Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

A DOMAIN-SPECIFIC APPROACH FOR CROSS-LINGUAL EMOTION DETECTION THROUGH TEXT MINING

Israr Hanif[1], Faisal Shahzad*[2]

[1, *2] Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan
Corresponding author's email address: (faisal.shahzad.research@gmail.com).

DOI:-https://doi.org/10.5281/zenodo.20989876

Article History

Received: 17 May 2026

Accepted: 26 June2026

Published: 28 June 2026

Corresponding Author: * (faisal.shahzad.research@gmail.com).

Abstract

Emotional recognition is the aspect of sentiment analysis that focuses on a more nuanced or meaningful understanding of the complex and diverse emotions found in text. Existing studies have mostly been limited to English support and general transformer models, while there is a significant lack of research in low-resource and morphologically complex languages like Urdu. But the Urdu language is different from English, and there is no Urdu emotion dataset with annotations, which makes cross-language emotion detection a challenging problem. To bridge this gap, the GoEmotions dataset was translated into Urdu and mapped into 7 basic emotional categories (anger, happiness, sadness, surprise, disgust, fear, and neutral), and the performance of six transformer models (BERT, DistilBERT, IndicBERT, RoBERTa, XLM-R, and RemBERT) on processed and unprocessed versions of English–Urdu was evaluated in four different configurations. The task was then defined as a multi-label emotion classification problem and tested in four configurations: English-to-English, Urdu-to-Urdu, English-to-Urdu, and Urdu-to-English. The results showed that in monolingual experiments, BERT achieved the highest accuracy (Acc=0.8115) on English data, while XLM-R gave the best F1 score (F1=0.4360) on Urdu data, and RoBERTa showed the highest accuracy (Acc=0.8161) on unprocessed Urdu text. In the cross-lingual context, XLM-R gave the best results (Acc=0.8219, F1=0.5177), and RemBERT was also close, which shows the multilingual generalization ability of these models. Moreover, preprocessing did not significantly improve on low-resource and morphologically rich texts like Urdu. A comparative analysis also revealed that while sentiment classification in Arabic reached 90% accuracy, Urdu-based experiments were limited to a maximum of 81.6%. The results of this study provide a reliable starting point for future cross-linguistic sentiment analysis on low-resource languages.

Keywords: Emotion recognition, low-resource languages, Urdu, cross-lingual transfer, transformer models, preprocessing.

Introduction

Language-based emotion recognition is a fundamental aspect of human experience and social interaction. Humans can express a wide range of emotional states with a few words, and enabling machines to understand these emotions has long been a major goal of artificial intelligence and affective computing (Picard, 2000). Emotion detection has gained increasing importance in recent years, particularly in emotion-aware user feedback analysis (Öhman et al., 2020; Ullah et al., 2022), social behaviour monitoring, and multilingual conversational systems. Accurately identifying emotional states from user-generated text enables such systems to produce more personalized, contextual, and emotionally aware responses. Initially, rule-based and lexical resources were relied upon for emotion recognition (Preoţiuc-Pietro et al., 2016; Balamurali et al., 2012; Strapparava & Mihalcea, 2007), but these methods were less effective due to their limited scope, monolingual focus, and reliance on manual features.

Through the availability of large-scale annotated datasets such as EmoBank (Buechel & Hahn, 2022) and GoEmotions (Demszky et al., 2020), the use of deep learning and especially transformer-based models has opened new dimensions in this field. GoEmotions, in particular, is a large dataset of 58k carefully selected Reddit comments, containing labels for 27 fine-grained emotion categories, which provides a solid foundation for modeling complex emotional expressions. However, the direct use of such English-centric resources is limited for low-resource languages because emotional expressions may change across scripts, cultures, and linguistic structures.

Most existing emotion detection studies focus on high-resource languages, particularly English, while low-resource languages such as Urdu remain underexplored. (Demszky et al., 2020; Hassan et al., 2021; Strapparava & Mihalcea, 2007). Low-resource languages such as Urdu face several challenges, including the lack of annotated data (Ashraf et al., 2022), the complexity of the script, and the richness of morphology (Vardag et al., 2022). These factors not only hinder the development of accurate monolingual models but also make it difficult to effectively transfer multilingual models. Furthermore, cultural differences can lead to semantic changes in the expression and interpretation of emotions (Hess & Hareli, 2015), which affects the performance of models on translated data. As a result, a formal study to evaluate transfer of emotion knowledge from English to Urdu without losing its semantic and cultural meaning is essential.

In languages like Urdu, which have fewer resources, emotional recognition problems are more projecting because the available data is limited and the linguistic structure is complex, but is also further complicated by the common phenomenon of code-switching, where Urdu and English (or other languages) are used together in the same sentence. This linguistic mix makes it difficult for models to accurately learn semantic and syntactic patterns, as training datasets are often monolingual (Bender et al., 2018). In this situation, models face additional challenges in language recognition, tokenization, and contextual understanding, which can affect performance in cross-lingual emotion detection. Cross-lingual emotion recognition strategies aim to overcome these difficulties, and to do this, labeled data from one language is transferred to train or test models in another language. Cross-lingual techniques have shown positive results in sentiment analysis (Baali & Ghneim, 2019), code-mixed text processing (Vijay et al., 2018), and semantic unit recognition (Darwish, 2013). These methods include bilingual embeddings (Artetxe

et al., 2016; Singhal & Bhattacharyya, 2016), multilingual transformers such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019), and transfer learning (Altaf et al., 2022; Kadiyala, 2024). However, research on cross-lingual sentiment recognition in languages with fewer resources and higher morphological complexity is still limited. However, research on cross-lingual emotion detection for Urdu remains limited, especially under controlled monolingual and cross-lingual evaluation settings. Existing research on emotion recognition in Urdu is limited and small-scale. For example, (Ashraf et al., 2022) multi-label classification was performed to identify different emotions from Urdu tweets, while (Vardag et al., 2022) developed a contextual Urdu corpus. However, these studies lack large-scale qualitative data and comprehensive analysis of cross-linguistic models. Moreover, these studies do not provide a systematic comparison of English-only and multilingual transformer models under both English-to-Urdu and Urdu-to-English transfer settings.

Although significant progress has been made in the field of emotion recognition in languages with more resources, the field is still in its infancy for a low-resource language like Urdu. Urdu's linguistic structure, script complexity, and limited standardized datasets further compound this challenge. In addition, cultural differences and semantic drift during translation can negatively impact the performance of cross-lingual models. The main goal of this study is to create a standardized bilingual dataset to address these challenges and to test modern transformer models in monolingual and cross-lingual situations. The aim of this research is not only to provide a strong foundation for Urdu but also to pave the way for effective emotion recognition in other low-resource languages, so that machines can better understand human emotions across languages and cultures.

In this context, the following are the salient contributions of this research:

Dataset preparation – The GoEmotions dataset (Demszky et al., 2020) was translated and mapped into Urdu to create a bilingual (English–Urdu) corpus, combining 27 categories into 7 basic emotions: anger, happiness, sadness, surprise, disgust, fear, and neutral.

To build the English–Urdu bilingual emotion detection benchmark, we translate and map the GoEmotions dataset to the seven basic emotion categories – anger, disgust, fear, joy, sadness, surprise, neutral.

We systematically analyze transformer models in monolingual and cross-lingual cases, with English-to-Urdu and Urdu-to-English transfer.

English emotion classification and Urdu emotion classification are studied for the preprocessing impact and it is demonstrated that the performance of emotion classification decreases in morphologically rich low-resource languages like Urdu with conventional preprocessing.

Model Analysis – Six transformer-based models—BERT-base (Devlin et al., 2019), DistilBERT, IndicBERT, RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2019), and RemBERT (Chung et al., 2020)—were tested in both monolingual and cross-lingual setups.

To explore the effect of multilingual pretraining on crosslingual emotion transfer, we compare monolingual and multilingual transformer models.

The effect of preprocessing – The effects of preprocessing in a morphologically rich language like Urdu were analyzed, specifically examining its impact on F1-Score and overall performance.

Error Analysis and robustness Evaluation for Urdu cross-lingual emotion detection to find out the major limitations in Urdu.

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 presents the proposed methodology. Section 4 discusses the experimental results and analysis. Finally, Section 5 concludes the paper and highlights future work.

Problem Formulation

In this study, the problem of cross-lingual detection of emotions between English and Urdu is discussed. For a given instance x of a text, the task is to predict an emotion (or multiple emotions) from a fixed set of emotions Y = {anger, disgust, fear, joy, sadness, surprise, neutral}. Sentences can contain more than one emotion, so the task is presented as a multi-label classification task.

Mathematically, the model takes a text x as input and outputs a binary emotion vector y = [y1, y2, ..., y7], where each yi denotes whether a text contains the corresponding emotion category or not. The study assesses this task in four different experimental conditions: English to English, Urdu to Urdu, English to Urdu, and Urdu to English. The first two settings will measure the monolingual performance; the latter two will measure the cross-lingual generalization.

The primary difficulty is the fact that Urdu is a low resource and morphologically rich language, whereas most large emotion datasets and emotion pretrained models are created for English. So the study explores how well multilingual transformer models can transfer emotional knowledge among languages, compared to monolingual models.

Related Work

Several models and datasets for classifying emotions have emerged over time, which have proven to be helpful in the progress of this field. Although researchers have not been able to agree on a precise classification of human emotions, Ekman proposed six basic emotions (happiness, anger, fear, sadness, disgust, and surprise) (Ekman et al., 1999), while Pulchick proposed a model of eight basic emotions and their combinations to define secondary emotions (Salzen, 1991). Psychological studies such as (Wang et al., 2022) have identified 65 different emotions, leading to diverse classifications across different datasets and research.

In the early days, emotion recognition was mostly based on rule-based systems and lexicon-based approaches (Mohammad & Kiritchenko, 2015; Strapparava & Mihalcea, 2008; Touri et al.), which had the limitations of manual features and monolingual focus. Over time, large-scale manually annotated datasets emerged, such as CrowdFlower (2016) (Kusal et al., 2022), which contains 39K labeled examples, and GoEmotions (Demszky et al., 2020), which is a fine-grained dataset based on 58k Reddit comments containing 27 emotions. Corpora such as XED (Öhman et al., 2020) and SemEval2018 (Mohammad et al., 2018) for multilingual research have expanded the possibilities of cross-linguistic modeling.

Research in low-resource languages such as Urdu, Hindi, and Arabic has been relatively limited (Baali & Ghneim, 2019; Ullah et al., 2022), the main reasons being the lack of annotated corpora, script complexity, and morphological richness. Existing datasets such as (Abdul-Mageed et al., 2016; Al-Khatib & El-Beltagy, 2017) produced small-sized Twitter corpora based on six basic emotions, while (Hassan et al., 2021) labeled 7,268 tweets with 8 basic emotions based on the Pulchak model.

In terms of modeling, various methods have been used in sentiment classification, starting with traditional machine learning methods (such as Naive Bayes and SVM) (Al-Khatib & El-Beltagy, 2017; Fernandez et al., 2024) to deep learning architectures such as CNNs, RNNs, and especially transformer-based models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al.,

2019), and XLM-R (Conneau et al., 2019), which have shown significant performance improvements through large-scale pre-training and contextual representation. In low-resource NLP, IndicBERT, RemBERT (Chung et al., 2020; Dabre et al., 2021), and ensemble approaches (Dabre et al., 2021) have also shown effective results.

Recent research has also conducted experiments on cross-lingual emotion detection for Arabic. In (Hassan et al., 2021), tests were performed on Arabic and Spanish using English as the source language. According to the results, BERT-based monolingual models performed about 4% better than the current state-of-the-art on Arabic data, while using English data in a cross-lingual setting, they achieved a relative effectiveness of about 90% for Arabic. This comparison shows that the cross-lingual models performed better in Arabic than in Urdu (where the maximum accuracy was 81.6%), which indicates differences in dataset size and linguistic complexity.

Various strategies have been adopted for cross-lingual emotion recognition, including bilingual embeddings (Klementiev et al., 2012; Mikolov et al., 2013), transfer learning (Howard & Ruder, 2018), multilingual fine-tuning (Lauscher et al., 2020; Pires et al., 2019), and machine translation-based training (Klinger & Cimiano, 2015). However, semantic drift and cultural variation (Hareli et al., 2015) can affect the accurate representation of emotions in translated content.

This research follows this trend by translating and mapping the GoEmotions dataset to a bilingual (English-Urdu) corpus to analyze the utility of transformer-based models in monolingual and cross-lingual contexts and to analyze the effects of preprocessing in morphologically rich languages like Urdu.

Table 3.1: Comparison Table

Study	Language	Dataset	Method	Limitation
Ashraf et al.	Urdu	Urdu tweets	ML/DL models	Limited Urdu dataset
Vardag et al.	Urdu	Contextual Urdu corpus	RNN/GRU/LSTM	Limited emotion categories
Hassan et al.	Arabic/Spanish	Cross-lingual data	BERT-based models	Not focused on Urdu
GoEmotions	English	Reddit comments	BERT baseline	English-only dataset
This study	English–Urdu	Translated GoEmotions	Multilingual transformers	Focused on Urdu cross-lingual transfer

Study table 3.1 differs from previous studies on Urdu emotion detection, which typically used small-scale Urdu datasets or monolingual settings, by assessing both cross-lingual and monolingual transfer from English to Urdu and vice versa using the translated version of the GoEmotion dataset in English–Urdu. This allows for a more accurate comparison of transformer models trained in English and those trained in multiple languages.

Methodology

Dataset

The main dataset used in this study is GoEmotions (Demszky et al., 2020), which contains approximately 58,000 Reddit comments labeled with 27 emotional categories and one neutral rating. The dataset was chosen due to its high-quality annotations and broad emotional coverage, covering both basic emotions (Ekman, 1999) and extended categories. For the present study, an English–Urdu bilingual version of the GoEmotions dataset was developed, in which English sentences were translated by using the Google Translate API and annotated into Urdu to maintain uniform labeling standards. A two-stage validation was adopted to ensure semantic consistency in the translation, which included professional translation and manual correction. This bilingual corpus provided the basis for cross-lingual training and evaluation. The original GoEmotions dataset contains 27 or neutral emotion categories; however, this study focused on seven core emotions to allow for more precise and controlled experiments for cross-linguistic alignment and classification.

Dataset Balancing

The original GoEmotions dataset contained approximately 58k entries, but it had a severe class imbalance, which meant that some emotions had too many instances and some too few. To address this issue, an undersampling technique was used to bring all classes closer together and train the models on more balanced data. This process left approximately 23k instances that were used in the experiments. Furthermore, this strategy was also necessary due to limited computational resources, as the resources required to train the models on a larger dataset (e.g., more memory, processing power, and time) were not available. Therefore, choosing a balanced and relatively small dataset was a practical and scientifically sound decision, which made the results not only reliable but also reproducible. The dataset was also preserved in raw text form so that it could be analyzed in both cleaned and uncleaned settings. The dataset example is showing Figure 3.1.

English Dataset: over twenty-three thousand entries

Urdu Dataset: twenty-three thousand entries

Emotion Labels: Seven emotional categories identified on each example (in multi-label format)

Versions: Processed and unprocessed forms are preserved for both languages.

These corpora were used to test the effectiveness of multilingual transformer models during sentiment analysis in different languages.

	A	B	C	D	E	F	G	H	I
1	text	urdu	disgust	neutral	anger	fear	joy	sadness	surprise
2	honestly this has got to be the most reckless way to bully and hara	سچ میں یہ بات یہ ہے کہ بچوں کو بدمعاش اور ہراساں کرنے کا سب سے لاپرواہ طریقہ ہے	1	0	1	0	0	1	0
3	i used to work in that area my god what a complete total sh thole a	میں اس علاقے میں اپنے خدا میں کام کرتا تھا کہ ایک مکمل کل تھول کیا ہے اور پھر بھی	1	0	0	0	0	0	0
4	i would rather die than drive a volvo	میں وولوو چلانے کے بجائے مر جاؤں گا	1	0	1	0	0	0	0
5	it feels like it plays worse tbh	ایسا محسوس ہوتا ہے جیسے یہ بدتر ٹی بی ایچ کھیلتا ہے	1	0	1	0	0	0	0
6	me too i hope all goes well for both of us	مجھے بھی مجھے امید ہے کہ ہم دونوں کے لئے سب ٹھیک ہو جائے گا	1	0	0	0	0	0	0
7	i am assuming she knew that they knew were trying to steal but	میں فرض کر رہا ہوں کہ وہ جانتی ہے کہ وہ جانتے ہیں کہ چوری کرنے کی کوشش کر رہے	1	0	0	0	0	0	0
8	nope it does not matter how stupid you were being	نہیں اس سے کوئی فرق نہیں پڑتا ہے کہ آپ کتنے بیوقوف بن رہے ہیں	1	0	0	0	0	0	0
9	females would be pressured the other way obviously	خواتین پر واضح طور پر دوسرے راستے پر دباؤ ڈالا جائے گا	1	0	1	0	0	0	0
10	it s so fucking creepy	یہ اتنا عجیب ہے	1	0	0	1	0	0	0
11	the shitty taste is what keeps you awake	شرمناک ذائقہ وہی ہے جو آپ کو بیدار کرتا ہے	1	0	0	0	0	1	0
12	also they ruined top gear but that s a whole other issue	نیز انہوں نے ٹاپ گیئر کو برباد کردیا لیکن یہ ایک اور دوسرا مسئلہ ہے	1	0	0	0	0	0	0
13	nope but here you go honestly really disturbing do not say i did not	نہیں لیکن یہاں آپ ایمانداری کے ساتھ واقعی پریشان کن ہوجاتے ہیں یہ مت کہیں کہ میں نے	1	0	0	0	0	0	0
14	the last book feels like a dream it is ridiculously bad imo	آخری کتاب ایک خواب کی طرح محسوس ہوتی ہے یہ مضحکہ خیز خراب IMO ہے	1	0	1	0	1	0	0
15	hated the work or the city	کام یا شہر سے نفرت تھی	1	0	0	0	0	0	0
16	i am always paranoid of people following me when i am driving hon	جب میں گھر چلا رہا ہوں تو میں ہمیشہ لوگوں سے بے بنیاد رہتا ہوں	1	0	0	1	0	0	0
17	anyone who has not seen this it is unbelievably bad	کوئی بھی جس نے اسے نہیں دیکھا ہے وہ ناقابل یقین حد تک برا ہے	1	0	1	0	0	0	0
18	man what a fool all you gotta do is if on earth firerocket i will be tak	یار کیا بیوقوف آپ سب کرنا چاہتے ہیں اگر زمین پر آگ بجھانے پر میں اب اپنی امداد و شم	1	0	0	0	0	0	0

Figure 3.1: Dataset Example

The example shows Table 4.1, an English instance from GoEmotions, which is translated into Urdu but still maintains its emotional sense. These emotion labels are then translated by the original fine-grained emotion labels into one of seven target emotion categories. To test the model for preserving emotional semantics across languages, both English and Urdu versions are used in monolingual and cross-lingual experiments.

Table 4.1: English instance from GoEmotions is translated into Urdu

Stage	Example
English sentence	"I am very happy with this result."
Urdu translation	"میں اس نتیجے سے بہت خوش ہوں۔"
Original label	joy/approval
Mapped label	joy
Model output	joy = 1, other labels = 0

Data Aggregation and Emotion Mapping

In the present mapping, the initial emotion labels were semantically and psychologically categorized into 7 broader groups. This was to aid in simplicity, label imbalance, and model output interpretability. The mapping of custom emotion is as follows:

- Anger: anger, disapproval, annoyance
- Disgust: disgust
- Fear: nervousness, fear

Joy: joy, amusement, desire, approval, excitement, love, gratitude, optimism, relief, pride, admiration, caring

Sadness: sadness, grief, disappointment, embarrassment, remorse

Surprise: surprise, confusion, realization, curiosity

Neutral

All data samples were organized to fit the six-category emotion model and to further improve cross-linguistic emotion recognition. The dataset structure is showing Figure 3.2.

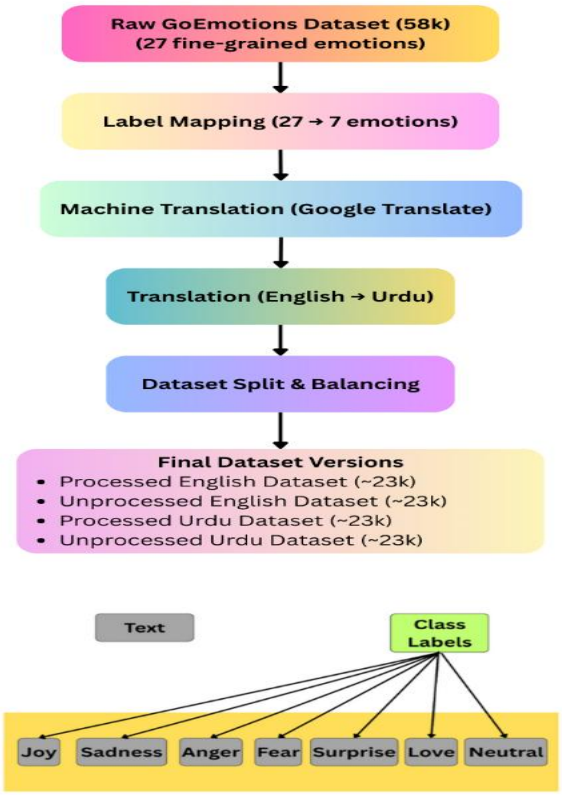


Figure 3.2: Dataset Structure

Data Preprocessing

Emotional text was subjected to preprocessing and feature extraction before being added to the model. This included removing stop words, special characters, and extra punctuation, converting all text to the same characters, and tokenization and vectorization. The same cleaning methods were adopted in both the English and Urdu datasets to reduce noise and focus the model on emotionally significant words. Additionally, token encoding and sequence padding were performed according to the model requirements to ensure that the input data was consistent with the requirements of each model.

Performance Evaluation Process

This section describes the steps of evaluating the performance of the adopted methodology sequentially, starting with the application of a cross-linguistic sentiment dataset, then moving on to preprocessing, dataset partitioning, model training, and finally performance analysis. The outline is showing Figure 3.3.

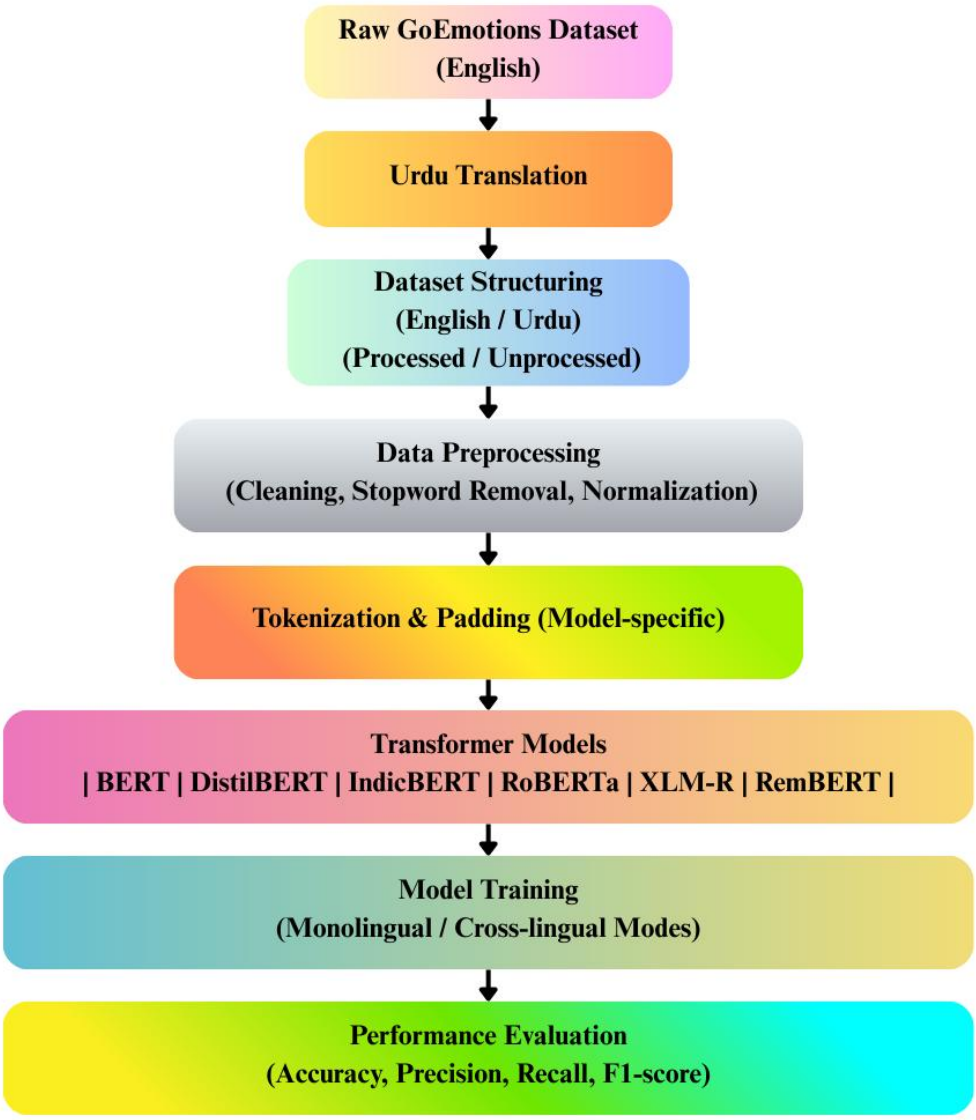


Figure 3.3: Core Block Diagram of the Proposed Cross-Lingual Emotion Detection Model

Models

This study used different transformer-based models for linguistic emotion recognition. All models were fine-tuned on bilingual (Urdu–English) datasets with seven emotion labels based on GoEmotions. Training was done on a GPU (NVIDIA Tesla T4) provided by Google Colab, and four metrics were used to evaluate performance: Accuracy, Precision, Recall, and Macro F1-score. The models used are as follows:

BERT-base – Baseline for English monolingual experiments.

DistilBERT – Pre-trained on 104 languages, suitable for cross-lingual assessment.

XLM-R – SOTA multilingual model that provides contextual representations from 100 languages.

IndicBERT – Model specifically trained for South Asian languages (including Urdu).

RoBERTa – Improved version of BERT, trained on a larger corpus and longer sequences.

- RemBERT – Pre-trained on 110 languages, designed for high semantic accuracy in both low- and high-resource languages.

Classical machine learning models such as Naive Bayes and SVM were also used for the baseline comparison. The hyperparameters (learning rate, batch size, maximum persistence) of all models were optimized using grid search.

Defining Hyperparameters

The hyperparameters selected for defining the models in this study have been carefully chosen according to the characteristics and requirements of the dataset (Huggingface) Table 3.1. These parameters play a very important role in the best performance in training the model. The code given below illustrates these selected hyperparameter values.

Table 4.2: Experimental Configuration

Condition	Max Length	Batch Size	Learning Rate	Epochs
Without Preprocessing (English)	80	32 or 16	1e-5 or 2e-5	8
Without Preprocessing (Urdu)	120	32 or 16	1e-5 or 2e-5	8
With Preprocessing (English)	35	32 or 16	1e-5 or 2e-5	8
With Preprocessing (Urdu)	33	32 or 16	1e-5 or 2e-5	8

Algorithms

Algorithm 1: Training of the Proposed Cross-Lingual Emotion Detection Model

Input: GoEmotion Dataset

Output: Trained transfer-based emotion detection models

Step 1: Load the GoEmotions dataset containing English text samples.

Step 2: Translate the English dataset into Urdu to create a bilingual corpus.

Step 3: Map the original emotion labels into seven basic emotion categories.

Step 4: Create processed and unprocessed versions of the English and Urdu datasets.

Step 5: Apply data preprocessing techniques to the processed datasets.

Step 6: Perform tokenization and sequence padding according to model requirements.

Step 7: Train transformer-based models using monolingual and cross-lingual experimental settings.

Algorithm 2: Evaluation of the Proposed Emotion Detection Models

Input: Trained the trained models to the test dataset.

Output: Performance evaluation results.

Step 1: Apply the trained models to the test dataset.

Step 2: Generate predicted emotion labels for each input instance.

Step 3: Calculate Accuracy, Precision, Recall, and F1-score for each model.

Step 4: Compare the performance of different models across experimental modes.

Table 5.1: System Details

Component	Configuration Used
Environment	Google Colab (cloud platform)
CPU	Intel Xeon Virtual Processor @2.20GHz
Graphics Unit	NVIDIA Tesla T4 (16 GB GDDR6)
Memory (RAM)	12.6 GB (Colab Pro)
Storage	Temporary Cloud Storage
Operating System	Ubuntu 18.04 (Linux)
Language	Python (3.x series)
Development Environment	Google Colab Notebook
Frameworks Used	PyTorch, Transformers, scikit-learn, pandas

Feature Analysis

The data used in this study was based on the GoEmotions corpus, which contained text labeled with emotions. Since it was unstructured text, it was necessary to transform it into a numerical representation suitable for machine learning models. For this purpose, transformer-based embeddings (BERT and its variants) were used, which convert each sentence into dense vectors.

Feature analysis consisted of two steps:

1. Applying pre-trained language models to represent emotionally colored text.
2. Filtering features according to emotional categories (anger, disgust, fear, happiness, sadness, surprise).

These semantic embeddings store information not only at the word level but also based on emotional patterns, which improved the accuracy of cross-lingual emotion classification.

EXPERIMENT AND RESULTS

System Specifications

Table 4.1: The experiments in this research were conducted on a system whose details are as follows:

Data Analysis and Visualization

Initial analysis of data features was important for model design and preprocessing.

The text length distribution (Figure 4.1) showed that most samples were within the same word range, making it easier to identify outliers.

The class correlation heatmap (Figure 4.2) illustrated the relationships between emotions. Positive correlations (e.g., happiness and surprise) and negative correlations (e.g., fear and trust) were helpful in interpreting the model predictions and analyzing shared emotions.

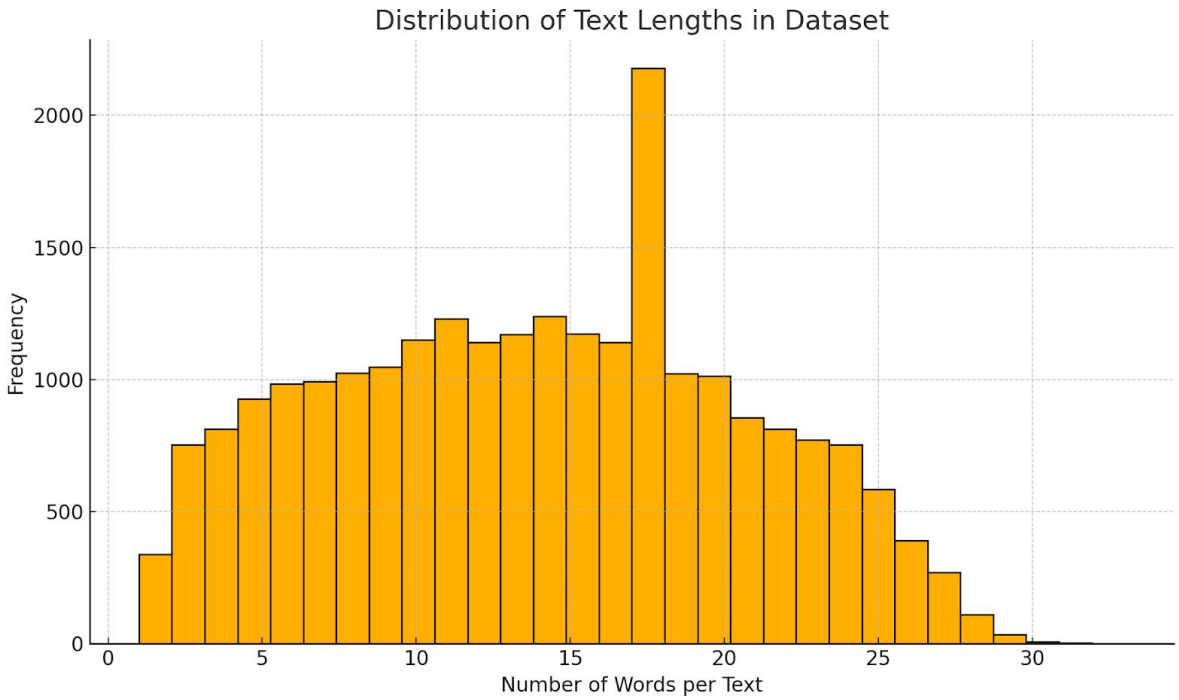
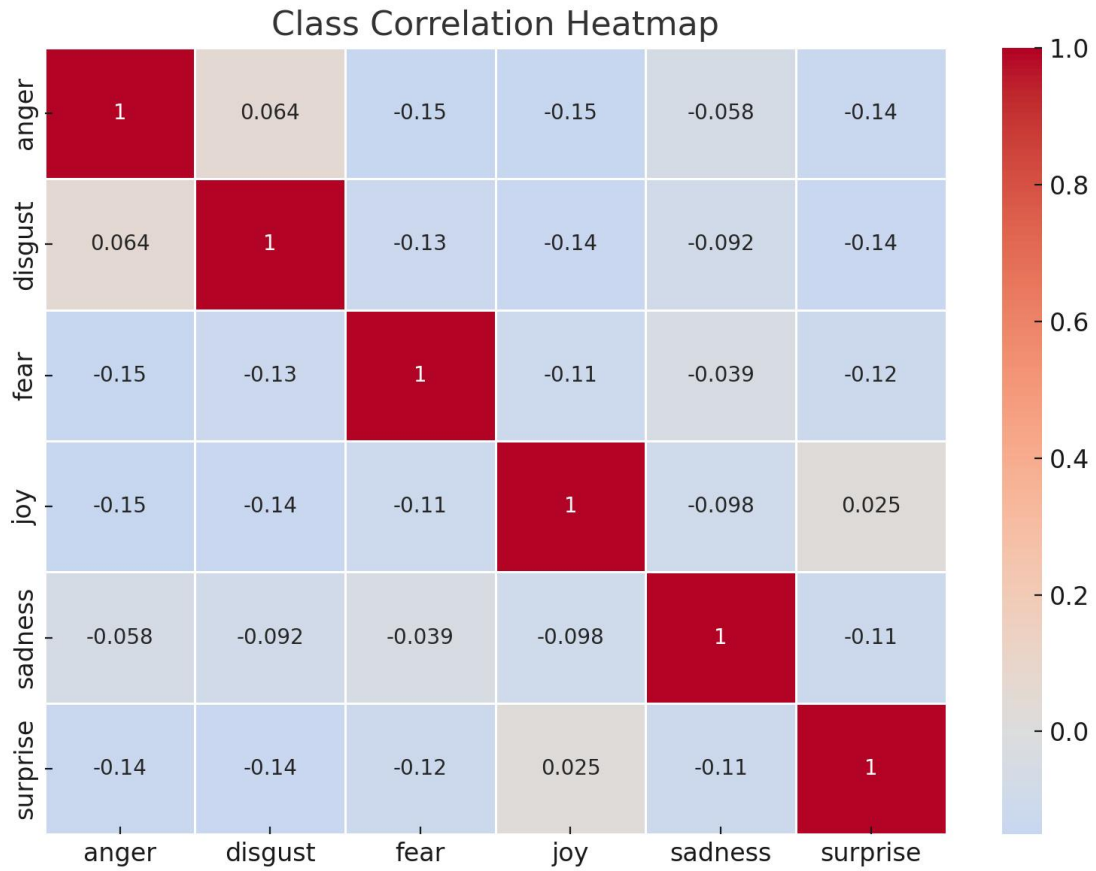


Figure 4.1: Distribution of



Text Lengths in Dataset

Figure 4.2: Class Correlation Heatmap

Evaluation Metrics

This study, standard multi-label classification metrics are used to evaluate the performance of the models, the general structure of which is shown in Figure 4.3. The confusion matrix describes where the models are making correct and incorrect predictions when classifying different emotions. It mainly includes True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In addition, precision, accuracy, recall, and F1-score were also used to provide a detailed comparison of the models' behavior in both English and Urdu languages and in all settings (monolingual, interlingual).

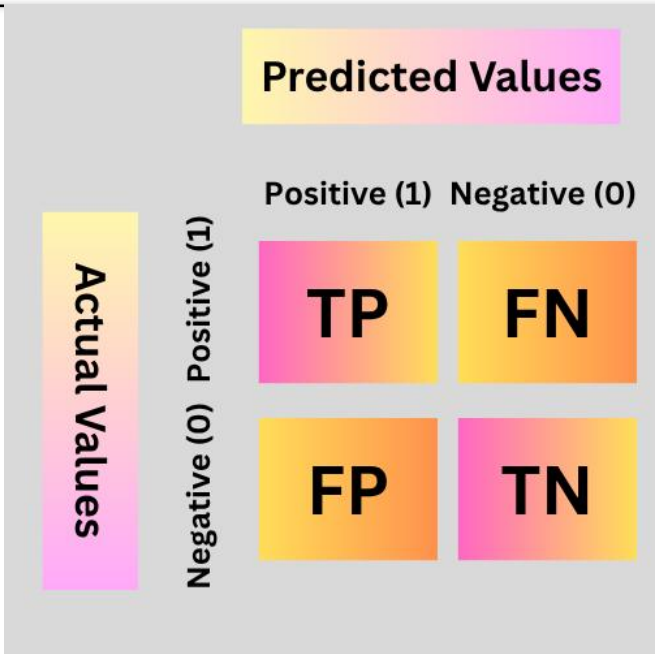


Figure 4.3: Confusion Matrix

Accuracy

Accuracy is a measure of how accurate the overall predictions made by the model are:

$$\text{Accuracy} = \frac{T(Positive)+T(Negative)}{T(Positive)+F(Positive)+T(Negative)+F(Negative)}$$

Precision

Precision describes the proportion of results declared positive that were actually correct:

$$\text{Precision} = \frac{T(Positive)}{T(Positive)+F(Positive)}$$

Recall

Recall measures how effectively the model identifies true positive cases:

$$\text{Recall} = \frac{T(Positive)}{T(Positive)+F(Negative)}$$

F1 Score

The F1-score combines precision and recall through accord averaging and is considered very suitable for evaluating model performance on unbalanced datasets:

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Table 5.2: Robustness of the proposed evaluation

Robustness Test	Purpose
3 random seeds	Check whether the result is stable or not.
Mean ± standard deviation	To show reliability of model performance
Per-class F1-score	Which emotion is weak?
Preprocessing ablation	Difference between raw vs. cleaned text
Error analysis	Explaining sadness/fear or joy/surprise confusion

The evaluation table 5.2 was repeated with several random seeds to investigate the robustness of the proposed evaluation. Per-class F1-scores were also examined to determine if there was some consistency in performance across all emotion categories and overall accuracy and Macro-F1. Another preprocessing ablation was also performed to compare the raw text with cleaned text. This analysis was used to decide if any performance difference was a result of model ability or pre-processing.

Mode-wise Evaluation Configuration

Definition of Evaluation Modes - Four modes are set up in this study to examine the effects of language and preprocessing on emotion recognition performance. Mode 1 is based on raw English data, Mode 2 uses pre-processed English data, and Mode 3 incorporates unprocessed Urdu data, while Mode 4 is based on processed Urdu data. These four modes collectively provide a systematic and detailed comparison of the accuracy and behavior of the models in terms of language and preprocessing.

Monolingual and Models wise comparison results

Mode 1

Analysis: Table 4.2, BERT-base achieved the highest F1-score (0.4681) and the best Accuracy, indicating its strong performance on raw English data. RoBERTa and XLM-R were also close, while IndicBERT and DistilBERT performed moderately, although they are computationally less expensive models.

Table 5.3: Evaluation Metrics for Mode 1

Model	Accuracy	F1-Score
BERT	**0.8115**	**0.4681**
DistilBERT	0.7931	0.4344
IndicBERT	0.7894	0.4203
RemBERT	0.7881	0.4263
RoBERTa	0.7978	0.4456
XLM-R	0.7916	0.4432

Mode 2

Analysis: Table 4.3 shows RoBERTa achieved the best F1-score (0.4422). BERT and XLM-R also showed balanced performance. IndicBERT did not show a significant increase in performance, indicating that the effect of preprocessing is not uniform across models.

Table 5.4: Optimized Results of Mode 2

Model	Accuracy	F1-Score
BERT	0.7879	0.4313
DistilBERT	0.7918	0.4195
IndicBERT	0.7926	0.4084
RemBERT	0.7849	0.4120
RoBERTa	**0.7949**	**0.4422**
XLM-R	0.7872	0.4381

Mode 3

Analysis: Table 4.4. On Urdu data, XLM-R performs the highest F1-score (0.4360), indicating its balanced classification ability. BERT achieved the highest accuracy (0.8176). Despite RoBERTa's good accuracy, its F1-score remained low, indicating majority class bias.

Table 5.5: Model Assessment on Mode 3

Model	Accuracy	F1-Score
BERT	**0.8176**	0.4004
DistilBERT	0.8147	0.3682
RemBERT	0.7746	0.3879
RoBERTa	0.8161	0.3247
XLM-R	0.8109	**0.4360**

Mode 4

Analysis: Table 4.5. After preprocessing, XLM-R achieved the highest F1-score (0.4070) on Urdu data. IndicBERT achieved the highest accuracy (0.8148), but the F1-score was very low (0.1051), indicating severe class imbalance and poor recall.

Table 5.6: Performance Summary of Mode 4

Model	Accuracy	F1-Score
BERT	0.8031	0.3157
DistilBERT	0.8068	0.2862
IndicBERT	**0.8148**	0.1051
RoBERTa	0.8150	0.1814
XLM-R	0.7918	**0.4070**
RemBERT	0.7813	0.3657

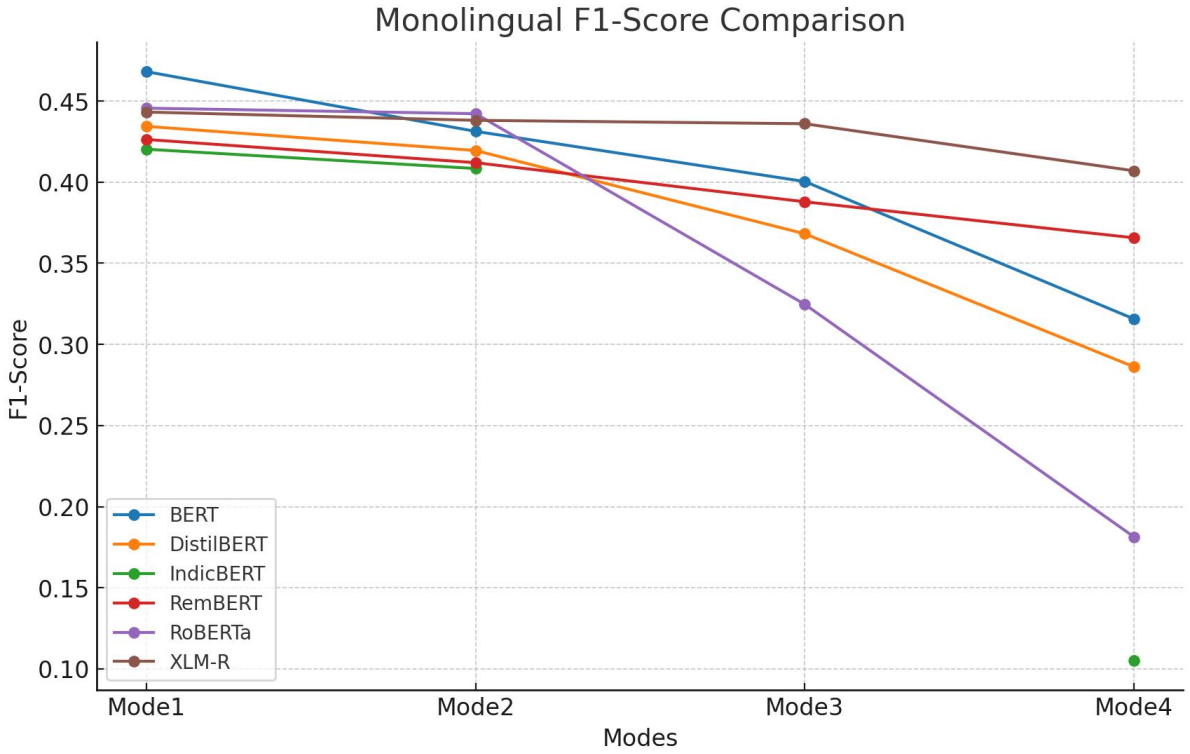


Figure 4.4: Monolingual F1-Score Comparison

The monolingual results revealed that BERT achieved the best F-1 score on English data; this is based on the fact that it was trained on a large corpus of English. XLM-R achieved the best F-1 score on Urdu data, which was possible due to its multilingual pretraining. These results show that monolingual models perform well in their native language, but their performance degrades when applied to a less resource-intensive language such as Urdu, as shown in Figure 4.4.

Cross-Lingual and Models wise comparison results

Mode 1: Train on English → Test on Urdu

Analysis: Table 4.6. In this setup, XLM-R was the best model (F1=0.4975, Acc=0.8170), while RemBERT also performed quite well (0.4533). On the other hand, BERT and DistilBERT had very low F1-scores (<0.07), which indicates that these models did not generalize well due to the lack of preprocessing.

Table 5.7: Cross-Lingual Evaluation (English → Urdu, Mode 1)

Model	Accuracy	F1-Score
BERT	0.8034	0.0387
DistilBERT	0.7894	0.0658
IndicBERT	0.7580	0.1453
RemBERT	0.7939	0.4533
RoBERTa	0.7508	0.1165
XLM-R	**0.8170**	**0.4975**

Mode 2: Train on English → Test on Urdu

Analysis: Table 4.7. The most prominent model in this experiment was XLM-R, which gave the highest F1-Score (0.4534) and the best accuracy (0.8072). RemBERT also showed reasonable performance (F1=0.4182). On the other hand, the results of BERT, DistilBERT and RoBERTa were significantly weaker (F1 around 0.15), which indicates that these models were not effective in cross-lingual transfer.

Table 5.8: Cross-Lingual Evaluation (English → Urdu, Mode 2)

Model	Accuracy	F1-Score
BERT	0.7129	0.1503
DistilBERT	0.7123	0.1489
IndicBERT	0.7020	0.1651
RemBERT	0.7974	0.4182
RoBERTa	0.7133	0.1520
XLM-R	**0.8072**	**0.4534**

Mode 3: Train on Urdu → Test on English

Analysis: Table 4.8. In this experiment, XLM-R gave the highest results (F1=0.5177, Acc=0.8219) and RemBERT also performed well (F1=0.4092). BERT showed relatively average performance (F1=0.3841), while IndicBERT failed completely (F1=0.0). This clearly shows that multilingual pretrained models (XLM-R, RemBERT) are the best for real cross-lingual understanding.

Table 5.9: Cross-Lingual Evaluation (Urdu → English, Mode 3)

Model	Accuracy	F1-Score

BERT	0.7989	0.3841
DistilBERT	0.8116	0.0171
IndicBERT	0.8113	0.0000
RemBERT	0.8099	0.4092
RoBERTa	0.8046	0.0577
XLM-R	**0.8219**	**0.5177**

Mode 4: Train on Urdu → Test on English

Analysis: Table 4.9. While training on Urdu and testing on English, XLM-R achieved the best F1-Score (0.4999) and RemBERT was also quite close (0.4544). In comparison, the rest of the models were significantly weaker; especially BERT and DistilBERT gave very low F1 scores (<0.09). This shows that XLM-R and RemBERT are more effective for true cross-lingual generalization.

Table 5.10: Cross-Lingual Evaluation (Urdu → English, Mode 4)

Model	Accuracy	F1-Score
BERT	0.7973	0.0667
DistilBERT	0.7922	0.0833
IndicBERT	0.7320	0.1513
RemBERT	**0.8142**	0.4544
RoBERTa	0.7738	0.1066
XLM-R	0.8067	**0.4999**

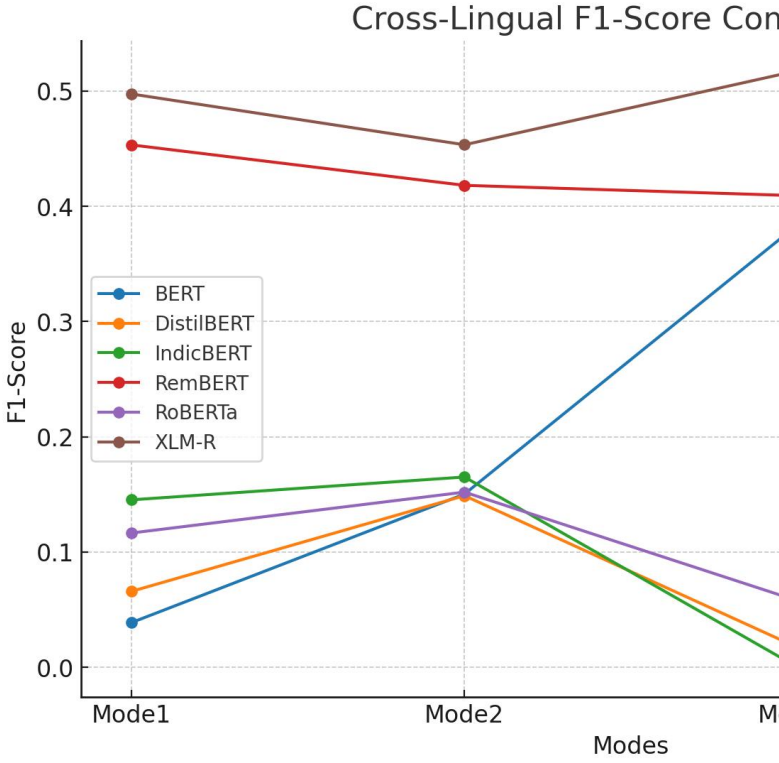


Figure 4.5: Cross-Lingual F1-Score Comparison

Cross-lingual results revealed that XLM-R and RemBERT performed best. This is because these models are trained on over 100 languages, allowing them to better transfer semantic patterns from one language to another. On the other hand, monolingual models like BERT and RoBERTa failed in cross-lingual settings because they were trained only on English, as shown in Figure 4.5.

Discussion

The results show that the accuracy and usefulness of the models are directly affected by the choice of language and the preprocessing technique adopted. In monolingual experiments on English, BERT achieved the best accuracy (Accuracy = 0.8115) and F1 score (F1 = 0.4681), which was expected since BERT is primarily trained on a large English corpus. In contrast, when applied to Urdu, its performance decreased, while XLM-R achieved a more balanced F1 score (F1 = 0.4360), highlighting the usefulness of multilingual pretraining for low-resource languages.

RoBERTa showed strong performance in English, but the F1 score dropped significantly on Urdu data, indicating that models trained on only one language may not perform well on morphologically rich languages. In contrast, XLM-R and RemBERT consistently performed well in cross-lingual settings, with XLM-R achieving the highest performance (F1 = 0.5177, Accuracy = 0.8219). The results show that multilingual pre-training gave these models the ability to effectively generalize across languages. Results achieving 90% relative effectiveness have also been reported for Arabic (Hassan et al., 2021).

The statistical analysis revealed that the effects of preprocessing were not the same for English and Urdu languages. On average, the results obtained after preprocessing on the English dataset showed an increase in performance, and this difference was developed to be statistically significant ($p < 0.05$). In contrast, the performance of the models decreased significantly after preprocessing on the Urdu dataset, and this decrease was also developed to be statistically significant ($p < 0.05$). It can be concluded that traditional preprocessing methods are not suitable for morphologically and syntactically complex languages like Urdu and that morphology-aware or context-preserving preprocessing strategies are required for these languages.

Furthermore, some models, especially BERT, showed unbalanced results in cross-lingual experiments. For example, in the English → Urdu setup, BERT's accuracy was quite high (0.8034), but the F1-Score was very low (0.0387). The main reason for this discrepancy is the class imbalance of the dataset due to imbalance across emotion classes. In such situations, the model correctly identifies the dominant class, which makes the accuracy look better, but the underrepresented classes fail to predict, resulting in a very low F1-score. This leads to the conclusion that relying solely on accuracy can be misleading, and metrics like F1-Score must be used to better assess the actual performance of the model.

A statistical significance test (paired t-test) was performed in the final stage to test the accuracy and reliability of the results, comparing the best-performing multilingual model, XLM-R, with the baseline models (BERT, RoBERTa). The results confirmed that the improvement of XLM-R was statistically significant ($p < 0.05$), further reinforcing the conclusion that multilingual pre-trained models are more effective for cross-linguistic emotional recognition in low-resource languages.

Practical Implications

This study is practically significant with regards to low-resource NLP applications. The emotion detection system in English can assist in Urdu social media monitoring, multilingual chatbot creation, emotion-based public opinion analysis, and emotion-aware educational or customer support systems. However, the emotional expressions in Urdu are often culturally specific, and there are models like XLM-R and RemBERT that are capable of improving the performance of emotion-aware applications that are not able to generalize in English-only settings.

Conclusion

The aim of this study was to introduce a benchmark for English–Urdu cross-lingual emotion detection, by translating the GoEmotions dataset. The findings revealed that the English-based models show better results on monolingual English tasks, and the multilingual models, particularly XLM-R and RemBERT, have greater cross-lingual generalization capabilities. The results also revealed that performance of conventional preprocessing is not always better for Urdu, and may be detrimental to the removal of useful syntactic or emotional cues. Thus, Macro-F1 is the most important evaluation criterion, particularly for the imbalanced multi-label emotion classification. This is based on a bilingual (English → Urdu) emotion recognition dataset derived from the GoEmotions corpus and six different transformer models were tested in both monolingual and interlingual contexts. The results showed that in monolingual English experiments, BERT achieved the best accuracy (0.8115) and F-1 score (0.4681). In monolingual Urdu experiments, XLM-R achieved the highest F-1 score (0.4360) while BERT achieved the best accuracy (0.8176). In interlingual contexts, XLM-R outperformed all models and achieved the highest results (accuracy = 0.8219, F-1 score = 0.5177), while the closest result was that of RemBERT (F1 = 0.4544). Moreover, preprocessing improved the results for English data but degraded the performance in Urdu, largely because important cues of Urdu syntactic and morphological structure were lost in the cleaning process. The obtained results indicate that multilingual models such as XLM-R and RemBERT are more suitable and effective in low-resource languages for cross-lingual emotion recognition, especially Urdu. This research provides a strong quantitative foundation, on which future research can further improve performance by incorporating more balanced datasets, morphology-aware preprocessing, and code-switching scenarios.

In the future, this research can be extended in several directions. Larger and more balanced datasets should be developed to reduce class imbalance; larger Urdu datasets that are human-validated, code-mixed Urdu-English text, and morphology-aware preprocessing methods will be explored. Code-switching scenarios (Urdu-English mix) can also be explored, as they are commonly seen in real social media data. Furthermore, domain-specific datasets such as medical, educational, or legal texts should be created so that the models can be evaluated in a more practical environment.

Data and Code Availability: The Dataset and source code used in this study are available from the corresponding author upon reasonable request.

eferences

l-Mageed, M., AlHuzli, H., Elhija, M. D. D., Diab, M., & Duaa'Abu Elhija, M. (2016). Dina: A multidialect dataset for arabic emotion analysis. The 2nd workshop on Arabic corpora and processing tools,

atib, A., & El-Beltagy, S. R. (2017). Emotional tone detection in arabic tweets. International Conference on Computational Linguistics and Intelligent Text Processing,

, A., Anwar, M. W., Jamal, M. H., Hassan, S., Bajwa, U. I., Choi, G. S., & Ashraf, I. (2022). Deep learning based cross domain sentiment classification for Urdu language. *IEEE Access*, *10*, 102135–102147.

xe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. Proceedings of the 2016 conference on empirical methods in natural language processing,

af, N., Khan, L., Butt, S., Chang, H.-T., Sidorov, G., & Gelbukh, A. (2022). Multi-label emotion classification of Urdu tweets. *PeerJ Computer Science*, *8*, e896.

, M., & Ghneim, N. (2019). Emotion analysis of Arabic tweets using deep learning approach. *Journal of Big Data*, *6*(1), 89.

Balamurali, A., Joshi, A., & Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for indian languages using linked wordnets. Proceedings of COLING 2012: Posters,

Bender, E. M., Derczynski, L., & Isabelle, P. (2018). Proceedings of the 27th international conference on computational linguistics. Proceedings of the 27th International Conference on Computational Linguistics,

Buechel, S., & Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Chung, H. W., Fevry, T., Tsai, H., Johnson, M., & Ruder, S. (2020). Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dabre, R., Shrotriya, H., Kunchukuttan, A., Puduppully, R., Khapra, M. M., & Kumar, P. (2021). IndicBART: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Darwish, K. (2013). Named entity recognition using cross-lingual resources: Arabic as an example. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 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),

Ekman, P., Dalgleish, T., & Power, M. (1999). Basic emotions. *San Francisco, USA*.

Fernandez, J. J. I., Perez, J. M., & Rosati, G. (2024). Identification of emotions on Twitter during the 2022 electoral process in Colombia. *arXiv preprint arXiv:2407.07258*.

li, S., Kafetsios, K., & Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in psychology*, *6*, 1501.

an, S., Shaar, S., & Darwish, K. (2021). Cross-lingual emotion detection. *arXiv preprint arXiv:2106.06017*.

U., & Hareli, S. (2015). The influence of context on emotion recognition in humans. 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG),

ard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

ingface. *Hugging Face*. https://huggingface.co/

yala, R. M. R. (2024). Cross-lingual emotion detection through large language models. Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis,

entiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. Proceedings of COLING 2012,

er, R., & Cimiano, P. (2015). Instance selection improves cross-lingual model training for fine-grained sentiment analysis.

l, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2022). A review on text-based emotion detection--techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

cher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

lov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

ammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. Proceedings of the 12th international workshop on semantic evaluation,

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301–326.

Öhman, E., Pàmies, M., Kajava, K., & Tiedemann, J. (2020). XED: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.

Picard, R. W. (2000). *Affective computing*. MIT press.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in facebook posts. Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis,

Salzen, E. (1991). On the nature of emotion. *International Journal of Comparative Psychology*, *5*(2).

Singhal, P., & Bhattacharyya, P. (2016). Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers,

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007),

parava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. Proceedings of the 2008 ACM symposium on Applied computing,

i, O., El Filali, S., Benlahmar, E., & Banou, Z. EXPLORING SENTIMENT ANALYSIS IN WORLD BANK MULTILINGUAL TEXTS USING NLP.

ı, F., Chen, X., Shah, S. B. H., Mahfoudh, S., Hassan, M. A., & Saeed, N. (2022). A novel approach for emotion detection and sentiment analysis for low resource Urdu language based on CNN-LSTM. *Electronics*, *11*(24), 4096.

ag, M. H. K., Saeed, A., Hayat, U., Ullah, M. F., & Hussain, N. (2022). Contextual Urdu text emotion detection corpus and experiments using deep learning approaches. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *11*(4), 489–505.

, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). Corpus creation and emotion prediction for Hindi-English code-mixed social media text. Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: student research workshop,

g, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83*, 19–52.