UPR at SemEval-2026 Task 9: Polarization Detection in Urdu with Language-Specific Transformer and Data Augmentation

Alishba Wazir[1], Muhammad Asad Khan[1], Junaid Rashid[2]*
Shamaila Hayat[1], Samira Kanwal[1]

[1]University of Poonch Rawalakot, Rawalakot, Pakistan
[2]Sejong University, Seoul, Republic of Korea

alishbawazir18@gmail.com, masadkhanek@gmail.com
junaid.rashid@sejong.ac.kr, shamailahayat@upr.edu.pk
samirakanwal09@upr.edu.pk

Abstract

This paper addresses polarization detection in Urdu, a low-resource language characterized by complex morphology and insufficient annotated data. We formulate the task as a binary classification problem of social media posts into polarized and non-polarized categories. Our approach is based on Urdu-BERT, a language-specific transformer model combined with language-specific preprocessing, duplicate removal, and data augmentation to mitigate class imbalance and improve generalization. Experimental results show that the fine-tuned Urdu-BERT outperforms TF-IDF-based lexical machine learning baselines and achieves strong performance relative to multilingual transformer baselines. The findings indicate that language-specific pretrained transformers, when combined with appropriate preprocessing and augmentation strategies, provide an effective and generalizable framework for low-resource Urdu polarization detection.

1 Introduction

Social media has emerged as a prominent platform for the expression of opinions and ideological positions, making the detection of polarized content an increasingly important research problem. Numerous studies have employed deep learning techniques for text classification in Urdu, particularly in sentiment and hate speech analysis. Ahmed et al. (2024) analyzed reviews and reported that Long Short-Term Memory (LSTM) networks demonstrated superior performance compared to Convolutional Neural Networks (CNNs). These findings highlight the effectiveness of advanced neural architectures and motivate further exploration for polarization detection in low-resource languages.

Despite these advances, polarization detection in Urdu presents challenges beyond conventional sentiment analysis, including class imbalance, limited annotated data, and the informal nature of social media text (Khattak et al., 2021). Recent research has increasingly emphasized the potential of multilingual and transformer-based models to address such challenges. Multilingual BERT (mBERT), which has shown effectiveness in cross-lingual settings (Devlin et al., 2019; Wu and Dredze, 2019), and hybrid deep learning approaches have achieved strong results in Urdu text classification (Singh and Jaiswal, 2023). However, the effectiveness of language-specific pretrained models for binary polarization detection remains limited.

Therefore, this work focuses on binary polarization detection in Urdu, where each post is classified as polarized or non-polarized based on its overall meaning and contextual cues. This task introduces challenges such as inherent class imbalance and noisy samples, which can negatively impact model generalization. To overcome these challenges, we use a transformer architecture built upon the pretrained Urdu-BERT model, *eshaaftab900/urdu-bert-base-2* (eshaaftab900, 2024).

While prior work mainly employs multilingual transformers, this study investigates the effectiveness of an Urdu-specific pretrained model combined with imbalance-aware training strategies. The framework integrates language-specific preprocessing, duplicate removal, and hybrid data augmentation techniques, including synonym replacement, to enhance data diversity and mitigate class imbalance. The main contributions of this work are summarized as follows:

- We formulate polarization detection in Urdu as a binary classification task to distinguish between polarized and non-polarized content.
- We use the pretrained Urdu-BERT model combined with language-specific preprocessing and duplicate removal for low-resource Urdu polarization detection.

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2635–2641

- We apply hybrid data augmentation techniques, including synonym replacement and structural variation, to improve generalization under class imbalance.
- We demonstrate that combining a language-specific transformer with appropriate preprocessing and augmentation strategies improves performance for low-resource Urdu polarization detection.

2 Literature Review

Sentiment analysis is a text-mining approach used to identify opinions and attitudes in textual data. However, polarization detection extends beyond conventional sentiment classification by focusing on ideologically or socially polarized content (Mohammad et al., 2016). Unlike high-resource languages such as English, low-resource languages such as Urdu face significant challenges due to the scarcity of labeled corpora. Recent studies on stance and polarization-related tasks indicate that transformer-based models are effective at encoding contextual relationships and implicit opinions (Pangtey et al., 2025; Gera and Neal, 2025).

Ahmed et al. (2024) demonstrated that an LSTM model outperformed a CNN model and achieved 96% accuracy and 91% F1-score on a dataset of 25,000 Urdu reviews, illustrating the effectiveness of deep neural architectures for Urdu language processing. Similarly, Naqvi et al. (2021) employed FastText and Word2Vec embeddings within CNN-LSTM and BiLSTM architectures and showed improved performance over conventional machine learning models in low-resource Urdu sentiment analysis.

Stance detection has also been extended to cross-lingual settings to address low-resource limitations, where lack of labeled data and poor generalization across unseen targets remain significant challenges (Zhang et al., 2023). Beyond standalone deep learning models, hybrid architectures have also been investigated. Muhammad and Burney (2023) reported an accuracy of 85.8% using hybrid architectures (RNN+CNN and LSTM+GRU), which demonstrated the potential of combining recurrent and convolutional structures. In addition, dataset expansion has contributed to performance improvements. The large-scale Urdu review dataset introduced by Ashraf et al. (2024) contains over 65,000 samples and evaluated transformer-based models such as XLM-R and GPT-2, which achieved up to 95% accuracy and outperformed earlier low-resource models.

Recent work has further incorporated attention-based and transformer-driven approaches for Urdu text classification. Aziz et al. (2024) introduced a sentiment classification framework for Urdu news headlines, integrating Graph Attention Networks (GAT) with multilingual BERT. Their manually labeled dataset and comprehensive preprocessing contributed to improved performance and expanded NLP resources for low-resource languages. Early work on social media text classification also explored alternative lexical and semantic approaches. Chong et al. (2014) analyzed tweet sentiment using subjectivity detection, semantic association, and polarity classification and outperformed traditional text-based methods, demonstrating the importance of semantic features in short informal text. Similarly, Shabbir and Majid (2024) developed deep learning models for Urdu text classification using 1D-CNN, LSTM, and Multilingual-MiniLM on an IMDB dataset translated into Urdu. Among these approaches, the transformer-based model achieved the highest accuracy of 89.36%, which reinforces the effectiveness of transformer architectures for Urdu language tasks.

While prior studies demonstrate the effectiveness of deep and transformer-based architectures for Urdu sentiment analysis, limited attention has been given to binary polarization detection using language-specific pretrained models under class imbalance conditions. This study therefore uses a fine-tuned Urdu-BERT model combined with language-specific preprocessing and hybrid data augmentation for low-resource Urdu polarization detection.

3 Methodology

Figure 1 presents the overall architecture of the proposed framework.

3.1 Task Definition

The objective of the task is binary classification of Urdu social media posts into polarized (1) and non-polarized (0) categories based on their overall meaning and contextual cues.

3.2 Dataset Description

The experiments were conducted on the Urdu polarization dataset released for SemEval-2026 Task 9, Subtask 1 (Naseem et al., 2026b,a). After pre-

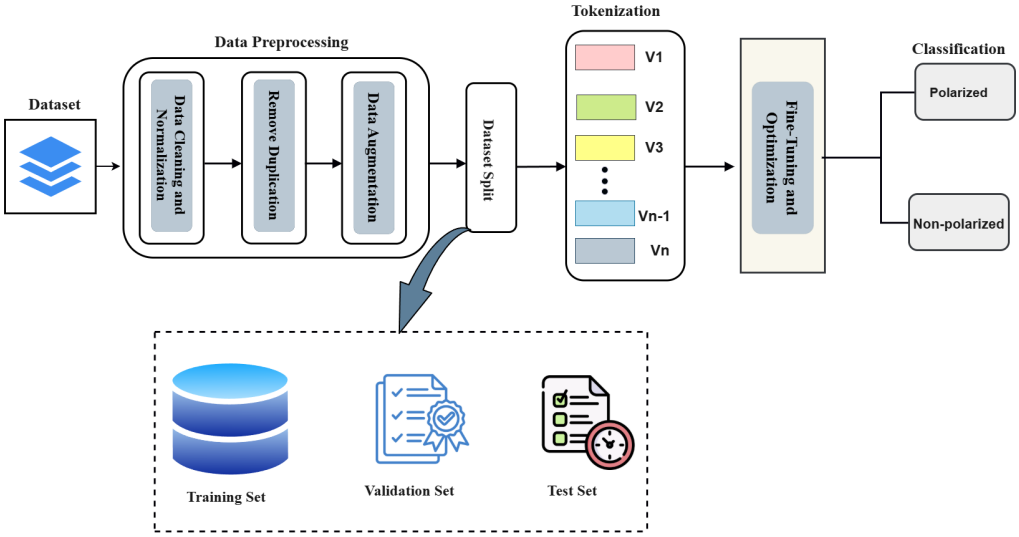

Figure 1: Overall architecture of the fine-tuned Urdu-BERT-based polarization detection framework.

processing and duplicate removal, the dataset contained 3,561 posts, including 2,475 polarized posts and 1,086 non-polarized posts, with an imbalance ratio of approximately 2.28:1. The official training and development splits were used for model training and hyperparameter tuning. The evaluation set provided by the organizers was used for final performance assessment. The dataset exhibits moderate class imbalance, which was addressed through data augmentation and imbalance-aware training strategies.

3.3 Data Cleaning and Normalization

Due to the noisy and informal nature of social media Urdu text, systematic preprocessing was applied prior to model training. Text normalization was performed using the UrduHack library to standardize orthographic variations and ensure consistent Unicode encoding. The cleaning process included the removal of URLs and hyperlinks, elimination of non-Urdu characters outside the standard Urdu Unicode range, reduction of excessive character repetition, normalization of visually similar characters, and whitespace standardization. Posts shorter than five characters and entries with missing labels were excluded. These preprocessing steps improved semantic consistency and reduced noise.

3.4 Duplicate Removal

To reduce redundancy and potential information leakage, duplicate and near-duplicate posts were identified using TF-IDF vectorization with unigram and bigram features, followed by cosine similarity computation. Posts with similarity greater than or equal to 0.95 were removed. This process reduced the dataset size from 3,563 to 3,561 posts and minimized the risk of model memorization of highly similar training examples.

3.5 Data Augmentation for Class Imbalance

To address class imbalance and improve generalization, we employed a hybrid data augmentation strategy for Urdu text, following prior work showing that such strategies improve text classification performance (Rahman et al., 2023). Synonym replacement (Wei and Zou, 2019) was used to introduce semantic variation, while controlled word reordering provided syntactic diversity. Random word deletion was applied as a fallback strategy when synonym replacement was not feasible for a given token. Augmentation was performed more extensively on minority-class samples to improve representation learning. Importantly, all data augmentation was applied exclusively to the training set after dataset splitting to ensure strict separation between training and evaluation data.

3.5.1 Synonym-Based Augmentation

A manually curated Urdu synonym dictionary was used to replace one to three randomly selected words in minority-class posts. The augmentation ratio was aligned with the observed class imbalance to achieve a more balanced class distribution during training.

3.5.2 Structural Variation Augmentation

In addition to synonym-based augmentation, a lightweight structural augmentation strategy was

Table 1: Class distribution and data split after data augmentation.

Dataset Class / Split	Number of Samples
Class 1	5,611
Class 0	4,768
Total	10,379
Training set	8,822
Validation set	1,557
Evaluation set	177

applied. Internal word reordering was performed within semantically coherent segments to introduce syntactic diversity while preserving label semantics. After augmentation, the resulting dataset contained 10,379 posts, including 5,611 polarized and 4,768 non-polarized samples, as shown in Table 1. The final training set contained 8,822 posts, the validation set contained 1,557 posts, and the evaluation set of 177 posts was reserved exclusively for final performance assessment.

3.6 Tokenization and Input Representation

Before input to the Urdu-BERT model, each input sentence was tokenized using the WordPiece tokenizer. The cleaned and preprocessed Urdu text was converted into a sequence of subword tokens represented as:

$$X = [[\text{CLS}], v_1, v_2, \ldots, v_n, [\text{SEP}]] \quad (1)$$

where $v_1, v_2, \ldots, v_n$ denote WordPiece tokens, [CLS] is the classification token, and [SEP] marks the end of the sequence. Sequences exceeding 256 tokens were truncated, while shorter ones were padded to ensure consistent input dimensions. This subword-level representation enables Urdu-BERT to capture contextual semantic and syntactic information effectively.

3.7 Model Architecture

We fine-tuned the pretrained Urdu-BERT model, *eshaaftab900/urdu-bert-base-2*, available on Hugging Face (eshaaftab900, 2024), which follows the BERT-base architecture (Devlin et al., 2019). The model consists of 12 transformer layers, 12 attention heads, and a hidden size of 768. The contextual representation of the [CLS] token was passed to a fully connected classification layer that produces two output logits for binary classification. The model was trained using cross-entropy loss as the classification objective.

3.8 Training Details

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16. Training was conducted for a maximum of 20 epochs, and early stopping with a patience of 3 epochs was applied based on the validation Macro-F1 to mitigate overfitting.

4 Results and Evaluation

This section presents the quantitative evaluation of the proposed framework using performance metrics. Table 2 presents the performance comparison between TF-IDF-based machine learning baselines and the fine-tuned Urdu-BERT model. The classical baselines include Naive Bayes (Rish et al., 2001), Linear SVM (Cortes and Vapnik, 1995), SVM with RBF kernel (Cortes and Vapnik, 1995), and Logistic Regression (Hosmer Jr et al., 2013), while the transformer-based baselines include XLM-RoBERTa (Conneau et al., 2020) and mBERT (Devlin et al., 2019), included for comparison as multilingual transformer models. Model performance was evaluated using precision, recall, and Macro-F1. Given the moderate class imbalance in the dataset, Macro-F1 was selected as the primary evaluation metric, as it provides a balanced assessment of precision and recall across classes. Validation loss was additionally monitored to analyze optimization stability and convergence behavior.

Table 2: Performance comparison of TF-IDF-based baselines and transformer-based models.

Model	Accuracy	Precision	Recall	Macro-F1
Naive Bayes	0.727	0.700	0.580	0.573
Linear SVM	0.736	0.690	0.670	0.674
SVM (RBF)	0.751	0.760	0.610	0.618
Logistic Regression	0.736	0.720	0.600	0.598
XLM-RoBERTa	0.782	0.794	0.777	0.780
mBERT	0.979	0.979	0.979	0.979
Urdu-BERT	0.983	0.987	0.988	0.989

Figure 2 illustrates the learning behavior of the model across epochs. The training loss steadily decreased and stabilized, which indicates effective optimization and convergence. The validation loss decreased during the early epochs before stabilizing, while the training loss continued to decrease, which suggests mild overfitting in later epochs that was effectively controlled by early stopping. Analysis of precision and recall trends shows consistent improvement across epochs, which indicates balanced performance between false positives and false negatives. The fine-tuned Urdu-BERT model exhibited stable convergence during training. The highest validation Macro-F1 of 0.989 was achieved at Epoch 13. Early stopping, with a patience of 3

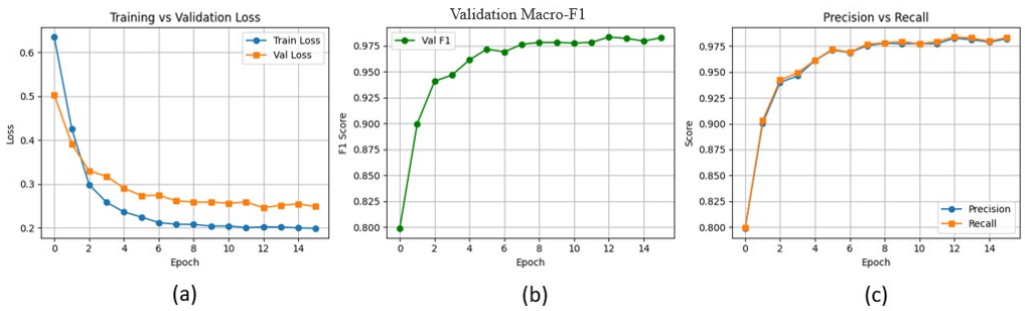

Figure 2: Training progress of the fine-tuned Urdu-BERT model across epochs, including training and validation loss curves (a), validation Macro-F1 progression (b), and precision-recall trends (c).

epochs, was applied to mitigate overfitting. Precision and recall remained well balanced across epochs, which reflects stable and discriminative classification behavior.

Urdu-BERT achieved the highest validation Macro-F1 of 0.989, compared with 0.979 for mBERT and 0.780 for XLM-RoBERTa, which indicates that language-specific pretraining provides improved in-domain performance.

4.1 Lexical Overlap Analysis

To analyze potential data leakage and dataset distribution consistency, we computed vocabulary overlap and Jaccard similarity between the training and test sets. The results show a training vocabulary size of 8,224 and a test vocabulary size of 2,964, with 2,107 shared tokens. The Jaccard similarity was 0.232, indicating limited but non-trivial lexical overlap between training and test sets. Additionally, 71% of the test vocabulary appeared in the training set, indicating that the model was evaluated in a realistic setting with adequate vocabulary coverage and sufficient lexical diversity.

5 Limitations

Despite strong validation performance, several limitations remain. First, the dataset is relatively small and may not fully capture dialectal variation and contextual diversity in Urdu polarization discourse. Second, the noisy and informal nature of social media text further limits generalization to unseen data. Third, although transformer models capture contextual dependencies, they may struggle with implicit or pragmatically nuanced polarization expressions. Fourth, data augmentation strategies, such as synonym replacement and structural variation, may introduce distributional artifacts that affect generalization to unseen evaluation data. Finally, only a single transformer architecture was explored, and future work may benefit from larger models, ensemble approaches, or domain-adaptive pretraining to produce further improvements.

6 Conclusion

This study presented a transformer-based approach for binary polarization detection in Urdu social media text. We fine-tuned the pretrained Urdu-BERT model and incorporated language-specific preprocessing, duplicate removal, and hybrid data augmentation to address data scarcity and class imbalance. Experimental results demonstrated strong validation performance and showed that language-specific pretrained models can effectively capture contextual cues in low-resource Urdu text. Furthermore, Urdu-BERT outperformed XLM-RoBERTa and achieved slightly stronger performance than mBERT, which suggests that language-specific pretrained models provide an advantage for low-resource Urdu polarization detection. Overall, these findings show the potential of Urdu-specific transformers combined with appropriate preprocessing and augmentation strategies for polarization detection.

References

Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Muhammad Saeed, Muhammad Bilal, and Houbing Song. 2024. A novel approach for sentiment analysis of a low resource language using deep learning models. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Muhammad Rehan Ashraf, Muzammal Hussain, M Arfan Jaffar, Waheed Yousuf Ramay, and Muhammad Faheem. 2024. Revolutionizing urdu sentiment analysis: Harnessing the power of xlm-r and gpt-2. *IEEE Access*, 12:99779–99793.

Kamran Aziz, Donghong Ji, Bobo Li, Fei Li, and Jun Zhou. 2024. Advancing urdu nlp: Aspect-based sentiment analysis with graph attention networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Wei Yen Chong, Bhawani Selvaretnam, and Lay-Ki Soon. 2014. Natural language processing for sentiment analysis: an exploratory analysis on tweets. In *2014 4th international conference on artificial intelligence with applications in engineering and technology*, pages 212–217. IEEE.

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 *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

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

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.

eshaaftab900. 2024. urdu-bert-base-2. https://huggingface.co/eshaaftab900/urdu-bert-base-2. Hugging Face model repository. Accessed: 2026-01-25.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Computing Surveys*, 58(1):1–37.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Asad Khattak, Muhammad Zubair Asghar, Anam Saeed, Ibrahim A. Hameed, Syed Asif Hassan, and Shakeel Ahmad. 2021. A survey on sentiment analysis in urdu: A resource-poor language. *Egyptian Informatics Journal*, 22(1):53–74.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Khalid Bin Muhammad and SM Aqil Burney. 2023. Innovations in urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets. *Symmetry*, 15(5):1027.

Uzma Naqvi, Abdul Majid, and Syed Ali Abbas. 2021. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094.

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.

Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. 2025. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *arXiv preprint arXiv:2505.08464*.

AM Muntasir Rahman, Wenpeng Yin, and Guiling Wang. 2023. Data augmentation for text classification with ease. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 324–332.

Irina Rish and 1 others. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Seattle, USA.

Mamoona Shabbir and Muhammad Majid. 2024. Sentiment analysis from urdu language-based text using deep learning techniques. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–5. IEEE.

Neha Singh and Umesh Chandra Jaiswal. 2023. Sentiment analysis based on urdu reviews using hybrid deep learning models. *Appl. Comput. Syst.*, 28(2):258–265.

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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 833–844.

Ruike Zhang, Hanxuan Yang, and Wenji Mao. 2023. Cross-lingual cross-target stance detection with dual knowledge distillation framework. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 10804–10819.