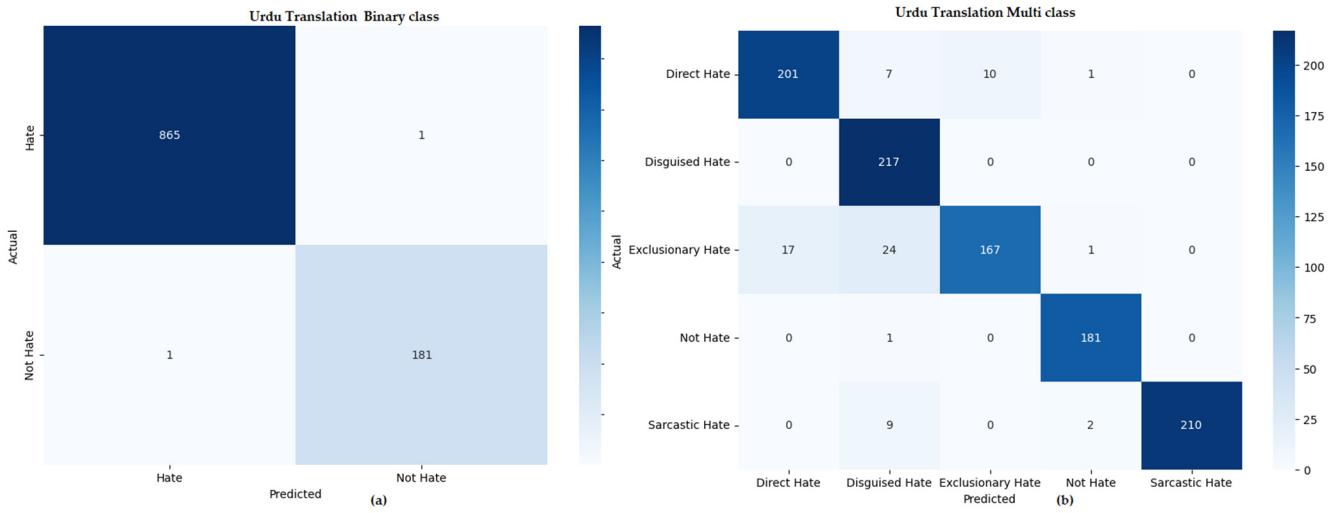


Figure 9. Confusion matrix for binary and multi-class hate speech detection in the Urdu translation setting: (**a**) Urdu translation (binary classification), and (**b**) Urdu translation (multi-class classification).

In the machine learning category, Decision Trees (DT) and Support Vector Machines (SVM) are used. For binary classification, the best results are seen in Urdu using DT, with perfect or near-perfect scores across all metrics (95%). SVM also performs decently on Arabic with an F1-score of 0.84. However, performance drops in the multi-class setup, where DT and KNN yield modest results, particularly in multi-lingual and Arabic datasets, reflecting the challenge of distinguishing between multiple hate speech categories using traditional ML models.

Deep learning methods show a significant performance boost. In binary classification, BiLSTM models excel across all three datasets, achieving near-perfect scores (98–99%), highlighting their strength in capturing sequence patterns in text. Even in the more complex multi-class setting, models like CNN and BiLSTM maintain strong performance, particularly for Urdu and Arabic, with F1-scores consistently above 0.86.

Transfer learning outperforms both traditional and deep learning methods. Models like XLM-RoBERTa (XLM-R) achieve exceptional results in both binary and multi-class settings, with F1-scores and accuracy as high as 0.99 in binary classification and around 0.95 in multi-class tasks. These results affirm the superior ability of pre-trained multilingual transformer models to generalize across languages and nuanced hate speech categories.

Overall, the table demonstrates that while traditional models can be effective for simpler binary classification—especially in high-resource languages like Urdu—deep learning and transfer learning approaches dominate in both binary and multi-class tasks, particularly when working with multilingual datasets.

5. Discussions

The results of this study present a comparative evaluation of traditional machine learning, deep learning, and transformer-based models for hate speech detection across multilingual, Arabic, and Urdu datasets under both binary and multi-class classification tasks. Traditional machine learning models such as Decision Trees and SVMs show competitive performance in binary classification, particularly in Urdu, where the Decision Tree achieves the highest F1-score (0.95). However, their effectiveness declines in multi-class tasks, especially in multilingual and Arabic datasets, highlighting the limitations of these models in capturing complex semantic distinctions across languages. Deep learning models like CNN and BiLSTM, trained with FastText embeddings, offer notable improvements over traditional models by leveraging subword information to capture nuanced linguis-

tic patterns. BiLSTM consistently outperforms CNN, particularly in binary classification (e.g., F1-score of 0.99 for Arabic), and performs competitively in multi-class settings with an F1-score of 0.87 for Arabic, demonstrating its strength in modeling sequential dependencies in language. The transformer-based models, especially XLM-RoBERTa, exhibit superior performance across both binary and multi-class tasks and all language settings, achieving near-perfect scores (0.99 F1) in binary classification and maintaining high performance (0.95 F1) in the more complex multi-class scenario. These results collectively underscore the robustness and scalability of transformer models for hate speech detection in morphologically rich and low-resource languages, while also revealing the relative strengths of deep learning over traditional machine learning in handling multilingual text. Overall, the findings affirm that as model complexity and contextual understanding increase—from traditional ML to deep learning to transformers—so does the accuracy and reliability of hate speech detection across diverse linguistic contexts.

6. Limitations

Despite the robustness of our dataset and methodology, several limitations exist in the process of annotating and detecting hate speech. First, the annotation of hate speech is inherently subjective, especially for ambiguous cases like sarcasm and hidden hate. For example, the Arabic post "ما شاء الله، دائمًا أفكارهم عبقرية جدًا! 😂" ("MashaAllah, their ideas are always so brilliant! 😂") could be interpreted as sarcasm or a genuine compliment, leading to inconsistencies in labeling. Similarly, in Urdu, a phrase like "کتني اچھي ہیں، ہمیشہ ہماری مدد کرتي ہیں! 😎" ("How great they are, always helping us! 😎") could be a sincere compliment or sarcastic, making it difficult to determine the true intent.

Second, the reliance on textual data without contextual information, such as user intent or historical interactions, can significantly affect classification accuracy. A tweet like "هم دائمًا الأفضل، لا أحد يضاهيهم!" ("They are always the best, no one can match them!") might be positive in one context but sarcastic in another. Similarly, an Urdu post like "ہمیشہ بہترین کام کرتي ہیں، بس دنیا کو دکھاني کي لیي!" ("Always doing the best work, just to show the world!") might be interpreted as either genuine praise or a sarcastic remark depending on the context in which it is used.

Third, the role of emojis and figurative language adds further complexity. Emojis, such as the 😂 emoji, can signal humor or mockery, altering the meaning of a sentence. For example, in the Arabic phrase "ما شاء الله، دائمًا أفكارهم عبقرية جدًا! 😂" ("MashaAllah, their ideas are always so brilliant! 😂"), the inclusion of the laughing emoji could turn a seemingly positive statement into a sarcastic one. Similarly, in Urdu, a statement like "سب کچھ بہت اچھا ہي! 😎" ("Everything is great! 😎") could either indicate genuine praise or mockery, depending on the tone conveyed by the emoji.

Fourth, our initial tweet collection relied heavily on keyword-based filtering, which may introduce sampling bias. While we attempted to create a broad keyword list across both Arabic and Urdu, implicit or context-dependent expressions of hate speech that do not contain explicit keywords might have been missed.

Fifth, our dataset is solely sourced from Twitter. Although Twitter provides a rich and publicly accessible stream of sociopolitical discourse, the style and frequency of hate speech may vary significantly across platforms such as Facebook, Reddit, or YouTube. As such, our findings may not generalize well to other platforms or media.

Sixth, during the manual annotation process, we applied strict inclusion criteria to ensure high-quality data. Only tweets that explicitly or implicitly exhibited hate speech were retained, resulting in a reduction from 30,000 initially collected tweets to 5240 final

samples. This high discard rate (>80%) reflects the importance we placed on annotation precision, but also raises concerns about potential over-filtering or an over-inclusive initial keyword list.

Finally, deep learning models require substantial computational resources, making real-time deployment challenging in resource-limited environments. The complexity of sarcasm, hidden hate, and figurative language means that real-time processing becomes even more demanding. This is particularly true for resource-limited platforms like mobile applications, where the computational cost could be prohibitive.

7. Conclusions and Future Work

This study addresses the underexplored challenge of multilingual hate speech detection in Arabic and Urdu, two linguistically rich yet low-resource languages. By introducing the UA-HSD-2025 dataset with both binary and multi-class annotations and implementing rigorous annotation guidelines, we provide a valuable resource for future research. Our comparative analysis of traditional machine learning, deep learning, and advanced transfer learning techniques—across 54 experiments—demonstrates the superior performance of contextual language models, particularly XLM-R, in both binary and multi-class classification tasks. The joint multilingual and translation-based approaches further highlight effective strategies for handling multilingual data. Overall, our findings emphasize the importance of language-specific datasets and contextual embeddings in enhancing hate speech detection systems, paving the way for more inclusive and accurate NLP solutions in low-resource settings.

For future work, we aim to expand the dataset by incorporating more dialectal variations and sources beyond Twitter to enhance generalizability. Additionally, we plan to explore large language models (LLMs) to further improve classification performance. Investigating adversarial attacks and bias mitigation techniques will also be crucial in ensuring the robustness and fairness of Arabic and Urdu hate speech detection tasks. Furthermore, we intend to incorporate more context-aware modeling approaches by leveraging user history, conversational context, and meta-information to better capture disguised, sarcastic, and context-dependent hate speech. We also plan to evaluate the impact of various preprocessing steps on model performance to better preserve important contextual cues. Finally, conducting qualitative error analyses on misclassified or ambiguous cases will provide valuable insights to refine both the dataset and model architectures.

Author Contributions: Conceptualization, M.A. and M.W.; methodology, M.A.; software, M.A. and A.H.; validation, M.A., A.H., G.S. and I.B.; formal analysis, I.B., S.U. and M.A.; investigation, M.A., S.U. and G.S.; resources, M.W.; data curation, M.W. and M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A. and S.U.; visualization, A.H. and M.A.; supervision, G.S.; project administration, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset utilized in this study is not publicly available due to ongoing research, but can be provided upon reasonable request. Interested researchers should contact the first author at mahmad2024@cic.ipn.mx, Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC-PN), Mexico City 07738, Mexico. Requests must include a detailed description of the intended use and the requester's institutional affiliation.

Acknowledgments: The work was completed with partial support from the Mexican Government through the grant A1-S-47854 of CONAHCYT, Mexico, grants 20241816, 20241819, and 20240951 of

the Secretaría de Investigación y Posgrado of the InstitutoPolitécnicoNacional, Mexico. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de AprendizajeProfundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Conflicts of Interest: The authors declare no conflicts of interest in this study.

References

1. Ullah, F.; Zamir, M.T.; Ahmad, M.; Sidorov, G.; Gelbukh, A. Hope: A multilingual approach to identifying positive communication in social media. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, 24 September 2024.
2. Ahmad, M.; Usman, S.; Farid, H.; Ameer, I.; Muzammil, M.; Hamza, A.; Batyrshin, I. Hope speech detection using social media discourse (Posi-Vox-2024): A transfer learning approach. *J. Lang. Educ.* **2024**, *10*, 31–43. [CrossRef]
3. Ahmad, M.; Ameer, I.; Sharif, W.; Usman, S.; Muzamil, M.; Hamza, A.; Sidorov, G. Multilingual hope speech detection from tweets using transfer learning models. *Sci. Rep.* **2025**, *15*, 9005. [CrossRef] [PubMed]
4. Meghazi, H.M.; Mostefaoui, S.A.; Maaskri, M.; Aklouf, Y. Deep Learning-Based Text Classification to Improve Web Service Discovery. *Comput. Y Sistemas* **2024**, *28*, 529–542. [CrossRef]
5. Zaman-Khan, H.; Naeem, M.; Guarasci, R.; Bint-Khalid, U.; Esposito, M.; Gargiulo, F. Enhancing text classification using BERT: A transfer learning approach. *Comput. Y Sistemas* **2024**, *28*, 5290. [CrossRef]
6. Mullah, N.S.; Zainon, W.M.N.W. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access* **2021**, *9*, 88364–88376. [CrossRef]
7. Yuan, L.; Rizoiu, M.A. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Comput. Speech Lang.* **2025**, *89*, 101690. [CrossRef]
8. Daouadi, K.E.; Boualleg, Y.; Guehairia, O. Comparing pre-trained language models for Arabic hate speech detection. *Comput. Y Sistemas* **2024**, *28*, 681–693. [CrossRef]
9. Sultan, D.; Toktarova, A.; Zhumadillayeva, A.; Aldeshov, S.; Mussiraliyeva, S.; Beissenova, G.; Imanbayeva, A. Cyberbullying-related hate speech detection using shallow-to-deep learning. *Comput. Mater. Contin.* **2023**, *75*, 2115–2131. [CrossRef]
10. Ibrahim, Y.M.; Essameldin, R.; Saad, S.M. Social media forensics: An adaptive cyberbullying-related hate speech detection approach based on neural networks with uncertainty. *IEEE Access* **2024**, *12*, 59474–59484. [CrossRef]
11. Mahmud, T.; Ptaszynski, M.; Eronen, J.; Masui, F. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Inf. Process. Manag.* **2023**, *60*, 103454. [CrossRef]
12. Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; Prasad, M. Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics* **2021**, *10*, 2810. [CrossRef]
13. Singh, V.K.; Ghosh, S.; Jose, C. Toward multimodal cyberbullying detection. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 2090–2099.
14. Al-Harigy, L.M.; Al-Nuaim, H.A.; Moradpoor, N.; Tan, Z. Building towards automated cyberbullying detection: A comparative analysis. *Comput. Intell. Neurosci.* **2022**, *2022*, 4794227. [CrossRef] [PubMed]
15. Bayari, R.; Bensefia, A. Text mining techniques for cyberbullying detection: State of the art. *Adv. Sci. Technol. Eng. Syst. J.* **2021**, *6*, 783–790. [CrossRef]
16. Demilie, W.B.; Salau, A.O. Detection of fake news and hate speech for Ethiopian languages: A systematic review of the approaches. *J. Big Data* **2022**, *9*, 66. [CrossRef]
17. Ameur, M.S.H.; Aliane, H. Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset. *Procedia Comput. Sci.* **2021**, *189*, 232–241.
18. Liu, M.; Liu, Y.; Fu, R.; Wen, Z.; Tao, J.; Liu, X.; Li, G. Exploring the role of audio in multimodal misinformation detection. In Proceedings of the 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), Beijing, China, 7–10 November 2024; pp. 204–208.
19. Abdali, S.; Shaham, S.; Krishnamachari, B. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Comput. Surv.* **2024**, *57*, 1–29. [CrossRef]
20. Cinelli, M.; Pelicon, A.; Mozetič, I.; Quattrociocchi, W.; Novak, P.K.; Zollo, F. Dynamics of online hate and misinformation. *Sci. Rep.* **2021**, *11*, 22083. [CrossRef]
21. Shang, L.; Kou, Z.; Zhang, Y.; Wang, D. A multimodal misinformation detector for covid-19 short videos on tiktok. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 899–908.

22. Bade, G.; Kolesnikova, O.; Sidorov, G.; Oropeza, J. Social media hate and offensive speech detection using machine learning method. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, St. Julian's, Malta, 22 March 2024; pp. 240–244.
23. Nghiem, H.; Daumé, H., III. HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models. *arXiv* **2024**, arXiv:2403.11456.
24. Krak, I.; Zalutska, O.; Molchanova, M.; Mazurets, O.; Bahrii, R.; Sobko, O.; Barmak, O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. *COLINS* **2024**, *3*, 16–28.
25. Sekkate, S.; Chebbi, S.; Adib, A.; Jebara, S.B. A deep learning framework for offensive speech detection. In Proceedings of the 2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC), Marrakech, Morocco, 21–23 May 2024; pp. 1–6.
26. Liu, G.; Yang, X.; Shi, X.; Li, Y. Unsupervised offensive speech detection for multimedia based on multilingual BERT. *Int. J. Sens. Netw.* **2024**, *46*, 186–196. [CrossRef]
27. Alghamdi, J.; Luo, S.; Lin, Y. A comprehensive survey on machine learning approaches for fake news detection. *Multimed. Tools Appl.* **2024**, *83*, 51009–51067. [CrossRef]
28. Khan, K.A.; Anser, M.K.; Pala, F.; Barut, A.; Zafar, M.W. Promoting sustainability in developing Countries: A Machine Learning-based approach to understanding the relationship between green investment and environmental degradation. *Gondwana Res.* **2024**, *132*, 136–149. [CrossRef]
29. Pradhan, J.; Verma, R.; Kumar, S.; Sharma, V. An efficient sarcasm detection using linguistic features and ensemble machine learning. *Procedia Comput. Sci.* **2024**, *235*, 1058–1067. [CrossRef]
30. Ganganwar, V.; Manvainder Singh, M.; Patil, P.; Joshi, S. Sarcasm and Humor Detection in Code-Mixed Hindi Data: A Survey. In *International Conference on Computing and Machine Learning*; Springer Nature: Singapore, 2024; pp. 453–469.
31. Khazeni, M.; Heydari, M.; Albadvi, A. Persian Slang Text Conversion to Formal and Deep Learning of Persian Short Texts on Social Media for Sentiment Classification. *arXiv* **2024**, arXiv:2403.06023.
32. Sundaram, A.; Subramaniam, H.; Ab Hamid, S.H.; Nor, A.M. A three-step procedural paradigm for domain-specific social media slang analytics. In Proceedings of the 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, Pune, India, 22–23 March 2024; pp. 1–7.
33. Embassy of Sri Lanka, Saudi Arabia. Arabic, Spoken by over 450 Million People and Holding Official Status in Nearly 25 Countries, Is a Global Language with Immense Cultural Significance. 2024. Available online: https://slemb.org.sa/2024/arabic-spoken-by-over-450-million-people-and-holding-official-status-in-nearly-25-countries-is-a-global-language-with-immense-cultural-significance/ (accessed on 12 April 2025).
34. Elteir, M.K. Cost-effective time-efficient subnational-level surveillance using Twitter: Kingdom of Saudi Arabia case study. *Discov. Appl. Sci.* **2025**, *7*, 60. [CrossRef]
35. Ethnologue. List of Languages by Total Number of Speakers. 2025. Available online: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers (accessed on 12 April 2025).
36. Alshaabi, T.; Dewhurst, D.R.; Minot, J.R.; Arnold, M.V.; Adams, J.L.; Danforth, C.M.; Dodds, P.S. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *arXiv* **2020**, arXiv:2003.03667. [CrossRef]
37. Ni, W.; Wang, T.; Wu, Y.; Chen, L.; Zeng, M.; Yang, J.; Yang, Z. Robust odor detection in electronic nose using transfer-learning powered Scentformer model. *ACS Sens.* **2025**, *10*, 3704–3712. [CrossRef]
38. Alkomah, F.; Ma, X. A literature review of textual hate speech detection methods and datasets. *Information* **2022**, *13*, 273. [CrossRef]
39. Albladi, A.; Islam, M.; Das, A.; Bigonah, M.; Zhang, Z.; Jamshidi, F.; Seals, C. Hate Speech Detection using Large Language Models: A Comprehensive Review. *IEEE Access* **2025**, *13*, 20871–20892. [CrossRef]
40. Hashmi, E.; Yayilgan, S.Y.; Yamin, M.M.; Abomhara, M.; Ullah, M. Self-supervised hate speech detection in norwegian texts with lexical and semantic augmentations. *Expert Syst. Appl.* **2025**, *264*, 125843. [CrossRef]
41. Chavinda, K.; Thayasivam, U. A Dual Contrastive Learning Framework for Enhanced Hate Speech Detection in Low-Resource Languages. In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), Abu Dhabi, United Arab Emirates, 19 January 2025; pp. 115–123.
42. Thapa, S.; Rauniyar, K.; Jafri, F.A.; Adhikari, S.; Sarveswaran, K.; Bal, B.K.; Naseem, U. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), Abu Dhabi, United Arab Emirates, 19 January 2025; pp. 71–82.
43. Khadka, P.; Bk, A.; Acharya, A.; Kc, B.; Shrestha, S.; Thapa, R. Nepali Transformers@ NLU of Devanagari Script Languages 2025: Detection of Language, Hate Speech and Targets. In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), Abu Dhabi, United Arab Emirates, 19 January 2025; pp. 314–319.

44. Finch, W.H.; French, B.F. *Educational and Psychological Measurement*; Routledge: Abingdon, UK, 2018.
45. Falotico, R.; Quatto, P. Fleiss' kappa statistic without paradoxes. *Qual. Quant.* **2015**, *49*, 463–470. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.