IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 September 2023, accepted 30 September 2023, date of publication 4 October 2023,
date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322101

RESEARCH ARTICLE

BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language

MUHAMMAD REHAN ASHRAF[1,2], YASMEEN JANA[2], QASIM UMER[2,3], M. ARFAN JAFFAR[1], SUNGWOOK CHUNG[4], AND WAHEED YOUSUF RAMAY[5]

[1]Department of Computer Science, Superior University, Lahore 54000, Pakistan
[2]Department of Computer Sciences, COMSATS University Islamabad, Vehari 61000, Pakistan
[3]Department of Computer Science, Hanyang University, Seoul 04763, South Korea
[4]Department of Computer Engineering, Changwon National University, Changwon 51140, South Korea
[5]Department of Computer Science, Air University, Multan 60000, Pakistan

Corresponding author: Qasim Umer (qasimumer@hanyang.ac.kr)

This work was supported by the Korea Meteorological Administration Research and Development Program under Grant KMI 2021-01310.

ABSTRACT Sentiment analysis holds significant importance in research projects by providing valuable insights into public opinions. However, the majority of sentiment analysis studies focus on the English language, leaving a gap in research for other low-resourced languages or regional languages, e.g., Persian, Pashto, and Urdu. Moreover, computational linguists face the challenge of developing lexical resources for these languages. In light of this, this paper presents a deep learning-based approach for Urdu Text Sentiment Analysis (USA-BERT), leveraging Bidirectional Encoder Representations from Transformers and introduces an *Urdu Dataset for Sentiment Analysis-23* (UDSA-23). USA-BERT first preprocesses the Urdu reviews by exploiting *BERT-Tokenizer*. Second, it creates BERT embeddings for each Urdu review. Third, given the BERT embeddings, it fine-tunes a deep learning classifier (BERT). Finally, it employs the Pareto principle on two datasets (the state-of-the-art (UCSA-21) and UDSA-23) to assess USA-BERT. The assessment results demonstrate that USA-BERT significantly surpasses the existing methods by improving the accuracy and f-measure up to 26.09% and 25.87%, respectively.

INDEX TERMS Natural language processing, Urdu, BERT, classification, sentiment analysis.

I. INTRODUCTION

In today's digital age, where blogs, forums, and Facebook have become the go-to platforms for communication, the Internet plays a pivotal role in global connectivity, online commerce, education, and sharing personal experiences. Consequently, extracting valuable insights from textual data, i.e., people's opinions, emotions, and attitudes, becomes crucial. Sentiment analysis allows us to analyze the sentiment expressed in various text forms, including customer reviews, social media posts, and news articles. This analysis helps us understand how individuals perceive products, services, events, and public figures. Such understanding is paramount for businesses, enabling them to assess customer satisfaction, make data-driven decisions, and enhance their offerings. Moreover, sentiment analysis has far-reaching implications in areas, i.e., brand management, reputation monitoring, market research, and tracking public sentiment. It also plays a significant role in social and political analysis by providing insights into public opinion, sentiment trends, and potential outcomes. Overall, sentiment analysis empowers governments, organizations and individuals to harness the power of language data, leading to improved decision-making, enhanced customer experiences, and a deeper understanding of human behavior [1], [2]. Sentiment analysis can be conducted in both monolingual and multilingual settings, and

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

it can be applied at different levels of analysis, including document-level, sentence-level, and aspect-level [3], [4], [5].

Recent technological advancements have opened up communication channels among people from different regions and countries, each with its own unique social values, cultures, and languages. Consequently, international companies receive product reviews in multiple languages, presenting a challenge for sentiment analysis. While sentiment analysis has garnered significant attention from researchers, particularly in English, other languages, especially regional ones, have been neglected or received less focus due to limited resources for conducting sentiment analysis in those languages [6], [7]. The key resources for developing a sentiment analysis, i.e., corpora, tokenizers, POS taggers, and lemmatizers, are often lacking for low-resourced languages. Additionally, cultural differences and linguistic complexity complicate sentiment analysis for such languages. Urdu, the language spoken by approximately 230 million people in Pakistan and comprised of around 100 million words [8], faces similar challenges. Urdu's complex character set, intricate grammatical structure, ambiguous word boundaries, and difficult stemming make it challenging to comprehend. Moreover, Roman Urdu (R-Urdu), which combines the English alphabet with Urdu script, adds an extra layer of complexity to the sentiment analysis of Urdu text.

Several studies have explored various techniques and approaches to analyze the sentiment of Urdu text effectively. Researchers have investigated using machine learning algorithms, i.e., Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF) for sentiment classification tasks [9], [10], [11], [12]. Feature engineering techniques, including n-grams, lexical features, and sentiment lexicons, have been employed to capture the linguistic characteristics of Urdu text. Additionally, researchers have also explored deep learning models [13], [14], i.e., Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs), to capture the contextual information and semantic relationships in Urdu text. Transfer learning approaches [15], i.e., pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) and Embedding from Language Model (ELMo), have also been utilized to improve the performance of sentiment classification. In conclusion, the literature on Urdu text sentiment classification showcases exploring various techniques and models to accurately analyze and classify sentiments in Urdu text. Despite being an active area of research, Urdu text sentiment classification is still in its early stages and requires substantial performance improvements.

This paper introduces a deep learning-based approach (USA-BERT) for Urdu Text Sentiment Analysis, leveraging Bidirectional Encoder Representations from Transformers. It also introduces a dataset called *Urdu Dataset for Sentiment Analysis-23* (UDSA-23). The proposed method preprocesses Urdu reviews using BERT-Tokenizer and generates implicit BERT embeddings for each review using *BERT-base-uncased* BERT tokenizer. A deep learning classifier (BERT) is then fine-tuned using these embeddings. The performance of USA-BERT is assessed using the Pareto principle on two datasets: the state-of-the-art UCSA-21 and UDSA-23. The assessment results demonstrate that USA-BERT significantly surpasses existing methods by improving accuracy and f-measure up to 26.09% and 25.87%, respectively.

The contributions of this paper can be summarized as follows:

- An Urdu dataset (UDSA-23) for sentiment analysis is introduced. The UDSA-23 is created for sentiment analysis for Urdu. It can serve as a foundation for future research by providing a standardized reference point for evaluating sentiment analysis models in the context of the Urdu language.
- An implicit BERT embeddings are used to propose a deep learning model (USA-BERT) for multi-class sentiment analysis of Urdu text.
- The proposed method (USA-BERT) significantly surpasses existing methods by improving the accuracy and f-measure up to 26.09% and 25.87%, respectively.

The subsequent sections of the paper are organized as follows: Section IV provides an explanation of USA-BERT. Section V outlines the evaluation methodology, including a performance comparison with baseline approaches and a discussion of potential threats. Section II presents an overview of the related work in the field. Finally, Section VI concludes the paper.

II. RELATED WORK

This section provides a comprehensive overview of the available and popular methods used for Urdu Sentiment Analysis (SA), which holds significant importance in digital marketing and customer reviews [16]. While SA is often perceived as a straightforward task, Erik and Soujanya [17] argue that it is, in fact, a complex problem. Huifeng et al. [18] highlight some crucial applications of sentiment detection, including product comparison, opinion summarization, and opinion reason mining. Additionally, they mention other valuable tasks facilitated by SA, i.e., political discussion group posting, message sentiment filtering, email sentiment classification, attitude analysis, and SA with search engines. Traditionally, textual information in SA has been represented using one-hot vectors, which suffer from high dimensionality and poor correlation. Bengio et al. [19] introduced a technique that replaces the one-hot vector with a low-dimensional distributed representation, which has become a standard approach to address these challenges. Several pre-trained word embedding techniques, i.e., Word2Vec [20] and Glove [21], are commonly employed to capture syntactic and semantic information from text.

A. TRADITIONAL URDU/ROMAN-URDU (R-URDU) DATASETS

A machine-readable gold-standard dataset is a basic requirement for all SA-related applications. The studies describe

TABLE 1. State-of-the-art datasets.

Dataset	Description	Available
Urdu Tweet [22]	10,000 Tweets	Yes
R-Urdu [23]	4065 Facebook Reviews	No
UCL R-Urdu [24]	20,229 Reviews	Yes
Urdu & R-Urdu [14] R-Urdu [25]	Urdu Tweets 24,000 Indo-Pak Music Industry's Reviews	No Yes
Roman Urdu E-commerce [26]	26,824 Daraz User Reviews	Yes
RUSA-19 [27], [24], [28]	10,021 R-Urdu Sentences	Yes
Sports/Smartphone [29]	2,200 Sports/Smartphone Tweets	Yes
Urdu [30]	10,008 Public Reviews	Yes
Urdu [31]	6,000 Urdu Blog/News Sentences	No
Urdu [32]	1,140,824 Tweets	Yes
R-Urdu [33]	1600 R-Urdu Translated Documents	No

three corpus construction approaches: manual, automatic, and bi-lingual [34], [35], [36], [37]. Table 1 shows the summary of the State-of-the-art Datasets.

Mahmood et al. [28] developed a benchmark corpus with human annotation RUSA-19 for the R-Urdu language. The 10,021 R-Urdu sentences in the corpus come from 566 internet threads in sports, software, food & recipes, drama, and politics. Mukhtar et al. [9] retrieved Urdu data from 151 Urdu blogging websites and classified them into positive ($+ve$), negative ($-ve$), and neutral (*neu*) classes. Sharf et al. [11] gathered a large R-Urdu dataset from social media websites. Sana et al. [29] developed two datasets (sports and smartphones) from Twitter via the API Tweepy, each containing 1200 and 1000 tweets, respectively. Noor et al. [12] developed an R-Urdu Dataset of 20.286K reviews from Daraz website.[1] Total reviews have been classified into three categories by various annotators.

Khan et al. [38] developed an Urdu Corpus for SA (UCSA) containing 9,601 reviews, of which 4,843 are $+ve$ and 4,758 $-ve$. The reviews contain information about Urdu drama, politics, movies, sports, and TV talk shows. Ahmad and Edalati [22] developed a manually annotated Urdu Tweet dataset containing 10,000 samples and classified them as $+ve$ and $-ve$ for Urdu SA. Qutab et al. [23] collected 4065 Facebook reviews to increase the size of the public R-Urdu dataset, containing four categories of politics, sports, education, and religion. Arif et al. [33] converted an existing English hotel reviews dataset to R-Urdu containing 1600 total reviews with an equal number of $+ve$ and $-ve$ reviews. Mehmood et al. [39] provided a public dataset in R-Urdu containing 3241 reviews and classified them into $+ve$, $-ve$, and *neu* classes. Ullah et al. [40] developed an R-Urdu sentence corpora of 5K from different domains and annotated four emotion classes.

[1]https://www.daraz.pk, accessed on March 24, 2023

Rehman et al. [14] developed a dataset by scraping Nastaleeq and R-Urdu tweets from Twitter using the Tweepy API. Stopwords for R-Urdu were manually retrieved from the dataset, and the stopword list for the Urdu script was collected from three distinct sources. Kumhar et al. [41] developed a monolingual Urdu dataset by converting R-Urdu into Urdu and translating English into Urdu. Khan et al. [15] proposed a dataset containing 9312 reviews manually labeled into three classes: $+ve$, $-ve$ and *neu*. Altaf et al. [42] collected Tweets from cricket and football domains to develop a dataset and categorized them as $+ve$, $-ve$, and *neu* sentiments tweets. Qureshi et al. [25] creates an R-Urdu Dataset (DRU) containing 24000 R-Urdu text reviews scraped from the reviews of 20 songs from the Indo-Pak Music Industry and used it for the binary classification. Chandio et al. [26] extracted R-Urdu comments of users on Daraz against several products and created a RUECD dataset containing 26,824 user reviews having 8252 $-ve$, 9833 $+ve$, and 8739 *neu*. Nagra et al. [27] evaluated SA on R-Urdu (RUSA-19) corpus of 10,021 sentences containing 3778 $+ve$ reviews, 2941 $-ve$ reviews, and 3302 *neu* reviews.

Safder et al. [30] developed an Urdu dataset of 10,008 reviews from 566 online threads on software, sports, food, entertainment, and politics and used it for binary and ternary classification. Naqvi et al. [31] created a dataset based on Urdu blogs and news websites such as BBC, Express, Dunya, and humsub. Literature, religion, sports, humor, the economy, health, technology, and politics are all covered in dataset reviews. The dataset contains 6000 sentences and 117685 words and is divided into $+ve$ and $-ve$ classes. Majeed et al. [43] created a dataset of 18k hand-labeled sentences gathered from various domains and labeled with six different classes, including sadness, anger, fear, happy, *neu* and love.

Mehmood et al. [10] developed a RUSA dataset of 11,000 reviews collected from six domains. Annotation guidelines were developed, and multiple analysts labeled the

dataset. Mehmood et al. [10] suggested an R-Urdu dataset of 779 reviews containing 412 $+ve$ and 367 $-ve$ reviews. A semi-automatic method was used to collect data from five different domains. Batra et al. [32] created a collection of 1,140,824 Urdu-language tweets gathered from Twitter between September and October 2020. This dataset contains the tweet-id, text, emoji extracted from the text, and a sentiment score.

B. MACHINE LEARNING BASED SENTIMENT ANALYSIS OF URDU TEXT

Mukhtar et al. [9] explored sentiment analysis (SA) on an Urdu blog dataset, which included classes for positive ($+ve$), negative ($-ve$), and neutral (neu) sentiments. They employed various classifiers, such as Decision Trees (DT), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) in the WEKA environment using 10-fold cross-validation. The experimental results demonstrated that KNN achieved the highest accuracy of 67.01%.

Mehmood et al. [10] introduced an SA system for R-Urdu, analyzing unigram, bigram, and uni-bigram features with five classification algorithms: KNN, Naive Bayes (NB), SVM, DT, and Logistic Regression (LR). The dataset was split into a 60% training and 40% testing ratio. Evaluation outcomes revealed that NB and LR performed better than other classifiers. Sharf et al. [11] conducted discourse-based SA on R-Urdu datasets from social media websites. They applied POS tagging, normalization, and tokenization techniques to identify discourse elements.

Sana et al. [29] proposed an emotion detection system for Urdu. They employed Support Vector Classifier (SVC), NB, KNN, and Random Forest (RF) algorithms to categorize Urdu tweets based on emotions. The evaluation results indicated that SVC achieved the highest accuracy of 80.05% on the smartphone dataset and 81.09% on the sports dataset.

Noor et al. [12] investigated R-Urdu reviews from the Daraz website. After performing feature extraction, they employed SVM for sentiment classification using bag-of-words and vector space models. Experimental results with different SVM kernels demonstrated that the Cubic Kernel achieved the highest accuracy on the given dataset.

C. DEEP LEARNING BASED SENTIMENT ANALYSIS OF URDU TEXT

Manzoor et al. [13] proposed a Self-attention Bidirectional LSTM (SA-BiLSTM) approach to deal with varying patterns of text representation, achieving an accuracy of 68.4% and 69.3%. Khan et al. [38] created a dataset for Urdu sentiment analysis and compared various machine learning and deep learning algorithms, achieving an accuracy rate of 81.94%. Jan et al. [22] developed a hand-labeled Urdu Tweet dataset and analyzed five distinct lexicon and rule-based algorithms, with Flair outperforming the others with an accuracy of 70%. Qutab et al. [23] analyzed the emotions in R-Urdu comments using TF-IDF and count vectorization with Multinomial Logistic Regression (MLR), achieving the highest accuracy of 94%. Arif et al. [33] converted an English hotel reviews dataset into R-Urdu and analyzed machine learning techniques for classification. SVM achieved the highest accuracy of 96% using TF-IDF vectorization.

Mehmood et al. [39] proposed a multi-channel hybrid strategy using recurrent and convolutional neural networks, achieving better results than other machine learning and deep learning approaches. Ullah et al. [40] studied the emotional polarity of R-Urdu sentences using RF, SVM, DT, and KNN algorithms with ten-fold cross-validation, achieving better F-measure. Khan et al. [24] presented English and R-Urdu sentiment analysis models based on CNN-LSTMs using various embedding models, i.e., FastText, Word2Vec, TF-IDF, and GloVe, achieving high accuracy, precision, recall, and F1-scores. Rehman et al. [14] performed sentiment analysis on Nastaleeq and R-Urdu tweets extracted from Twitter using 31 different classifiers and found Sequential Minimal Optimization (SMO) and RF to be the most effective machine-learning algorithms for Nastaleeq and R-Urdu tweets, respectively. Tehreem et al. [44] evaluated an R-Urdu dataset containing people's comments on different Pakistani dramas and TV shows, exploring five different supervised learning algorithms, and found SVM to be the best-performing algorithm with an accuracy of 64%. These studies highlighted the effectiveness of various machine learning and deep learning techniques for sentiment analysis in the R-Urdu language.

Kumhar et al. [41] developed a monolingual Urdu dataset and used the skip-gram model to create Urdu word embeddings and LSTM for SA. Khan et al. [15] fine-tuned Multilingual BERT and used pre-trained FastText, character N-grams, word embeddings from BERT, and word N-grams from words for SA. Sehar et al. [45] proposed a hybrid system that combines deep neural network techniques with dependency-based Urdu grammatical rules to analyze sentiment in Urdu. Habiba et al. [46] proposed a rule-based classifier for SA in complicated R-Urdu feelings. Altaf et al. [42] used machine and deep learning methods, including linear SVC, Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), LR, RF, RNN, LSTM, and GRU, for sentiment analysis on a Twitter dataset. BNB outperformed all classifiers on the balanced and unbalanced dataset. At the same time, GRU served as the foundation for future studies on Urdu as a resource-constrained language suitable for cross-domain sentiment analysis.

Chandio et al. [47] created RU-BiLSTM with attention mechanism and embeddings to analyze sentiment in R-Urdu. Masood et al. [48] developed a deep learning-based LSTM architecture with an 830-word manual dictionary for stemming in Urdu and achieved an accuracy rate of 86.8% and F1-Score of 89%. Qureshi et al. [25] compared nine machine learning algorithms using a dataset named R-Urdu (DRU) and found that LR performed better than others with an accuracy of 92.25% for binary classification. Chandio et al. [26] proposed a fine-tuned SVM powered by R-Urdu Stemmer and achieved 68% accuracy. Nagra et al. [27] developed a

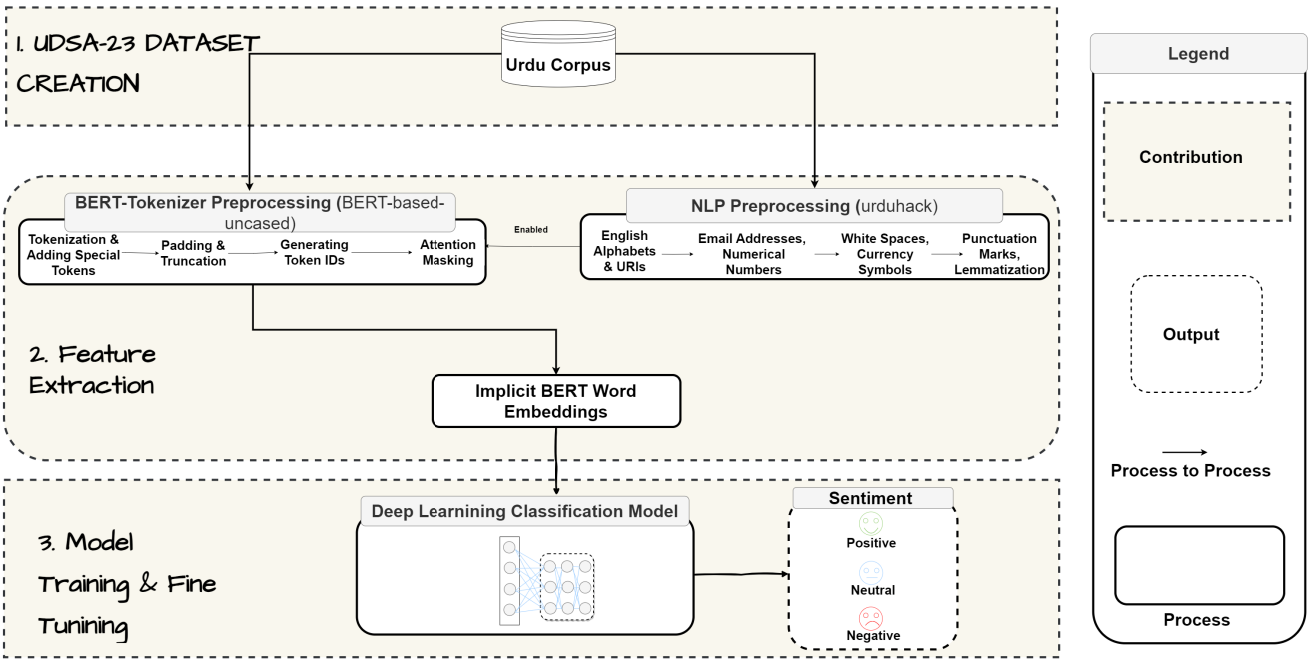


FIGURE 1. The overview of USA-BERT.

deep-learning-based model using Faster RCNN and achieved the highest accuracy of 91.73% for binary classification and 89.94% for tertiary classification.

Safder et al. [30] proposed a deep learning model and an open-source corpus of 10,008 Urdu reviews, where RCNN outperformed others with a ternary classification accuracy of 68.56% and a binary classification accuracy of 84.98%. Naqvi et al. [31] developed an architecture for analyzing the sentiment of Urdu text using four deep learning approaches and found that BiLSTM with attention is the most efficient model with a 77.9% accuracy rate and 72.7% F1 score. Majeed et al. [43] analyzed the emotional polarity of R-Urdu sentences using machine learning algorithms and achieved a 69.4% accuracy rate with the SVM model. Mukhtar et al. [49] presented an effective Urdu sentiment analyzer for Urdu blogs with an accuracy rate of 89.03% and an F-measure of 0.88. Mahmood et al. [28] presented a deep learning model to extract emotions and attitudes of people expressed in R-Urdu, where RCNN outperformed baseline approaches with accuracy scores of 0.652% and 0.572% for binary and ternary classifications, respectively.

Muhammad and Burney [50] applied machine and deep learning algorithms for Urdu text Classification of symmetric datasets. They found that machine learning algorithms showed good results, while most deep learning algorithms improved the results further. They applied a combination of machine learning algorithms to the Urdu text data, leading to further improvement of results. Ahmed et al. [51] proposed an attention-based neural network for the review level Urdu sentiment analysis. They visualized attention weights to uncover the black box of the models and confirm their intuition. They archived 91% accuracy and 88% F1 score, respectively.

In conclusion, researchers have proposed many approaches for Urdu sentiment classification. However, only a few studies [15] focus on the pre-train word embeddings for Urdu sentiment classification. The proposed approach differs from the existing approaches in that we apply BERT embeddings as a feature extraction method with a deep learning classifier for sentiment classification.

III. PRELIMINARIES

A. BERT

Bidirectional Encoder Representations from Transformers (BERT) was originally presented by a team of Google researchers [52]. BERT employs a transformer architecture, a deep learning model that utilizes self-attention mechanisms to capture the contextual relationships between words. The transformer architecture allows BERT to model the bidirectional context of words, taking both left and right contexts into account. With just a single additional layer, BERT can be fine-tuned for tasks, i.e., sentiment analysis or question answering. The pre-training of BERT involved two unsupervised tasks: masked language modeling and next sentence prediction [53], [54]. BERT can be exploited in two ways: BERT word embeddings and fine-tuned BERT classifier.

1) WORD EMBEDDINGS

BERT word embeddings are dense vector representations of words that capture semantic and contextual information. BERT can be utilized to generate word embeddings for a given text. By passing a review through BERT, we can obtain high-quality word embeddings that capture the contextual meaning of the words. These embeddings can be used as input features for downstream deep learning classifiers. Note that we employ BERT word embeddings for USA-BERT as the performance of USA-BERT with BERT word embeddings is significant in contrast to the fusion of USA-BERT with other word embedding methods, i.e., word2vec [20] and FastText [55].

2) FINE-TUNED BERT CLASSIFIER

BERT can be fine-tuned using a specific corpus and can be leveraged as a classifier, i.e., BERT classifier is leveraged for the classification task using the generated dataset (UDSA-23). Pre-trained BERT models are trained on large-scale datasets and can learn general language representations. However, to adapt BERT for a specific task, you can fine-tune it on a smaller labeled dataset. This involves training the BERT on task-specific data, which helps it learn task-specific patterns and improve its performance for that particular classification task. The pre-trained BERT serves as a powerful feature extractor, capturing rich contextual information, while the classification layer enables the model to make predictions based on the specific task requirements.

IV. APPROACH

A. OVERVIEW

Fig. 1 depicts an overview of USA-BERT that classifies Urdu reviews into three categories: *positive*, *negative*, and *neutral*. Given the constructed Urdu review dataset (UDSA-23), USA-BERT predicts sentiments of Urdu reviews as follows:

- First, we preprocess each review by exploiting BERT-Tokenizer and split the reviews into tokens.
- Then, we convert each review into word embeddings using BERT.
- Next, we collect the generated word embeddings of each review as input for the fine-tuning BERT classifier.
- Finally, we input reviews to the trained classifier to predict their labels, i.e., positive, negative, or neutral, to check the performance of USA-BERT.

Each of the essential steps of USA-BERT is presented in the following sections.

B. PROBLEM DEFINITION

A function f can assign the sentiment of a new Urdu review ur to a specific category sentiment s.

$$s = f(ur) \quad s \in \{positive, negative, neutral\}, \quad ur \in UR \tag{1}$$

$$ur =< ut, s > \tag{2}$$

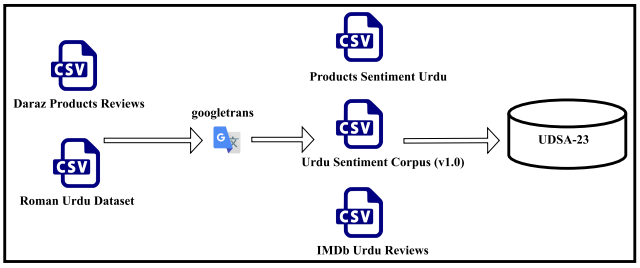


FIGURE 2. An overview of UDSA-23 dataset creation.

where, s represents the sentiment of ur, which can be categorized as *positive*, *negative*, or *neutral*. The function f denotes the sentiment classification function, where ur represents an individual review as the function's input and UR represents a set of Urdu reviews. However, each ur contains Urdu text ut and its sentiment status s.

C. UDSA-23 DATASET CREATION

Figure 2 illustrates the procedure for creating the proposed Urdu Dataset for Sentiment Analysis-23 (UDSA-23). The UDSA-23 dataset is generated by converting Roman Urdu (R-Urdu) datasets, including Daraz Product Review[2] and Roman Urdu Dataset,[3] into Urdu using the Googletrans[4] library. Note that the converted reviews from R-Urdu into Urdu are verified by the Ph.D. scholars and professionals of the Urdu language, Institute of Languages & Linguistics, Punjab University, Pakistan. They first verified the reviews individually. Then, a Zoom meeting is scheduled for them to resolve the divergent viewpoints. For example, *googletrans* tranforms the R-Urdu reviews "shukriya daraz apka product orignal tha" and "audio achi hai base k sath" into آپ کا شکریہ ، آپ کی مصنوعات اصل تھی and ساتھ اچھا ےھآڈیو یبس کے, respectively. Finally, the converted datasets are then combined with Urdu dataset, including Product Sentiment Urdu,[5] Urdu Sentiment Corpus (v1.0),[6] and imdb_urdu_reviews[7] to create a standardized benchmark dataset resource.

The UDSA-23 dataset comprises 34,270 negative reviews, 13,665 neutral reviews, and 36,808 positive reviews. The data distribution for each review type is depicted in Figure 3. Additionally, Figure 4 showcases the maximum and minimum lengths of reviews within the UDSA-23 dataset. An overview

[2]https://github.com/mirfan899/Urdu/blob/master/sentiment/daraz_products_reviews.csv.tar.gz, accessed on March 24, 2023
[3]https://github.com/mirfan899/Urdu/blob/master/sentiment/roman.csv.tar.gz, accessed on March 24, 2023
[4]https://pypi.org/project/googletrans/, accessed on March 24, 2023
[5]https://github.com/mirfan899/Urdu/blob/master/sentiment/products_sentiment_urdu.csv.tar.gz, accessed on March 24, 2023
[6]https://github.com/MuhammadYaseenKhan/Urdu-Sentiment-Corpus/blob/master/urdu-sentiment-corpus-v1.tsv, accessed on March 24, 2023
[7]https://github.com/mirfan899/Urdu/blob/master/sentiment/imdb_urdu_reviews.csv.tar.gz, accessed on March 24, 2023

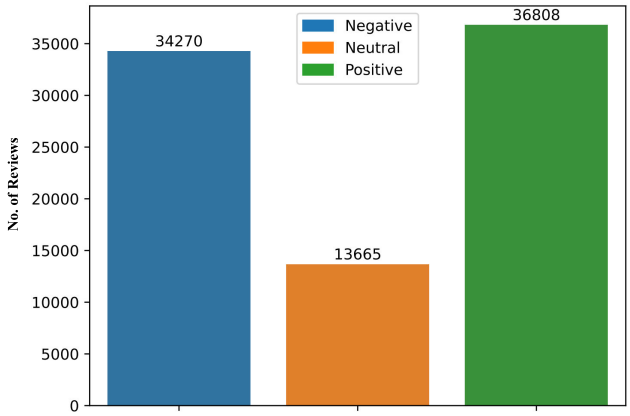


FIGURE 3. Distribution of UDSA-23 reviews.

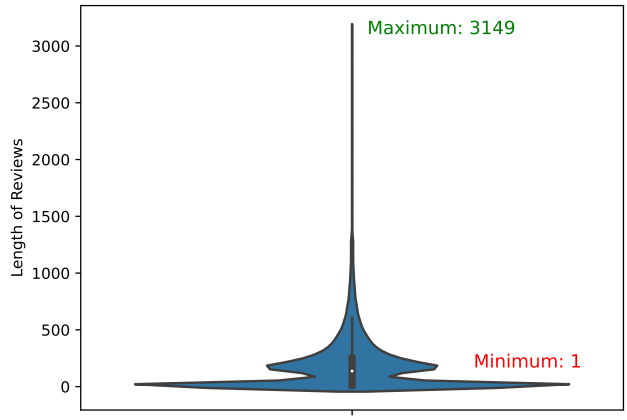


FIGURE 4. Minimum and maximum length of UDSA-23 reviews.

TABLE 2. UDSA-23 statistics.

No. of Reviews	84743
Words Count	15182055
Maximum Words in a Review	3149
Minimum Words in a Review	1
No. of Sentiments	3
No. of Positive Reviews	36808
No. of Negative Reviews	34270
No. of Neutral Reviews	13665

TABLE 3. Positive, negative, and neutral reviews.

Positive Reviews	Negative Reviews	Neutral Reviews
یہ بہت اچھا فون ہے۔ واہ کیا بات ہے۔	مجھے بہت افسوس ہو رہا ہے۔ میرا دل ٹوٹ گیا ہے۔	میں نہیں جانتا کہ کیا کروں۔ کچھ بھی ہو سکتا ہے۔

of the UDSA-23 corpus statistics and examples of each class are provided in Table 2 and Table 3, respectively.

D. PREPROCESSING

The sentiment analysis of Urdu text strongly relies on representing necessary words to BERT, a powerful NLP model pre-trained on a large text corpus. Although recent research explores various word representation techniques in Natural Language Processing (NLP), i.e., Word2Vec [20] and FastText [55], BERT's ability to learn highly contextualized word and phrase representations makes it effective for various tasks without NLP preprocessing. We utilize the *Bert-Tokenizer* from the *Transformers* library in this context. The preprocessing for the Urdu text (ur) is as follows:

- The ur text is tokenized using the Bert-Tokenizer text preprocessing, which splits it into a sequence of subwords. These subwords are then mapped to integer IDs using a pre-defined vocabulary. Special tokens (CLS and SEP) are added to mark the beginning and end of the ur text, respectively. Notably, the tokenizer breaks down Urdu text into words and subword units, effectively combining word and subword levels. Unlike English, Urdu can be tokenized directly using the language-agnostic BPE tokenizer, eliminating the need for English adaptation. The tokenization process can be represented as:

$$W = [CLS] + w_1 + w_2 + \ldots + w_n + [SEP] \quad (3)$$

  where, W represents the sequence of subwords in ur, w_i denotes a subword, and n is the total number of subwords in ur.
- Each subword w_i in the sequence W is converted to a numerical ID using a pre-defined vocabulary. This conversion can be expressed as:

$$T = 101 + t_1 + t_2 + \ldots + t_n + 102 \quad (4)$$

  In this equation, 101 and 102 are special tokens added by the tokenizer (CLS and SEP, respectively), T represents the sequence of token IDs for ur, t_i is the ID corresponding to w_i, and n is the total number of token IDs for each ur.
- The token IDs are then padded and truncated to a maximum length of the longest review from the dataset. If the total number of IDs for a given ur is less than 256 tokens, it is padded with the special token '0'. If the number of tokens exceeds 256, the extra tokens are truncated as follows:

$$T'_{1:256} = \begin{cases} T_{1:m}\,[PAD]^{256-m} & \text{if } m < 256 \\ T_{1:256} & \text{if } m > 256 \end{cases} \quad (5)$$

  where, T' represents the final IDs after padding and truncation for ur, m is the total number of token IDs, and $T_{1:m}$ denotes the original token IDs.
- Finally, attention masks distinguish between actual and padding tokens in the input sequence. This is crucial because the attention mechanism in the transformer architecture utilizes these masks to focus on the real tokens and ignore the padding tokens. The attention mask MT' for the input sequence T' can be represented as:

$$MT' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad (6)$$

In this matrix, 1's indicates the actual tokens in the input sequence, while 0's indicates the padding tokens.

To elaborate on the preprocessing step, consider the following example:

- The Urdu review *ur* "میں بہت خوش ہوں" is passed to BERT-Tokienizer.
- Then, BERT-Tokenizer splits *ur* into token as ["ہوں", "خوش", "بہت", "میں"].
- Next, the special tokens are added to the tokenized review by the BERT-Tokenizer as ["[CLS]", "ہوں", "خوش", "بہت", "میں", "[SEP]"].
- After that, BERT-Tokenizer applies padding and truncation if required, and generates the Token IDs as [101, 1001, 2002, 3003, 4004, 102]
- Finally, the attention masks are employed to differentiate between the actual tokens and the padding tokens, e.g., if two additional tokens are padded for this *ur* then the attention masks can be represented as [1, 1, 1, 1, 1, 1, 0, 0].

Note that we also apply NLP techniques to the Urdu reviews to check its influence on USA-BERT as discussed and evaluated in Section V-D2.

E. WORD EMBEDDING

Word embedding is a method for representing words as numerical vectors in machine learning. The primary idea behind word embedding is to capture semantic and syntactic information of words in a high-dimensional vector space, where similar words are close together in contrast to different words. Unlike traditional word embedding models, i.e., Word2Vec and GloVe, BERT provides contextualized word embeddings. It considers the surrounding words in a sentence and generates word representations that capture the contextual meaning of the word. BERT utilizes a transformer architecture, effectively capturing long-range dependencies and context information. It is trained on a large corpus of text data, using masked language modeling and next-sentence prediction objectives. In BERT embeddings, the same word can have different embeddings in different contexts. e.g., the word "talk" will have different embeddings in the sentences "I want to talk" and "I will attend a talk". Consequently, there is not a singular vector of embeddings for each word, distinguishing BERT from models like word2vec or GloVe. The application of a masked language model on a pre-trained BERT is considered as this research deals with a small dataset, serving to fine-tune the BERT model accordingly.

To obtain word embeddings, we pass uw_t to BERT to generate the contextualized representations for each word in the sequence. These representations capture the meaning of the word within its context. One of the advantages of using BERT word embeddings is their ability to handle polysemy (multiple meanings) and homonymy (same word with different meanings) effectively. Since BERT considers the context in which a word appears, it can differentiate between different senses of a word based on its surrounding words. Notably, BERT-base-uncased[8] word-piece model, supplied by Google, was employed in our research. Table 4 presents the generated vectors from BERT-base-uncased.

F. TRAINING AND TESTING

For the sentiment classification of Urdu reviews as positive, negative, and neutral, the pre-trained BERT model is fine-tuned using the proposed UDSA-23 dataset. Its ability to capture contextualized representations takes advantage of pre-trained knowledge and refines task-specific data, making it a highly efficient model for analyzing Urdu sentiments.

Table 5 summarizes the hyper-parameter used for BERT fine-tuning. The reviews are randomly split into the 80%:20% ratio for fine-tuning and prediction. We employ the BERT-Tokenizer (BERT-based-uncased) and a BERT classifier with 12 transformer layers, 12 attention heads, 768 hidden layers, an epsilon of 1e-8, 10 epochs on UDSA-23 and 15 epochs on UCSA-21 [15], and a learning rate of 2e-5 is fine-tuned using Google Colab. A batch size of 16 is used for training, where the optimizer updates parameters for each batch during each epoch. In the final step, validation loss and accuracy are computed on a validation split to evaluate the model's performance. The training and testing accuracy of USA-BERT on UDSA-23 and UCSA-21 are (75.45% and 89.53%) and (83.64% and 94.81%), respectively.

V. EVALUATION

In this section, we assess the effectiveness of USA-BERT by comparing it to state-of-the-art methods.

A. RESEARCH QUESTIONS

The evaluation of USA-BERT focuses on addressing the following research questions:

- RQ1: What is the performance comparison between USA-BERT and state-of-the-art methods?
- RQ2: Does the application of NLP preprocessing techniques enhance the performance of USA-BERT?
- RQ3: To what extent does re-sampling improve the performance of USA-BERT?
- RQ4: Can USA-BERT, utilizing BERT-base-uncased generated word embeddings, effectively identify sentiment in Urdu text?

To address RQ1, we compare USA-BERT with baseline approach [15] approaches to assess its performance improvement. The baseline approach is the recent approach with significant results in sentiment prediction and compares with the UCSA-21 benchmark dataset. Therefore, we select the baseline [15] as a benchmark for comparison with USA-BERT.

To investigate the impact of NLP preprocessing on the performance of USA-BERT, as mentioned in Section IV-D, we address RQ2. Preprocessing techniques are applied to clean the provided dataset, as textual datasets often contain punctuation and other inconsistencies.

[8]https://github.com/google-research/bert, accessed on March 24, 2023

TABLE 4. An example of word embedding.

Input	Text	Word Embeddings
BERT-Tokenizer Preprocessed Text	شمسی یہاں ادا کریں	[0.7124861478805542 . . . 0.049222107976675034]
NLP Preprocessed Text	شمسی ادا	[-0.7124861478805542 . . . 0.049222107976675034]

TABLE 5. Hyperparameter settings.

Hyper-parameter	Value
Parameters	110 M
Batch Size	16
No. of Epochs	10 (UDSA-23) & 15 (UCSA-21)
Learning Rate	2E-05
Gradient Accumulation Steps	16
Hidden Size	768
Hidden Layers	12
Maximum Sequence Length	128

Algorithm 1 USA-BERT Evaluation

1: **procedure** Evaluate-USA-BERT
2: UDSA-23 ← ProposeUrduDataset()
3: RE ← GetReviews()
4: preprocessed_reviews ← PreprocessReviews(RE)
5: word_embeddings ← GenerateWordEmbeddings(preprocessed_reviews)
6: TrainModel(word_embeddings)
7: ComputeEvaluationMetrics()
8: **end procedure**

To examine the influence of re-sampling on USA-BERT, RQ3 focuses on our imbalanced dataset. We employ the under-sampling and over-sampling techniques to balance the dataset and compare the performance of USA-BERT on both the balanced and imbalanced datasets to observe any changes in performance.

For RQ4, we evaluate the performance of USA-BERT by comparing the performance results of different deep learning algorithms with embeddings generated by BERT.

B. PROCESS

The process of USA-BERT is presented in Algorithm 1. The proposed algorithm outlines the process for evaluating USA-BERT for sentiment analysis using dataset UDSA-23. It first involves creating the UDSA-23 dataset as indicated in Line 2. This dataset is specifically designed for sentiment analysis tasks in the Urdu language. Next, the algorithm retrieves a set of reviews from a specified source, as mentioned in Line 3. The sources used to build UDSA-23 are explained in Section IV-C. Once the reviews are obtained, the algorithm proceeds to preprocess each review individually. This preprocessing step (discussed in Line 4) involves tokenization and adding special characters, generating token ids, padding and truncation, and attention masking. With the preprocessed reviews, the algorithm generates word embeddings. Line 5 specifies using BERT-base-uncased, a popular language model, to generate the word embeddings. This step aims to transform the text data into numerical representations that can be utilized by machine/deep learning classifiers. Following the generation of word embeddings, the algorithm trains the deep-learning classifier. The training process, denoted in Line 6, employs the generated word embeddings as inputs to the machine/deep learning classifiers. The hyperparameter settings for the training process are presented in Table 5. Finally, the algorithm computes the evaluation metrics for each classifier. This step, as mentioned in Line 7, aims to assess and compare the performance of the different classifiers utilized in sentiment analysis. The evaluation metrics employed for this purpose are explained in the following section.

C. METRICS

We calculate several sentiment evaluation metrics based on Urdu reviews *UR* to evaluate USA-BERT. These metrics include sentiment accuracy (*Acc*), precision (*Pre*), recall (*Rec*), and f-measure (*FM*). These metrics have been widely used in previous studies [56], [57]. The formal definitions of *Acc*, *Pre*, *Rec*, and *FM* are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$Rec = \frac{TP}{TP + FN} \tag{9}$$

$$FM = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} \tag{10}$$

where, *Acc*, *Pre*, *Rec*, and *FM* represent the accuracy, precision, recall, and f-measure of the classifiers used for sentiment prediction of *UR*. In Eq. 7 - Eq. 10, *TP* represents the number of *ur* in *UR* that are correctly predicted, while *TN* denotes the number of reviews correctly predicted as negative. On the other hand, *FP* corresponds to the number of falsely predicted *ur*, and *FN* represents the number of *ur* that were not predicted as sentiment *s* but were actually *s*.

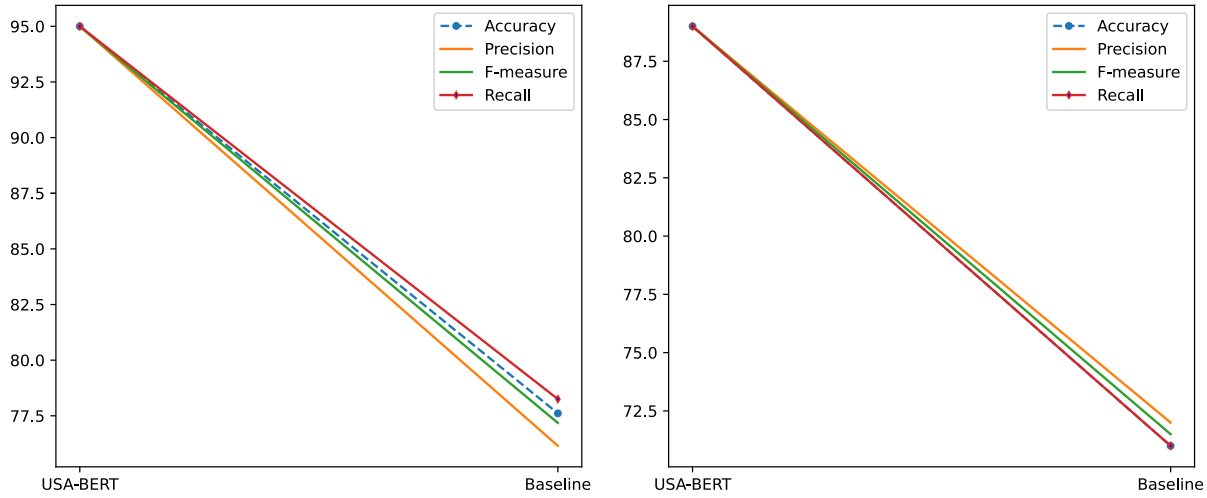
D. RESULTS

1) RQ1: COMPARISON OF USA-BERT AGAINST BASELINE APPROACH

To address RQ1, we conducted a comparative analysis between USA-BERT and the baseline approach using two datasets: UDSA-23 (our proposed dataset) and UCSA-21 (a state-of-the-art dataset). The results of these approaches can be found in Table 6, where the columns represent *Acc*, *Pre*, *Rec*, and *FM*. The rows of the table indicate the performance of each approach. Additionally, Figure 5 provides a

TABLE 6. Performance of USA-BERT and the baseline approach.

Approach \ Dataset	UDSA-23				UCSA-21			
	Acc	Pre	Rec	FM	Acc	Pre	Rec	FM
USA-BERT	89.53%	90%	90%	90%	94.81%	94.82%	94.81%	94.81%
Baseline [15]	71%	72%	71%	71.50%	77.61%	76.15%	78.25%	77.18%
Improvement	**26.09%**	**25.00%**	**26.76%**	**25.87%**	**22.16%**	**24.52%**	**21.16%**	**22.84%**



(a) Performance of USA-BERT and Baseline Approach on UDSA-23

(b) Performance of USA-BERT and Baseline Approach on UCSA-21

FIGURE 5. Performance of USA-BERT and baseline approach.

visual representation of the performance comparison between the approaches using the UDSA-23 and UCSA-21 datasets. Note that we use 15 epochs for the comparison of approaches with UCSA-21 dataset, as the results in the baseline paper are reported with 15 epochs.

Based on the findings presented in Table 6 and Fig. 5, the following observations can be made:

- USA-BERT exhibits superior performance compared to the baseline approach regarding *Acc* and *FM*. On the UDSA-23 dataset, USA-BERT improves *Acc* by 26.09% and *FM* by 25.87%, calculated as the percentage increase from the baseline values of 71% and 71.50% respectively.
- Similarly, on the UCSA-21 dataset, USA-BERT achieves an *Acc* improvement of 22.16% and an *FM* improvement of 22.84%, relative to the baseline values of 77.61% and 77.18%, respectively.
- Several factors contribute to the performance enhancement of USA-BERT [52]. Firstly, BERT generates contextualized word embeddings, allowing it to capture rich contextual information, including word order and sentence semantics, which is vital for accurate classification. Secondly, the pre-training of BERT on a large unlabeled text corpus helps it learn general language patterns and semantic relationships, resulting in comprehensive and informative representations. Lastly, the fine-tuning process adapts the pre-trained BERT model to the specific classification task, enabling it to capture task-specific features and enhance overall classification performance.

However, despite the accuracy of USA-BERT, it still exhibits a significant number of false positives and false negatives. This misclassification issue can be attributed to the incorrect translation of reviews from R-Urdu to Urdu using the "googletrans" API. For example, the translation of the R-Urdu review "mil gya hai acha hai aur toota howa b nai hai" as "نئی ہے چلا گیا اچھا اور ٹوٹا ہوا ہووا" by googletrans is incorrect. To mitigate such misclassifications, manual translation of R-Urdu reviews into Urdu by a native speaker should be considered. Further investigation is needed to understand the reasons behind these misclassifications and develop effective measures to reduce them.

In conclusion, based on the analysis presented above, it can be concluded that USA-BERT outperforms the baseline approach in sentiment classification of Urdu reviews.

2) RQ2: INFLUENCE OF NLP PREPROCESSING ON THE PERFORMANCE OF USA-BERT

Urdu is a complex language with unique linguistic characteristics, i.e., right-to-left script, complex morphology, and

TABLE 7. Influence of NLP preprocessing on the performance of USA-BERT.

NLP Preprocessing \ Dataset	UDSA-23				UCSA-21			
	Acc	Pre	Rec	FM	Acc	Pre	Rec	FM
Enabled	89.23%	89.28%	89.23%	89.22%	**95.43%**	**95.44%**	**95.43%**	**95.42%**
Disabled	**89.53%**	**90%**	**90%**	**90%**	94.81	94.82	94.81	94.81

TABLE 8. An example of NLP preprocessing.

	Urdu Text	English Translation
Original Review	!!!!!! ہولناک ہے ..... کیا .... موی ..... یہ	This ..... movie .... what .....is a terrifying!!!!
After Punctuation Removal	یہ موی کیا ہولناک ہے	What a terrifying movie is this
After Stop-words Removal	موی ہولناک	Terrifying movie

various diacritical marks. NLP preprocessing helps to handle these complexities and ensure proper handling of the text during subsequent analysis. NLP preprocessing tasks, e.g., tokenization, stemming, and removing stop-words help in reducing the dimensionality of the text data and improving the efficiency of downstream tasks. By breaking the text into smaller units and removing irrelevant or redundant words, preprocessing enables more effective analysis and classification of Urdu text.

Moreover, preprocessing aids in addressing noise and inconsistencies in the data. It involves tasks, e.g., removing punctuation and normalizing text by converting it to lowercase. According to the baseline paper, these steps improve data quality and consistency. Therefore, Urdu text NLP preprocessing is applied to USA-BERT to check its influence on the proposed classification method. Note that *urduhack*[9] Python library is used for the preprocessing of Urdu reviews.

Once the preprocessing step is completed, an Urdu review ur can be formalized as follows:

$$ur' =< uw_t, s > \tag{11}$$

$$uw_t =< uw_1, uw_2, \ldots, uw_n > \tag{12}$$

where, uw_t represents the preprocessed tokens of ur as presented in Table 8.

To address RQ2, we conducted a comparative analysis between NLP preprocessing enabled/disabled USA-BERT using two datasets: UDSA-23 and UCSA-21. The results of preprocessing enabled/disabled USA-BERT can be found in Table 6, where the columns represent *Acc*, *Pre*, *Rec*, and *FM*. The rows of the table indicate the performance of preprocessing enabled/disabled USA-BERT.

Based on the findings presented in Table 7, the following observations can be made:

- The preprocessing-disabled USA-BERT exhibits superior performance compared to preprocessing-enabled USA-BERT in terms of *Acc* and *FM*. On the UDSA-23 dataset, the preprocessing-enabled USA-BERT reduces *Acc* and *FM* by 0.33% and 0.87%, calculated as the percentage increase from the preprocessing-enabled USA-BERT of 89.23% and 89.22%, respectively. Note that USA-BERT eliminates the NLP preprocessing step for the sentiment classification of Urdu text.
- However, the preprocessing-enabled USA-BERT exhibits superior performance compared to preprocessing-disabled USA-BERT in terms of *Acc* and *FM*.On the UCSA-21 dataset, the preprocessing-disabled USA-BERT reduces *Acc* and *FM* by 0.65% and 0.64%, calculated as the percentage increase from the preprocessing-disabled USA-BERT of 94.81% and 94.81%, respectively.
- The possible reason for such improvement is that USA-BERT uses a subword tokenization method (WordPiece) to split words into subword units. WordPiece allows BERT to handle Out-Of-Vocabulary (OOV) words by representing them as a combination of subwords. In contrast, fastText requires predefined word embeddings and may struggle with OOV words that are not present in its embedding vocabulary [52].

In conclusion, based on the analysis presented above, it can be concluded that NLP preprocessing of the Urdu text does not require USA-BERT in the sentiment classification of Urdu reviews.

3) RQ3: INFLUENCE OF RE-SAMPLING ON THE PERFORMANCE OF USA-BERT

To investigate RQ3, we applied two re-sampling techniques to address the class imbalance in the dataset: over-sampling and under-sampling. Over-sampling involves generating additional samples for the minority class using the RandomOverSampler technique while under-sampling removes excess records from the majority class using the RandomUnderSampler technique. We evaluated the performance of preprocessing-enabled and preprocessing-disabled USA-BERT models with and without re-sampling, and the results are presented in Table 9. The table columns correspond to *Acc*, *Pre*, *Rec*, and *FM*, while the rows represent the performance of the different re-sampling. Note that we only re-sample the UDSA-23 dataset to check the impact of re-sampling on the performance of USA-BERT.

[9]https://pypi.org/project/urduhack/

TABLE 9. Influence of re-sampling on the performance of USA-BERT.

Re-sampling	Acc	Pre	Rec	FM
No	89.53%	90%	90%	90%
Undersampling	91%	91%	91%	91%
Oversampling	93%	93%	92%	92.45%

Based on the findings presented in Table 9, the following observations can be made:

- Both re-sampling methods improve the performance of USA-BERT. The undersampling improves *Acc* and *FM* by 1.64% and 1.11%, calculated as the percentage increase from the re-sampling disabled USA-BERT of 89.53% and 90%, respectively.
- Similarly, the oversampling improves *Acc* and *FM* by 3.87% and 2.72, calculated as the percentage increase from the re-sampling disabled USA-BERT of 89.53% and 90%, respectively.
- The possible reason for such improvement is that re-sampling techniques help reduce bias towards the majority class. This bias reduction can result in a more balanced decision boundary and improved generalization performance. The USA-BERT becomes less inclined to favor the majority class and can better capture the nuances and specific features of the minority class, leading to better classification results [52].

In conclusion, based on the analysis presented above, it can be concluded that undersampling and oversampling of the proposed dataset (UDSA-23) significantly improve the performance of USA-BERT in the sentiment classification of Urdu reviews.

4) INFLUENCE OF BERT EMBEDDINGS ON DEEP LEARNING CLASSIFIERS

BERT embeddings have had a significant impact on deep learning classifiers, particularly in the field of natural language processing. BERT embeddings provide powerful contextualized representations of words and sentences, capturing syntactic and semantic information. To investigate the RQ4, we generate BERT word embeddings and pass them to USA-BERT and other deep-learning classifiers to confirm the performance of USA-BERT. Table 10, Fig. 6 and Fig. 7 illustrate the performance comparison of deep learning classifiers on UDSA-23 and UCSA-21 for the sentiment classification of Urdu text, respectively.

Based on the findings in Table 10, Fig. 6 and Fig. 7, the following observations can be made:

- The performance (*Acc*, *Pre*, *Rec*, and *FM*) of AdaBoost, Ensemble (RNN and LSTM), LSTM, Multi-Layer Perceptron (MLP), RNN, USA-BERT on UCSA-21 are (69%, 68%, 69%, and 68.50%), (71%, 55%, 71%, and 61.98%), (74%, 75%, 74%, and 74.50%), (74%, 74%, 74%, and 74%), (70%, 66%, 70%, and 67.94%), and (95.43%, 95.44%, 95.43%, and 95.42%), respectively.

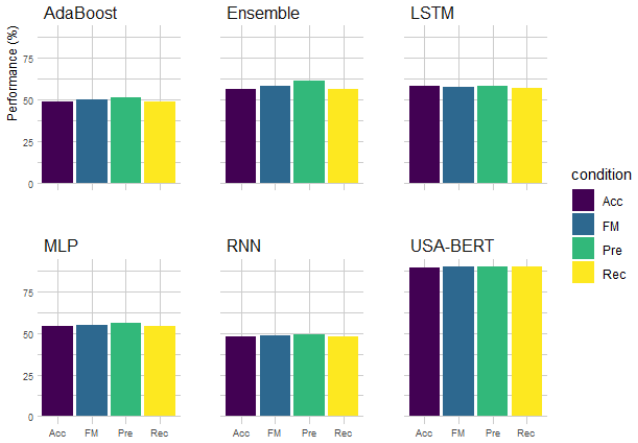


FIGURE 6. Performance of deep learning classifier on UDSA-23.

- Similarly, the performance (*Acc*, *Pre*, *Rec*, and *FM*) of AdaBoost, Ensemble (RNN and LSTM), LSTM, Multi-Layer Perceptron (MLP), RNN, USA-BERT on UDSA-23 are (49%, 51%, 49%, and 49.98%), (56%, 61%, 56%, and 58.39%), (58%, 58%, 57%, and 57.50%), (54%, 56%, 44%, and 54.98%), (48%, 49%, 48%, and 48.50%), and (89.23%, 89.28%, 89.23%, and 89.22%), respectively.
- USA-BERT outperforms AdaBoost, Ensemble (RNN and LSTM), LSTM, Multi-Layer Perceptron (MLP), and RNN. The possible reason for such improvement is that BERT is a transformer-based model that captures the contextual information of words in a sentence. It considers both the preceding and following words to understand the meaning of a word in a specific context. This contextual understanding helps BERT to grasp the nuances and intricacies of sentiment in Urdu text.
- While RNN, AdaBoost, LSTM, Ensemble (RNN, LSTM), MLP, and RNN are all valuable models in their own right, BERT's contextual understanding, pre-training on a large corpus, fine-tuning, and transfer learning capabilities give it an edge in sentiment classification tasks, particularly for languages like Urdu where contextual understanding is crucial for accurate sentiment analysis.

In conclusion, based on the analysis presented above, it can be concluded that the performance of USA-BERT with BERT embeddings is significant and yields the performance of other deep learning classifiers in the sentiment classification of Urdu reviews.

E. THREATS TO VALIDITY

We are aware of potential threats to the construct validity of USA-BERT due to the choice of evaluation metrics. We employ widely adopted metrics in the research community, namely accuracy, precision, recall, and F-measure, as highlighted by Illahi et al. [58]. However, it is important to

TABLE 10. Performance comparison of deep learning classifiers.

Approach \ Dataset	UDSA-23				UCSA-21			
	Acc	Pre	Rec	FM	Acc	Pre	Rec	FM
USA-BERT	89.53%	90%	90%	90%	94.81%	94.82%	94.81%	94.81%
LSTM	74%	75%	74%	74.50%	58%	58%	57%	57.50%
Ensemble	71%	55%	71%	61.98%	56%	61%	56%	58.39%
MLP	74%	74%	74%	74%	54%	59%	48%	54.98%
RNN	70%	66%	70%	67.94%	48%	49%	48%	48.49%
AdaBoost	69%	68%	69%	68.50%	49%	51%	49%	49.98%

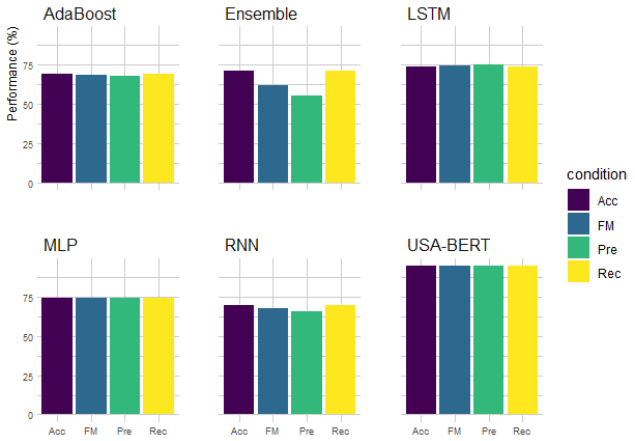


FIGURE 7. Performance of deep learning classifier on UCSA-21.

acknowledge that relying heavily on these metrics may have limitations in terms of construct validity.

We also recognize that classification algorithms can be susceptible to validity threats stemming from parameter values. To mitigate this, we conduct experiments to identify optimal parameter settings rather than relying on default values. Nonetheless, it should be noted that adjustments to the selected parameters have the potential to impact the results.

Using the BERT model for generating word embeddings from Urdu Reviews introduces a potential threat to construct validity. Although alternative tools are available, BERT was chosen due to its superior performance compared to other models at the time of selection. Nevertheless, it is crucial to consider that the lack of comprehensive embedding calculation tools for sentiment analysis of Urdu text may influence the overall performance of USA-BERT.

Concerns regarding internal validity arise from the implementation of USA-BERT. To address this, we conduct cross-checks to ensure the accuracy of USA-BERT. However, there is still a possibility that some errors may have been overlooked.

External validity is another area of concern regarding the generalizability of USA-BERT. Our analysis is limited to reviews included in UDSA-23 and UCSA-21, and the performance of USA-BERT may vary when applied to other Urdu reviews.

VI. CONCLUSION

This paper delves into the sentiment analysis task targeted at low-resource languages, i.e., Urdu. To effectively perform sentiment classification in these scenarios, the study introduces a new dataset named UDSA-23, and leverages the Bert base-uncase to create word embeddings. Various deep learning-based classifiers are then trained on these generated word embeddings. The proposed approach (USA-BERT) tackles sentiment analysis for Urdu text by first preprocessing the provided reviews. This involves tokenization and subsequent vector generation for the preprocessed reviews. These vectors are used for both training and evaluating the USA-BERT model. The evaluation process employs the Pareto principle on two distinct datasets: the state-of-the-art UCSA-21 dataset and the newly introduced UDSA-23 dataset. This comparative assessment illustrates the superiority of USA-BERT. It outperforms existing methods by demonstrating remarkable enhancements in accuracy and f-measure, achieving improvements of up to 26.09% and 25.87%, respectively. There are plans to extend the validation of the USA-BERT approach to cover other low-resource languages as well.

ACKNOWLEDGMENT

The authors would like to thank the Ph.D. scholars and professionals of the Urdu Language, Institute of Languages and Linguistics, Punjab University, Pakistan, and also would like to thank Faiza Anwar, Govt. Graduate College, Islampura, Lahore, Pakistan, for their arrangements, evaluation, and constructive suggestions for creating the UDSA-23 dataset.

REFERENCES

[1] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[2] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study," *Appl. Sci.*, vol. 11, no. 9, p. 3986, Apr. 2021.

[3] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 845–863, Apr. 2022.
[4] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 805–831, Dec. 2017, doi: 10.1007/s10115-017-1055-z.
[5] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing*, vol. 2, N. Indurkhya, F. J. Damerau, Eds. Boca Raton, FL, USA: CRC Press, 2010, pp. 627–666.
[6] G. Pergola, L. Gui, and Y. He, "TDAM: A topic-dependent attention model for sentiment analysis," *Inf. Process. Manag.*, vol. 56, no. 6, Nov. 2019, Art. no. 102084.
[7] A. Khattak, M. Z. Asghar, A. Saeed, I. A. Hameed, S. A. Hassan, and S. Ahmad, "A survey on sentiment analysis in Urdu: A resource-poor language," *Egyptian Informat. J.*, vol. 22, no. 1, pp. 53–74, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866520301171
[8] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.
[9] N. Mukhtar and M. A. Khan, "Urdu sentiment analysis using supervised machine learning approach," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 02, Feb. 2018, Art. no. 1851001.
[10] K. Mehmood, D. Essam, and K. Shafi, "Sentiment analysis system for Roman Urdu," in *Intelligent Computing* (Sentiment Analysis System for Roman Urdu), vol. 858. 2019, pp. 29–42.
[11] Z. Sharf, D. Saif, and U. Rahman, "Performing natural language processing on Roman Urdu datasets," *Int. J. Comput. Sci. Netw. Secur.*, vol. 18, pp. 141–148, Jan. 2018.
[12] F. Noor, M. Bakhtyar, and J. Baber, "Sentiment analysis in e-commerce using SVM on Roman Urdu text," in *Emerging Technologies in Computing*, vol. 285. Berlin, Germany: Springer, 2019, pp. 213–222.
[13] M. A. Manzoor, S. Mamoon, S. Kei, A. Zakir, M. Adil, and J. Lu, "Lexical variation and sentiment analysis of Roman Urdu sentences with deep neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, 2020. [Online]. Available: www.ijacsa.thesai.org
[14] I. Rehman and T. R. Soomro, "Urdu sentiment analysis," *Appl. Comput. Syst.*, vol. 27, no. 1, pp. 30–42, Jun. 2022.
[15] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, "Multi-class sentiment analysis of Urdu text using multilingual BERT," *Sci. Rep.*, vol. 12, no. 1, p. 5436, 2022.
[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, NY, NY, USA, Aug. 2004, p. 168, doi: 10.1145/1014052.1014073.
[17] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov. 2017.
[18] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Exp. Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417409001626
[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003.
[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162
[22] W. Ahmad and M. Edalati, "Urdu speech and text based sentiment analyzer," 2022, *arXiv:2207.09163*.
[23] I. Qutab, K. I. Malik, and H. Arooj, "Sentiment classification using multinomial logistic regression on Roman Urdu text," *Int. J. Innov. Sci. Technol.*, vol. 4, no. 2, pp. 323–335, Apr. 2022. [Online]. Available: https://academic.oup.com/bjps/article-abstract/XVI/62/102/1473834?redirectedFrom=PDF
[24] L. Khan, A. Amjad, K. M. Afaq, and H.-T. Chang, "Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media," *Appl. Sci.*, vol. 12, no. 5, p. 2694, Mar. 2022.
[25] M. A. Qureshi, M. Asif, M. F. Hassan, A. Abid, A. Kamal, S. Safdar, and R. Akbar, "Sentiment analysis of reviews in natural language: Roman Urdu as a case study," *IEEE Access*, vol. 10, pp. 24945–24954, 2022.
[26] B. Chandio, A. Shaikh, M. Bakhtyar, M. Alrizq, J. Baber, A. Sulaiman, A. Rajab, and W. Noor, "Sentiment analysis of Roman Urdu on e-commerce reviews using machine learning," *Comput. Model. Eng. Sci.*, vol. 131, no. 3, pp. 1263–1287, 2022.
[27] A. A. Nagra, K. Alissa, T. M. Ghazal, S. Saigeeta, M. M. Asif, and M. Fawad, "Deep sentiments analysis for Roman Urdu dataset using faster recurrent convolutional neural network model," *Appl. Artif. Intell.*, vol. 36, Dec. 2022, Art. no. 2123094, doi: 10.1080/08839514.2022.2123094.
[28] Z. Mahmood, I. Safder, R. M. A. Nawab, F. Bukhari, R. Nawaz, A. S. Alfakeeh, N. R. Aljohani, and S.-U. Hassan, "Deep sentiments in Roman Urdu text using recurrent convolutional neural network model," *Inf. Process. Manag.*, vol. 57, no. 4, Jul. 2020, Art. no. 102233.
[29] L. Sana, K. Nasir, A. Urooj, Z. Ishaq, and I. A. Hameed, "BERS: Bussiness-related emotion recognition system in Urdu language using machine learning," in *Proc. 5th Int. Conf. Behav., Econ., Socio-Cultural Comput. (BESC)*, Nov. 2018, pp. 238–242.
[30] I. Safder, Z. Mahmood, R. Sarwar, S. Hassan, F. Zaman, R. M. A. Nawab, F. Bukhari, R. A. Abbasi, S. Alelyani, N. R. Aljohani, and R. Nawaz, "Sentiment analysis for Urdu online reviews using deep learning models," *Exp. Syst.*, vol. 38, no. 8, Dec. 2021, Art. no. e12751.
[31] U. Naqvi, A. Majid, and S. A. Abbas, "UTSA: Urdu text sentiment analysis using deep learning methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021.
[32] R. Batra, Z. Kastrati, A. S. Imran, S. M. Daudpota, and A. Ghafoor, "A large-scale tweet dataset for Urdu text sentiment analysis," Mendeley Data, V1, Tech. Rep., 2021, doi: 10.17632/rz3xg97rm5.1.
[33] H. Arif, K. Munir, A. S. Danyal, A. Salman, and M. M. Fraz, "Sentiment analysis of Roman Urdu/Hindi using supervised methods," in *Proc. ICICC*, vol. 8, 2016, pp. 48–53.
[34] M. Ijaz and S. Hussain, "Corpus based Urdu lexicon development," in *Proc. Conf. Language Technol.*, 2007.
[35] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Associating targets with SentiUnits: A step forward in sentiment analysis of Urdu text," *Artif. Intell. Rev.*, vol. 41, no. 4, pp. 535–561, Apr. 2014.
[36] S. Mukund and R. Srihari, "A vector space model for subjectivity classification in Urdu aided by co-training," in *Proc. Coling*. Beijing, China, Aug. 2010 pp. 860–868. [Online]. Available: https://aclanthology.org/C10-2099
[37] S. Mukund and R. K. Srihari, "NE tagging for Urdu based on bootstrap POS learning," in *Proc. 3rd Int. Workshop Cross Lingual Inf. Access Addressing Inf. Need Multilingual Societies*, 2009, pp. 61–69.
[38] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97803–97812, 2021.
[39] F. Mehmood, M. U. Ghani, M. A. Ibrahim, R. Shahzadi, W. Mahmood, and M. N. Asim, "A precisely xtreme-multi channel hybrid approach for Roman Urdu sentiment analysis," *IEEE Access*, vol. 8, pp. 192740–192759, 2020.
[40] K. Ullah, I. Mumtaz, M. A. Zia, and A. Razzaq, "Text based emotion detection by using classification and regression model," in *Proc. 16th Int. Conf. Manag. Sci. Eng. Manag.*, vol. 1, J. Xu, F. Altiparmak, M. H. A. Hassan, F. P. García Márquez, and A. Hajiyev, Eds. Cham, Switzerland: Springer, 2022, pp. 414–419.
[41] S. H. Kumhar, J. Sheetlani, and M. Hassan, "Sentiment analysis of Urdu language on different social media platforms using word2vec and LSTM," *Turkish J. Comput. Math. Educ.*, vol. 11, no. 3, pp. 1439–1447, 2020.
[42] A. Altaf, M. W. Anwar, M. H. Jamal, S. Hassan, U. I. Bajwa, G. S. Choi, and I. Ashraf, "Deep learning based cross domain sentiment classification for Urdu language," *IEEE Access*, vol. 10, pp. 102135–102147, 2022.
[43] A. Majeed, H. Mujtaba, and M. O. Beg, "Emotion detection in Roman Urdu text using machine learning," in *Proc. 35th IEEE/ACM Int. Conf. Automated Softw. Eng.*, Sep. 2020, pp. 125–130.
[44] T. Tehreem, "Sentiment analysis for Youtube comments in Roman Urdu," Feb. 2021, *arXiv:2102.10075*.
[45] U. Sehar, S. Kanwal, K. Dashtipur, M. Gogate, and F. Khan, "A hybrid dependency-based approach for Urdu sentiment analysis," Res. Square, Version 1, 2022, doi: 10.21203/rs.3.rs-1835013/v1.
[46] R. Habiba, D. M. Awais, and D. M. Shoaib, "A technique to calculate national happiness index by analyzing Roman Urdu messages posted on social media," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 6, pp. 1–16, Nov. 2020.

[47] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, "Attention-based RU-BiLSTM sentiment analysis model for Roman Urdu," *Appl. Sci.*, vol. 12, no. 7, p. 3641, Apr. 2022.
[48] M. Masood, F. Azam, M. W. Anwar, and J. Ur Rahman, "Deep-learning based framework for sentiment analysis in Urdu language," in *Proc. 2nd Int. Conf. Digit. Futures Transformative Technol. (ICoDT2)*, 2022, pp. 1-7. [Online]. Available: https://api.semanticscholar.org/CorpusID:249474819
[49] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2521–2548, Apr. 2020.
[50] K. B. Muhammad and S. M. A. Burney, "Innovations in Urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets," *Symmetry*, vol. 15, no. 5, p. 1027, May 2023. [Online]. Available: https://www.mdpi.com/2073-8994/15/5/1027
[51] N. Ahmed, R. Amin, H. Ayub, M. Iqbal, M. Saeed, and M. Hussain, "Urdu sentiment analysis using deep attention-based technique," May 2023. [Online]. Available: https://www.researchgate.net/publication/370750678_Urdu_Sentiment_Analysis_Using_Deep_Attention-based_Technique
[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[53] Y. Liu, "Fine-tune BERT for extractive summarization," 2019, *arXiv:1903.10318*.
[54] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data augmentation for BERT fine-tuning in open-domain question answering," 2019, *arXiv:1904.06652*.
[55] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Jun. 2017.
[56] N. Q. K. Le and T.-T. Huynh, "Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation," *Frontiers Physiol.*, vol. 10, p. 1501, Dec. 2019, doi: 10.3389/fphys.2019.01501.
[57] N. Q. K. Le, T. N. K. Hung, D. T. Do, L. H. T. Lam, L. H. Dang, and T.-T. Huynh, "Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104320. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521001141
[58] I. Illahi, H. Liu, Q. Umer, and N. Niu, "Machine learning based success prediction for crowdsourcing software projects," *J. Syst. Softw.*, vol. 178, Aug. 2021, Art. no. 110965.

MUHAMMAD REHAN ASHRAF received the M.Sc. and M.Phil. degrees from Quaid-i-Azam University Islamabad, Pakistan. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus, Pakistan. His research interests include machine learning, data mining, and digital image processing.

YASMEEN JANA received the M.C.S. degree from COMSATS University Islamabad, Vehari, Pakistan, in 2020, where she is currently pursuing the M.S. degree in computer science. She is also a Laboratory Engineer with COMSATS University Islamabad. Her research interests include machine learning and natural language processing.

QASIM UMER received the B.S. degree in computer science from Punjab University, Pakistan, in 2006, the M.S. degree in net distributed system development and the M.S. degree in computer science from the University of Hull, U.K., in 2009 and 2013, respectively, and the Ph.D. degree from the Beijing Institute of Technology, China. He is currently a Postdoctoral Researcher with the Department of Computer Science, Hanyang University, Seoul, South Korea, and an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus, Pakistan. His research interests include machine/deep learning, NLP, the IoTs, and developing practical tools to assist software engineers.

M. ARFAN JAFFAR received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, in March 2003, and the M.S. and Ph.D. degrees in computer science from the FAST National University of Computer and Emerging Sciences, in 2006 and 2009, respectively. He received a Postdoctoral Research Fellowship from South Korea and carried-out research at the top ranking Korean University, Gwangju Institute of Science and Technology, Gwangju, South Korea, from 2010 to 2013. He was an Assistant Professor with Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia, from March 2013 to August 2018. He is currently the Dean of the Faculty of Computer Science and Information Technology, Superior University, Lahore, Pakistan. He is also the Director of Intelligent Data Visual Computing Research (IDVCR). He is a Reviewer of 30 reputed international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, *Pattern Recognition*, and *Knowledge and Information Sciences*. His research interests include image processing, data science, machine learning, computer vision, artificial intelligence, and medical images.

SUNGWOOK CHUNG received the B.S. degree in computer science from Sogang University, South Korea, in 2002, and the M.S. and Ph.D. degrees from the Computer and Information Science and Engineering (CISE) Department, University of Florida, USA, in 2005 and 2010, respectively. From 2010 to 2012, he was a Research Engineer with Korea Telecom (KT), developing the IPTV network architectures and IPTV services. He has been an Associate Professor with the Department of Computer Engineering, Changwon National University, South Korea, since 2012. His research interests include the IoT network architectures and services, high-quality real-time content delivery and distribution, and high-performance computing configurations and services.

WAHEED YOUSUF RAMAY received the Ph.D. degree from the University of Science and Technology Beijing (USTB), China. He is currently an Assistant Professor with Air University, Islamabad. His academic and clinical focus is the use of algorithms (deep learning, machine learning, and big data analysis), advanced text analysis techniques, and sentiment analysis.