ShahiEmotion: A Benchmark Dataset for Punjabi Shahmukhi Emotion Detection

Usman Nawaz[1], Muhammad Junaid Iqbal[2], Tahir Alyas[3]
Muhammad Asaf[4], Shumayla Yaqoob[5], Usman Ahmed Raza[6]
Muhammad Amin Nadim[7,8,9], Aftab Rafique[11], Faisal Rehman[10,12]

[1]University of Palermo, Italy [2]University of Rome Tor Vergata, Italy [3]Lahore Garrison University, Pakistan
[4]University of Calabria, Italy [5]Sohar University, Oman [6]University of Naples Federico II, Italy
[7]Macquarie University, Australia [8]University of Foggia, Italy [9]Pegaso Telematic University, Italy
[10]NUST, Islamabad, Pakistan [11]Muslim Youth University, Pakistan [12]University of Mianwali, Pakistan
usmannawaznlp@gmail.com

Abstract

Emotion detection is an important text classification task with applications in sentiment analysis, social media monitoring, human-computer interaction, and affective language understanding. However, Punjabi written in the Shahmukhi script remains severely under-resourced for emotion detection, with limited benchmark-style resources available for supervised evaluation. This paper introduces `ShahiEmotion`, a new Punjabi Shahmukhi emotion detection dataset containing 30,379 sentence-level instances annotated with seven emotion categories: sadness, surprise, happiness, anger, neutral, fear, and disgust. The dataset is designed to support research in a low-resource setting characterized by script-specific challenges, lexical variation, and substantial class imbalance. We establish baseline results using several pretrained transformer-based models and formulate emotion detection as a sentence-level classification task. In particular, we fine-tune multilingual BERT, multilingual DistilBERT, XLM-RoBERTa, and Urdu RoBERTa under the same training and evaluation setting using standard cross-entropy loss. Experimental results show that XLM-RoBERTa provides the strongest overall performance among the compared models. The best model achieves 77.95% accuracy, 58.47% macro-F1, and 77.60% weighted-F1 on the test set. The dataset, evaluation protocol, and baseline results introduced in this work are intended to support future research on Punjabi Shahmukhi emotion analysis and low-resource NLP.

1 Introduction

Emotion detection aims to identify the affective state expressed in text, such as happiness, sadness, anger, fear, disgust, surprise, or neutrality (Plutchik, 1980; Mohammad et al., 2018; Demszky et al., 2020). It is an important task in natural language processing (NLP) because emotion-bearing language appears widely in social media (Mohammad et al., 2018, 2013), online reviews (Chutia and Baruah, 2024; Javed et al., 2022), conversational systems (Zhong et al., 2019; Poria et al., 2019), news comments and public-opinion discussions (Li and Li, 2023; Miao, 2023), and other forms of user-generated text (Demszky et al., 2020). Reliable emotion detection can support applications such as sentiment analysis (Mohammad et al., 2013; Singh et al., 2021; Hussain et al., 2025; Irfan et al., 2019), opinion mining (Liu, 2012), affect-aware dialogue systems (Zhong et al., 2019; Poria et al., 2019), public opinion monitoring (Li and Li, 2023; Miao, 2023), and mental-health-related text analysis (Zhang et al., 2022). Compared with coarse-grained sentiment classification, which usually distinguishes positive, negative, and neutral polarity, emotion detection provides a more fine-grained representation of affective meaning (Mohammad et al., 2018; Demszky et al., 2020).

Despite progress in emotion classification for high-resource languages, many low-resource languages remain underrepresented in available datasets and benchmark evaluations (Joshi et al., 2020). Punjabi written in the Shahmukhi script has a large speaker base in Pakistan: the 2023 Population and Housing Census of Pakistan reports Punjabi as the most widely spoken mother tongue in the country, with 36.98% of the population, corresponding to about 88.9 million speakers (Pakistan Bureau of Statistics, 2024). Shahmukhi is the dominant script for Punjabi in Pakistan, and prior work notes that it is used by nearly three-fourths of Punjabi speakers worldwide (Ahmad et al., 2020). However, computational resources for Shahmukhi remain limited compared with languages that have larger annotated corpora and more mature NLP pipelines. This lack of benchmark resources makes it difficult to evaluate models consistently across shared splits, label

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 211–220

inventories, and evaluation metrics.

Punjabi Shahmukhi emotion detection presents several challenges. First, Shahmukhi is written in a Perso-Arabic script, which creates script-specific processing requirements and limits direct reuse of resources developed for Gurmukhi Punjabi or other language varieties. This script difference is important because Gurmukhi and Shahmukhi differ in alphabets, writing direction, word formation, and punctuation usage, and speakers familiar with one script may not be able to read the other script (Ahmad et al., 2020; Lehal, 2009). Second, Shahmukhi text often contains lexical and orthographic variation, including informal spelling, borrowed vocabulary, and variation in written forms. Third, emotion classes are naturally imbalanced: neutral or frequent affective categories often occur much more often than rarer emotions such as surprise, fear, or anger. In such settings, models may achieve reasonable overall accuracy while still performing poorly on minority classes (Sokolova and Lapalme, 2009). This makes balanced evaluation across emotion categories important for this task.

To address these challenges, we introduce `ShahiEmotion`, a new sentence-level Punjabi Shahmukhi emotion detection dataset containing 30,379 annotated instances. Each sentence is assigned one of seven emotion labels: sadness, surprise, happiness, anger, neutral, fear, and disgust. We formulate Punjabi Shahmukhi emotion detection as a seven-class sentence classification task.

The main contributions of this paper are as follows:

- We introduce `ShahiEmotion`, a new benchmark dataset for Punjabi Shahmukhi emotion detection with 30,379 sentence-level instances and seven emotion categories.
- We establish baseline results using five pretrained transformer models under a fixed fine-tuning setup. We show that XLM-RoBERTa provides the strongest overall performance, while Urdu RoBERTa small is the strongest alternative in terms of macro-F1.

The remainder of this paper is organized as follows. Section 2 reviews related work on emotion detection, low-resource NLP, and Punjabi Shahmukhi language resources. Section 3 describes the construction of the `ShahiEmotion` dataset, including the label inventory, dataset statistics, and train-development-test splits. Section 4 presents the task formulation, transformer-based classification framework, and training objective. Section 5 describes the experimental setup, pretrained models, hyperparameters, and evaluation metrics. Section 6 reports the main results and per-class performance. Section 7 presents the error analysis. Section 8 discusses the main findings, and Section 9 concludes the paper.

2 Related Work

Emotion detection has been widely studied as a fine-grained affective text classification task. Unlike sentiment analysis, which typically assigns text to broad polarity categories such as positive, negative, and neutral, emotion detection aims to identify more specific affective states such as joy, sadness, anger, fear, disgust, and surprise (Ekman, 2004; Plutchik, 1980). This fine-grained formulation is useful for applications that require deeper understanding of user attitudes, including social media analysis, opinion mining, affect-aware dialogue systems, and human-computer interaction (Mohammad et al., 2018; Demszky et al., 2020). Several benchmark datasets have been introduced for emotion classification in high-resource languages, including SemEval affect-related tasks and large-scale English emotion datasets such as GoEmotions (Mohammad et al., 2018; Demszky et al., 2020). These resources have supported systematic model comparison and have played an important role in the development of modern emotion classification systems.

Early approaches to emotion detection relied on lexical resources, manually designed features, and conventional supervised classifiers. Emotion lexicons such as the NRC Emotion Lexicon have been used to associate words with affective categories and to build feature-based classifiers for emotion analysis (Mohammad et al., 2013). Traditional machine learning methods such as support vector machines, logistic regression, and Naive Bayes have also been applied to emotion classification using word n-grams, character n-grams, lexicon features, and part-of-speech information (Alm et al., 2005; Aman and Szpakowicz, 2007; Mohammad et al., 2013). These methods can be effective when labeled data is limited and when lexical cues are highly predictive, but they often struggle with contextual ambiguity, informal language, and cross-domain generalization.

Neural models have substantially changed the

modeling landscape for emotion classification. Recurrent neural networks, convolutional neural networks, and attention-based architectures have been used to learn distributed representations for affective text classification (Kim, 2014; Cho et al., 2014; Bahdanau et al., 2014). More recently, pretrained transformer models have become the dominant approach for many text classification tasks. Models such as BERT, RoBERTa, mBERT, and XLM-R learn contextualized representations from large-scale unlabeled corpora and can be fine-tuned effectively for downstream classification tasks (Devlin et al., 2019; Liu et al., 2019; Pires et al., 2019; Conneau et al., 2020; Maryam et al., 2025). In multilingual and low-resource settings, multilingual encoders are particularly attractive because they provide cross-lingual transfer and support many languages and scripts without requiring task-specific pretraining data for each language (Pires et al., 2019; Conneau et al., 2020).

Low-resource emotion detection remains challenging because many languages lack annotated datasets, standardized label inventories, and reproducible benchmark splits. Prior work on low-resource NLP emphasizes that the absence of high-quality annotated resources is often a primary bottleneck for progress (Joshi et al., 2020; Pakray et al., 2025). This limitation is especially relevant for affective language processing, where emotion expression can be culturally and linguistically specific. Recent work on culturally aware LLM optimization highlights the need to evaluate NLP systems beyond accuracy, including fairness, robustness, informativeness, and efficiency (Iqbal et al., 2026). Direct transfer from high-resource languages may not capture local lexical choices, idiomatic expressions, script conventions, and emotion-specific discourse patterns.

The situation is especially challenging for Punjabi written in the Shahmukhi script. Shahmukhi uses a Perso-Arabic writing system and differs substantially from Gurmukhi Punjabi in script, orthographic conventions, and available NLP resources. Existing Punjabi NLP work has addressed tasks such as part-of-speech tagging (Tehseen et al., 2023a), named entity recognition (Ahmad et al., 2020; Tehseen et al., 2023b), machine translation (Ambreen and Rauf, 2023; Lehal, 2008), sentiment analysis (Hussain et al., 2025; Singh et al., 2021), and lexical normalization (Kaur and Saini, 2016), but benchmark-style resources for Punjabi Shahmukhi remain comparatively limited. This resource gap limits the development of robust models for downstream tasks such as emotion detection, where script-specific representation and language-specific affective expressions are important.

Research on Punjabi emotion analysis has generally been more limited compared to English, Arabic, Urdu, Hindi, or other higher-resource languages. However, many available resources either focus on the Gurmukhi script or use coarse-grained sentiment polarity labels. Emotion detection is more fine-grained than sentiment classification and requires distinguishing affective categories that may be semantically close, such as anger and disgust or fear and surprise. This makes a dedicated Shahmukhi emotion dataset important for systematic progress.

3 Dataset Construction

We introduce `ShahiEmotion`, a sentence-level Punjabi Shahmukhi emotion detection dataset. The Shahmukhi sentences were derived from the English-Punjabi Shahmukhi Parallel Sentences Corpus available through the Mozilla Data Collective (Mozilla Data Collective, 2026). We used the Punjabi Shahmukhi text of the corpus and manually annotated each sentence with one gold emotion label. After removing empty or unusable rows, the final dataset contains 30,379 annotated sentences across seven emotion categories: sadness, surprise, happiness, anger, neutral, fear, and disgust. In Shahmukhi, these labels correspond to اداسی, حیرانی, خوشی, غصہ, غیرجانبدار , ڈر, and گھن, respectively.

3.1 Preprocessing and Label Inventory

The dataset is stored in a sentence-level format with two fields: the Shahmukhi sentence and its emotion label. During preprocessing, we remove empty rows and normalize whitespace. The text is kept in Unicode Shahmukhi script, and the label mapping is fixed across all splits to ensure reproducible training and evaluation. The sentences were manually annotated according to the fixed seven-label inventory. Annotators assigned the emotion that was most clearly expressed by each sentence. Sentences without a clear affective state were labeled as غیرجانبدار . Ambiguous cases were resolved using the dominant expressed emotion, following the single-label classification setup used in this work.

Table 1: Label distribution in the Punjabi Shahmukhi emotion dataset.

Label	Count	English
غیرجانبدار	17,459	Neutral
خوشی	6,687	Happiness
گھن	2,516	Disgust
اداسی	1,355	Sadness
ڈر	1,008	Fear
غصہ	826	Anger
حیرانی	528	Surprise
Total	30,379	-

Table 2: Label distribution across train, development, and test splits.

Label	Train	Dev	Test
غیرجانبدار	13,967	1,746	1,746
خوشی	5,350	669	668
گھن	2,013	252	251
اداسی	1,084	135	136
ڈر	806	101	101
غصہ	661	82	83
حیرانی	422	53	53
Total	24,303	3,038	3,038

3.2 Dataset Statistics

The dataset is substantially imbalanced, with غیرجانبدار and خوشی being the largest classes, while حیرانی, غصہ, and ڈر contain fewer examples. Table 1 shows the label distribution of the full dataset.

We split the dataset into fixed train, development, and test sets using an 80/10/10 stratified split. The training split contains 24,303 sentences, while the development and test splits contain 3,038 sentences each. The development set is used for model selection, and the test set is reserved for final evaluation.

4 Methodology

We formulate Punjabi Shahmukhi emotion detection as a sentence-level multi-class classification task. Given an input sentence

$$S = (w_1, w_2, \ldots, w_n), \tag{1}$$

where w_i denotes the i-th token in the sentence, the goal is to predict an emotion label

$$y \in \mathcal{Y}, \tag{2}$$

where $\mathcal{Y}$ is the set of seven emotion categories: اداسی, حیرانی, خوشی, غصہ, غیرجانبدار , ڈر, and گھن. Each sentence receives exactly one label, so the task is treated as single-label classification.

4.1 Sentence Classifier

We use pretrained transformer encoders as sentence classification models. Given a Shahmukhi sentence, the tokenizer converts the input into a sequence of subword units, which are passed through the encoder to obtain contextualized representations. The encoder output is then passed to a task-specific classification head that produces logits over the seven emotion categories:

$$\mathbf{o} = \text{Classifier}(\text{Encoder}(S)), \tag{3}$$

where $\mathbf{o} \in R^7$ is the output logits. The predicted label is obtained by selecting the class with the highest logit:

$$\hat{y} = \arg\max_{k \in \mathcal{Y}} o_k. \tag{4}$$

This architecture is used as a standard fine-tuning framework for comparing pretrained transformer encoders on the proposed Punjabi Shahmukhi emotion dataset. All models are trained using standard cross-entropy loss. Since the dataset is imbalanced, we report macro-averaged and weighted metrics in addition to accuracy.

5 Experimental Setup

We evaluate Punjabi Shahmukhi emotion detection using the fixed train, development, and test splits described in Section 3. All systems are trained on the training split, evaluated on the development split during training, and reported on the held-out test split.

5.1 Models and Training Settings

We compare five pretrained transformer models under the same fine-tuning setup. The multilingual BERT models are included as multilingual contextual baselines (Devlin et al., 2019; Pires et al., 2019). Multilingual DistilBERT is included as a smaller distilled multilingual encoder (Sanh et al., 2019). XLM-RoBERTa is included as a strong multilingual transformer trained on large-scale multilingual data (Conneau et al., 2020). Urdu RoBERTa small is included because Urdu and Punjabi Shahmukhi share a related Perso-Arabic script environment (UrduHack, 2021).

Each model is fine-tuned with a sequence classification head over the seven emotion labels. We keep the dataset splits, preprocessing, maximum sequence length, batch size, number of epochs, learning rate, and evaluation metrics fixed across all experiments. All experiments are implemented

in PyTorch using the Hugging Face Transformers library. Table 3 summarizes the main hyperparameters.

Table 3: Hyperparameter settings for fine-tuning.

Hyperparameter	Value
Compared models	5
Maximum sequence length	128
Batch size	16
Epochs	10
Learning rate	2×10^{-5}
Weight decay	0.01
Optimizer	AdamW

5.2 Evaluation Metrics

We report accuracy, macro-averaged precision, macro-averaged recall, macro-F1, and weighted-F1. Since the dataset is imbalanced, accuracy alone is not sufficient: a model may obtain high accuracy by performing well on the majority classes while failing on minority emotions. We therefore use macro-F1 as the primary comparison metric, because it gives equal importance to each emotion category regardless of its frequency.

Accuracy is defined as the proportion of correctly classified sentences:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\hat{y}_i = y_i\}, \tag{5}$$

where $\hat{y}_i$ is the predicted label, y_i is the gold label, and N is the number of test instances.

For each class, precision, recall, and F1 are computed in the standard way. Macro-F1 is the unweighted average of class-level F1 scores:

$$\text{Macro}-\text{F1} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c, \tag{6}$$

where C is the number of emotion classes. Weighted-F1 is also reported to show performance while accounting for class frequency.

6 Results and Analysis

The results show that the choice of pretrained encoder has a clear effect on performance for Punjabi Shahmukhi emotion detection. Table 4 presents the test-set results for the evaluated pretrained transformer models.

XLM-RoBERTa base achieves the strongest overall performance among the compared models. It obtains 77.95% accuracy, 58.47% macro-F1, and 77.60% weighted-F1 on the test set. This result shows that XLM-RoBERTa provides the most effective pretrained representation for this dataset under the fixed fine-tuning setup. Its advantage is especially clear in macro-F1, which is the primary metric for this imbalanced emotion classification task.

Table 4: Test-set results for different pretrained models.

Model	Acc.	Macro-F1	Weighted-F1
mBERT cased	77.72	54.14	76.90
mBERT uncased	77.42	54.71	76.74
DistilBERT multilingual	76.07	52.02	75.31
XLM-RoBERTa base	**77.95**	**58.47**	**77.60**
Urdu RoBERTa small	77.45	56.32	76.84

Urdu RoBERTa small obtains the second-best macro-F1 score, with 56.32% macro-F1. This result suggests that pretraining on a related Perso-Arabic script language can be useful for Punjabi Shahmukhi emotion detection. Although Urdu RoBERTa small does not outperform XLM-RoBERTa, it performs better than both multilingual BERT variants in macro-F1. This indicates that related-script pretraining may help with some emotion categories, even when the model is smaller than the multilingual baselines.

The multilingual BERT models also provide competitive results. The cased version achieves 77.72% accuracy and 76.90% weighted-F1, while the uncased version achieves 77.42% accuracy and 76.74% weighted-F1. However, their macro-F1 scores remain lower than XLM-RoBERTa and Urdu RoBERTa small. The uncased model slightly improves macro-F1 compared with the cased model, but the cased model obtains slightly higher accuracy and weighted-F1. This shows that the two multilingual BERT variants behave similarly overall, with small differences across metrics.

Multilingual DistilBERT obtains the lowest performance among the evaluated models, with 76.07% accuracy, 52.02% macro-F1, and 75.31% weighted-F1. Since DistilBERT is a smaller distilled model, this result suggests that reduced model capacity may affect performance on this low-resource and imbalanced emotion detection task. The decrease is most visible in macro-F1, indicating that the lighter model has more difficulty with minority classes.

Overall, the results show that pretrained transformer models can provide strong baselines for Punjabi Shahmukhi emotion detection, but the choice of pretrained encoder matters.

XLM-RoBERTa base is the best-performing model across all three main metrics, while Urdu RoBERTa small is the strongest alternative in terms of macro-F1. These findings support the use of large multilingual pretrained encoders for Punjabi Shahmukhi emotion classification, while also showing that related-script models can be useful for this setting.

6.1 Per-Class Performance

The model performs best on the two largest classes, غیرجانبدار and خوشی. It obtains an F1 score of 85.66% for غیرجانبدار and 82.37% for خوشی. These results indicate that the model can learn reliable representations for frequent classes. Table 5 reports per-class precision, recall, and F1 for the best-performing model, XLM-RoBERTa base with standard cross-entropy.

Table 5: Per-class test performance of XLM-RoBERTa base with standard cross-entropy.

Label	Prec.	Rec.	F1
اداسی	61.11	64.71	62.86
حیرانی	45.00	33.96	38.71
خوشی	80.54	84.28	82.37
غصہ	41.77	39.76	40.74
غیرجانبدار	85.12	86.20	85.66
ڈر	55.17	47.52	51.06
گھن	51.13	45.02	47.88

The minority classes remain more difficult. The lowest F1 scores are observed for حیرانی, غصہ, and گھن. This is expected because these classes contain fewer training examples and may overlap semantically with other affective categories. For example, anger and disgust can share similar negative lexical cues, while surprise may overlap with fear or neutral event descriptions depending on context. These results show that class imbalance and semantic overlap remain central challenges for Punjabi Shahmukhi emotion detection.

The model obtains a relatively stronger F1 score for اداسی, reaching 62.86%, even though this class is smaller than خوشی and غیرجانبدار . This suggests that sadness may contain more consistent lexical or contextual cues than some other minority classes. In contrast, حیرانی has the lowest F1 score, indicating that surprise is particularly difficult to detect at the sentence level. This may be because surprise can be expressed indirectly or may require wider context beyond a single sentence.

Overall, the per-class results show that XLM-RoBERTa base provides the strongest overall baseline, but performance is still uneven across emotion categories. Frequent classes are recognized more reliably, while minority and semantically overlapping emotions remain harder to classify.

7 Error Analysis

To better understand the behavior of the best-performing model, we analyze the confusion matrix of XLM-RoBERTa base with standard cross-entropy. Table 6 presents the test-set confusion matrix. Rows correspond to gold labels and columns correspond to predicted labels.

Table 6: Test-set confusion matrix for XLM-RoBERTa base. Rows are gold labels and columns are predicted labels.

Gold/Pred.	اداسی	حیرانی	خوشی	غصہ	غیرجانبدار	ڈر	گھن
اداسی	88	0	8	2	30	5	3
حیرانی	0	18	2	3	19	7	4
خوشی	5	1	563	2	84	2	11
غصہ	4	5	2	33	25	3	11
غیرجانبدار	28	9	102	19	1505	13	70
ڈر	7	3	6	5	23	48	9
گھن	12	4	16	15	82	9	113

The largest number of errors comes from the غیرجانبدار class, which is also the largest class in the dataset. In particular, 102 gold غیرجانبدار sentences are predicted as خوشی, and 70 are predicted as گھن. This suggests that some neutral sentences contain words or phrases that are also common in affective contexts. Since neutral examples are diverse and often lack explicit emotional markers, the boundary between neutral and weakly emotional sentences can be difficult for the model to learn.

Another frequent confusion occurs between گھن and غیرجانبدار , where out of 251 gold گھن examples, 82 are predicted as غیرجانبدار . This indicates that disgust is difficult to identify when the sentence does not contain strong lexical cues. Some disgust expressions may be implicit, indirect, or expressed through evaluative wording that overlaps with neutral criticism.

The model also confuses حیرانی with غیرجانبدار and ڈر, where out of 53 gold حیرانی examples, 19 are predicted as غیرجانبدار and 7 are predicted as ڈر. This pattern may arise because surprise can be expressed with limited affective wording and may depend heavily on context. In some cases, surprising events may also imply fear or uncertainty, making the distinction between these classes difficult at the sentence level.

For غصہ, the model correctly predicts 33 out of 83 test examples, but it also confuses anger with گھن , غیرجانبدار, and حیرانی. This pattern suggests that negative affective classes share overlapping lexical and semantic features. Anger and disgust are particularly close because both can express negative evaluation, rejection, or disapproval.

The model performs more reliably on خوشی and غیرجانبدار . For خوشی, 563 out of 668 test examples are correctly classified, while for غیرجانبدار , 1505 out of 1746 examples are correctly classified. These results are consistent with the per-class scores and show that the model benefits from larger training support for frequent classes.

Overall, the error analysis shows that the main difficulty is not only class imbalance but also semantic overlap among emotion categories. The model performs strongly on frequent and lexically clearer classes such as خوشی and غیرجانبدار , but it remains challenged by minority classes and by emotions whose expressions are indirect or context-dependent. These findings suggest that future improvements may require larger annotated data for minority classes, clearer annotation guidelines for borderline cases, and richer modeling of contextual and pragmatic cues.

8 Discussion

The results show that Punjabi Shahmukhi emotion detection is a challenging task even when using pretrained transformer encoders. All evaluated models achieve reasonable overall accuracy, but performance varies considerably across emotion categories. The models perform better on frequent classes such as غیرجانبدار and خوشی, while minority classes such as حیرانی, غصہ, ڈر, and گھن remain more difficult. This pattern reflects the combined effect of class imbalance, limited examples for rare emotions, and semantic overlap among affective categories.

The comparison between pretrained encoders shows that model choice has a clear effect on performance. XLM-RoBERTa base achieves the strongest overall result, with the highest accuracy, macro-F1, and weighted-F1 among the compared models. This suggests that its multilingual pretraining and subword representation are effective for Punjabi Shahmukhi emotion classification. Since Shahmukhi is a low-resource script setting, broad multilingual pretraining appears to provide useful transfer for sentence-level emotion detection.

Urdu RoBERTa small obtains the second-best macro-F1 score. This result suggests that related-script pretraining can be useful for Punjabi Shahmukhi, even when the pretrained model is smaller than some multilingual alternatives. Urdu and Punjabi Shahmukhi share a Perso-Arabic script environment, and this may help the model represent script-specific patterns more effectively. However, Urdu RoBERTa small still does not outperform XLM-RoBERTa, which indicates that broader multilingual pretraining remains highly useful for this task.

The multilingual BERT models perform competitively but remain below XLM-RoBERTa in macro-F1. The cased version achieves slightly higher accuracy and weighted-F1, while the uncased version achieves slightly higher macro-F1. This shows that both mBERT variants are useful baselines, but their performance is less balanced across emotion categories than XLM-RoBERTa. Multilingual DistilBERT obtains the lowest macro-F1 among the compared models, suggesting that a smaller distilled encoder may have less capacity to model minority classes and fine-grained affective distinctions in this dataset.

These findings also indicate that accuracy alone is not a sufficient measure for Punjabi Shahmukhi emotion detection. Several models achieve similar accuracy, but their macro-F1 scores differ more clearly. Since emotion analysis often requires reliable recognition of minority emotions such as fear, anger, and surprise, macro-F1 provides a more informative evaluation criterion. The stronger macro-F1 of XLM-RoBERTa shows that it provides a better balance across the full emotion inventory.

Overall, the experiments establish a benchmark-style evaluation setting for Punjabi Shahmukhi emotion detection. The results show that pretrained transformer models can be effective for this low-resource script setting, but they also reveal important limitations related to class imbalance and semantic overlap. The dataset and model comparison introduced in this work therefore provide a foundation for future research on more robust Punjabi Shahmukhi affective language understanding.

9 Conclusion

This paper introduced `ShahiEmotion`, a new sentence-level dataset for Punjabi Shahmukhi emo-

tion detection. The dataset contains 30,379 annotated sentences across seven emotion categories: اداسی, حیرانی, خوشی, غصہ, غیرجانبدار, ڈر, and گھن. By providing fixed train, development, and test splits, the dataset establishes a reproducible benchmark setting for supervised emotion classification in Punjabi written in the Shahmukhi script.

We evaluated five pretrained transformer models under the same fine-tuning setup using standard cross-entropy loss. The compared models include multilingual BERT cased, multilingual BERT uncased, multilingual DistilBERT, XLM-RoBERTa base, and Urdu RoBERTa small. The results show that XLM-RoBERTa base provides the strongest overall performance among the evaluated models. It achieves 77.95% accuracy, 58.47% macro-F1, and 77.60% weighted-F1 on the test set.

The comparison also shows that Urdu RoBERTa small is a strong alternative, obtaining the second-best macro-F1 score. This suggests that related-script pretraining can be useful for Punjabi Shahmukhi emotion detection. However, XLM-RoBERTa remains the best model overall, indicating that large-scale multilingual pretraining provides effective transfer for this low-resource setting.

The per-class results show that the main remaining challenges involve minority-class recognition and semantic overlap among emotion categories.

Overall, this work contributes a new benchmark-style resource and baseline model comparison for Punjabi Shahmukhi emotion detection. We hope that the dataset and results presented here will support future research on Punjabi Shahmukhi affective computing, low-resource text classification, and multilingual emotion analysis.

Limitations

This study has several limitations. First, the experiments are conducted on a single Punjabi Shahmukhi emotion dataset. Although the dataset provides a useful benchmark setting, additional evaluation on external domains would be needed to measure how well the models generalize to other forms of Shahmukhi text, such as social media posts, news comments, dialogue, or literary text.

Second, the dataset is substantially imbalanced. Minority classes such as حیرانی, غصہ, and ڈر contain far fewer examples than غیرجانبدار and خوشی. This affects model performance and makes some emotion categories more difficult to learn. Expanding the dataset with additional examples for under-represented emotions would likely improve macro-level performance.

Third, emotion annotation is inherently subjective. Some sentences may express weak, implicit, or overlapping emotions, and a single-label classification setup may not fully capture such cases. For example, a sentence may express both fear and surprise, or may contain negative evaluation that could be interpreted as anger or disgust. Future versions of the dataset could include multiple annotators, agreement analysis, emotion intensity scores, or multi-label annotations for ambiguous cases.

Fourth, the modeling experiments compare a limited set of pretrained encoders and use a single standard fine-tuning objective. Although the selected models provide useful multilingual and related-script baselines, future work should evaluate additional multilingual, regional, and script-specific models, including larger transformer variants and models trained on Urdu, Punjabi, or other related languages written in Perso-Arabic scripts.

Finally, the present work establishes baseline results rather than proposing a novel model architecture. The main contribution is the dataset and benchmark setting. Future work can build on this resource by exploring class-weighted loss, focal loss, data augmentation, contrastive learning, parameter-efficient fine-tuning, domain adaptation, and methods that better capture contextual or pragmatic cues in emotion expression.

Ethics Statement

This work uses publicly available Punjabi Shahmukhi text and manually annotates it for sentence-level emotion detection. Emotion labels may involve subjective judgments and may reflect biases in the source text or annotation process. The dataset and models are intended for research use and should not be used for high-stakes decision-making.

Data Availability

The `ShahiEmotion` dataset and documentation will be released for non-commercial research use under the CC BY-NC 4.0 license at: github.com/usmannawaz01/ShahiEmotion.

References

Muhammad Tayyab Ahmad, Muhammad Kamran Malik, Khurram Shahzad, Faisal Aslam, Asif Iqbal,

Zubair Nawaz, and Faisal Bukhari. 2020. Named entity recognition and classification for punjabi shahmukhi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–13.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International conference on text, speech and dialogue*, pages 196–205. Springer.

Shehzadi Ambreen and Sadaf Abdul Rauf. 2023. Neural machine translation system for pakistani languages. In *2023 20th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 220–225. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1724–1734.

Tulika Chutia and Nomi Baruah. 2024. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57(8):203.

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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.

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.

Paul Ekman. 2004. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation: Proceedings of the sixth international colloquium on cognitive science (ICCS-99)*, pages 39–50. Springer.

Muzammal Hussain, Saddam Ali, Hina Sattar, Ali Raza, Muhammad Hamza Akbar, and Muhammad Ahsan Rafiq. 2025. Urdu-punjabi code switched sentiment analysis empowered by a deep learning framework integrating xlm-r, and gpt. *VAWKUM Transactions on Computer Sciences*, 13(2):01–20.

Muhammad Junaid Iqbal, Muhammad Asghar Khan, Tahir Alyas, Sagheer Abbas, Arif Jawaid, and Fabio Massimo Zanzotto. 2026. A multi-objective reinforcement learning approach to prompt optimization in nlp. *Procedia Computer Science*, 275:966–974.

Rizwana Irfan, Osman Khalid, Muhammad Usman Shahid Khan, Faisal Rehman, Atta Ur Rehman Khan, and Raheel Nawaz. 2019. Socialrec: A context-aware recommendation framework with explicit sentiment analysis. *IEEE Access*, 7:116295–116308.

Arslan Javed, Faisal Rehman, Nadeem Sarfraz, Hanan Sharif, Rashid Khan, and Abdul Manan Khan. 2022. Movie recommendation system with sentimental analysis using cosine similarity technique. In *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, pages 1–8. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

Jasleen Kaur and Jatinderkumar R Saini. 2016. Punjabi stop words: a gurmukhi, shahmukhi and roman scripted chronicle. In *Proceedings of the ACM Symposium on Women in Research 2016*, pages 32–37.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751.

Gurpreet Singh Lehal. 2008. Shahmukhi to gurmukhi transliteration system: A corpus based approach. *Research on computing science*.

Gurpreet Singh Lehal. 2009. A gurmukhi to shahmukhi transliteration system. In *proceedings of ICON-2009: 7th international conference on Natural Language Processing*, pages 167–173.

Chen Li and Fanfan Li. 2023. Emotion recognition of social media users based on deep learning. *PeerJ Computer Science*, 9:e1414.

Bing Liu. 2012. Sentiment analysis and opinion mining.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zahra Maryam, Faisal Rehman, Ummar Ashraf, Muhammad Sarmad Shakil, Muhammad Yousif, and 1 others. 2025. Sentiment analysis on social media posts using roberta: A deep learning approach for text classification. *Journal of Computing & Biomedical Informatics*, 9(01).

Ruomu Miao. 2023. Emotion analysis and opinion monitoring of social network users under deep convolutional neural network. *Journal of Global Information Management (JGIM)*, 31(1):1–12.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Mozilla Data Collective. 2026. English–punjabi (shahmukhi) parallel sentences corpus. https://mozilladatacollective.com/datasets/cmkh9rso90076nv076jwxgjv3. Mediamen Archives; steward: MEDIAMEN; CC-BY-NC-4.0.

Pakistan Bureau of Statistics. 2024. 7th population and housing census 2023: National census report. Technical report, Pakistan Bureau of Statistics, Government of Pakistan.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

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*.

Jaspreet Singh, Gurvinder Singh, Rajinder Singh, and Prithvipal Singh. 2021. Morphological evaluation and sentiment analysis of punjabi text using deep learning classification. *Journal of King Saud University-Computer and Information Sciences*, 33(5):508–517.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Amina Tehseen, Toqeer Ehsan, Hannan Bin Liaqat, Amjad Ali, and Ala Al-Fuqaha. 2023a. Neural pos tagging of shahmukhi by using contextualized word representations. *Journal of King Saud University-Computer and Information Sciences*, 35(1):335–356.

Amina Tehseen, Toqeer Ehsan, Hannan Bin Liaqat, Xiangjie Kong, Amjad Ali, and Ala Al-Fuqaha. 2023b. Shahmukhi named entity recognition by using contextualized word embeddings. *Expert Systems with Applications*, 229:120489.

UrduHack. 2021. roberta-urdu-small. https://huggingface.co/urduhack/roberta-urdu-small. Urdu RoBERTa small language model.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. 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 165–176.