Multilingual BERT-Based Emotion Recognition: A Comparative Study in Urdu and Dari Texts

1st Malika Muradi
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
malika.muradi@g.bracu.ac.bd

2nd Basit Hussain
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
basit.hussain@g.bracu.ac.bd

3rd Ehsanur Rahman Rhythm
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
ehsanur.rahman.rhythm@g.bracu.ac.bd

4th Mehnaz Ara Fazal
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
mehnaz.ara.fazal@g.bracu.ac.bd

5th Annajiat Alim Rasel
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
annajiat@gmail.com

Abstract—This study explores the application of a multilingual approach to emotion detection using BERT (Bidirectional Encoder Representations from Transformers) in Urdu and Dari languages. Emotion detection is a complex task, especially when dealing with diverse linguistic contexts. Leveraging the power of BERT modal, we conducted analysis of emotion classification across multiple languages through the BERT model. The study involves the utilization of BERT to capture intricate contextual information from users' tweets, allowing for a nuanced understanding of emotions expressed in text. Our approach aims to enhance the accuracy and versatility of emotion detection systems by incorporating the multilingual capabilities of BERT. We present experimental results that demonstrate the effectiveness of our model, showcasing improved performance and adaptability in the challenging task of multilingual emotion detection. This research contributes to the broader field of natural language processing and emotion analysis by providing insights into the benefits of BERT in a multilingual context.

Index Terms—Emotion Detection, Natural language processing, Multilingual BERT model

I. Introduction

Emotion detection is a crucial field in Natural Language Processing (NLP). It involves determining the emotional tone or sentiment expressed in text data.It is valuable in various applications, including social media analysis, product reviews, customer feedback analysis and so on which understanding the emotional tone of text can help businesses make data-driven decisions. The field of emotion detection has witnessed significant improvement in recent years, with the advancement of NLP techniques in so many languages, specifically English. Though, we witnessed the good improvement of other languages including Dari and Urdu in NLP too where the researchers have built and worked on word embedding, sentiment analysis and NER model which are quite good improvements. However we won't find research on emotion detection in both languages using the Multilingual Bert model which can be designed to handle multiple languages within a single modal. Dari and Urdu are two languages that have their own scripts and nuances, and our goal is to adapt our emotion detection system to work effectively in these specific linguistic contexts. This includes considerations for script differences, cultural nuances, and local sentiment expressions to ensure accurate and culturally sensitive emotion analysis for Urdu and Dari.

II. Literature Review

This section provides a thorough description of the most widely used approaches for emotion recognition and sentiment analysis in Dari and Urdu text. These approaches cover a wide range of applications, including market analysis, personality analysis, healthcare, human-computer interaction and many more [1].

In the study [2], the researcher's emphasis is on the crucial role of emotion detection, particularly in addressing the research gap for low-resource languages such as Urdu. The paper introduces the Urdu Nastalique Emotions Dataset (UNED) and presents a deep learning-based technique for emotion classification. Through an evaluation of the UNED corpus, the research showcases the superior performance of the deep learning model compared to generic machine learning approaches. With an impressive F1 score of 85% on sentence-based and 50% on paragraph-based corpora, the paper underscores the significance of UNED and its potential applications in emotion classification for Urdu text.

In another study [3], the researchers demonstrate the challenge of emotion identification in Roman Urdu (RU), a resource-poor oriental language prevalent in Asia. The study introduces a varied RU text dataset annotated with six basic emotions following Paul Ekman's theory. Utilizing dense word embeddings, specifically BERT, the proposed method surpasses baseline algorithms and existing state-of-the-art techniques, achieving an average accuracy of 91% in empirical assessments.

In the other study [4], the researcher examined two different methods to analyze the emotion in Dari text. The first approach, FastText and the Bi-LSTM methods to classify the emotion and the second approach,a deep bidirectional transformer to extract features from text. resulting in notable improvement of multilingual classification.

In the other study [5], the researchers introduce the Urdu sentiment Analysis Bidirectional Encoder Representation from Transformers (USA-BERT) approach for sentiment analysis in Urdu, a low-resource language. Leveraging the UDSA-23 dataset and BERT embedding, the study demonstrates enhanced sentiment classification through re-sampling techniques and deep learning classifiers.

In on other study [6], the researchers introduce two new human-labeled emotion datasets for the Dari language, EmoPars and ArmanEmo. The study evaluate these dataset and addressing the imbalance problem using data augmentation, resampling, and class-weights with Transformer-based Pretrained Language Models (PLMs).

Additionally, feature selection enhance the model performance by finding the specific text features.The proposed model achieves a F1-score of 0.81 and 0.76 on ArmanEmo and EmoPars, establishing new state-of-the-art results in these benchmarks.
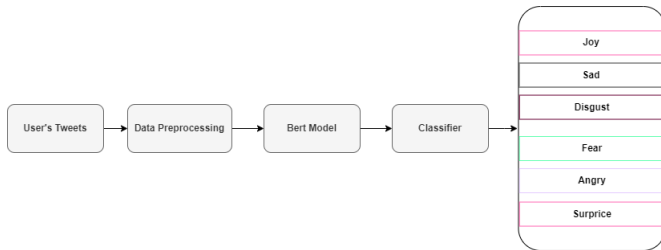
III. Methodology

Fig. 1. Flow chart of Methodology

References

[1] S. Kaur and Dr. N. Kulkarni, "Emotion Recognition - A Review," SSRN Electronic Journal, 2020, doi: https://doi.org/10.2139/ssrn.3647958

[2] M. F. Bashir, A. R. Javed, M. U. Arshad, T. R. Gadekallu, W. Shahzad, and M. O. Beg, "Context Aware Emotion Detection from Low Resource Urdu Language using Deep Neural Network," ACM Transactions on Asian and Low-Resource Language Information Processing, Apr. 2022, doi: https://doi.org/10.1145/3528576

[3] N. Ali et al., "Towards Enhanced Identification of Emotion from Resource-Constrained Language through a novel Multilingual BERT Approach," ACM Transactions on Asian and Low-Resource Language Information Processing, Apr. 2023, doi: https://doi.org/10.1145/3592794

[4] Azadeh Khodaei, Azam Bastanfard, Hadi Saboohi et al. Deep Emotion Detection Sentiment Analysis of Persian Literary Text, 05 July 2022, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1796157/v1]

[5] "BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language — IEEE Journals & Magazine — IEEE Xplore," ieeexplore.ieee.org https://ieeexplore.ieee.org/abstract/document/10271297

[6] Abaskohi, A., Sabri, N., Bahrak, B. (2022). Persian Emotion Detection using ParsBERT and Imbalanced Data Handling Approaches. arXiv preprint arXiv:2211.08029.