IECE

IECE Transactions on Intelligent Systematics
http://dx.doi.org/10.62762/TIS.2024.585616

RESEARCH ARTICLE

Check for updates

In-depth Urdu Sentiment Analysis Through Multilingual BERT and Supervised Learning Approaches

Muhammad Saeed[1], **Naeem Ahmed**[1,*], **Danish Ali**[2,*], **Muhammad Ramzan**[3], **Muzamil Mohib**[4], **Kajol Bagga**[5], **Atif Ur Rahman**[6] and **Ikram Majeed Khan**[7]

[1] School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] School of Computer Science, Wuhan University, Wuhan 430070, China
[3] Department of Information technology, the University of Haripur, 22620, Pakistan
[4] School of Business, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5] Department of Computer Science, Technical Hochshule (THWS) Wuerzburg, Germany
[6] Department of Computer Science, IQRA National University, Swat Campus Pakistan
[7] Coventry University, Priory St, Post code:CV1 5FB, Coventry, England,UK

Abstract

Sentiment analysis is the process of identifying and categorizing opinions expressed in a piece of text. It has been extensively studied for languages like English and Chinese but still needs to be explored for languages such as Urdu and Hindi. This paper presents an in-depth analysis of Urdu text using state-of-the-art supervised learning techniques and a transformer-based technique. We manually annotated and preprocessed the dataset from various Urdu blog websites to categorize the sentiments into positive, neutral, and negative classes. We utilize five machine learning classifiers: Support Vector Machine (SVM), K-nearest neighbor (KNN), Naive Bayes, Multinomial Logistic Regression (MLR), and the transformer-based multilingual BERT (mBERT) model. This model was fine-tuned to capture deep contextual embeddings specific to Urdu text. The mBERT model was pre-trained on 104 languages and optimized for Urdu-specific sentiment classification by fine-tuning it on the dataset. Our results demonstrated that the mBERT model significantly outperformed traditional classifiers, achieving an accuracy of 96.5% on the test set. The study highlights the effectiveness of transfer learning via mBERT for low-resource languages such as Urdu, making it a highly promising approach for sentiment analysis.

Keywords: machine learning, sentiment analysis, Urdu language, natural language processing (NLP), computational linguistics.

Academic Editor:
Habib Khan

Submitted: 03 October 2024
Accepted: 27 October 2024
Published: 09 November 2024

Vol. 1, **No.** 3, 2024.
10.62762/TIS.2024.585616

***Corresponding authors:**
Naeem Ahmed
naeem.uoh@gmail.com
Danish Ali
danishalikhan545@gmail.com

1 Introduction

Social media platforms, such as Facebook, Twitter, Tumblr, and Reddit, have become ubiquitous tools for communication, experience sharing, and opinion expression. A recent report by MediaKix reveals that people spend more time on social media than on

Citation
Saeed, M., Ahmed, N., Ali, D., Ramzan, M., Mohib, M., Bagga, K., Rahman, A. U., & Khan, I. M. (2024). In-depth Urdu Sentiment Analysis Through Multilingual BERT and Supervised Learning Approaches. *IECE Transactions on Intelligent Systematics*, 1(3), 161–175.

essential activities like eating, drinking, and socializing combined [1]. Further, a survey by Smart Insights indicates that an astonishing 3.3 million posts are published on Facebook and 4.5 million tweets are sent on Twitter every minute [2]. These statistics underscore the profound impact of social media on modern communication and engagement [3]. Blogs and social media websites contain various information types, such as product reviews, messages, news, etc. [2]. According to a survey, 65 billion messages are sent and received on WhatsApp per day [4], while approximately 9000 tweets per second are shared on Twitter. Another survey indicates that about 4.4 million blog posts are posted on the internet daily [5]. The data on social media platforms and blogs are in various languages, such as English, Urdu, Chinese, etc [6]. Text classification is used in multiple applications, including text filtering, document organization, news article categorization, web searching for useful data, etc [7]. These are language-specific systems [8], most of which are built for English. However, less work has been done on sentiment analysis of the Urdu language [9]. Developing a text classification model for the Urdu language is challenging due to its complex morphology [10]. Sentiment analysis (SA) can be done on three levels, i.e., document level, sentence level, and aspect level sentiment analysis [11]. The whole document is categorized as positive, negative, or neutral at the document level. In contrast, sentences are categorized in sentence-level SA [12].

On the other hand, the process of recognizing fine-grained opinion polarity towards a specific aspect associated with a given class is known as aspect-level SA [13]. Sentiment analysis can be done in two ways, i.e., machine learning and lexicon-based techniques. A dataset is created and used to train a machine-learning model [14]. After training, the model is used to predict the given data category. In the second technique, sentiment lexicons of the language are created. Every word in the lexicon is given a degree of positivity and negativity [15]. This degree indicates their class, i.e., positive, negative, and neutral [15]. This degree indicates their class, i.e., positive, negative, and neutral. Any document or sentence is classified into these classes based on the sum degrees of all words used in the sentences or document [16].

In the last decade, sentiment analysis has gained the attention of researchers. The applications of SA are everywhere and increasing day by day [6]. Researchers develop a lot of tools and techniques to analyze various languages [16]. SA models are facing multiple challenges. These challenges include sarcasm detection, negations, compound phrases, repetition of words, etc. [16]. Like some other languages, the Urdu language is widely used by individuals for data sharing on the internet [17, 18].

It is clear from the literature study that techniques used in SA for other languages cannot be used to deal with the Urdu language. Urdu sentiment analysis (UrSA) has been becoming popular for a few years because of its increasing rate on the internet [19]. There are a lot of applications of text analysis. One example of this is [20], which presents a novel approach to automatically detect political hate speech in Roman Urdu, a variant of Urdu written using the Latin alphabet [21]. To facilitate this task, a comprehensive dataset of Roman Urdu texts labeled for political hate speech (RU-PHS) was developed, containing 5002 instances along with city-level information [22]. To address the challenges posed by Roman Urdu's extensive lexical structure, a novel lexical unification algorithm was developed specifically for this language [11]. Furthermore, three vectorization techniques were employed to represent Roman Urdu text for machine learning: TF-IDF, word2vec, and fastText [23]. The results demonstrate that a random forest classifier and the proposed feedforward neural network achieved an accuracy of 93% when employing fastText word embeddings to distinguish between neutral and politically offensive speech. We have used different supervised ML models for sentiment analysis [24]. Various techniques that can be used for UrSA are illustrated in Figure 1.

We used supervised ML-based techniques for the sentiment analysis of Urdu text. The main contributions of the proposed article are listed below:

- We implemented supervised machine learning and transformer-based techniques specifically for sentiment analysis of Urdu text, utilizing probabilistic and linear classifiers.
- We used a dataset annotated by native Urdu speakers that ensures reliable sentiment labeling for enhanced analysis accuracy.
- We conducted comprehensive data preprocessing and feature extraction to prepare Urdu text for accurate analysis
- We evaluate the proposed models using various standard metrics for reliable model evaluations.

2 Related Works

. Exploring the sentiments expressed in text is very meaningful because the text is one of the easiest

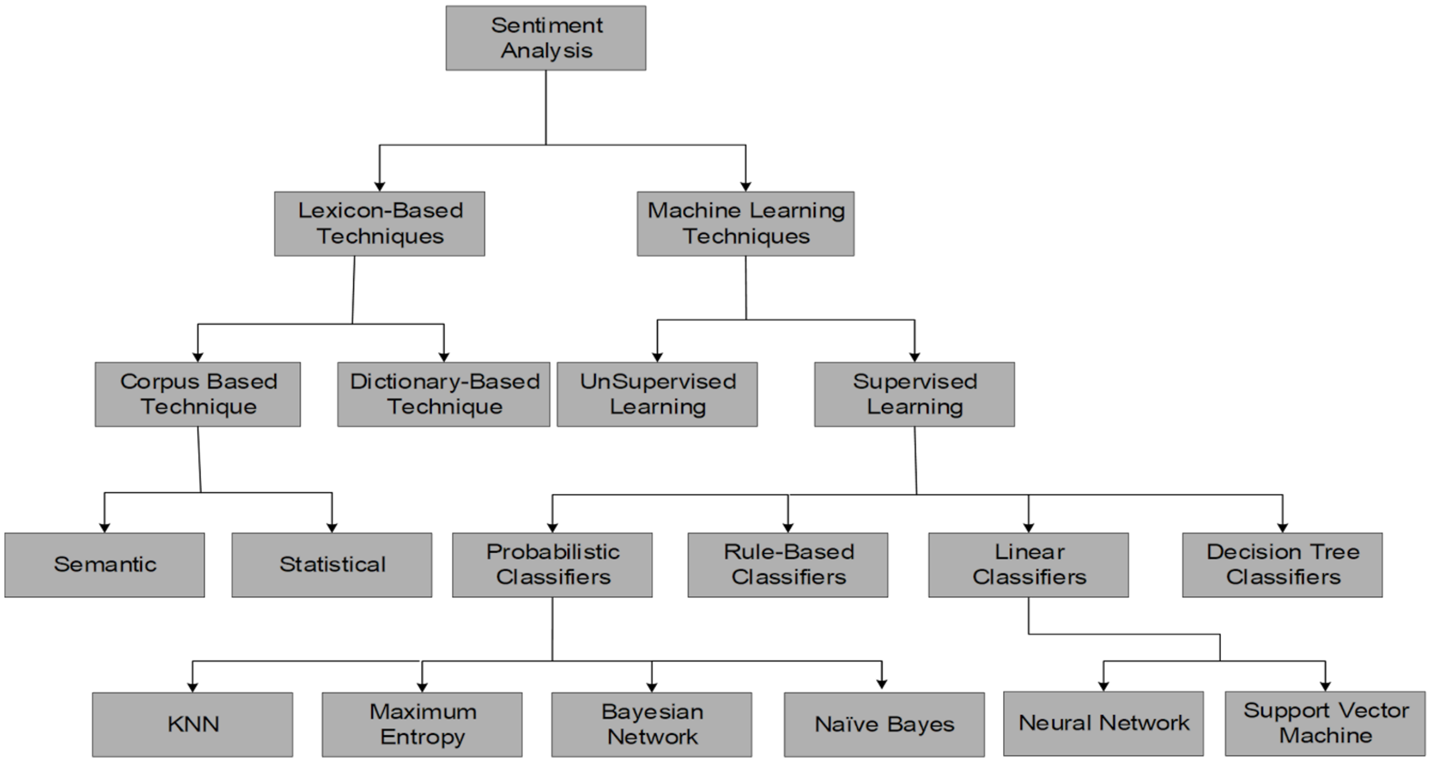


Figure 1. Techniques used for sentiment analysis.

and most effective ways to interpret and express emotions. This paper summarizes two types of recent research studies for textual sentiment analysis, i.e., traditional and machine learning techniques. Various conventional approaches have been employed to analyze textual data [20].

2.1 Roman Urdu Script

Mehmood et al. [16] aimed at enhancing the accuracy and effectiveness of sentiment analysis in Roman Urdu. To this end, they introduced a publicly available dataset consisting of 9006 features and 3241 sentiments in Roman Urdu. Pre-trained embeddings were also proposed for Roman Urdu using Word2vec, FastText, and Glove models. These embeddings can be utilized to improve the performance of various deep learning-based tasks related to Roman Urdu processing. The paper proposed a novel approach for Roman Urdu sentiment analysis, combining multiple neural network predictions, recurrent neural networks, and convolutional neural networks in an extreme-multi-channel hybrid methodology.
Khan et al. [21] presented a deep learning approach, combining CNN-LSTM, for sentiment analysis in both Roman Urdu and English languages. The analysis is conducted on user-generated text data from social media platforms. The study evaluated the performance of different word embedding models, namely Word2Vec (CBOW and skip-gram), GloVe, Fast Text, and TF-IDF words-to-vectors models, specifically for Roman Urdu text classification. Similarly, Chandio et al. [25] provided a new publicly available dataset consisting of over 26,824 labeled instances obtained from Daraz. pk and Twitter, annotated by field experts. They exploited three neural word embeddings—word2vec, Glove, and FastText to propose a novel attention-based BiLSTM model for sentiment analysis specifically designed for Roman Urdu.
Li et al. [26] used the combination of attention mechanisms and transfer learning to improve the accuracy and effectiveness of sentiment analysis in Roman Urdu. By emulating the attention mechanism observed in human brains, they aimed at prioritizing important words while disregarding less significant ones. Their approach involves using a convolutional neural network (CNN) with attention specifically designed for sentiment analysis in Roman Urdu. Transfer learning is applied to enhance the performance of sentiment analysis models when dealing with small datasets.

2.2 Urdu Script

Khan et al. [27] work presents a study on the sentiment analysis of Urdu text using ML. Firstly, a new dataset for Urdu sentiment analysis is introduced, comprising user reviews from diverse domains such as food and beverages, movies and plays, software and apps, politics, and sports. Secondly, a multilingual BERT model is fine-tuned specifically for Urdu sentiment classification. This model has been trained on 104 languages, including Urdu. It is based on a BERT base architecture with 12 layers, 768 hidden heads, and 110M parameters. Ahmed et al. [28] establishes a set of baseline results by evaluating rule-based approaches, various machine learning algorithms, and deep learning models, thereby creating a benchmark for multi-class sentiment analysis using different text representations. Another similar work [29], contributes by collecting a new sentiment analysis corpus in Urdu, manually annotated by experts. Baseline results are provided for state-of-the-art machine learning and deep learning models using two text representations. Additionally, the effectiveness of pre-trained word embeddings in resource-deprived languages like Urdu is examined, filling a gap in existing research.

Ahmed et al. [30] proposed a novel Urdu sentiment analysis mechanism that combines ML and DL models into a 2-tier ensemble model. Multiple deep learning models are trained using benchmark datasets with varying architectures, and their performance is compared to the proposed ensemble model. The study also investigates the impact of different types of deep learning predictions. It evaluates the efficacy of the approach in low-resource languages like Urdu, filling a gap in prior research. Another work by Sehar et al. [31]introduces novel applications of sentiment analysis. (SA) framework for the Urdu language, utilizing a combination of CNN and LSTM to extract unimodal and multimodal features. A multimodal Urdu dataset from YouTube is developed, enabling sentiment analysis implementation. The proposed framework is applied to determine sentiment polarity in Urdu videos and compared with a test-based SA approach.

Naqvi et al. [27] used a publicly available labeled dataset consisting of 6000 sentences of Urdu language. Their proposed framework explores various deep learning techniques for sentiment classification, with a focus on LSTM models' robustness in handling Urdu text. A Urdu sentiment analysis system using the RUSA data set was proposed by Mehmood et al. [32]. The data set contained 11,000 reviews of products. They presented three distinct techniques to achieve text normalization. They utilized six well-known phonetic algorithms and TERUN to optimize the RUSA data set. The resulting data was used for the training of machine learning models. According to the empirical review, the results obtained by TERUN were statistically significant and comparable to those obtained by well-known phonetic algorithms. The study concludes that text normalization enhances machine learning algorithms' accuracy rate. Another result was that a phonetic algorithm designed for one language would not generalize well to other languages unless it becomes properly updated to fulfill its target languages' phonological needs.

Ahmed et al. [33] Used the generative adversarial networks and recurrent neural networks (RNN), this article suggested a multi-task model called slogan generative adversarial network systems (Slo-GAN) to improve coherence and diversity in slogan production. Slo-GAN filtered the text with the higher reporting score level, achieving an 87.2% classification accuracy. Nasim et al. [34] present an article on Urdu sentiment analysis by combining various linguistic and lexical features. Their work focuses on developing the UrSA system for Urdu tweets. A Markov chain model was used to design the approach in this paper. They gathered the data with the help of Twitter's API. The proposed model was trained on that data, and the model was able to predict people's attitudes based on their tweets. They also discussed the challenges and limitations of UrSA systems. A word-level translation framework was proposed by Asghar et al. [35]to enhance the UrSA lexicon. The framework was developed by combining different linguistic and lexicon resources, such as the English word list, SentiWord Net, the bilingual English-to-Urdu dictionary, Urdu grammar improvements, and a novel scoring mechanism. Their model consisted of three major modules, i.e., the collection of words in English for an opinion, the translation of English words into Urdu, and sentiment scoring using SentiWordNet and manual scoring. Figure 2 illustrates the process of lexicon-based technique for sentiment analysis.

Mukhtar et al. [37] presented a supervised machine-learning technique for Urdu SA. The data of various blogs and categories, i.e., sports, politics, products, etc., were used for classification purposes. This study trained three supervised ML models, namely KNN, LibSVM, and J48, to classify Urdu data. After the successful training and testing phases, all models were compared with accuracy, precision, and recall. Lib SVM was very similar to KNN in terms of efficiency and its accuracy is on par with KNN on average, but it is the slowest of the three classifiers used in the analysis. Table 1 presents the comparison of previous studies.

Table 1. Comparison of previous studies.

Authors	Study	Models	Accuracy
Mehmood et al. [16]	Roman Urdu	ML, DL and Hybrid Learning	84%
Khan et al. [21]	English and Roman Urdu	LSTM and SVM	90%
li et al. [26]	Roman Urdu	Transfer Learning	94%
Khan et al. [38]	Urdu	ML and DL	82%
Ahmed et al. [33]	Urdu	RNN	87%
Ahmed et al. [30]	Urdu	BERT base multilingual model	84%
Sehar et al. [31]	Urdu	Multimodal Sentiment Analysis	95%
Asghar et al. [35]	Urdu	Statement lexicon	Not Mentioned
Mukhtar et al. [36]	Urdu	Sentiment Analyzer	89%
Mukhtar et al. [37]	Urdu	Sentiment Ananlysis	83%
khan et al. [38]	Urdu	Sentiment Classification Techniques	Not Mentioned
Rehman et al. [10]	Urdu	Sentiment Ananlysis	66%

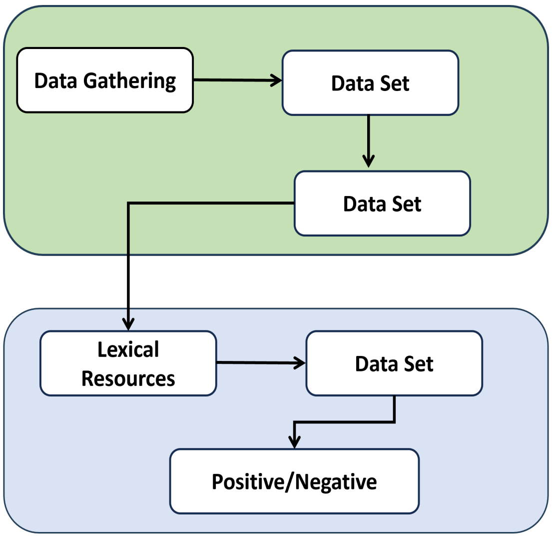


Figure 2. Lexicon-based technique for sentiment analysis.

Mukhtar et al. [36] discussed that handling intensifiers is crucial to maintain higher accuracy in Urdu SA. Urdu intensifiers were gathered and saved in a different place. Based on the experiments conducted, accuracy improved by 5%, indicating a statistically significant improvement in sentence classification accuracy.

Rehman et al. [10]proposed a novel framework for Urdu sentiment analysis using Urdu comment data. The polarities were assigned to Urdu sentence-generated tokens. The lexicon has 7335 entries, of which 2607 were negative and 4728 positive. The absolute polarity of the sentence was calculated by summing up the contradictions in all the respective words. Their proposed model got an accuracy of 66on the testing data set.

Khan et al. [38] discussed that many potential methodologies are available for SA. Still, little work has been done on the analysis of Urdu sentiments. This paper examines the increasing rate of Urdu language on the internet and the need for UrSA systems.Their article outlines and summarizes the most recent SA updates and classification techniques used in the Urdu language. Various suggestions and improvements were suggested in this article for UrSA.

3 Proposed System

We present a comprehensive approach for sentiment analysis of Urdu text employing a diverse set of machine and deep learning models. The methodology encompasses both traditional supervised

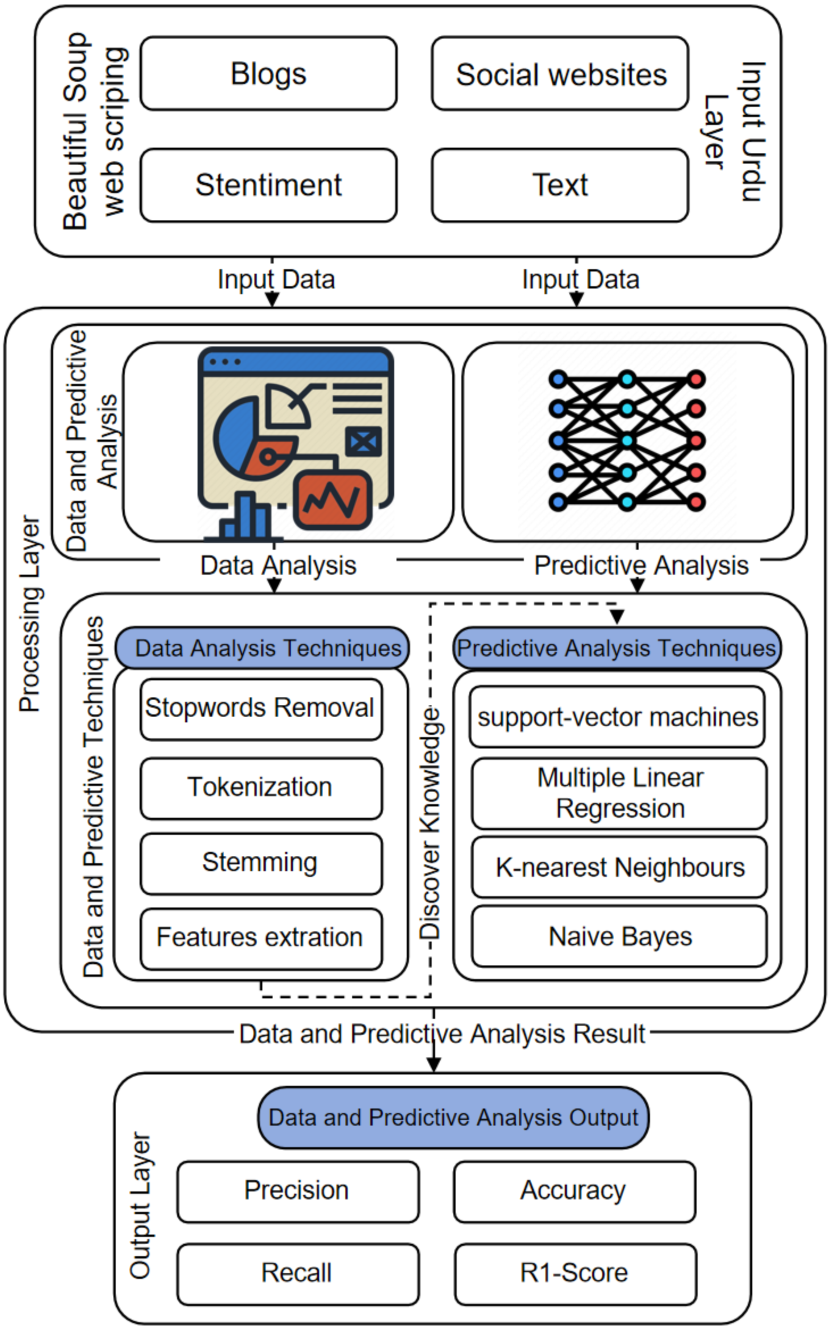


Figure 3. Process of Proposed Model.

machine learning classifiers, including Support Vector Machine(SVM), K-Nearest Neighbor (KNN), Naive Bayes, Multinomial Logistic Regression (MLR), and a state-of-the-art deep learning model using pre-trained multilingual BERT (mBERT). Feature extraction of data was done using bag-of-words representations for traditional classifiers and fine-tuning mBERT for capturing deep contextual embeddings. The proposed methodology aims to provide a robust framework for sentiment analysis for Urdu text, addressing linguistic nuances and offering insights into the efficacy of different models. Figure 3 illustrates the proposed process of the study.

The proposed approach used for UrSA consists of three layers. Using a beautiful soup web scraping tool, the first layer of data was gathered from various Urdu blogs. After collecting data, it was not in a standard form. The second layer is data preprocessing; it is performed in which unnecessary data, like stop words, URLs, white spaces, etc., are removed from the data. Various steps were performed in the preprocessing phase. These layers include filtering URL links, tokenizing words, removing stop words and blank spaces, removing redundant letters such as non-Urdu

Algorithm 1: Preprocessing of Dataset

Input: Raw Dataset
Output: Cleaned Dataset
Function `Preprocess(`*Raw Dataset*`)`:
  Dataset ← Read_Dataset_Reviews;
  for *sentence* **in** *Dataset* **do**
    Reviews ← sentence.split(" ") ;   `// Split sentence into words`
    Cleaned data ← `Remove_Noise(`*Reviews*`)` ; `// Cleaning`
    Cleaned data ← `Remove_English_Alphabets(`*Cleaned data*`)`;
    New_Data ← `Append_data(`*Cleaned data*`)`;
  end

letters and numbers, and normalization of Urdu text. Algorithm 1 shows the process for preprocessing of the dataset.

After that, a contiguous sequence of "n" terms from a given sequence of text was constructed as a unigram (n=1), bigram (n=2), and trigram (n=3). Urdu stop words were used to remove those stop words in this study. A list of 650 stop words was created for this purpose. Some of the stop words are shown in Figure 4.

Some Urdu stopwords			
آئی	آئے	آپ	آپکی
آپکے	آپ کو	آپکے	آپ کی
آپکے	ہم	اُس	اُسکا
اُسکی	اُسکو	اُسکے	اِس
اِسکا	اِسکی	اِسکو	اِسکے
اُن	اُنکا	اُنکی	اُنکو
اُنکے	اُنہوں	اُنکے	اُنہیں
تمہارے	تمہیں	اُسکی	اُسکو
تجھ	تجھے	اِسکا	اِسکی
تم	تمہارا	اُن	اُنکا

Figure 4. Some Stop Words of Urdu Language.

The data set was divided into training and testing sets using the sklearn Standard Library. Seventy percent of the data was used for training of proposed models, while 30 percent was used for testing purposes. In the next step, various features were extracted from the cleaned data to train the proposed models. Data cleaning, feature extraction, tokenization, and stemming are done using the Urdu Hack Library. After feature extraction, model training was done on the training data set and tested on the testing dataset. The data set is divided into three categories. From the data, 37% of data was labeled as positive, 35% as Negative, and 28% as neutral. Figure 5 presents the division of the dataset.

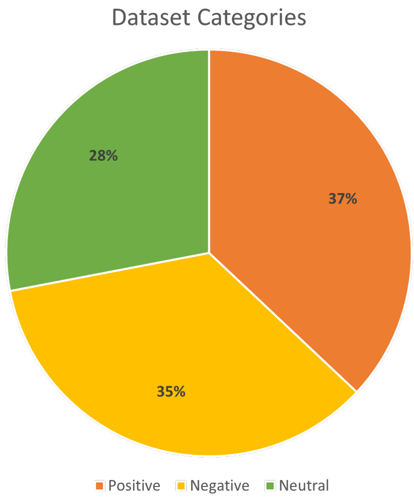


Figure 5. Division of Dataset.

For the classification of reviews as positive, negative, or neutral, various supervised ML models, i.e., Support Vector Machine (SVM), Naive Bayes, KNN, Multinomial logistic regression, and transformer-based model, i.e., mBERT are applied. For the testing of classifiers, the cross-testing technique was also used for reliable results. The k-10 fold cross-validation method is used for the evaluation. We searched thoroughly for all parameters. Algorithm 2 presents the training and testing process of the proposed models.

3.1 Proposed Models

We have applied various machine learning and a transformer-based model Urdu text sentiment analysis. The description of all models is discussed below.

3.1.1 *Support Vector Machine*

SVM is a flexible algorithm that searches for the best hyperplane in the feature space to divide multiple sentiment classes. It works well for Urdu sentiment analysis because it draws a distinct line between neutral, negative, and positive opinions. The decision function for the proposed problem can be explained by equation (1).

$$f(x) = \text{argmax}(w_i \cdot X + b_i) \tag{1}$$

where $f(x)$ represents the decision function of SVM, w_i and b_i are the weight vector and bias for each sentiment

Algorithm 2: Training of Proposed Models (Traditional Models and mBERT)

Input: Dataset (Reviews)
Output: Labels
Function `Preprocess()`:

```
    Transform(Dataset) ;   // Preprocessing step
    X_train, X_test, train_y, test_y ←
      train_test_split(data['review'],
      data['sentiment'], 0.3) ;      // Split dataset
    // Model initialization
    ML_models ← Model.fit(X_train, train_y) ;
      // Fitting traditional models
    tokenizer ←
      BertTokenizer.from_pretrained('bert-base-
    multilingual-cased');
    model ←
      BertForSequenceClassification.from_pretrained
    ('bert-base-multilingual-cased',
      num_labels=num_classes);
    input_ids, attention_masks, labels ←
      Tokenize_and_Prepare_Data(X_train, y_train,
      tokenizer);
    dataloader ← Create_Dataloader(input_ids,
      attention_masks, labels, batch_size);
    optimizer, loss_function ←
      Initialize_Optimizer_and_Loss_Function(model);

    // Training mBERT model
    model.train();
    for epoch to epochs do
        for batch to tqdm(dataloader, desc="Training
         mBERT") do
            optimizer.zero_grad();
            // Train mBERT model
        end
    end
    // Evaluate mBERT model
    model.eval();
    bert_accuracy ← accuracy_score(y_test,
      predictions);
    print("mBERT Evaluation:", bert_accuracy,
      precision, recall, F1-score);
    // Evaluation for traditional models
    Model_evaluate ← Evaluate_Model(X_test,
      test_y);
    Print("Evaluation of ML models: ", accuracy,
      precision, recall, F1-score);
```

class. The model predicts the sentiment class that maximizes the decision function.

3.1.2 K-Nearest Neighbors

The KNN algorithm is a popular and flexible machine-learning method that is mostly employed for its simple design. The K-NN algorithm locates the K closest neighbors, using a distance metric like Euclidean distance, to a given data point. The majority of votes or the average K neighbors is then used to establish the class or value of the data item. KNN can be utilized for sentiment analysis since it uses the sentiments of its closest neighbors to classify a given Urdu text. The classification is based on the opinion that the majority of the k-nearest neighbors hold. This process can be represented by equation (2).

$$y = \text{maj\_class} = \text{k\_nearest\_neighbors}(x, X, y) \tag{2}$$

where y is the predicted class while x represents the input text. X and y represent the nearest neighbors.

3.1.3 Naïve Bayes

Naive Bayes is a probabilistic model that calculates the probability of each sentiment class given the features of the Urdu text. It's particularly useful for sentiment analysis due to its simplicity and efficiency in handling multiple classes. It can be used to calculate the probability of each sentiment class given the features of the Urdu text. The process of Urdu sentiment analysis can be explained by equation (3).

$$P(\mathrm{S} = i \mid \mathrm{U}) \propto P(\mathrm{S} = i) \prod_{j=1}^{n} P(\mathrm{W}_j \mid \mathrm{S} = i) \tag{3}$$

where $P(Sentiment = i|UrduText)$ is the posterior probability of sentiment given the text while $P(Wordj|Sentiment = i)$ is the likelihood of word "j" given sentiment "i". The model calculates probabilities and selects the sentiment class with the highest probability.

3.1.4 Multinomial Logistic Regression (MLR)

MLR is a generalized linear model suitable for multi-class sentiment analysis. It models the probability of each sentiment class using a softmax function, making it well-suited for Urdu sentiment classification with three classes. The process of Urdu sentiment analysis using MLR can be explained using equation (4).

$$P(\text{Sentiment} = i \mid \text{Urdu Text}) = \frac{e^{w_i \cdot x + b_i}}{\sum_{j=1}^{3} e^{w_j \cdot x + b_j}} \tag{4}$$

where, $P(Sentiment = i|UrduText)$ is the probability of sentiment i given the text, and w_i and b_i are the weight vector and bias for sentiment class i. The

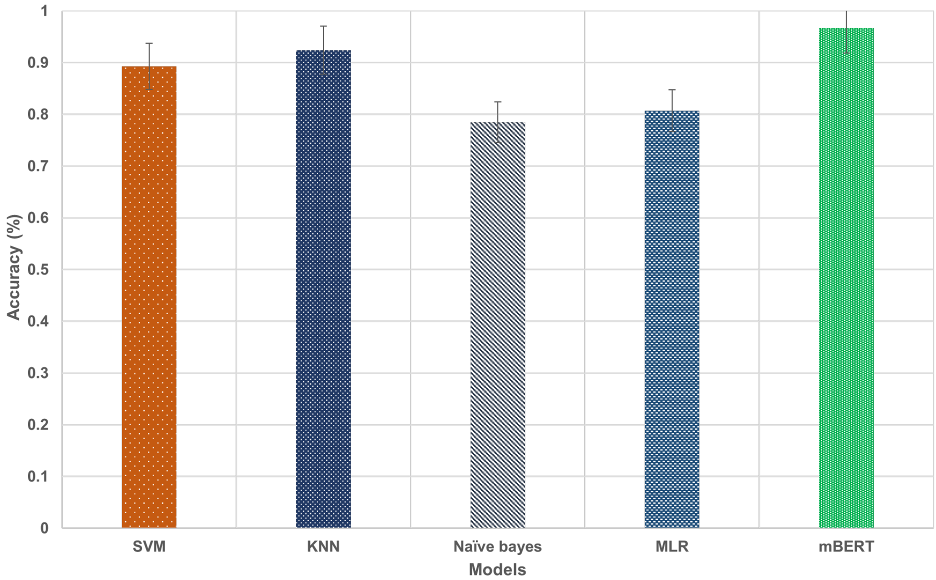


Figure 6. Accuracy Comparison of Proposed Models.

softmax function is used to normalize the probabilities, ensuring they sum to 1.

3.1.5 *Multilingual BERT (mBERT)*

mBERT is a powerful contextual embedding model that can be used to capture the contextual information of words in Urdu text. Fine-tuning mBERT for sentiment analysis enables it to understand the nuanced relationships between words in the given text. The mBERT model was fine-tuned for sentiment classification using the preprocessed Urdu dataset. Transfer learning was employed to leverage the knowledge captured during the multilingual pre-training phase. Pre-trained multilingual BERT (mBERT) embeddings were used for training the mBert Model. mBERT has 12 transformer layers, whereas BERT large has 24 transformer layers. The underlying architecture, known as the Transformer, is a paradigm in natural language processing designed for sequence-to-sequence tasks with long-range dependencies. Transformers consist of encoders and decoders, with each encoder comprising two components: Multi-Head Attention and Feed Forward Neural Network. The Decoder includes Masked Multi-Head Attention along with Multi-Head Attention Feed Forward Neural Network. Encoders and decoders are stacked on top of each other in the implementation. The Transformer heavily relies on attention mechanisms, particularly self-attention, to capture the contextual understanding of a word in a text based on neighboring words in the sentence. The sentiment classification mBERT model underwent a two-phase training process. The initial phase involved pre-training the mBERT language model, and the subsequent phase focused on fine-tuning the outermost classification layer. Fine-tuning was conducted using the training set of the proposed dataset, comprising labeled user reviews. Specifically, the fully connected classification layer underwent training, employing categorical cross-entropy as the loss function during the training process. The output of mBERT model can be described as in equation (5).

$$\text{Output} = \text{BERT}(\text{Tokenized Urdu Text}) \tag{5}$$

where, the output is the contextual embedding obtained from mBERT. The Tokenized Urdu text represents the input text tokenized for mBERT. Table 2 presents the hyperparameters for the proposed mBERT model.

Table 2. Hyperparameters and their Values.

Hyperparameters	Value
Learning rate	0.00002
Batch size	32
Number of epochs	10
Attention heads	8
Gradient accumulation steps	4
Hidden size	512
Hidden layers	6
Maximum sequence length	256
Parameters	85 M

4 Results and Discussion

The features were chosen using the Xlm Roberta framework. The Urdu data set was trained with unigrams and bigrams, SVM, NB, Multinomial logistic regression, and KNN models. Figure 6 represents the accuracies of the proposed models for UrSA. The results are analyzed and compared with previous studies on UrSA using machine learning models. The results clearly show that the proposed model mBert outperforms other models and studies discussed in the literature review section. One reason for the better accuracy of the model is the manual annotation of data, and the second is the usage of both manual and automatic preprocessing techniques for the Urdu language. The results can further be improved by using deep learning techniques for UrSA.

Evaluating the classification model just based on accuracy is not enough especially in the case of an imbalanced dataset. We evaluate the models based on the confusion matrix so that the performance of the proposed models can be judged easily. A confusion matrix is a matrix that shows how well a classification model ("classifier") performs on a set of test data for which the correct labels are known. The confusion matrix is a simple and easy way to understand the performance of machine learning models. This matrix contains four elements true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). Figure 7 presents the confusion matrix for the SVM model.

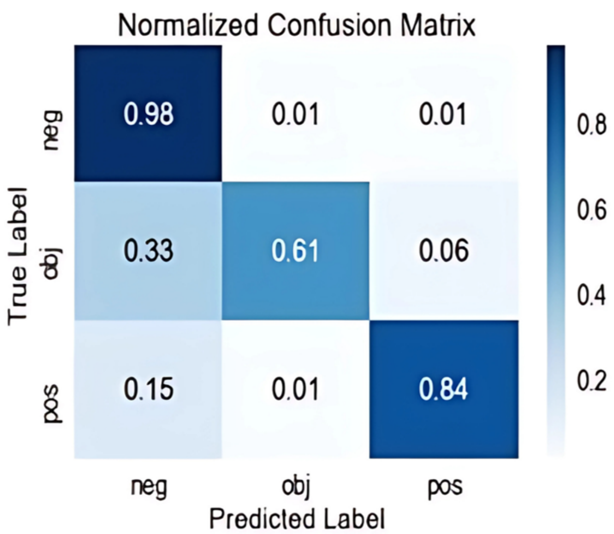


Figure 7. Confusion matrix of SVM.

Figure 7 illustrates the SVM classifier's confusion matrix for the proposed Urdu data set. Figure 8 presents the confusion matrix of the KNN model.

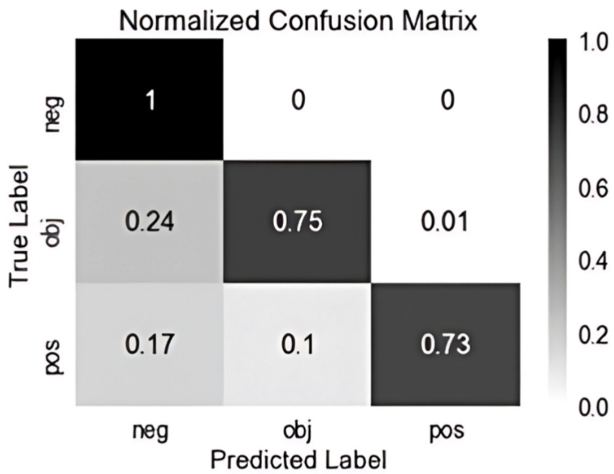


Figure 8. Confusion matrix of KNNs model.

Off-diagonal elements show classification errors, with 0.33 percent of label objective misclassified as negative and 0.15 percent of label positive misclassified as negative. Figure 9 presents the confusion matrix for the Naïve Bayes classifier. Nave Basie outperformed SVM in classifying the majority mark, as seen in the first element of Figure 9. Since Naïve Bayes is dependent on probabilities, the majority class's prior probabilities exceeded the minority classes. Therefore, for the most part, all data points were categorized as members of the majority class. Figures 10 and 11 demonstrate the major deficiencies of Multinomial Logistic Regression and KNN incorrectly classifying minority instances. SVM also outperforms Naive Bayes, Multinomial Logistic Regression, and KNN to classify minority cases correctly. The KNN's low performance, shown in Figure 8, is due to its propensity to overgeneralize the majority instance, specifically when there is a large class imbalance. Neighbors often surround minority data points from the majority class, and the chances of being classified as a majority are high.

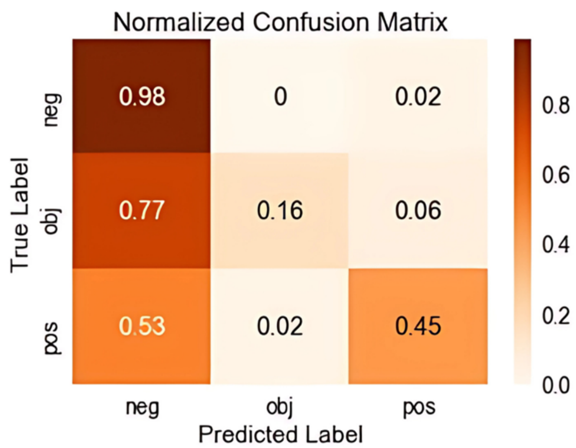


Figure 9. Confusion matrix for the Naïve Bayes.

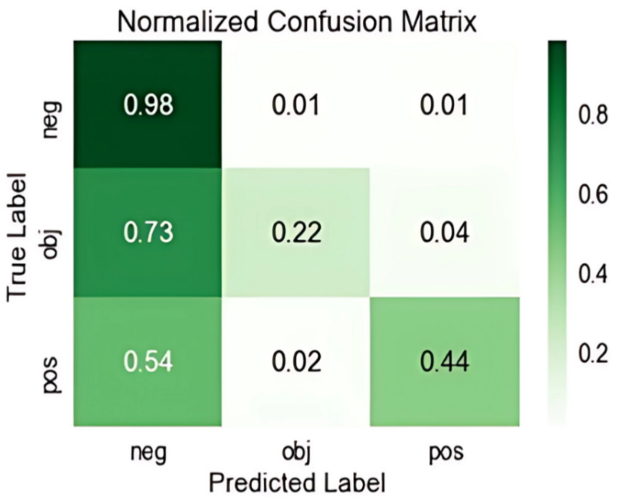


Figure 10. Confusion matrix for MLR model.

Our proposed mBERT model achieved state-of-the-art results on UrSA, outperforming all other models on the proposed dataset. The mBERT model achieved an accuracy of 96.5%, a precision of 95.35%, and a recall of 94.42% on the test dataset. The results demonstrate the effectiveness of the mBERT model for Urdu sentiment analysis. Figure 11 presents the normalized confusion matrix for the proposed mBERT model.

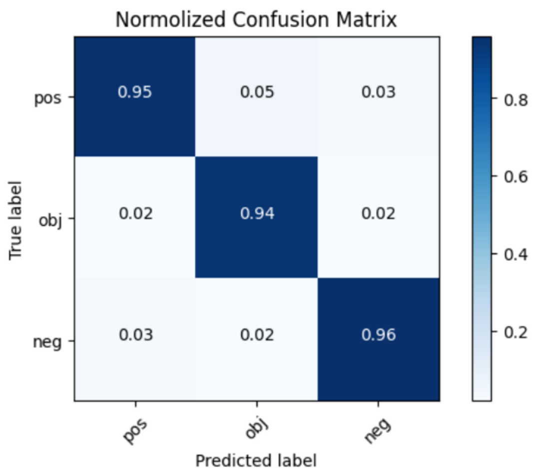


Figure 11. Confusion matrix for mBERT model.

5 Discussion

Sentiment analysis can be used for a number of issues, including evaluating campaign performance, analyzing stock market patterns, monitoring brand reputation, and studying consumer feedback. The increasing significance of sentiment analysis in today's information era has led to research on the subject, particularly in Urdu, which is spoken by over 100 million people globally. With the internet becoming a ubiquitous platform for expressing opinions and sharing reviews, the demand for robust sentiment analysis tools in Urdu has become apparent. However, existing sentiment analysis techniques, primarily developed for English, encounter difficulties when applied to Urdu due to its complex morphological structure and script disparities. The significant role of sentiment analysis in guiding decision-making processes for producers, service providers, and organizational leaders further propels this research. User-generated information from blogs and social media sites is a useful tool for determining public opinion. For example, in election conditions, social media sentiment provides information about the popularity of political leaders and parties, which can help with strategic preparation. The proposed study aims to harness this resource by conducting sentiment analysis in Urdu, thereby facilitating informed decision-making processes.

As Urdu becomes increasingly prevalent on social media platforms and blogging websites, UrSA can serve the purposes mentioned earlier effectively.

Experimental results of the proposed ML models demonstrate their superiority over traditional models for UrSA. These models adeptly capture semantic nuances in text, using word embeddings to incorporate word meanings in a specific context.

The study's data is sourced from diverse channels, as detailed in Section 3. Annotation, a crucial aspect of model efficiency, involves manual annotation by three annotators. The proposed mBERT model demonstrates superior performance compared to traditional machine learning models, making it suitable for analyzing Urdu sentiment across multiple tasks such as assessing customer feedback, monitoring brand reputation, and tracking campaign effectiveness. However, to enhance the accuracy of UrSA (SA), further research utilizing advanced techniques is imperative, particularly to address challenges like detecting sarcasm which can undermine automated model accuracy. Strategies like ensemble learning and hybrid models hold promise for achieving enhanced accuracy, especially when dealing with larger datasets. Data acquisition spans numerous websites catering to various services and products.

6 Comparison With Previous Studies

Different techniques have been proposed to understand and interpret the sentiments expressed in the Urdu language. This section compares various ML and DL-based techniques with the most efficient

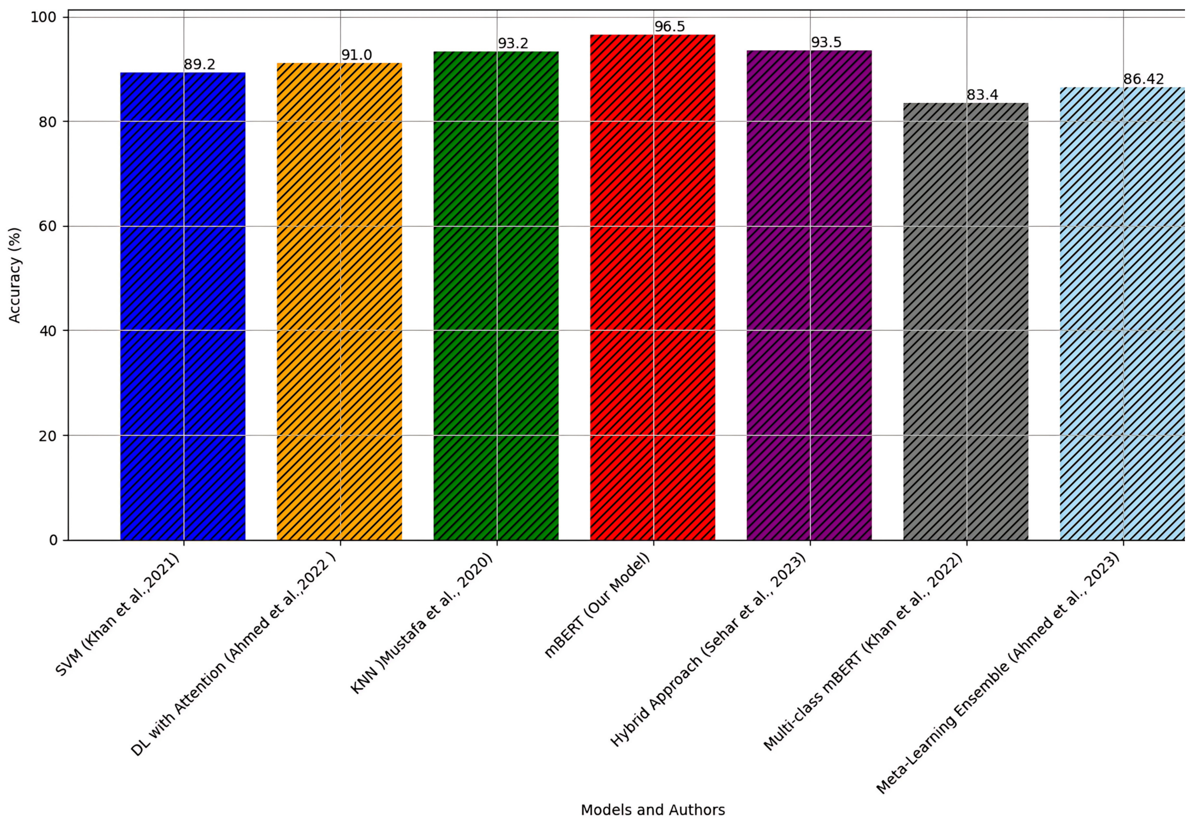


Figure 12. Comparison with Previous Studies.

proposed model (mBERT) for the task of UrSA. Traditional machine learning classifiers have been widely used in Urdu text analysis. For instance, Khan et al. [27] reported an accuracy of 89% using SVM, while Ahmad et al. [15] achieved a 91% accuracy with a deep attention model. Khan et al. [29] also evaluated and demonstrated the effectiveness of mBERT for UrSA. They achieved an F1 score of 81.49% and claimed that mBERT outperformed other classifiers due to its BERT pre-trained word embeddings. The finding highlights the capability of transformer-based models to manage the complicated processes involved in UrSA. Sehar et al. [31] presented a hybrid dependency-based technique for Urdu sentiment analysis that offers a new perspective on the processing of Urdu text. The approach emphasizes the possibility of integrating different techniques for improved accuracy. Ahmed et al. [30] also presented a contextually enriched meta-learning ensemble model that combines the strengths of several classifiers to increase performance. Mustafa et al. [39] focused on sentiment analysis in Urdu texts using SVM, NB, and KNN classifiers designed on a dataset of 6,000 tweets. The results of this study showed that SVM outperformed NB and KNN in terms of accuracy and other metrics. In our research, we used these classifiers alongside mBERT and found that, although traditional techniques are useful, they fall short in terms of effectiveness as compared to mBERT. This is in line with the results of Hussain et al. [39], who pointed out that deep learning techniques are more effective at capturing semantic depth than traditional ML techniques. The performance of mBERT in our work not only establishes a new standard for Urdu sentiment analysis but also highlights the ability of multilingual models for low-resource language processing. This finding encourages further exploration of transfer learning and multilingual models for languages with limited computational resources. Figure 12 shows the comparison of the proposed mBERT model with previous studies.

7 Conclusion and Future Work

In this article we focused on Urdu sentiment analysis using supervised machine-learning techniques. The data set from various blogs and e-commerce websites consists of 4712 reviews. These reviews were manually labeled and annotated by native speakers for better accuracy. The data was preprocessed using the Urdu Hack Python Library and divided into training and testing data sets. The supervised ML algorithms, i.e., Naïve Bayes, SVM, KNN, MLR, and transformer-based mBERT were trained and validated within the test data set. The KNN model performed well by obtaining 92.0% KNN (k=3), while the mBERT model outperformed all models by achieving 96.5% accuracy. In future work, we will work on sarcasm detection from Urdu data to enhance the models' accuracy. Domain adaptation and cultural sensitivity are an issue in sentiment analysis because the same text might have a different meaning or connotation based on culture. Deep learning techniques should also be considered for UrSA.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work was supported without any funding.

References

[1] Mukhtar, N., & Khan, M. A. (2018). Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02), 1851001.[CrossRef]

[2] Ghulam, H., Zeng, F., Li, W., & Xiao, Y. (2019). Deep learning-based sentiment analysis for roman urdu text. *Procedia computer science*, 147, 131-135.[CrossRef]

[3] Ali, D., Huque, M. T., Godhuli, J. J., & Ahmed, N. (2022). Detection of Face Emotion and Music Recommendation System using Machine Learning. *International Journal of Research and Innovation in Applied Science*, 7(11), 05-08.

[4] Ali, D., Iqbal, S., Mehmood, S., Khalil, I., Ullah, I., Khan, H., & Ali, F. (2024). Unleashing the Power of AI in Communication Technology: Advances, Challenges, and Collaborative Prospects. In *Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies* (pp. 211-226). Singapore: Springer Nature Singapore.

[5] Amin, R., Gantassi, R., Ahmed, N., Alshehri, A. H., Alsubaei, F. S., & Frnda, J. (2024). A hybrid approach for adversarial attack detection based on sentiment analysis model using Machine learning. *Engineering Science and Technology, an International Journal*, 58, 101829.[CrossRef]

[6] Khan, H., Ullah, I., Shabaz, M., Omer, M. F., Usman, M. T., Guellil, M. S., & Koo, J. (2024). Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset. *Image and Vision Computing*, 149, 105195.[CrossRef]

[7] Dar, G., Bhagat, A., Ansarullah, S., Othman, M., Hamid, Y., Alkahtani, H., Ullah, I. & Hamam, H. (2023). A novel framework for classification of different Alzheimer's disease stages using CNN model. *Electronics*, 12, 469.

[8] Ali, D., Younis, B., & Iqbal, S. (2024). A DEEP LEARNING TECHNIQUE FOR CLASSIFYING IMAGES OF BRAIN TUMOR. *Exceed Journal of Biological and Computer Sciences*, 1(1).

[9] Mukhtar, N., Khan, M. A., & Chiragh, N. (2017). Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis. *Cognitive Computation*, 9, 446-456.

[10] Rehman, Z. U., & Bajwa, I. S. (2016, August). Lexicon-based sentiment analysis for Urdu language. In *2016 sixth international conference on innovative computing technology (INTECH)* (pp. 497-501). IEEE.[CrossRef]

[11] Khan, H., Ullah, M., Al-Machot, F., Cheikh, F. A., & Sajjad, M. (2023). Deep learning based speech emotion recognition for Parkinson patient. *Electronic Imaging*, 35, 298-1.

[12] Saeed, M., Ahmed, N., Mehmood, A., Aftab, M., Amin, R., & Kamal, S. (2023). Sentiment analysis for covid-19 vaccine popularity. *KSII Transactions on Internet and Information Systems (TIIS)*, 17(5), 1377-1393.[CrossRef]

[13] Khan, H., Hussain, T., Khan, S. U., Khan, Z. A., & Baik, S. W. (2024). Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, 237, 121288.[CrossRef]

[14] Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2019). Discriminative feature spamming technique for roman urdu sentiment analysis. *IEEE Access*, 7, 47991-48002.[CrossRef]

[15] Ahmed, N., Amin, R., Ayub, H., Iqbal, M. M., Saeed, M., & Hussain, M. (2022). Urdu Sentiment Analysis Using Deep Attention-Based Technique. *Foundation University Journal of Engineering and Applied Sciences (HEC Recognized Y Category, ISSN 2706-7351)*, 3(1), 1-12. [CrossRef]

[16] Mehmood, F., Ghani, M. U., Ibrahim, M. A., Shahzadi, R., Mahmood, W., & Asim, M. N. (2020). A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access*, 8, 192740-192759. [CrossRef]

[17] Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2010). Lexicon based sentiment analysis of Urdu text using SentiUnits. In *Advances in Artificial Intelligence: 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, November 8-13, 2010, Proceedings, Part I 9* (pp. 32-43). Springer Berlin Heidelberg.

[18] Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges. *Security and Communication Networks*, 2022(1), 1862888. [CrossRef]

[19] Pal, R., Adhikari, D., Heyat, M. B. B., Ullah, I., & You, Z. (2023). Yoga meets intelligent internet of things: recent challenges and future directions. *Bioengineering*, 10(4), 459. [CrossRef]

[20] Rajalakshmi, S., Asha, S., & Pazhaniraja, N. (2017, March). A comprehensive survey on sentiment analysis. In *2017 fourth international conference on signal processing, communication and networking (ICSCN)* (pp. 1-5). IEEE. [CrossRef]

[21] Khan, L., Amjad, A., Afaq, K. M., & Chang, H. T. (2022). Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. *Applied Sciences*, 12(5), 2694. [CrossRef]

[22] Ahmed, N., Amin, R., Aldabbas, H., Saeed, M., Bilal, M., & Song, H. (2024). A Novel Approach for Sentiment Analysis of a Low Resource Language Using Deep Learning Models. *ACM Transactions on Asian and Low-Resource Language Information Processing*.[CrossRef]

[23] Ahmed, N., Amin, R., Ayub, H., Iqbal, M. M., Saeed, M., & Hussain, M. (2022). Urdu Sentiment Analysis Using Deep Attention-Based Technique. *Foundation University Journal of Engineering and Applied Sciences (HEC Recognized Y Category, ISSN 2706-7351)*, 3(1), 1-12.[CrossRef]

[24] Rajalakshmi, S., Asha, S., & Pazhaniraja, N. (2017, March). A comprehensive survey on sentiment analysis. In *2017 fourth international conference on signal processing, communication and networking (ICSCN)* (pp. 1-5). IEEE. [CrossRef]

[25] Chandio, B. A., Imran, A. S., Bakhtyar, M., Daudpota, S. M., & Baber, J. (2022). Attention-based RU-BiLSTM sentiment analysis model for roman Urdu. *Applied Sciences*, 12(7), 3641. [CrossRef]

[26] Li, D., Ahmed, K., Zheng, Z., Mohsan, S. A. H., Alsharif, M. H., Hadjouni, M., ... & Mostafa, S. M. (2022). Roman Urdu sentiment analysis using transfer learning. *Applied Sciences*, 12(20), 10344. [CrossRef]

[27] Khan, L., Amjad, A., Ashraf, N., Chang, H. T., & Gelbukh, A. (2021). Urdu sentiment analysis with deep learning methods. *IEEE access*, 9, 97803-97812. [CrossRef]

[28] Ahmed, M. J., Afridi, U., Shah, H. A., Khan, H., Bhatt, M. W., Alwabli, A., & Ullah, I. (2024). CardioGuard: AI-driven ECG authentication hybrid neural network for predictive health monitoring in telehealth systems. *SLAS technology*, 29(5), 100193. [CrossRef]

[29] Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. *Scientific Reports*, 12(1), 5436.

[30] Ahmed, K., Nadeem, M. I., Li, D., Zheng, Z., Al-Kahtani, N., Alkahtani, H. K., ... & Mamyrbayev, O. (2023). Contextually enriched meta-learning ensemble model for Urdu sentiment analysis. *Symmetry*, 15(3), 645. [CrossRef]

[31] Sehar, U., Kanwal, S., Dashtipur, K., Mir, U., Abbasi, U., & Khan, F. (2021). Urdu sentiment analysis via multimodal data mining based on deep learning algorithms. *IEEE Access*, 9, 153072-153082. [CrossRef]

[32] Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2019). Sentiment analysis for a resource poor language—Roman Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1), 1-15. [CrossRef]

[33] Ahmad, P. N., Liu, Y., Ullah, I., & Shabaz, M. (2024). Enhancing coherence and diversity in multi-class slogan generation systems. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8), 1-24. [CrossRef]

[34] Nasim, Z., & Ghani, S. (2020). Sentiment analysis on Urdu tweets using Markov chains. *SN Computer Science*, 1(5), 269.

[35] Asghar, M. Z., Sattar, A., Khan, A., Ali, A., Masud Kundi, F., & Ahmad, S. (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems*, 36(3), e12397.[CrossRef]

[36] Mukhtar, N., Khan, M. A., Chiragh, N., & Nazir, S. (2018). Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. Expert Systems, 35(6), e12317. [CrossRef]

[37] Mukhtar, N., & Khan, M. A. (2020). Effective lexicon-based approach for Urdu sentiment analysis. *Artificial Intelligence Review*, 53(4), 2521-2548.

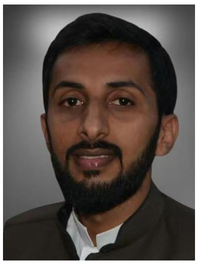
[38] Khan, K., Khan, W., Rahman, A. U., Khan, A., Khan, A., Khan, A. U., & Saqia, B. (2018). Urdu sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 9(9).

[39] ul Mustafa, F., Ashraf, I., Baqir, A., Ahmad, U., Malik, S., & Mehmood, S. (2020, October). Prediction of user's interest based on urdu tweets. In *2020 International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS)* (Vol. 5, pp. 1-6). IEEE. [CrossRef]

Muhammad Saeed is a computer scientist from Pakistan who has a strong academic background and expertise in a variety of computer science fields. He received the graduate degree in computer science from the University of Haripur, Haripur, Pakistan, in 2015, and the M.S. degree in computer science from the University of Engineering and Technology, Taxila, in 2021. Currently, he is a PhD. Scholar at Nanjing University of Information Science and Technology (NUIST) China. His research interests are in machine learning, information security, natural language processing, and the Internet of Things. Since January 2019, he has been working as a teacher in elementary and secondary education department KPK, Pakistan. His research interests are in Machine learning, IoMT and IOT. (Email: saeed.uet17@gmail.com)

Naeem Ahmed is a highly motivated computer scientist hailing from Pakistan, with a strong educational background and expertise in various fields of computer science. He received the bachelor's degree in computer science from the University of Haripur, Haripur, Pakistan, in 2019, and the M.S. degree in computer science from the University of and Technology, Taxila, in 2022. He has worked as Lecturer in Computer Science in University of Haripur and Government Post Graduate College Haripur. Currently, he is a PhD. Scholar at Nanjing University of Information Science and Technology (NUIST) China. His research interests are in machine learning, information security, natural language processing, and the Internet of Things. (Email: naeem.uoh@gmail.com)

Danish Ali is a highly motivated computer researcher from Pakistan, with a strong educational background and expertise in various fields of computer science. He received the bachelor's degree in computer science from the University of Haripur, Haripur, Pakistan, in 2023. Currently, he is a MS Scholar at Wuhan University China. His research interests are in machine learning, biomedical imaging, natural language processing, and the deep learning. (Email: danishalikhan545@gmail.com)



Muhammad Ramzan is a Computer Science professional from Pakistan. He has obtained a bachelor's degree in computer science from the University of Haripur, Pakistan. At present, he is a self-employed Researcher. (Email: muhamdramzan2023@outlook.com)

Muzamil Mohib is a highly motivated Big Data and business intelligence scientist hailing from Pakistan, with a strong educational background and expertise in various fields of Big data and business intelligence. He received the bachelor's degree in BBA marketing from the International Islamic university Islamabad, Pakistan, in 2021, and the M.S. degree in Marketing from the NUML University Islamabad, in 2024. He has worked in different corporate sector in stock exchange Islamabad and IT sector. Currently, he is a PhD. Scholar at Nanjing University of Information Science and Technology (NUIST) China. His research interests are Big data, AI, Business intelligence. (Email: muzamilmohib@outlook.com)

Kajol Bagga did her Diploma and Bachelor's in computer science from Sri Sai college of engineering (Punjab Technical University India) and Masters of Artificial Intelligence from Technical Hochshule (THWS) Wuerzburg, Germany. Reserch work and interest is step length estimation using Neural Network and Machine learning, deep learning and AI. (Email: kajol.bagga@study.thws.de)

Atif ur Rahman Born on February 1, 1996, in District Malakand, Pakistan, Atif ur Rahman is a driven and ambitious individual excelling in the field of computer science. He completed his BS Computer Science, Hazara University Mansehra (2019) and now Pursuing MS Computer Science, Iqra National University Swat (ongoing). His research area includes medical image processing, Deep learning, Artificial learning. (Email: atifurrahman.inu@gmail.com)

Ikram Majeed Khan earned his Bachelor's degree in Software Engineering from Islamia College University Peshawar and a Master's degree in Computer Science from Coventry University, England, UK. His research interests include Artificial Intelligence, Machine Learning, Deep Learning and Visual Intelligence. (Email: Khani72@coventry.ac.uk)