International Journal of Computational Intelligence in Control

Hate Speech Detection from Urdu language Tweets using Deep Learning Technique

Furqan Khan Saddozai[1*],Hussain Ahmad[1], Muhamamd Usama Asghar[1], Afnan Khan Saddozai[2]

furqan.dcma.dik@gmail.com,hussainicit@gmail.com,usama.asghar@yahoo.com, afnansaddozai1879@gmail.com

1. *Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, 29050, Pakistan.*
2. *Institute of Business Administration, Gomal University, Dera Ismail Khan, 29050, Pakistan.*

*Corresponding Author: (furqan.dcma.dik@gmail.com)

ABSTRACT

The growing popularity of social media sites has encouraged users to post messages about different topics and events. Users share their sentiments, opinions, and feelings using social media platforms such as Twitter and Facebook. These platforms allow people to share posts,and no major restrictions are imposed on these sites. However, some people misuse these sites by posting hateful and aggressive content. Manual identification of hate speech is impractical due to the huge volume of online data. Therefore, automated techniques are needed to detect and remove hateful content. Most of the existing research has been carried out in resource-rich languages such as English. However, in the case of resource-poor languages such as Urdu, more work is needed. In this study, a novel dataset is created to detect and classify hate speech from Urdu language tweets. Furthermore, the BiLSTM model is used to identify hateful tweets. We used the BiLSTM model as it is capable of processing contextual information from both forward and backward directions. We compared the performance of the proposed BiLSTM model with different ML and DL models. The results showed that the proposed model outperformed other comparing models by attaining an accuracy of 0.82, precision of 0.82, recall of 0.82, and F1 score of 0.82.

Keywords: Hate Speech, Urdu, BiLSTM, Tweets, Deep Learning, Machine Learning

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Tree
GRU	Gated Recurrent Unit
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning

MLP	Multilayer Perceptron
MNB	Multinomial Naïve Bayes
NB	Naïve Bayes
RF	Random Forest
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

1. Introduction

With the recent innovations in technology and the internet, the popularity of social networking sites is growing rapidly. A report [1] indicated that there are more than 300 million active users on Twitter. Moreover, 200 billion tweets are generated annually. In addition, the report uncovered that there are 2.88 million active users of Facebook, and approximately 10 billion messages are posted onthis platform. These platforms are popular among users as people are allowed to express their feelings about different topics such as religion, health, politics, and government policies. However, some people misuse these platforms by promoting hate speech, violence, prejudice, and abusive language.

Hate speech is defined as "a direct attack on people/group based on sex, colour, religious beliefs, disability, disease and race through writing, behavior or speech" [2]. It is considered as one of the major challenges faced by the modern society [3].Moreover, it promotes violence and leads to anti-social behavior [4]. The analysis of different terrorist attacks revealed that the suspects had posted hateful and aggressive content on the social media platform before committing the crime [5]. Moreover, another report indicated that hateful and blasphemous social media posts resulted in mob attacks [6]. Therefore, hateful and abusive content should be identified and removed to avoid violence and social unrest.

To detect and remove hateful content from social media platforms, dedicated employees are hired by these platforms. However, manual detection of hate speech is difficult and time-consuming due to the huge volume of online-generated content. Therefore, automated techniques are needed to identify such content. The existing work shows that different researchers have presented Machine Learning (ML) and Deep Learning (DL) models to detect hate speech from text messages. However, the majority of the work is focused on detecting hateful and offensive content from resource-rich languages, such as English [2]. There are only a few works that detect hate speech from resource-poor languages, such as Urdu [7,8,9].

This work aims to develop a DL-based BiLSTM model to detect and classify hate speech from Urdu language tweets. This model is used as it considers context information from both forward and backward directions. Moreover, the Urdu language is considered as it is the national and official language of Pakistan. In addition, it is widely spoken in different countries of the world, such as India, Afghanistan, and Bangladesh [5]. Furthermore, it is a low-resource language.

1.1 Research Questions

This work is going to address the following research questions: RQ1: How to classify Urdu language tweets into 'Hate' or 'No-Hate' categories using different DL models such as BiLSTM, LSTM, GRU, and CNN? RQ2:

What is the performance of the proposed BiLSTM model w.r.t traditional ML models? RQ3: What is the performance of the proposed BiLSTM model w.r.t existing models?

1.2 Research Contributions

The contributions of this work are as follows:(1). A novel dataset is created to detect and classify hate speech from Urdu language tweets, (2). Different DL models are used to classify hateful content from Urdu tweets, such as BiLSTM, LSTM, GRU, and CNN(3). Various ML models are used to detect hate speech from Urdu tweets(4). The performance of the proposed model is evaluated w.r.t existing models for the Urdu language.

2. Related Work

A number of research works have addressed the problem of unwanted or suspicious text detection and classification using natural language processing techniques. Authors have analyzed various different flavors of text-based controversial content, such as identifying offensive, hateful, controversial, abusive and toxic messages. The majority of the existing work considered languages which are rich in resources, such as English, Spanish, Italian and others[2,16,17, 18-21]. However, detecting and categorizing harmful content from resource-poor languages such as Urdu remained under-explored. In this section, we focus on the previous work in Urdu and Roman Urdu languages. Table 1 presents existing studies that detect unwanted text from Urdu and Roman Urdu languages.

Mustafa et al. [7] presented a novel system to detect and classify controversial text from Urdu language tweets. They used TF-IDF scheme to extract features, and different ML models were trained. Finally, the trained models are used to categorize tweets into corresponding categories. They obtained good results using the NB classification model. Rizwan et al. [3] categorized five different classes of abusive text using embedding techniques and designed an effective model to perform categorization. They considered religious hate, offensive, profane, sexist, and normal categories of the unwanted text in Roman Urdu. The results showed that top performing model attained an accuracy of 82%.In another work [8], a model to identify and categorize offensive text from Urdu and Roman Urdu language is introduced. They used character and word-level N-gram features to develop different ML classification models and achieved promising results. To detect anti-social behavior from Urdu text, the study [9] proposed a model using lexical features. The performance of the proposed model is compared with the baseline algorithms.

Kausar et al. [10] created an Urdu language corpus to recognize propaganda posts from online news. Various experiments were performed using sematic and word embeddings-based features. They reported that N-gram features showed good performance. To detect abusive words from Urdu language tweets, Haq et al. [11] presented a customized model to filter unwanted posts. They developed a lexicon consisting of slang and abusive words. Finally, the generated lexicon issued to classify tweets into abusive and non-abusive categories. However, a limited number of tweets are used in their work.

Table 1.Related Studies to Detect Unwanted Text in Urdu and Roman Urdu Languages

Author [Reference]	Year	Language (s)	Technique	Feature(s)	Source	Classes
Mustafa et al. [7]	2017	Urdu	SVM,NB, LR	TF-IDF	Twitter	Controversial, Not-controversial
Akhter et al. [8]	2020	Urdu & Roman Urdu	RF, KNN, SVM, NB	Word and Char N-grams	YouTube	Offensive, Not-offensive
Sohail et al. [9]	2020	Urdu	Customized Technique	Lexicon-based Features	Not Found	Sentiment Analysis
Kausar et al. [10]	2020	Urdu	CNN	LIWC, Word2Ve, NELA, BERT	News	Propaganda, Not-propaganda
Haq et al. [11]	2020	Urdu	Customized Technique	Lexicon-based Features	Twitter	Abusive, Not abusive
Khan et al. [12]	2021	Roman Urdu	LR,NB, SVM, Boosting, Bagging	Word2Ve, TF-IDF	Twitter	Hate, Simple-Complex, Neutral
Amjad et al. [13]	2021	Urdu	MLP,SVM,LR ,LSTM, CNN,	FastText, Word and Char N-grams	Twitter	Threatening, Not-threatening; Individual, Group
Das et al. [14]	2021	Urdu	BERT Classification Model	BERT	Twitter	Abusive, Not-abusive; Threatening, Not-threatening
Saeed at al. [15]	2021	Roman Urdu	BiLSTM, CNN, BGRU	Word Embeddings	Multiple	Toxic, Not-toxic

A system to identify hateful content from Urdu tweets is proposed by Ali et al. [5]. They used an opinion mining-based technique and handled a number of issues such as sparsity, imbalanced classes and high dimensional features to enhance the performance of hateful content detection. The features extracted using TF-IDF and other techniques are considered for training the MNB and the SVM models. They reported that the SVM model showed better performance than the MNB technique. However, posts from other social sites are not considered in their work. Moreover, DL models remained unexplored. In another study [12], a multiclass hateful content detection and classification model is introduced using word2vec and TF-IDF feature extraction techniques for Roman Urdu text. They used five different ML algorithms to perform experiments. The results showed that their proposed technique obtained satisfactory performance in terms of different evaluation metrics.

Some researchers addressed the task of threatening text detection and categorization. To identify threatening content from Urdu tweets, Amjad et al. [13] proposed a system using various feature-extracting techniques. The results showed that the MLP model outperformed comparing models using the word N-grams scheme. Das et al. [14] introduced a model to recognize threatening and abusive content from Urdu text. They reported that the BERT classifier showed good performance. Saeed et al. [15] introduced a novel system to identify and categorize toxic content from Roman Urdu language text. They used various embedding techniques and a DL-based ensemble model. Their proposed model achieved remarkable performance.

3. Methodology

This section presents a novel DL-based system to detect and classify hate speech from Urdu language tweets. The proposed technique consists of the following six phases. In the first phase, tweets are scrapped from Twitter using a list of keywords. In the second phase, a corpus is developed to train a corresponding model to detect and classify hateful content. The third phase is used to preprocess tweets from unwanted text. In the fourth phase, the corpus is divided into two subsets, i.e. training and test tweets. In the upcoming phase, different models will be trained using training tweets. In the final phase, the performance of the different models is evaluated using test tweets. The proposed system is presented in the figure. 1.

1. Scrapping Tweets
2. Urdu Hate Speech Corpus Development
3. Text Preprocessing
4. Corpus Partitioning
5. Training
6. Testing& Evaluation

3.1 Scrapping Tweets

This is the first phase of our proposed system. In this phase, a set of keywords was used to scrape Urdu language tweets from Twitter. The keywords related to religion, politics, national origins, race, sects, and other domains were considered for this purpose. A sample list of domains and associated words is presented in Table 2. A Python-based script was designed to acquire tweets using Twitter API.

Table 2.A Sample List of Domains & Associated Keywords

S#	Domain	Associated Keywords in Urdu (English)
1	Religion	مسلمان (Muslim), صیہونی (Zionist), ہندو (Hindu), عیسائی (Christian)
2	Politics	ڈیزل (Diesel), شوباز (Showman), زرداری (Zardari), عمران (Imran)
3	Race	سندھی (Sindhi), پنجابی (Panjabi), پشتون (Pashtoon), بلوچ (Baloch)
4	Sect	دیوبندی (Deobandi), شیعہ (Shia), وھابی (Wahabi), بریلوی (Barelvi)

5	National Origins	افغان (Afghan), امریکہ (America), بھارت (India), اسرائیل (Israel)

3.2 Urdu Hate Speech Corpus Development

Tweets scrapped in the previous phase were filtered to remove duplicate tweets. Moreover, tweets belonging to languages such as Punjabi, Arabic, and others were identified and removed. The cleaned Urdu language tweets were manually assigned a label of 'Hate' and 'No-Hate'. In addition, superfluous tweets were removed to balance the number of tweets in each category.

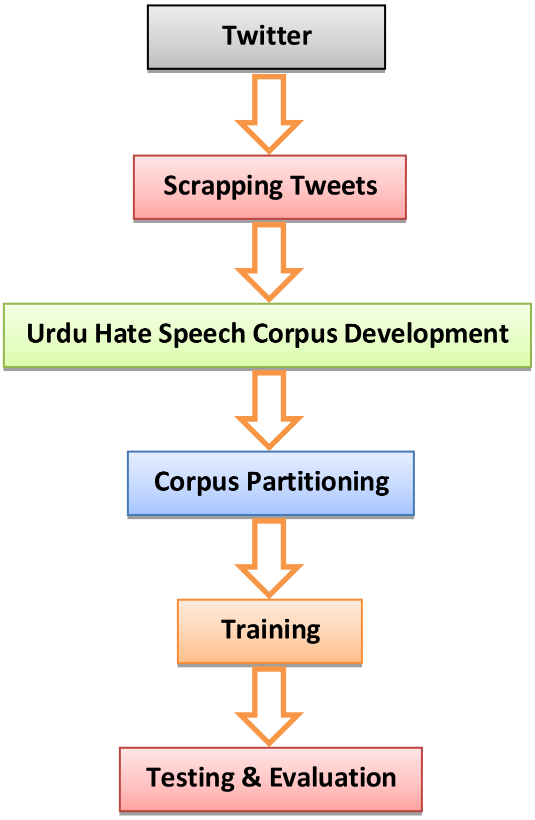

Figure.1. Proposed System to Detect Hate Speech from Urdu Tweets

3.3 Text Preprocessing

Text preprocessing is the third phase of the proposed hate speech detection system. Text preprocessing is necessary as it transforms the dataset into an appropriate format. Moreover, it plays a crucial role in optimizing the performance of the classification models. We applied a number of preprocessing techniques to detect and remove unwanted text.

- **Removing Usernames:** - All usernames were identified and removed from Urdu tweets.
- **Removing URLs:** - We have detected and removed all URLs from tweets.
- **Removing Emojis:** -Emojis are used to express feelings or emotions. The emojis were removed from the text.
- **Removing White Spaces:** - All extra white spaces were identified and removed from the tweets.

- **Removing Digits& Special Characters:** -We removed digits and special characters found inside the text.
- **Removing English Characters &Words:** -Tweets were cleaned from English language characters and words.
- **Removing Diacritics:** -Diacritics are frequently used in Urdu language and help readers to correctly pronounce the words. We identified and removed all diacritics from tweets.
- **Removing Stop words:** -Stop words are words that are considered meaningless while deciding class of corresponding text. Therefore, such words are removed from dataset. We prepared a manually compiled list of stop words. This list is used to detect and remove stop words from tweets.
- **Tokenization:** -The tweets were tokenized into individual words using different delimiters such as comma, white spaces and others.

3.4 Corpus Partitioning

This is the fourth phase of our proposed system. In this phase, the preprocessed tweets were partitioned into two subsets, i.e. training tweets and test tweets. In this work, we used a partitioning ratio of 70:30. Therefore, 70% tweets are used for training models, while remaining 30% tweets are used to evaluate the performance of different classification models. We used 'train_test_split' routine of python to divide the corpus.

3.5 Training

In this phase of the proposed system, training tweets are used to develop models. These tweets are used to learn hidden patterns/features in the dataset. In each epoch, the DL model is used to process training tweets multiple times. The model tries to extract features from given tweets.In our work, 70% of tweets are used to train various models. We considered different DL models to detect hateful content from Urdu tweets, such as BiLSTM, LSTM, GRU, and CNN. The training tweets were preprocessed, and task-based embeddings were generated. The resultant embeddings were used to develop DL models.

3.6 Testing& Evaluation

In this phase of the proposed hate speech detection system, the trained models were evaluated using test tweets. Test tweets were preprocessed and corresponding embeddings are generated.To evaluate the performance of different DL and ML models, various popular metrics were used such as Accuracy, precision, recall, and F1-measure.

4. Results& Discussion

This section presents the results of the different DL models such as BiLSTM to detect hate speech from Urdu language tweets. Moreover, the performance of different ML models is showed. In addition, the outcomes of proposed and existing models to detect hate speech is discussed.

4.1 Experimental Setting

Weused Jupyter Notebook to conduct experiments in this work. Different python language libraries were used during experiments such as NLTK, Keras, Numpy, Pandas, and Sklearn. The experiments were carried out on the CPU. Moreover, Python 3.10 was used to design corresponding experiments while training and testing of models. Table 3 is used to represent different system settings while performing experiments.

Table 3.System Specifications for Experiments

System Specifications	
CPU	Intel(R) Core (TM) i3-4200M CPU @ 2.50GH 2.50 GHz
OS	Windows 10
RAM (installed)	20GB
Storage	500GB

4.2 Urdu Hate Speech Corpus Specifications

We used a list of more than 400 keywords to scrape tweets from Twitter. Hence, more than 30,000 tweets were obtained. We filtered tweets belonging to languages such as Punjabi, Arabic, and Pashto. As a result, only Urdu tweets were obtained. In addition, duplicate tweets were identified and removed. The tweets were manually assigned a label of either 'Hate' or 'No-Hate' depending upon the corresponding textual content. Finally, superfluous tweets were removed to keep an even number of tweets in each of the categories. There are 10,000 tweets in our dataset. Hence, we have 5,000 tweets for each 'Hate' and 'No-Hate' class. Table 4 represents a sample of tweets in each category.

Table 4.Sample Tweets from Urdu Hate Speech Corpus

Tweet Text (Urdu)	Tweet Text (English Translation)	Label
@najamwalikhan @mushtaqminhas یہحرامیبھیصحافیھے۔ایسےصحافیکوٹھونسےسیدھامافیاکےدربارمیناکےگرتےہیں۔تیریاوقاتیھ ھے https://t.co/I5x7uCdSXd	@najamwalikhan @mushtaqminhas This bastard is also a journalist. Such journalists fall straight from the dungeons to the court of the mafia. This is your time https://t.co/I5x7uCdSXd	**Hate**
بلوچ ،پشتون ،سرائیکی اور ہزارہ سمیت تمام اقوام نیشنل پارٹی کو مظبوط بنائیں، نیشنل پارٹی کی واضح پالیسی اور مخلص قیادت کی بدولت نیشنل پارٹی میں تمام اقوام سے تعلق رکھنے والے شامل ہورہےہیں ۔ صوبائی خواتین سیکریٹری گودی کلثوم نیاز بلوچ KalsoomNiazNp https://t.co/Q5JacWRami@	All nations including Baloch, Pashtun, Saraiki and Hazara should strengthen the National Party, thanks to the clear	**No-Hate**

	policy and sincere leadership of the National Party, people belonging to all nations are joining the National Party. Provincial Women Secretary GodiKulsoom Niaz Baloch @KalsoomNiazNp https://t.co/Q5JacWRami	
روشنیوں کے شہر کراچی کو اندھیروں میں دھکیلنے کے لیے ہی ملٹری اسٹیبلشمنٹ نے MQM بنائی تھی- اور اب پھر سے کرپٹ کرنیل MQM کو کٹھا کرنے جارہے کراچی کو مزید برباد کرنے کے لیے--- کاش که ہم نے اِن جرنیلوں کی جگه اچھی نسل کے کتے پالے ہوتے--😠😠 TeamiPians@ #حقیقی_قائد_صرف_عمران https://t.co/tJBaFPcXSf	MQM was created by the military establishment to push Karachi, the city of lights, into darkness. And now the corrupt colonels are going to destroy MQM again to further ruin Karachi. I wish we had raised good breed dogs instead of these generals.😠😠 @TeamiPians #زرب_قاعد_حقيقي_ عمران https://t.co/tJBaFPcXSf	Hate
عظیم خان بلور و ثمر ہارون بلور این اے 31 کے ضمنی الیکشن کے سلسلے میں موچی لڑہ بازار میں کمپین کرتے ہوئے Khadimhussain4@ SamarHBilour@ @ https://t.co/5wWSQstrFg AsgharNawazKha2 https://t.co/tg8p2VNlXt	Azeem Khan Bilor and Samar Haroon Bilor campaigning in Mochi Lala Bazar in connection with NA 31 by-election. @Khadimhussain4 @SamarHBilour @AsgharNawazKha2 https://t.co/5wWS	No-Hate

	QstrFg https://t.co/tg8p2V NlXt	

4.3 Addressing RQ1

To address the*RQ1: How to classify Urdu language tweets into 'Hate' or 'No-Hate' categories using different DL models such as BiLSTM, LSTM, GRU, and CNN?* We performed various experiments using different DL models. We used BiLSTM, LSTM, GRU, and CNN models for this purpose. Table 5 is used to represent different hyperparameters considered to perform experiments.

Table 5.DL Models & Hyperparameters with Values

Model	Parameters & Values
BiLSTM	Embedding_scheme=Word2Vec, Max_features=4000, Input_length=100, Embedding_dimensions=100, Model_layers= 2 BiLSTM layers, Units=200, 100, Return_sequences=True, Recurrent_dropout=0.2, Activation_function= ReLu, Dropout_layer= 0.4, Dense_layer=1, Dense_layer_size=2, Activation_function=Sigmoid, Optimizer=Adam, Loss_function= Binary_crossentropy, Epochs=20, Batch_size=32, Early_stopping_criteria= (Monitor=Val_loss, Patience=3, Restore_best_weights=True)
LSTM	Embedding_scheme=Word2Vec, Max_features=4000, Input_length=100, Embedding_dimensions=100, Model_layers= 2 LSTM layers, Units=200, 100, Return_sequences=True, Recurrent_dropout=0.2, Activation_function= ReLu, Dropout_layer= 0.4, Dense_layer=1, Dense_layer_size=2, Activation_function=Sigmoid, Optimizer=Adam, Loss_function= Binary_crossentropy, Epochs=20, Batch_size=32, Early_stopping_criteria= (Monitor=Val_loss, Patience=3, Restore_best_weights=True)
GRU	Embedding_scheme=Word2Vec, Max_features=4000, Input_length=100, Embedding_dimensions=100, Model_layers= 2 GRU layers, Units=200, 100, Return_sequences=True, Recurrent_dropout=0.2, Activation_function= ReLu, Dropout_layer= 0.4, Dense_layer=1, Dense_layer_size=2, Activation_function=Sigmoid, Optimizer=Adam, Loss_function= Binary_crossentropy, Epochs=20, Batch_size=32, Early_stopping_criteria= (Monitor=Val_loss, Patience=3, Restore_best_weights=True)
CNN	Embedding_scheme=Word2Vec, Max_features=4000, Input_length=100, Embedding_dimensions=100, Model_layers= 2 Conv layers, Filter_size=128, 64, Max_pooling= 1, Activation_function= ReLu, Flatten_layer=1, Dense_layer=1, Dense_layer_size=2, Activation_function=Sigmoid, Optimizer=Adam, Loss_function= Binary_crossentropy, Epochs=20, Batch_size=32, Early_stopping_criteria= (Monitor=Val_loss, Patience=3, Restore_best_weights=True)

For the BiLSTM model, the preprocessed and tokenized words were forwarded to an embedding layer of size 100 to generate embedding using the Word2Vec technique. The embedding layer was followed by two BiLSTM layers of 200 and 100 units, respectively. We used the 'ReLu' activation function on these layers. The BiLSTM layers were followed by a Dropout layer of 0.4. Finally, a Dense layer with two units and a sigmoid function is used. The Adam optimizer is selected to optimize the weights of the DL model. Moreover, the loss function of 'binary_crossentropy' is used. The epoch and batch size were adjusted to 20 and 32, respectively.

In the LSTM model, an embedding layer is used to generate word embeddings. After the embedding layer, two LSTM layers were used with 200 and 100 units, respectively. The 'ReLu' activation function is used in both LSTM layers. Furthermore, a Dropout layer was used with a 0.4 value. Finally, a Dense layer with two units is added to the DL model with the 'Sigmoid' activation function. Moreover, the Adam optimizer is used with a loss function (i.e. Binary_crossentropy). For the GRU model, an embedding layer was followed by two GRU layers with 200 and 100 units. The 'ReLu' activation function is used in both GRU layers. Furthermore, a Dropout layer was used with a 0.4 value. Finally, a Dense layer with two units is added to the DL model with asigmoid activation function. Moreover, the Adam optimizer is used with a loss function (i.e. Binary_crossentropy).

Inthe CNN model, an embedding layer was placed to generate embeddings. After the embedding layer, two one-dimensional convolutional layers were used with filter sizes of 128 and 64, respectively. The 'ReLu' activation function was utilized on both layers. After convolutional layers, Maxpooling and Flatten layers were added to the DL model. Finally, a Dense layer with two units and a sigmoid activation function was used. Adam optimizer is selected to optimize the weights of the DL model. Moreover, the loss function of 'binary_crossentropy is used'. Table 6 shows the results obtained using different DL models.

Table 6.Performance of DL Models to Detect HateSpeech

S#	Model	Accuracy	Precision	Recall	F1 Score
1	**BiLSTM**	**0.82**	**0.82**	**0.82**	**0.82**
2	LSTM	0.77	0.74	0.82	0.78
3	GRU	0.76	0.72	0.80	0.78
4	CNN	0.78	0.77	0.78	0.78

The results reported in the above table show that the BiLSTM model outperformed other DL models in detecting hate speech from Urdu language tweets. It attained an accuracy of 0.82, precision of 0.82%, recall of 0.82, and F1 score of 0.82. The CNN model attained an accuracy of 0.78. Moreover, LSTM and GRU models showed accuracies of 0.77 and 0.76. Therefore, the BiLSTM model was the top-performing model when compared to the DL models. Figure 2 is used to graphically represent the comparative performance of different DL models to detect hate speech from Urdu language tweets.

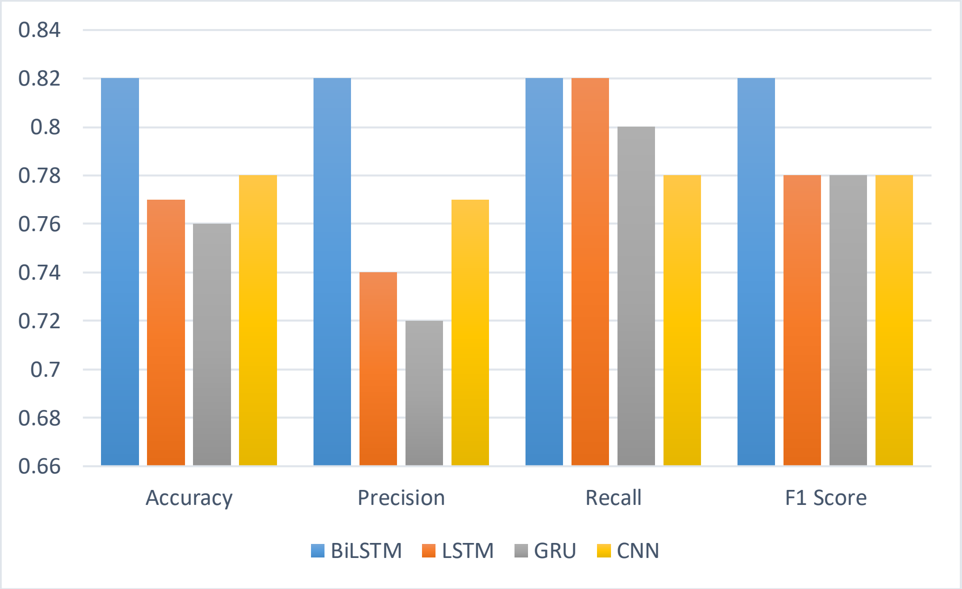

Figure. 2. Comparative Performance of DL Models to Detect Hate Speech from Urdu Tweets

4.4 Addressing RQ2

To address the *RQ2: What is the performance of the proposed BiLSTM model w.r.t traditional ML models?* A number of experiments were performed to investigate the performance of traditional ML models. We selected eleven models, including KNN, DT, RF, LR, MNB, AdaBoost, CatBoost, GradBoost, LGBoost, Extra Trees and SGD. We used the BOW technique to extract features for the training and testing of models. The performance of these models is reported inTable 7.

Table 7.Performance of ML Models to Detect Hate Speech

S#	Model	Accuracy	Precision	Recall	F1 Score
1	KNN	0.55	0.74	0.55	0.44
2	DT	0.71	0.71	0.71	0.71
3	RF	0.77	0.77	0.77	0.77
4	LR	0.75	0.75	0.75	0.75
5	MNB	0.78	0.78	0.78	0.78
6	AdaBoost	0.71	0.73	0.71	0.7
7	CatBoost	0.77	0.78	0.77	0.76
8	GradBoost	0.72	0.77	0.72	0.71
9	LGBoost	0.75	0.76	0.75	0.75
10	ExtraTrees	0.76	0.76	0.76	0.76
11	**SGD**	**0.79**	**0.79**	**0.79**	**0.79**

The results showed that the SGD model outperformed the ML models. It attained an accuracy of 0.79, precision of 0.79, recall of 0.79, and F1 score of 0.79. However, its performance was lower than the BiLSTM model.The MNB model attained an accuracy of 0.78, precision of 0.78, recall of 0.78, and F1 Score of 0.78. The RF and CatBoost models reached an accuracy of 0.77. Moreover, the Extra trees model showed an accuracy of 0.76. The LR and LGBoost models reached an accuracy of 0.75. In addition, the GradBoost model attained an accuracy of 0.72. The DT and AdaBoost models showed similar results and attained an accuracy of 0.71. Finally, the KNN model's performance was the lowest in detecting hate speech from Urdu tweets. It attained an accuracy of 0.55.Figure 3 is used to represent the performance of different ML models in detecting hate speech from Urdu language tweets.

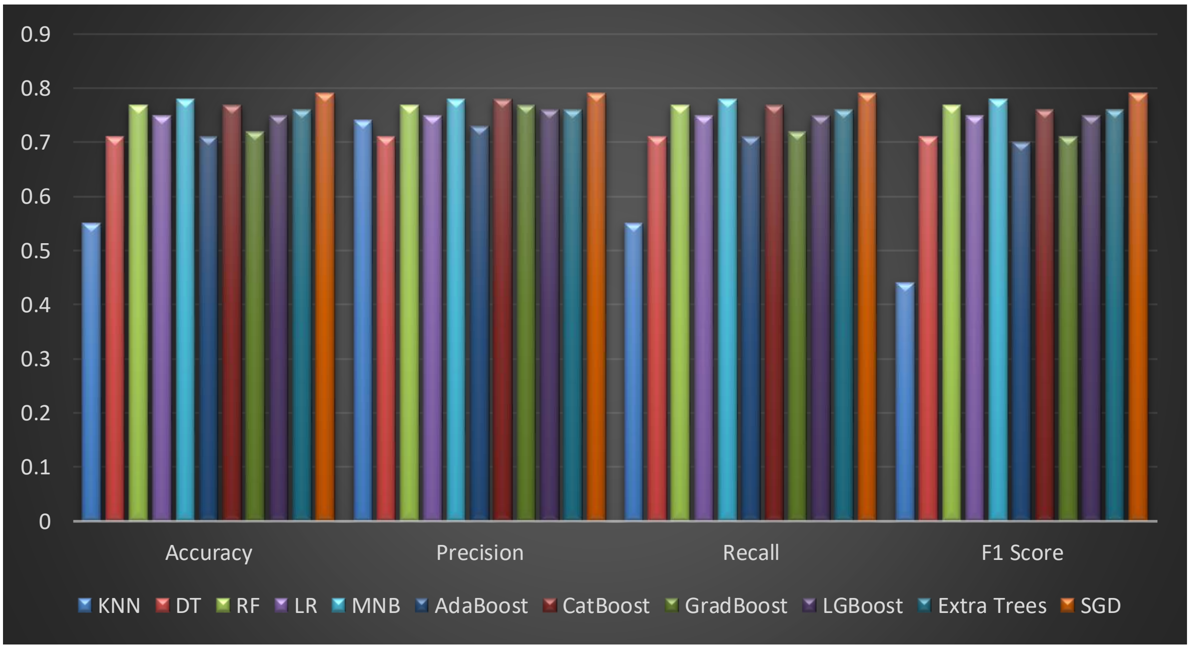

Figure. 3. Comparative Performance of ML Models to Detect Hate Speech from Urdu Tweets

4.5 Addressing RQ3

To address theRQ3: *What is the performance of the proposed BiLSTM model w.r.t existing models?* We used two model architectures presented in [5, 13] to compare the performance of our proposed system. Ali et al. [5] showed that their proposed SVM model outperformed comparing models in detecting hate speech using TF-IDF features. In another work, the MLP model was used to detect threatening and non-threatening language from Urdu text [13]. We applied both models to detect hateful content from Urdu tweets. The results are presented in Table 8. The results showed that our proposed BiLSTM model outperformed the existing models in terms of accuracy, precision, recall, and F1 score.

Table 8.Performance of Existing & Proposed Models to Detect Hate Speech

S#	Model	Accuracy	Precision	Recall	F1 Score
1	Ali et al. [5]	0.795	0.796	0.797	0.795
2	Amjad et al. [13]	0.80	0.80	0.80	0.80
3	**BiLSTM (Proposed)**	**0.82**	**0.82**	**0.82**	**0.82**

The model presented by Ali et al. [5] attained an accuracy of 0.795, precision of 0.796, recall of 0.797, and F1 score of 0.795. Moreover, an accuracy of 0.80, precision of 0.80, recall of 0.80, and F1 score of 0.80 was attained by model [13] in our experiments. Therefore, the results showed that the proposed BiLSTM model outperformed existing models' architectures in detecting hate speech. Figure 4 is used to represent the comparative performance of proposed and existing models in detecting hate speech from Urdu tweets.

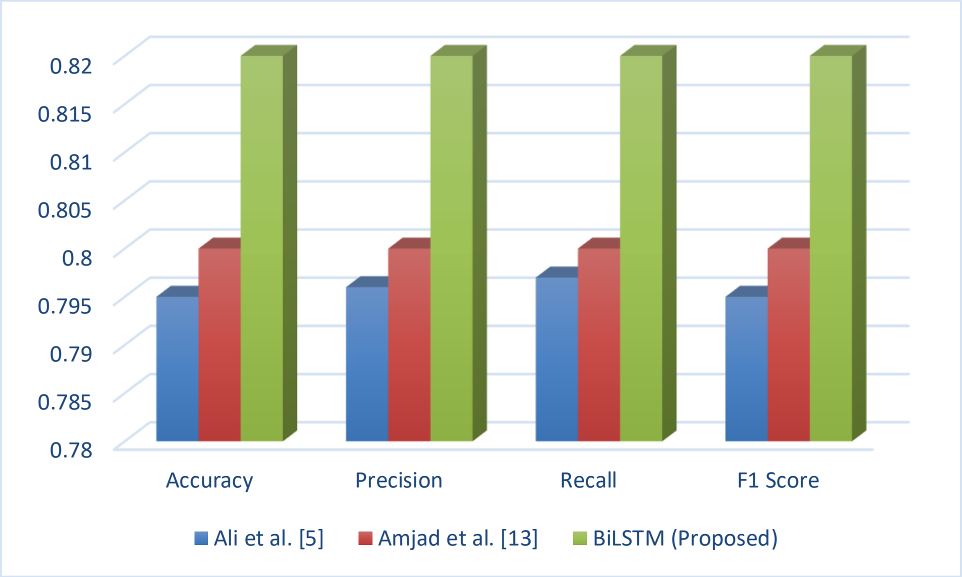

Figure.4. Comparative Performance of Proposed and Existing Models to Detect Hate Speech from Urdu Tweets

5. Conclusions & Future Work

Due to the increase in hateful and aggressive content online, there is a need to develop systems to detect and classify hate speech from Urdu language tweets. This work presents a BiLSTM model to detect and categorize hate speech from Urdu text. We collected Urdu tweets using a list of keywords. The scrapped tweets were filtered to remove redundant and non-Urdu tweets. Moreover, tweets were annotated. The labelled tweets were preprocessed and divided into training and test tweets. Different experiments were performed to categorize tweets into corresponding classes. The results showed that the proposed BiLSTM model outperformed the comparing DL and ML models in detecting hateful content from Urdu language tweets. The proposed work can be extended by considering online text from other social media sites such as Facebook. Moreover, using multilingual text to detect hate speech can be investigated in future to explore the ability of the proposed system. In addition, hybrid DL models such as BiLSTM+CNN and CNN+BiLSTM can be explored in future.

References

[1]. J. Clement, “Number of global social network user (2017-2025),” 2020. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[2]. N. S. Mullah, and W. M. N. W. Zainon, “Advances in machine learning algorithms for hate speech detection in social media: a review,” *IEEE Access*, vol. 9, 88364-88376, 2021.

[3]. H. Rizwan, M. H. Shakeel,and A. Karim, "Hate-speech and offensive language detection in roman Urdu," *In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 2512-2522, 2020.

[4]. J. A. Pálmadóttir, and I. Kalenikova, “Hate speech an overview and recommendations for combating it,” *Icelandic Human Rights Centre*, pp. 1-27, 2018.

[5]. M. Z. Ali, S. Rauf, K. Javedand S. Hussain, "Improving hate speech detection of Urdu tweets using sentiment analysis," *IEEE Access*, vol. 9, pp. 84296-84305, 2021.

[6]. H. Farhan and A. Akbar, “Mardan university student lynched by mob over alleged blasphemy: police,” *Dawn*, Available: https://www.dawn.com/news/1326729.

[7]. R. U. Mustafa, M. S. Nawaz, J. Farzund, M. I. Lali, B. Shahzad,and P. Viger, "Early detection of controversial Urdu speeches from social media," *Data Sci. Pattern Recognit.*, vol. 1, no. 2, pp. 26-42, 2017.

[8]. M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeedand M. T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu," *IEEE Access*, vol. 8, pp. 91213-91226, 2020.

[9]. M. Sohail, A. Imran, H. U. Rehmanand M. Salman, "Anti-social behavior detection in Urdu language posts of social media," *In 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1-7, 2020, IEEE.

[10]. S. Kausar, B. Tahirand M. A. Mehmood, "ProSOUL: a framework to identify propaganda from online Urdu content," *IEEE access*, vol. 8, pp. 186039-186054, 2020.

[11]. N. U. Haqet al., "USAD: an intelligent system for slang and abusive text detection in PERSO-Arabic-scripted Urdu," *Complexity*, vol. 2020, pp. 1-7, 2020.

[12]. M. M. Khan, K. Shahzadand M. K. Malik, "Hate speech detection in roman urdu," ACM *Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1-19, 2021.

[13]. M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, and A. Gelbukh, "Threatening language detection and target identification in Urdu tweets," *IEEE Access*, vol. 9, pp. 128302-128313, 2021.

[14]. M. Das, S. Banerjeeand P. Saha, "Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach," *arXiv preprint* arXiv:2111.14830, 2021.

[15]. H. H. Saeed, M. H. Ashraf, F. Kamiran, A. Karim, and T. Calders, "Roman Urdu toxic comment classification," *Language Resources and Evaluation*, pp. 1-26, 2021.

[16]. F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Urena-López, and M. T. Martín-Valdivia, “Comparing pre-trained language models for Spanish hate speech detection,” *Expert Systems with Applications*, vol. 166, 114120, 2021.

[17]. A. Duzha, C. Casadei, M. Tosi, and F. Celli, “Hate versus politics: detection of hate against policy makers in Italian tweets,” *SN Social Sciences*, vol. 1, no. 9, 223, 2021.

[18]. I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, B., ... and M. Alfawareh, “Intelligent detection of hate speech in Arabic social network: A machine learning approach,” *Journal of information science*, vol. 47, no. 4, pp. 483-501, 2021.

[19]. E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, “Detecting ethnicity-targeted hate speech in Russian social media texts,” *Information Processing & Management*, vol. 58, no. 6, 102674, 2021.

[20]. I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” *In 2017 international conference on advanced computer science and information systems (ICACSIS)*, pp. 233-238, 2017, IEEE.

[21]. A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, “Bangla hate speech detection on social media using attention-based recurrent neural network,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578-591, 2021.