UPR at SemEval-2026 Task 9: Multi-Label Polarization and Manifestation Detection in Urdu

Mtayyaba Shahzad[1], **Inzmam Khadam**[1], **Zaufishan Mahmood**[1], **Junaid Rashid**[2], **Shamaila Hayat**[1], **Fakhar Ayub**[1]
[1]University of Poonch Rawalakot, Rawalakot, Pakistan
[2]Sejong University, Seoul, Republic of Korea
mtayyabakhan2002@gmail.com,malikinzmam92@gmail.com
zaufishanmahmoodkhan@gmail.com,junaid.rashid@sejong.ac.kr
shamailahayat@upr.edu.pk,fakharayubstdcsit@upr.edu.pk

Abstract

Polarization detection in low-resource languages such as Urdu remains underexplored despite its importance for public discourse analysis. We address two complementary subtasks of polarization analysis in Urdu social media text. For social-dimension classification, we formulate the task as a multi-label problem across five dimensions: political, religious, racial/ethnic, gender/sexual, and other. For manifestation identification, we target six polarization expressions: stereotype, vilification, dehumanization, extreme language, lack of empathy, and invalidation. We fine-tune XLM-RoBERTa with language-specific preprocessing, duplicate filtering, and imbalance-aware augmentation and resampling for both subtasks. The proposed framework achieves a Macro-F1 of 0.75 on Subtask-2 and 0.72 on Subtask-3 on the validation set, which demonstrates the effectiveness of multilingual transformer models for multi-dimensional polarization analysis in low-resource Urdu text.

1 Introduction

Social media growth has facilitated the circulation of toxic and polarized content, posing significant challenges to social cohesion and democratic discourse. Consequently, automatic detection of polarized or harmful content has emerged as a critical research problem in natural language processing (NLP), with direct applications in content moderation, political analysis, and online safety (Poletto et al., 2021; Fortuna and Nunes, 2018). Unlike conventional sentiment analysis, polarization detection requires modeling multiple overlapping dimensions of bias or hostility that may co-occur within a single text.

In this task, polarization analysis is formulated as a multi-label problem, as a single message may simultaneously target multiple social groups, including political, religious, racial/ethnic, or gender-based communities. Multi-label classification therefore provides a more appropriate framework for capturing overlapping forms of polarization (Tsoumakas and Katakis, 2007).

This problem is further complicated in low-resource languages such as Urdu, where NLP resources remain limited due to scarce annotated data, rich morphology, flexible syntax, and frequent code-mixing. Informal writing styles and script variations in social media text add further complexity to automatic analysis. Prior work has shown that detecting harmful or abusive language in Urdu is particularly difficult due to implicit expressions and contextual dependencies (Bilal et al., 2023; Raza et al., 2021).

Recent advances in multilingual transformer models, such as mBERT and XLM-RoBERTa, have substantially improved performance in low-resource languages through cross-lingual transfer learning (Devlin et al., 2019; Conneau et al., 2020; Ruder et al., 2019). These models learn contextualized representations from large multilingual corpora, enabling stronger generalization for languages such as Urdu.

SemEval-2026 Task 9 provides a benchmark for polarization analysis in Urdu across two complementary subtasks: social-dimension classification and manifestation identification (Naseem et al., 2026a).

To address these challenges, we propose an XLM-RoBERTa-based framework for multi-label polarization analysis in Urdu social media text. The framework incorporates language-aware preprocessing, including Unicode normalization, removal of non-Urdu characters, TF-IDF-based duplicate filtering, and imbalance-aware data augmentation. The key contributions of this study are as follows:

- We present an XLM-RoBERTa-based framework for multi-label polarization analysis in

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2642–2647

Urdu, covering two complementary subtasks: social-dimension classification and manifestation identification.

- We develop a script-aware preprocessing pipeline for noisy Urdu social media text, including Unicode normalization, noise filtering, and duplicate removal.
- We evaluate the proposed framework against TF-IDF-based machine learning baselines and transformer-based baselines across both subtasks.

2 Related Work

The detection of harmful and polarized language has attracted significant attention in natural language processing, particularly in English and low-resource languages. Early approaches relied on lexical and syntactic features such as n-grams and sentiment lexicons combined with traditional machine learning classifiers (Davidson et al., 2017). While these methods were effective for explicit cases, they were less reliable for capturing implicit bias, sarcasm, and context-dependent hostility.

Subsequent work explored neural architectures such as CNNs and RNNs to learn distributed text representations. Hybrid architectures have also been explored for hate speech detection in low-resource languages, including Urdu and Roman Urdu (Ashiq et al., 2024). However, such models are limited in their ability to capture long-range dependencies and contextual semantics across entire sequences.

Transformer architectures substantially improved text classification by enabling contextualized language representations. Models such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) have demonstrated strong results across multilingual NLP tasks, while transfer learning has become central to low-resource language modeling (Ruder et al., 2019). These advances are particularly relevant for Urdu, where annotated resources remain limited.

Multi-label classification, where each instance may be assigned multiple simultaneous labels, is well-suited for polarization analysis, where a single post may express overlapping forms of hostility or bias (Tsoumakas and Katakis, 2007; Read et al., 2009; Vidgen et al., 2021). This formulation is commonly modeled using Binary Cross-Entropy loss with sigmoid activation. In low-resource settings, preprocessing quality further influences performance, and prior work on Urdu demonstrates the importance of normalization and script-aware cleaning (Bilal et al., 2023), though strong contextual encoders remain essential for robust multilingual classification.

3 Methodology

Figure 1 presents an overview of the proposed workflow.

3.1 Datasets

We use the official Urdu datasets released for SemEval-2026 Task 9 (Naseem et al., 2026a,b). The two subtasks are used separately within a shared multi-label learning framework.

For Subtask-2, the dataset consists of an annotated Urdu polarization corpus in which each text is associated with five binary labels: political, religious, racial/ethnic, gender/sexual, and other. After removing samples with missing text or incomplete annotations, the training split contains 3,561 samples.

For Subtask-3, the dataset is designed for polarization manifestation identification and contains 3,563 training samples annotated with six binary labels: stereotype, vilification, dehumanization, extreme language, lack of empathy, and invalidation. The dataset is written in Urdu using the Nastaliq script.

3.2 Label Characteristics

Both subtasks exhibit strong multi-label behavior and class imbalance. In Subtask-2, the five social-dimension labels are not mutually exclusive, and a single post may target multiple dimensions simultaneously. In Subtask-3, label overlap is also common. Among the 3,563 samples, 1,087 samples contain no manifestation label, 67 instances contain exactly one manifestation, and 2,409 samples contain multiple manifestation labels.

3.3 Text Preprocessing

Urdu social media text was normalized through Unicode standardization to ensure consistent character encoding. URLs, HTML tags, emojis, non-Urdu characters, and excessive whitespace were removed. Texts containing fewer than 3 tokens were filtered out during preprocessing. These preprocessing steps were applied consistently across both subtasks to reduce noise while preserving semantic content.

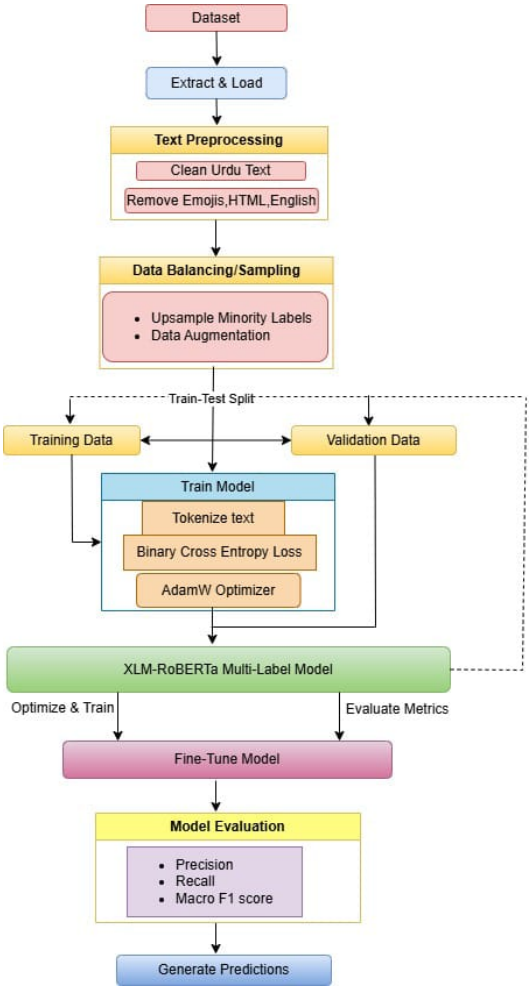

Figure 1: Overview of the proposed XLM-RoBERTa-based framework for multi-label polarization analysis in Urdu text.

3.4 Duplicate Removal

Duplicate and near-duplicate samples were removed prior to dataset splitting for both subtasks to prevent data leakage and ensure fair evaluation. TF-IDF representations were constructed using unigram and bigram features for each text instance. Pairwise cosine similarity was then computed, and any pair of samples with similarity greater than 0.95 was considered a near-duplicate. This threshold-based filtering, combined with exact duplicate removal, ensured that highly similar instances were not present across the training data.

3.5 Class Imbalance Handling

Both subtasks exhibit class imbalance, but the imbalance differs in structure. In Subtask-2, minority labels are underrepresented at the individual label level. Targeted augmentation was therefore applied to training instances containing minority labels. In Subtask-3, imbalance mainly appears in rare label combinations. Controlled resampling of infrequent label sets was applied together with targeted augmentation. All balancing operations were applied exclusively to the training split to prevent data leakage.

3.6 Data Augmentation

Data augmentation was performed using a synonym replacement strategy applied to selected training instances. A manually constructed Urdu synonym dictionary was used, where each word was mapped to a small set of contextually similar lexical alternatives. During augmentation, each non-stopword had a replacement probability of $p = 0.1$, and a maximum of two words per sentence were allowed to be replaced, which ensured controlled perturbation while preserving sentence semantics. Class imbalance was addressed using an upsampling strategy. Instances with no active labels were treated as the majority class, while instances with at least one positive label were

treated as the minority class. Minority-class instances were randomly oversampled with replacement to achieve a balanced 1:1 distribution between majority and minority classes. The augmentation process was applied independently for each subtask, resulting in a final training size of 2,183 samples for Subtask-2 and 2,176 samples for Subtask-3. All augmentation was strictly restricted to the training split to avoid data leakage. Prior work has shown that synonym-based data augmentation improves robustness and generalization in low-resource text classification tasks (Wei and Zou, 2019; Feng et al., 2021), and such techniques are particularly effective for morphologically rich languages such as Urdu.

3.7 Tokenization and Input Representation

Texts were encoded using the XLM-RoBERTa tokenizer, and input lengths were standardized to 128 tokens through truncation or padding. Each instance was represented by input token IDs and corresponding attention masks. Each sample in Subtask-2 is associated with a five-dimensional binary label vector, and each sample in Subtask-3 with a six-dimensional binary label vector.

3.8 Model Architecture

We use XLM-RoBERTa as the backbone encoder for both subtasks. A task-specific multi-label classification head is placed on top of the pretrained encoder. For Subtask-2, the classification head outputs five logits corresponding to the polarization dimensions. For Subtask-3, the classification head outputs six logits corresponding to the manifestation labels. Sigmoid activation is applied to convert logits into label-wise probabilities.

3.9 Training Setup

The models were fine-tuned using the Hugging Face Transformers framework with a PyTorch backend. For both subtasks, the model was optimized with AdamW for a maximum of 10 epochs, using a learning rate of 2×10^{-5}, batch size of 16, and weight decay of 0.01. The model was trained using Binary Cross-Entropy loss, and the best checkpoint was selected based on validation Macro-F1. A fixed random seed of 42 was used in both subtasks to ensure reproducibility.

3.10 Evaluation Metrics

Model performance was evaluated using precision, recall, and Macro-F1 for both subtasks. Given the multi-label nature of the tasks and label-combination imbalance, Macro-F1 was selected as the primary evaluation metric, as it provides a balanced measure across all labels.

3.11 Evaluation Setup and Baselines

After fine-tuning, the model generates label-wise predictions by applying a sigmoid function to the output logits, followed by a fixed decision threshold of 0.5 to obtain binary predictions for each label. We compare XLM-RoBERTa (Conneau et al., 2020) against TF-IDF-based machine learning baselines, including Linear SVM (Cortes and Vapnik, 1995) and XGBoost (Chen and Guestrin, 2016), and transformer-based baselines, including DistilBERT (Sanh et al., 2019) and mBERT (Devlin et al., 2019), across both subtasks.

4 Results

This section presents the experimental results for both subtasks. We first discuss Subtask-2, social-dimension classification, followed by Subtask-3, polarization manifestation identification.

4.1 Performance Comparison

Tables 1 and 2 show the detailed performance comparison of XLM-RoBERTa and the baselines. Among the evaluated models, XLM-RoBERTa achieves the best overall performance, with a Macro-F1 of 0.75 on Subtask-2 and 0.72 on Subtask-3. For Subtask-2, XLM-RoBERTa outperforms both TF-IDF-based machine learning models and transformer-based baselines. For Subtask-3, XLM-RoBERTa achieves the highest Macro-F1, precision, and recall, indicating stronger label-wise prediction performance. The comparatively lower Macro-F1 of DistilBERT relative to the Linear SVM baseline across both subtasks may be related to its reduced model capacity and limited multilingual coverage, which results in lower precision despite competitive recall.

All results in Tables 1 and 2 correspond to validation performance on the development sets, which were used for model selection and hyperparameter tuning. The final trained model was evaluated on the unseen test set for submission. These results collectively indicate that large-scale multilingual pretraining provides the strongest contextual representations for low-resource multi-label polarization detection in Urdu.

Table 1: Performance comparison of TF-IDF-based machine learning baselines and transformer-based models for Subtask-2.

Model	Macro-F1	Precision	Recall
Linear SVM	0.68	0.67	0.69
XGBoost	0.67	0.67	0.68
DistilBERT	0.66	0.64	0.73
mBERT	0.68	0.66	0.75
XLM-RoBERTa	**0.75**	**0.77**	**0.85**

Table 2: Performance comparison of TF-IDF-based machine learning baselines and transformer-based models for Subtask-3.

Model	Macro-F1	Precision	Recall
Linear SVM	0.70	0.70	0.74
XGBoost	0.71	0.70	0.77
DistilBERT	0.68	0.65	0.73
mBERT	0.70	0.68	0.74
XLM-RoBERTa	**0.72**	**0.74**	**0.82**

4.2 Error Analysis

In Subtask-2, misclassifications frequently occur in posts where multiple social dimensions overlap, particularly political and religious categories. Confusion is also observed between racial/ethnic and gender/sexual labels, especially in informal or sarcastic content with limited explicit lexical cues. These challenges reflect the limitations of both TF-IDF-based machine learning and transformer-based models in capturing implicit polarization cues.

In Subtask-3, errors mainly arise where multiple manifestation labels co-occur within a single text. Confusion is observed between closely related categories such as stereotype and vilification, and between dehumanization and extreme language, which indicates that distinguishing subtle differences between overlapping harmful expressions remains difficult even for advanced transformer models.

Across both subtasks, label distribution analysis reveals strong interdependencies. In Subtask-2, frequent co-occurrence is observed between political and religious dimensions, and between racial/ethnic and other categories. Similarly, Subtask-3 shows strong dependencies among stereotype, vilification, and dehumanization labels. These patterns suggest that polarization in Urdu social media is inherently multi-dimensional and context-dependent, which makes it challenging to model using isolated label predictions.

5 Limitations

Although the proposed system achieves strong performance, several limitations remain. First, the dataset size is relatively small, and annotation noise is possible in subjective tasks such as polarization manifestation identification and social-dimension classification.

Second, the model uses a fixed decision threshold of 0.5 for all labels, which is standard in multi-label classification with sigmoid outputs. While label-wise threshold tuning could improve class-specific performance, it introduces additional hyperparameter complexity. A unified threshold is therefore adopted to maintain simplicity and reproducibility.

Third, the study primarily focuses on XLM-RoBERTa, with comparisons limited to TF-IDF-based machine learning models such as Linear SVM and XGBoost and transformer-based baselines, DistilBERT and mBERT. Although XLM-RoBERTa achieves the best performance, lighter or language-specific models remain a promising direction for improved computational efficiency and domain adaptation.

Finally, the models do not explicitly incorporate discourse structure or external knowledge sources, which are often important for understanding implicit, sarcastic, and context-dependent polarization in Urdu social media text.

6 Conclusion

This paper presented an XLM-RoBERTa-based framework for multi-label polarization analysis in Urdu across social-dimension classification and manifestation identification. XLM-RoBERTa, fine-tuned with script-aware preprocessing and imbalance-aware augmentation, achieves Macro-F1 scores of 0.75 and 0.72 on Subtask-2 and Subtask-3, respectively, and consistently outperforms all TF-IDF-based machine learning baselines and transformer-based baselines. These results demonstrate the effectiveness of multilingual pretraining for polarization detection in a low-resource language and motivate future work on discourse structure and language-specific models for Urdu.

References

Waqar Ashiq, Samra Kanwal, Adnan Rafique, Muhammad Waqas, Tahir Khurshaid, Elizabeth Caro Mon-

tero, Alicia Bustamante Alonso, and Imran Ashraf. 2024. Roman urdu hate speech detection using hybrid machine learning models and hyperparameter optimization. *Scientific Reports*, 14(1):28590.

Muhammad Bilal, Atif Khan, Salman Jan, Shahrulniza Musa, and Shaukat Ali. 2023. Roman urdu hate speech detection using transformer-based model for cyber security applications. *Sensors*, 23(8):3909.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 968–988.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Muhammad Owais Raza, Qaisar Khan, and Ghulam Muhammad Soomro. 2021. Urdu abusive language detection using machine learning. In *FIRE (Working Notes)*, pages 774–783.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 254–269. Springer.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6382–6388.