Slur and Emoji Aware Models for Hate and Sentiment Detection in Roman Urdu Transgender Discourse

Muhammad Owais Raza
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul 34303, Turkey
6210024002@std.izu.edu.tr

Aqsa
Sindh Madressatul Islam University
City Campus, Karachi
aqsa.umar3505@gmail.com

Mehrub Awan
Gender Interactive Alliance
mehrubmoizawan@gmail.com

Abstract

The rise of social media has amplified both the visibility and vulnerability of marginalized communities, such as the transgender population in South Asia. While hate speech detection has seen considerable progress in high resource languages like English, under-resourced and code mixed languages such as Roman Urdu remain significantly understudied. This paper presents a novel Roman Urdu dataset derived from Instagram comments on transgender related content, capturing the intricacies of multilingual, code-mixed, and emoji-laden social discourse. We introduce a transphobic slur lexicon specific to Roman Urdu and a semantic emoji classification grounded in contextual usage. These resources are utilized to perform fine-grained classification of sentiment and hate speech using both traditional machine learning models and transformer-based architectures. The findings show that our custom-trained BERT-based models, Senti-RU-Bert and Hate-RU-Bert, show the best performance, with F1 scores of 80.39% for sentiment classification and 77.34% for hate speech classification. Ablation studies reveal consistent performance gains when slur and emoji features are included.

1 Introduction

In recent years, there has been a significant increase in social networks and content consumption. What people share on social media platforms has a direct impact on their daily lives (Aziz et al., 2023). Social media has become a powerful platform for sharing ideas and perspectives; however, it is also increasingly misused to spread hate against individuals, groups, and communities. Such content poses serious threats to social harmony, online safety, and mental health (Sharma et al., 2025), contributing to the alarming rise in hate speech.

Hateful speech is a type of language which conveys the negative sentiment that can shame users while also promoting radicalism and inciting violence (Gitari et al., 2015). Hate speech is a growing issue, especially in the comment sections of social media platforms such as Instagram, Facebook, and YouTube. Several approaches to categorizing hate speech on social media platforms were investigated in the study (Martins et al., 2018). They presented a combination of machine learning and lexicon-based methods for predicting hate speech. Notably, they employed emotional content in sentences to improve the detection accuracy of hate speech.

While hate speech detection in English (Shahi and Majchrzak, 2024; Pan et al., 2024; Gandhi et al., 2024) have been widely studied, low resource South Asian languages remain significantly underexplored, leaving millions of social media users vulnerable. In Pakistan, for instance, around 26% of the population is bilingual, leading to frequent code mixing in user generated content (Aziz et al., 2023). This linguistic diversity often results in text that blends English script with local languages such as Roman Urdu and Roman Sindhi posing unique challenges for automated detection systems. Addressing this gap requires the development of dedicated resources for these underrepresented languages.

Although sentiment analysis in Roman Urdu has received some attention in prior work, substantial limitations persist. Several Roman Urdu datasets are publicly available, and previous studies have focused on unsupervised lexical normalization (Mehmood et al., 2020b) and the impact of lexical variation on sentiment classification (Manzoor et al., 2020). Notable efforts include the Roman Urdu E-commerce Dataset (RUECD), which comprises 26,824 customer reviews from DarazPK (Chandio et al., 2022), a binary-labeled dataset of 11,000 reviews across six domains (Mehmood et al., 2019), and a third dataset with 3,241 annotated attitudes (Mehmood et al., 2020a). Despite these

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 131–139, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-014

efforts, Roman Urdu remains a low-resource language, and progress in sentiment and hate speech detection is hindered by a lack of standardized linguistic tools and comprehensive annotated datasets (Khan et al., 2024). This scarcity highlights the pressing need for further research and resource development to support natural language understanding in code-mixed, under-resourced settings.

Existing studies on Urdu language television dramas reveal a distorted portrayal of transgender characters, often exaggerating dominance and neglecting economic violence while emphasizing psychological abuse (Mehmood et al., 2020b; Manzoor et al., 2020; Ullah et al., 2024). Despite such media shaping cultural discourse, no curated Roman Urdu dataset exists that reflects the unique linguistic patterns and hate speech dynamics in social media discussions about transgender people in Pakistan. This underscores an urgent need to build specialized Roman Urdu datasets for hate speech detection that can more accurately capture the sociolinguistic realities of digital discourse in underrepresented languages and communities.

In this study, we address the underexplored issue of hate speech and sentiment analysis in Roman Urdu social media discourse, particularly focusing on transgender related content. We construct a novel dataset of Instagram comments that reflects real-world multilingual and code-mixed usage, enriched with emojis and varying sentiment. To capture the specific linguistic and multimodal nuances of online hate, we also develop a Roman Urdu transphobic slur lexicon and a context aware emoji classification. These resources enable deeper analysis of abuse patterns and facilitate fine-grained feature extraction. Building on these contributions, we perform a thorough evaluation of traditional machine learning models and transformer-based architectures, comparing their performance across different combinations of textual, lexical, and emoji-based features for both sentiment and hate speech classification tasks. This comprehensive approach highlights the complexities of detecting hate in low resource, code-mixed contexts and offers tools to improve detection and understanding in similar sociolinguistic settings. Following are three main contribution of this study:

- Creation of a novel, Roman Urdu dataset from Instagram comments on transgender related content, incorporating multilingual and emoji-inclusive text.
- Development of a Roman Urdu based transphobic slur lexicon and an emoji classification, enabling fine grained linguistic and multimodal feature enrichment.
- Conducted a comprehensive evaluation of ML and transformer models using text, emoji, and slur features for sentiment and hate speech classification.

2 Dataset

The dataset used in this study was collected from the Instagram comment sections of Binax Studio[1], a social media page focusing on transgender oriented content in Pakistan. The dataset comprises comments written in multiple languages, including Roman Urdu, Urdu (Arabic script), Hindi (Devanagari), Arabic, English and a high frequency of emojis. Given their dominance in the dataset, we retained only comments written in Roman Urdu or code-mixed Roman Urdu (i.e., Roman Urdu combined with English or emojis), as well as purely emoji based expressions, for further analysis. A zero-shot learning approach with human feedback was employed to label the comments. A typical prompt provided to the model was:

> “Analyze the following text and return JSON with exactly these keys: - sentiment: one of ’positive’, ’neutral’, or ’negative’ - negative_type: if sentiment is ’negative’, classify it as ’abusive’, ’threatening’, ’call for action’, or ’other’. Else null. - language: detect the language or script used, like ’Urdu’, ’Roman Urdu’, ’English’, etc. Consider emoji and transliteration.”

For reliability, two bilingual authors (fluent in Urdu, Roman Urdu, and English) independently examined 200 model labeled comments. Substantial agreement was achieved with the model’s assignments (Cohen’s $\kappa = 0.75$), and any differences were settled through discussion, which informed adjustments to the labeling guidelines.

Table 1 summarizes the distribution of content types in the dataset. A majority of the comments (1584) were written in Roman Urdu without emojis, followed closely by 1282 comments that combined

[1]https://www.instagram.com/binax.studio

Roman Urdu with emojis. Additionally, 294 comments consisted entirely of emojis. The dataset was annotated for two key tasks:

Table 1: Distribution of Content Types with Examples

Content Type	Count	Example
Roman Urdu	1584	Ab khusron ky bhi podcast hongy?
Roman Urdu + Emoji	1282	Apko dekh kr ma kya hi judge krunga 😭
Emoji Only	294	😂😂😂😂😂
Total	3160	–

2.1 Task 1: Sentiment Analysis

The first task involved classifying each comment into one of three sentiment categories: Positive, Neutral, or Negative. This task captures the overall emotional tone of public engagement with transgender content (Mao et al., 2024).

- **Positive:** Comments that express support, admiration, or encouragement.

  Example: ”Jaan ho 😍”

  Translation: ”You are my life”

- **Neutral:** Comments that are factual, descriptive, or unclear in tone.

  Example: “yaar yeh q ka mtlv kya hai”

  Translation: ”Hey, what does this mean?”

- **Negative:** Comments expressing disapproval, mockery, or hostility.

  Example: ”Saleh hijre”

  Translation: ”Damn eunuch”

2.2 Task 2: Hate Speech Classification

The second task involved classifying hate speech present in the comments into four categories: Abusive, Threat, Call for Action, and Other. This categorization aims to capture different intensities and types of harmful speech targeting transgender individuals (Kumar et al., 2025).

- **Abusive:** Insults, slurs, or dehumanizing language directed at transgender people.

  Example: “Tum dono ki MKC”

  Translation: ”You both are motherf***ers” (highly offensive)

- **Threat:** Explicit or implicit threats of violence or harm.

  Example: “The scream at the end will be u screaming when i will execute u. Fitnah phelate ho tum loug. JAHANAMMI!!”

  Translation: ”You spread chaos. The scream at the end will be yours when I execute you. You are hell-bound!!”

- **Call for Action:** Urging others to act against transgender individuals, including boycotts or violence.

  Example: “Report karo inko”

  Translation: ”Report them”

- **Other:** Includes misgendering, sarcastic derision, or religiously framed delegitimization that doesn’t fit the above but still contributes to a hostile environment.

  Example: “Ouch”

  Translation: Same in English; sarcastic or mocking tone.

Table 2 presents the distribution of sentiment and hate speech categories within the filtered dataset. Out of 3160 comments, the majority (1603) expressed negative sentiment, followed by 966 positive and 591 neutral comments. In terms of hate speech, 1225 comments were labeled as abusive, while a smaller number fell into the categories of other (339), threatening (22), and call for action (17).

Table 2: Distribution of Sentiment and Hate Speech Categories

Sentiment	Count	Hate Speech Type	Count
Negative	1603	Abusive	1225
Positive	966	Other	339
Neutral	591	Threatening	22
–	–	Call for Action	17

3 Linguistic Analysis

This section explores the linguistic features and affective cues used in online discourse targeting transgender individuals, particularly in Roman Urdu. We focus on two key dimensions: (1) lexical abuse through Roman Urdu slurs, and (2) the affective and rhetorical functions of emojis in these hostile or supportive comments. Together, they reveal how language and visual symbols convey support, mockery, identity assertion, or threats.

3.1 Emoji Classification Design

To analyze the role of emojis in transgender-related discourse, we developed a custom taxonomy based on contextual usage rather than Unicode semantics. As shown in Figure 1, we identified ten functional categories: Supportive/Affective (affection, solidarity), Mocking/Dismissive (ridicule, sarcasm), Identity Pride (queer or gender identity markers), Aggressive/Threatening (symbolic violence, hostility), Gesture/Emphasis (tone amplification), Religious/Moral (judgment) Humor/Ambiguous (irony, camp), Sadness/Vulnerability (grief, helplessness), Body/Gendered (physical or sexed features), and Mock Femininity (caricatured feminized traits). Although each emoji was placed in one main category based on its context, in rare cases, the same emoji could fit into more than one category depending on how it was used. Building on the taxonomy, Figure 2 presents the distribution of emoji usage across the dataset. The most frequent category was Mocking (38.76%), highlighting the prevalence of ridicule and sarcasm. This was followed by Supportive/Affective (27.33%), indicating a substantial presence of emotional solidarity. Mock Femininity (8.84%) and Aggressive/Threatening (7.51%) emojis also appeared prominently, often signaling coded transphobia or symbolic hostility.

3.2 Construction of Transphobic Slur Lexicon for Roman Urdu

We constructed a domain specific lexicon by identifying 124 unique slurs from the comment dataset. This lexicon includes both explicit terms and more implicit or coded expressions used in South Asian digital discourse (e.g., chakka, khusra, etc). The lexicon construction was in done in following 4 steps.

3.2.1 Orthographic Normalization:

To consolidate orthographic variants of abusive terms, we applied phoneme aware normalization rules that convert common Roman Urdu digraphs and vowel elongations into base forms (e.g., *"gandoo", "gaanduu"* $\rightarrow$ *gandu*). This step was implemented using regular expressions and phonetic substitution rules that we manually created to handle common Roman Urdu spelling variants (e.g.,"aa"/"a"). As no standardized resources exist for Roman Urdu normalization, our rules were iteratively refined through manual inspection of sample outputs. To group closely related variants and misspellings, tokens were clustered based on Levenshtein edit distance (threshold ≤ 2). This approach ensured that minor typographical variations were treated as a single lexeme.

3.2.2 Token Filtering and Frequency Thresholding:

We removed stopwords using an expanded bilingual stopword list covering both Roman Urdu and English function words. Tokens with a frequency less than three were discarded to focus on commonly used terms.

3.2.3 Manual Filtering:

After converting minor typographical variations into single lexemes, we obtained 664 tokens. These were manually reviewed to remove ambiguous or contextually irrelevant words, resulting in a final curated lexicon of 124 semantically abusive and transphobic slurs.

Table 3 shows the most common abusive and transphobic slurs found in our dataset. These slurs frequently appeared in user comments and were retained in our final lexicon after normalization and manual review. Many of the terms, such as `bc` and `bkl`, are abbreviations or phonetic spellings, but within the comment context, they clearly function as tools of verbal abuse. While this data is specific to our collected dataset, it reflects a wider trend of how such harmful language is casually and repeatedly used in online conversations related to transgender topics in the South Asian social media space.

Table 3: Most common slurs in the dataset based on our Roman Urdu Transphobic Slur lexicon

Normalized	**Count**	**Description**
`bc`	58	Abbreviation of incest-based Urdu profanity
`khusra`	57	Slur for transgender person (Roman Urdu for *Khusra*)
`gand`	44	Vulgar term for buttocks (Roman Urdu)
`bkl`	32	Abbreviation for "idiot" or "stupid" in abuse contexts
`gandu`	24	Used in the context of abuse to call someone F*gg*t

Table 4 shows the most common co-occurrences

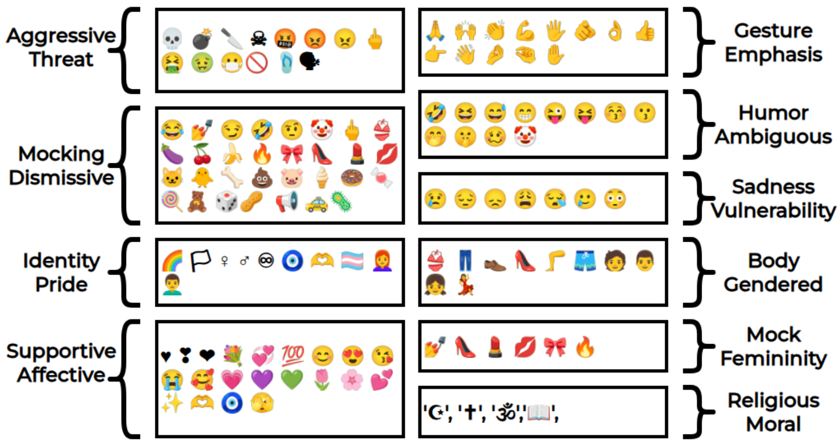

Figure 1: Proposed Emoji Classification

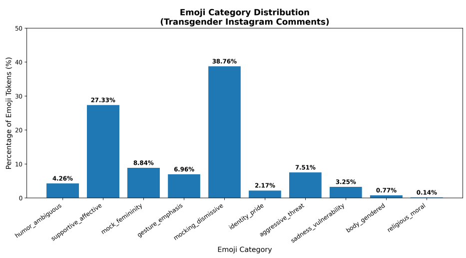

Figure 2: Emoji Distribution Based on Proposed Emoji Classification

between transphobic slurs and emojis within Roman Urdu comments. These combinations were extracted from our dataset and reveal how users often use affective emojis especially the 😂 (joy) emoji alongside abusive terms like *gand*, *bc*, and *chakka*. The use of such emojis tends to amplify mockery, express sarcasm, or downplay the abuse, making it seem more casual. This highlights an important multimodal layer of hate speech and shows why emoji signals should be included in models that detect implicit or indirect abuse.

Table 4: Top Slur Emoji Co-occurrences by Frequency

Slur (Description)	Emoji	Count
`gand` (buttocks)	😂	29
`bc` (incest profanity)	😂	19
`chakka` (trans slur)	😂	18
`lan` (penis)	😂	15
`chod` (f**k)	😂	14

Table 5 presents examples of how emojis frequently appear alongside transphobic or vulgar slurs in Roman Urdu social media comments. Each row includes a commonly used slur, its co-occurring emoji, the frequency of this pairing, and an anonymized usage example translated into English. Emojis like 😂 (laughter) or 💀 (mockery) often shift the emotional tone either by making the abuse sound humorous, sarcastic, or more intense. These combinations reflect a deliberate use of visual cues to strengthen or mask hateful intent, making the abuse seem more socially acceptable or performative. The findings emphasize the need to treat emojis not just as add-ons, but as important features in detecting and understanding online hate speech.

Table 5: Examples of Emoji Usage in Slur Contexts

Slur	Emoji	Count	Example
`gando`	😂	29	”*Gando log* 😂” ”F*gg*t people” with mocking laughter.
`lanat`	✋	5	”*Lanat* ✋✋” ”Shame/curse” with repeated hand gestures.
`chutiya`	😢	3	”*Chutiya promoting nudity* 😢” Sarcastic or mocking tone.
`bc`	💀	3	”*Chaako ki podcast* 💀 *are bc*” ”Chaako” is a derogatory slur for transgender people, and ”bc” is a strong abusive term

4 Experimental Settings

We conducted comprehensive experiments for two classification tasks: (i) Sentiment analysis (positive, neutral, negative), and (ii) Hate speech classification (e.g., abusive, threatening, etc.). Each task was evaluated using both traditional machine learning models and fine tuned transformer based model.

4.1 Dataset Preprocessing and Feature Engineering

We have dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each x_i is a instance of the dataset (i.e. comment on an Instagram post) and $y_i \in \mathcal{Y}$ is the corresponding label (for either sentiment or hate speech category), we augmented the inputs with structured linguistic and contextual features to capture additional social and semantic cues beyond raw text.

Each input x_i was decomposed into its base text t_i, a categorical emoji feature $e_i \in \mathcal{E}$, and a binary indicator $s_i \in \{0, 1\}$ for the presence of slurs. The slur flag s_i is computed based on a manually curated lexicon lex_{slur} of 124 normalized Roman Urdu transgender related slurs. The flag is defined as:

$$s_i = \begin{cases} 1 & \text{if } \exists\, w \in \text{lex}_{\text{slur}} w \in \text{Tok}(t_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In equation 1 $\text{Tok}(t_i)$ denotes the tokenized form of the comment t_i. The emoji feature e_i was derived by mapping each emoji in t_i to a predefined semantic category (e.g., *mocking*, *affective*, *gesture*, *threat*).

To systematically evaluate the contribution of each feature type, we defined four feature sets: (i) $\mathcal{F}_{\text{text}} = \{t_i\}$, representing raw text only; (ii) $\mathcal{F}_{\text{text+emoji}} = \{t_i, e_i\}$, which includes emoji category alongside text; (iii) $\mathcal{F}_{\text{text+slur}} = \{t_i, s_i\}$, incorporating slur presence; and (iv) $\mathcal{F}_{\text{all}} = \{t_i, e_i, s_i\}$, the full feature combination. These sets were used across experiments to assess the role of lexical and paralinguistic features under controlled comparisons.

4.2 Modeling and Evaluation

Once dataset preprocessing and feature engineering is done, the next step is perform machine learning modeling and evaluation across both sentiment analysis and hate speech type classification tasks. We implemented two modeling approaches: one based on traditional machine learning (ML) algorithms with bag of words representations, and another using a custom fine tuned BERT (Devlin et al., 2019) based transformer model. For consistency each model was trained and tested using a stratified 80/20 train-test split with fixed random seed to ensure reproducibility.

4.2.1 Traditional Machine Learning Models

For classical approaches, we trained a diverse set of supervised classifiers: Logistic Regression (LogReg), Linear Support Vector Machine (LinearSVC), Multinomial Naive Bayes (MultinomialNB), Random Forest, XGBoost, and LightGBM. In this modeling approach the textual input t_i was converted into numerical form using TF-IDF (De Santis et al., 2024) vectorization, capturing both unigrams and bigrams (n-gram range = (1, 2)) and restricted to a maximum of 5000 features. Emoji categories $e_i \in \mathcal{E}$ were encoded using one-hot encoding, and the slur flag $s_i \in \{0, 1\}$ was passed through as a numeric binary feature.

4.2.2 Transformer Based Model

To model context and semantic nuance directly from raw text, we trained a task specific transformer classifier based on BERT (bert-base-uncased). Input text was tokenized using the corresponding WordPiece tokenizer (Schönle et al., 2024), subword information, and special tokens. Each input sequence was truncated or padded to a maximum length of 128 tokens and passed to the BERT encoder. The transformer model was implemented using the HuggingFace Transformers library[2]. All experiments were conducted with a batch size of 8, 3 epochs of training with 50 logging steps.

4.2.3 Model Evaluation

Both traditional machine learning models and the transformer based model were evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1 Score. All metrics were computed using weighted averaging, which accounts for the distribution of samples across classes and ensures that minority classes are not ignored during evaluation. To ensure a fair and consistent comparison across all models and feature combinations, the dataset was partitioned into 80% training and 20% testing subsets using stratified sampling, preserving the original class proportions in both splits.

5 Results

This section presents the performance outcomes of various models applied to two primary tasks:

[2] https://huggingface.co/

Sentiment Classification and Negativity Type (hate speech) Classification. We evaluate both traditional machine learning classifiers and a transformer-based BERT model using four different feature configurations: text only (T), text with emoji features (T+E), text with slur features (T+S), and all features combined (T+E+S). For experiments including emojis (E) and slur flags (S), these features were concatenated as auxiliary inputs to the BERT classifier's final hidden layer before the classification head. The evaluation metrics include Accuracy, Precision, Recall, and F1 Score. All experiments were conducted on Google Colab with a 16 GB RAM and 128 GB disk environment.

Table 6 presents the comparative performance of multiple models across two classification tasks: Sentiment Detection and Hate speech Classification. All models were evaluated using a unified feature set comprising textual (T), emoji features (E), and slur features (S). Among traditional machine learning models, LinearSVC and Random Forest demonstrated relatively strong and consistent performance. However, BERT-based models (Senti-RU-BERT and Hate-RU-BERT) outperformed all baselines across both tasks, achieving the highest accuracy (80.54% for sentiment, 78.01% for negativity type) and F1 scores (80.39% and 77.34% respectively). Our domain specific variants, Senti-RU-BERT and Hate-RU-BERT, fine-tuned on Roman Urdu data with emoji and slur context, yielded the best results. These findings shows the effectiveness of contextual embeddings in capturing the nuanced affective and abusive signals in under-resourced, code-mixed languages especially when enhanced with multimodal cues.

Table 6: Performance (Accuracy, Precision, Recall, F1) for Sentiment and Hate Speech Classification

Task	Model	Acc.	Prec.	Rec.	F1
Sentiment	LinearSVC	74.37	74.77	74.37	74.54
	LogReg	74.21	73.51	74.21	73.09
	RandomForest	73.26	73.72	73.26	73.46
	XGBoost	71.99	73.62	71.99	72.60
	MultinomialNB	68.83	74.00	68.83	63.60
	LightGBM	68.04	70.11	68.04	68.77
	Senti-RU-BERT	**80.54**	**80.29**	**80.54**	**80.39**
Negativity	LinearSVC	72.15	71.21	72.15	71.19
	LogReg	71.68	71.11	71.68	70.17
	RandomForest	74.21	74.43	74.21	72.86
	XGBoost	71.20	69.85	71.20	70.36
	MultinomialNB	71.36	72.20	71.36	69.86
	LightGBM	68.67	67.99	68.67	67.83
	Hate-RU-BERT	**78.01**	**76.75**	**78.01**	**77.34**

Table 7 presents a detailed ablation study reporting F1 scores for different combinations of input features Text only (T), Text + Emoji features (T+E), Text + Slur features (T+S), and all combined (T+E+S) across various models for both sentiment and negativity classification. For sentiment classification, the best results were consistently achieved with the T+E+S combination. Traditional models like LinearSVC (74.54), Logistic Regression (73.09), and Random Forest (73.46) showed marked improvements over their text-only baselines (66.74, 64.94, and 66.79 respectively). Notably, Senti-RU-BERT attained the highest F1 score of 80.77 with the T+E setting, slightly outperforming the T+E+S score (80.39), suggesting that emojis alone contributed more than slurs in this transformer model. Among traditional models, XGBoost (72.60) and LightGBM (68.77) also benefited from feature fusion, while MultinomialNB saw the greatest relative gain jumping from 45.75 (T) to 63.60 (T+E+S). In the case of hate speech type classification, results were more nuanced. The highest F1 score was obtained by Hate-RU-BERT with 77.64 using T+S, showing that slur features were particularly informative for this task. Among classical models, Random Forest (73.88) and Logistic Regression (72.45) also achieved their top scores using T+S, indicating the value of slur-based lexical signals for distinguishing hate subtypes. Interestingly, unlike the sentiment task, adding emojis (T+E) showed limited or no improvement here. The overall best traditional performance came from Random Forest with T+S (73.88), while MultinomialNB, though less effective overall, reached 72.03 with T+S its peak performance across all configurations. These results validate that incorporating task specific linguistic signals such as emojis for sentiment and slurs for hate subtype detection improves classifier performance.

6 Limitations

Despite achieving strong performance on both sentiment and hate speech classification tasks, several limitations persist. First, the class imbalance, especially in the hate speech subtype categories (e.g., threatening, call for action), may bias the models toward more frequent classes like abusive. This limits generalizability and reduces sensitivity to underrepresented categories. Second, many instances of hate speech in our dataset exhibit subtle or context-dependent abuse, including sarcasm,

Table 7: Ablation Study: F1 Score Comparison Across Feature Sets and Models

Features	LinearSVC	LogReg	RandomForest	XGBoost	MultinomialNB	LightGBM	Senti-RU-BERT / Hate-RU-BERT
Sentiment							
T	66.74	64.94	66.79	64.41	45.75	54.09	72.65
T+E	73.25	71.10	71.36	68.76	62.52	63.28	**80.77**
T+S	68.00	66.82	68.62	66.75	46.83	57.85	74.31
T+E+S	74.54	73.09	73.46	72.60	63.60	68.77	**80.39**
Negativity Type							
T	70.70	69.54	70.21	68.58	69.22	60.71	76.14
T+E	70.76	69.44	69.53	68.49	68.36	61.21	76.82
T+S	72.42	72.45	73.88	69.46	72.03	68.06	**77.64**
T+E+S	71.19	70.17	72.86	70.36	69.86	67.83	**77.34**

code switching, or rhetorical phrasing, which remain challenging for both traditional models and BERT. Although auxiliary features like slur flags and emoji categories improve performance, they cannot fully capture nuanced socio-pragmatic cues.

7 Future Work

Future work can focus on curating more balanced and culturally grounded datasets for Roman Urdu and code-switched text, incorporating linguistic annotation informed by sociolinguistic cues to better detect subtle or indirect expressions of hate speech. Further to deal with class imbalance the dataset should be curated in such a way that there is balance between both overrepresented and underrepresented classes such as such as threatening or call for action. This can also be achieved through targeted data augmentation techniques like paraphrasing, back translation, or synthetic oversampling. Future work could also explore applying more recent large language models (e.g., GPT or open-source alternatives).

8 Conclusion

This study presents a novel computational approach to analyzing online discourse concerning transgender communities in Pakistan, with a particular emphasis on Roman Urdu—a low-resource, code-mixed language prevalent across social media platforms. We introduce a comprehensive Instagram-based dataset annotated for both sentiment and hate speech, and further enrich this resource through the development of a Roman Urdu transphobic slur lexicon and an emoji classification grounded in contextual semantics. Experimental evaluations reveal that transformer-based architectures, notably BERT, consistently outperform traditional machine learning models on both classification tasks, achieving F1 scores of 80.39% for sentiment and 77.34% for hate speech detection. Ablation analyses demonstrate that the integration of lexicon-based and emoji-derived features yields significant performance improvements, especially in identifying implicit or nuanced forms of hate speech. These findings highlight the critical role of culturally and linguistically informed resources in advancing hate speech detection in low-resource settings. By integrating domain-specific linguistic insights with state-of-the-art natural language processing techniques, this work establishes foundational tools, benchmarks, and methodologies for future research in socially-aware, multilingual, and inclusive NLP.

References

Samia Aziz, Muhammad Shahzad Sarfraz, Muhammad Usman, Muhammad Umar Aftab, and Hafiz Tayyab Rauf. 2023. Geo-spatial mapping of hate speech prediction in roman urdu. *Mathematics*, 11(4):969.

Bilal Chandio, Asadullah Shaikh, Maheen Bakhtyar, Mesfer Alrizq, Junaid Baber, Adel Sulaiman, Adel Rajab, and Waheed Noor. 2022. Sentiment analysis of roman urdu on e-commerce reviews using machine learning. *CMES-Comput. Model. Eng. Sci*, 131(3):1263–1287.

Enrico De Santis, Alessio Martino, Francesca Ronci, and Antonello Rizzi. 2024. From bag-of-words to transformers: A comparative study for text classification in healthcare discussions in social media. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Marwa Khan, Asma Naseer, Aamir Wali, and Maria Tamoor. 2024. A roman urdu corpus for sentiment analysis. *The Computer Journal*, 67(9):2864–2876.

Mohit Kumar et al. 2025. Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications*, pages 1–37.

Muhammad Arslan Manzoor, Saqib Mamoon, Song Kei Tao, Zakir Ali, Muhammad Adil, and Jianfeng Lu. 2020. Lexical variation and sentiment analysis of roman urdu sentences with deep neural networks. *Int. J. Adv. Comput. Sci. Appl*, 11(2).

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, 36(4):102048.

Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

Faiza Mehmood, Muhammad Usman Ghani, Muhammad Ali Ibrahim, Rehab Shahzadi, Waqar Mahmood, and Muhammad Nabeel Asim. 2020a. A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access*, 8:192740–192759.

Khawar Mehmood, Daryl Essam, Kamran Shafi, and Muhammad Kamran Malik. 2019. Sentiment analysis for a resource poor language—roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–15.

Khawar Mehmood, Daryl Essam, Kamran Shafi, and Muhammad Kamran Malik. 2020b. An unsupervised lexical normalization for roman hindi and urdu sentiment analysis. *Information Processing & Management*, 57(6):102368.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES-Computer Modeling in Engineering & Sciences*, 140(3).

Daniel Schönle, Christoph Reich, and Djaffar Ould Abdeslam. 2024. Linguistic-aware wordpiece tokenization: Semantic enrichment and oov mitigation. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 134–142. IEEE.

Gautam Kishore Shahi and Tim A Majchrzak. 2024. Hate speech detection using cross-platform social media data in english and german language. *arXiv preprint arXiv:2410.05287*.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: a survey of tasks, datasets and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–44.

Wali Ullah, Robina Saeed, and Nadir Saeed Sarhadi. 2024. Portrayal of violence against transgender in pakistani urdu dramas: A critical analysis. *Annals of Human and Social Sciences*, 5(4):250–263.