

# Urdu Text Sentiment Analysis Using Recurrent Neural Networks

Syed Usman Ali Shah

**Abstract**— This paper presents a sentiment analysis approach applied to Urdu movie reviews using bidirectional Gated Recurrent Unit (GRU) architectures. The dataset was acquired from Kaggle, comprising 50k movie reviews translated into Urdu. Preprocessing involved stop words removal, normalization, lemmatization, and tokenization using UrduHack and Spacy libraries. We experimented with various neural network architectures, including unidirectional and bidirectional LSTMs and GRUs, exploring different configurations for optimal performance. Our findings indicate that bidirectional GRU models outperform other architectures in terms of accuracy, F1 score, and training efficiency.

## I. INTRODUCTION

Sentiment analysis, a subfield of natural language processing (NLP), aims to identify and classify sentiments expressed in text data. With the increasing availability of digital content in multiple languages, sentiment analysis in non-English languages has gained importance. Urdu, being a widely spoken language, presents challenges and opportunities for sentiment analysis tasks. In this work, we focus on sentiment analysis of Urdu movie reviews, leveraging deep learning techniques.

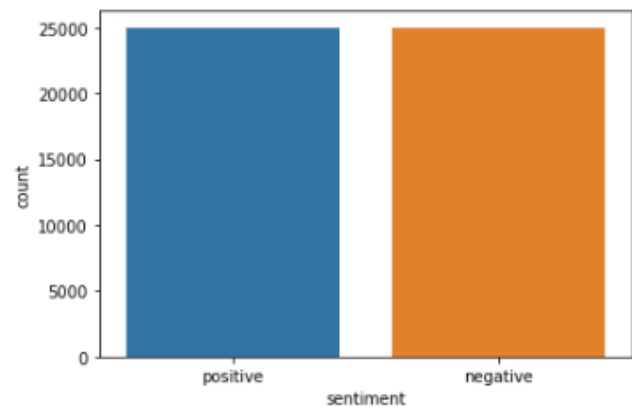
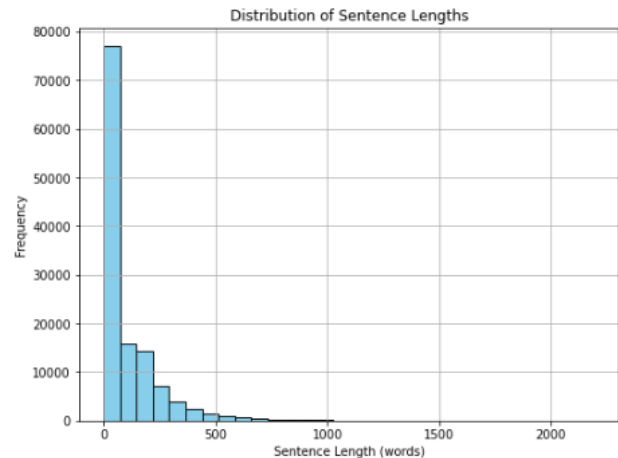
## II. DATASET ACQUISITION

The dataset used in our study was sourced from Kaggle, a popular platform for data science and machine learning datasets. Specifically, we accessed the "IMDB Dataset of 50k Movie Translated Urdu Reviews" [1]. This dataset is valuable for sentiment analysis tasks as it contains a large collection of movie reviews in Urdu, a language widely spoken in Pakistan and other regions.

The dataset comprises 50,000 rows, each representing a movie review. These reviews are labelled with sentiment polarity, indicating whether the review expresses positive or negative sentiment. The dataset is evenly balanced, with 50% of the reviews labelled as positive sentiment and the other 50% as negative sentiment. This balanced distribution is essential for training machine learning models effectively, ensuring that the model learns from a diverse set of examples representing different sentiments.

By leveraging this dataset, we were able to create a robust framework for sentiment analysis on Urdu text. The richness of the dataset, combined with its balanced sentiment labels, provided us with a solid foundation for training and evaluating our sentiment analysis models accurately.

```
Maximum sentence length: 2196 words
Minimum sentence length: 0 words
Average sentence length: 94.32 words
```

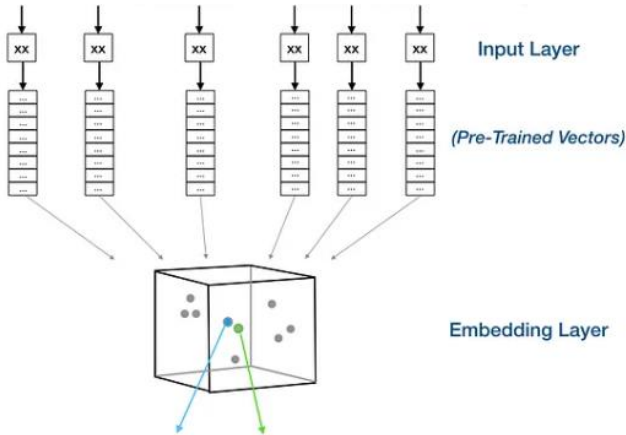


### III. PREPROCESSING

Preprocessing is a crucial step in natural language processing (NLP) tasks as it helps clean and transform raw text data into a format suitable for machine learning models. In our study, we utilized the UrduHack library, which is specifically designed for processing Urdu text, for the following preprocessing tasks:

1. **Stop Words Removal:** Stop words are common words in a language (like "آه" "آپ" "آء" etc.) that often do not carry significant meaning for analysis. By removing stop words, we reduce noise in the data and focus on more meaningful content.
2. **Normalization:** Normalization involves standardizing text by converting it to a consistent format. This may include converting numbers to words, replacing abbreviations with their full forms, and handling special characters or symbols.

3. **Lemmatization:** Lemmatization is the process of reducing words to their base or root form, known as the lemma. This helps in reducing the dimensionality of the data by grouping together different forms of the same word.
4. **Tokenization using Spacy:** Tokenization is the process of splitting text into smaller units, typically words or tokens. We used the Spacy library for tokenization, which provides robust support for various languages, including Urdu. This step is essential for breaking down text into manageable units for analysis.
5. **Word2Vec Embeddings:** Word2Vec is a popular technique for generating word embeddings, which are dense vector representations of words in a continuous vector space. These embeddings capture semantic relationships between words and are crucial for training deep learning models like neural networks. We generated word2vec embeddings to represent words in our text data, enabling the model to learn semantic similarities and patterns during training.



By performing these preprocessing steps, we ensure that the input data for our sentiment analysis model is clean, standardized, and structured in a way that facilitates effective learning and pattern recognition. This preprocessing pipeline lays the foundation for building accurate and robust sentiment analysis models on Urdu text.

#### IV. MODEL CREATION

##### Embedding Layer:

The model starts with an embedding layer, which is responsible for converting the input text data (word indices) into dense vectors of fixed dimensions. These vectors represent semantic meanings of words in a continuous vector space.

The embedding layer helps the model understand relationships and similarities between words based on their context within the dataset.

##### Bidirectional GRU Layer:

Following the embedding layer is a bidirectional GRU (Gated Recurrent Unit) layer. GRU is a type of recurrent neural network (RNN) that can capture long-term dependencies in sequential data.

The bidirectional aspect means that this GRU layer processes input sequences in both forward and backward directions. This bidirectional processing helps the model capture context from both past and future time steps, improving its understanding of the text's context.

The GRU layer has 256 units (neurons) and uses a hyperbolic tangent (tanh) activation function, which introduces non-linearity into the model's computations.

##### Dense Layers with Dropout:

After the bidirectional GRU layer, the model includes dense layers. Dense layers are fully connected layers where each neuron is connected to every neuron in the previous and subsequent layers.

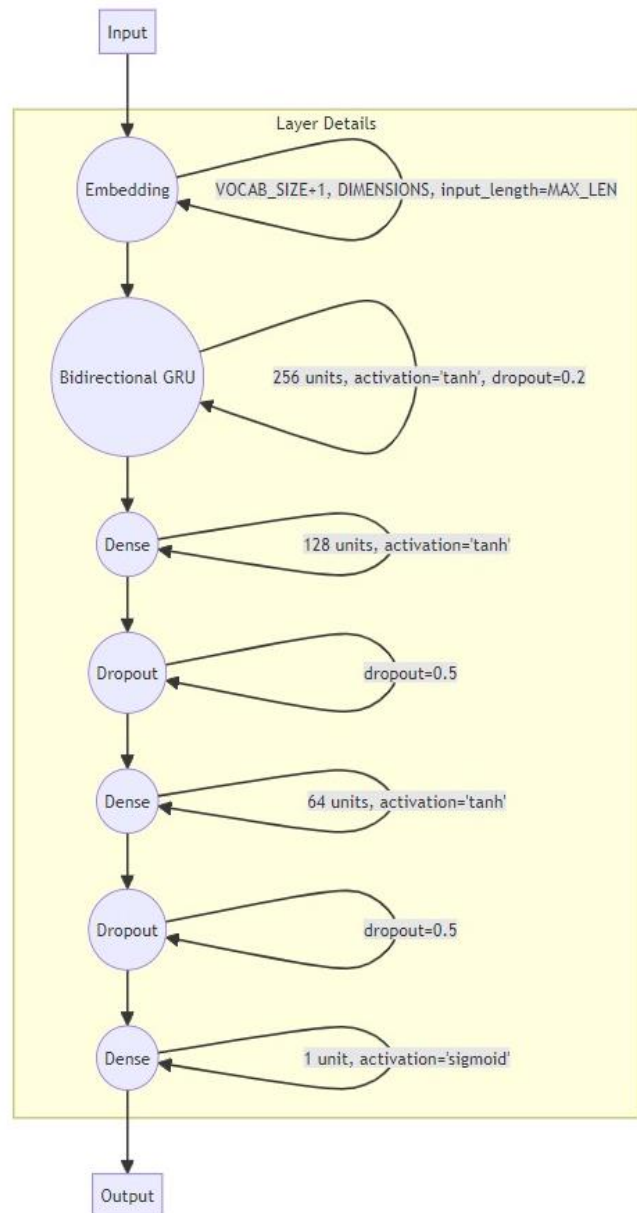
The first dense layer has 128 units with a hyperbolic tangent (tanh) activation function. This layer helps the model learn complex patterns and features from the output of the GRU layer.

Dropout layers are inserted after each dense layer with a dropout rate of 0.5 (50%). Dropout is a regularization technique that randomly "drops out" a fraction of units during training, preventing overfitting and improving model generalization.

##### Output Layer:

The final layer in the architecture is the output layer. It consists of a single neuron with a sigmoid activation function.

The sigmoid activation function outputs a value between 0 and 1, representing the predicted probability that the input text belongs to the positive sentiment class (1) or negative sentiment class (0). A threshold can be applied to this probability to make binary sentiment predictions.



Layer (type)	Output Shape	Param #
embedding_10 (Embedding)	(None, 1398, 128)	13839616
bidirectional_8 (Bidirection	(None, 512)	592896
dense_30 (Dense)	(None, 128)	65664
dropout_20 (Dropout)	(None, 128)	0
dense_31 (Dense)	(None, 64)	8256
dropout_21 (Dropout)	(None, 64)	0
dense_32 (Dense)	(None, 1)	65
Total params: 13,786,497		
Trainable params: 666,881		
Non-trainable params: 13,039,616		

## V. EXPERIMENTATION

The experimentation phase of our study involved systematically testing and evaluating different neural network architectures to determine the most effective model for sentiment analysis on Urdu movie reviews. We explored the following architectures:

1. **Unidirectional LSTM (Long Short-Term Memory):** LSTM networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. In unidirectional LSTMs, information flows only in one direction, making them suitable for tasks where past information is crucial for prediction.
2. **Unidirectional GRU (Gated Recurrent Unit):** GRU networks are like LSTMs but have a simpler architecture with fewer parameters. They also address the vanishing gradient problem and are faster to train compared to LSTMs. Unidirectional GRUs were included in our experimentation to compare their performance with LSTMs.
3. **Bidirectional LSTM:** Bidirectional LSTMs incorporate information from both past and future time steps by processing sequences in both forward and backward directions. This allows the model to capture context from both directions, potentially improving performance in tasks that require understanding context from the entire sequence.
4. **Bidirectional GRU:** Like bidirectional LSTMs, bidirectional GRUs process information from both directions. GRUs have fewer parameters than LSTMs and can be computationally more efficient while still capturing long-term dependencies.

During experimentation, we varied several parameters to optimize model performance:

- **Layer Sizes:** We tested different sizes (512, 256, 128, 64 etc.) for the LSTM and GRU layers to find the optimal balance between model complexity and learning capacity. Larger layers can capture more complex patterns but may also increase training time and risk overfitting.

- **Dropout Rates:** Dropout is a regularization technique used to prevent overfitting by randomly dropping neurons during training. We experimented with different dropout rates to control model complexity and generalization. We tried using 0.3 and 0.5 dropout rates and then settling on the rate that gave the best performance according to the various performance metrics.
- **Activation Functions:** Activation functions play a crucial role in neural network architectures by introducing non-linearity. We tested different activation functions such as tanh, sigmoid, and relu to assess their impact on model performance.

By systematically varying these parameters and evaluating the performance metrics (such as accuracy, F1 score, and training efficiency) for each architecture, we aimed to identify the most effective model configuration for sentiment analysis on Urdu movie reviews. This rigorous experimentation process ensures that our final model is well-tuned and optimized for the task at hand.

## VI. RESULTS AND ANALYSIS

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2 \times P \times R}{P + R},$$

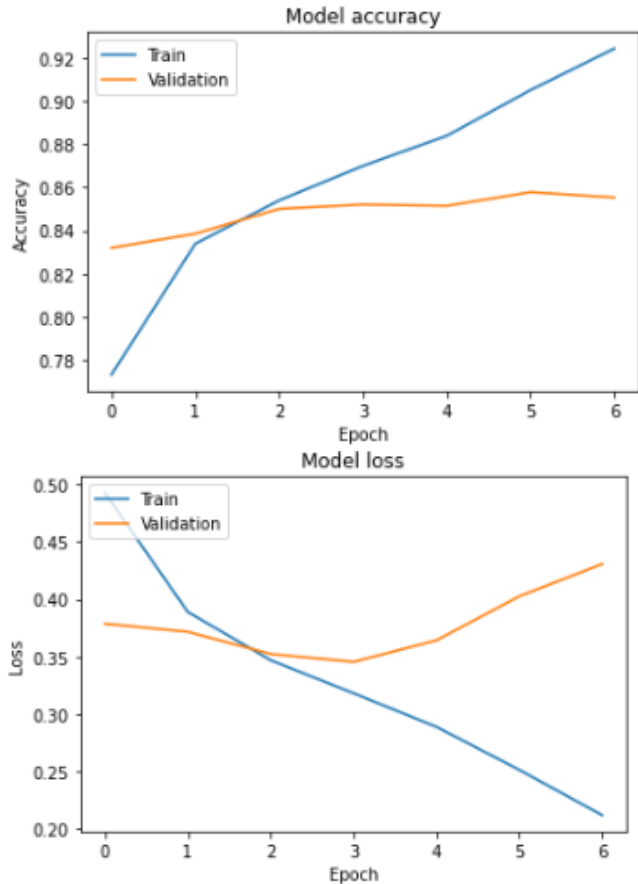
- In our study, we evaluated various neural network architectures for sentiment analysis on Urdu movie reviews. Among these architectures, the bidirectional GRU (Gated Recurrent Unit) model emerged as the top performer, demonstrating superior performance compared to other architectures. Here are the key findings from our analysis:
- **Superior Performance:** The bidirectional GRU model consistently outperformed other architectures in terms of multiple performance metrics, including validation accuracy, precision, recall and F1 score. These metrics are crucial for evaluating the model's ability to correctly classify sentiments in the reviews.
- **High Accuracy:** The bidirectional GRU model achieved high accuracy, 90% training accuracy and 85% validation accuracy. Bidirectional LSTM had 92% training accuracy, but the validation accuracy was lower, 84%.
- **F1 Score:** The F1 score is a metric that balances precision and recall, providing a comprehensive evaluation of the model's performance, especially in binary classification tasks like sentiment analysis. The bidirectional GRU model achieved a higher F1 score of 0.86 indicating its ability to achieve a good balance between identifying positive and negative sentiments while minimizing misclassifications. Compared to the bidirectional LSTM which had an F1 score of 0.85.
- **Efficient Training:** In addition to its superior performance, the bidirectional GRU model exhibited

efficient training compared to both unidirectional models (such as unidirectional LSTM and unidirectional GRU) and bidirectional LSTM. Bidirectional GRU was able to achieve the best results whilst training for 20 minutes. On the other hand, bidirectional LSTM gave nearly similar results but had a higher training time of around 26 minutes.

- **Effectiveness for Urdu Text:** The success of the bidirectional GRU model underscores its effectiveness specifically for sentiment analysis on Urdu text. Urdu, being a complex language with unique linguistic characteristics, benefits from architectures like bidirectional GRU that can capture contextual dependencies effectively from both past and future sequences.

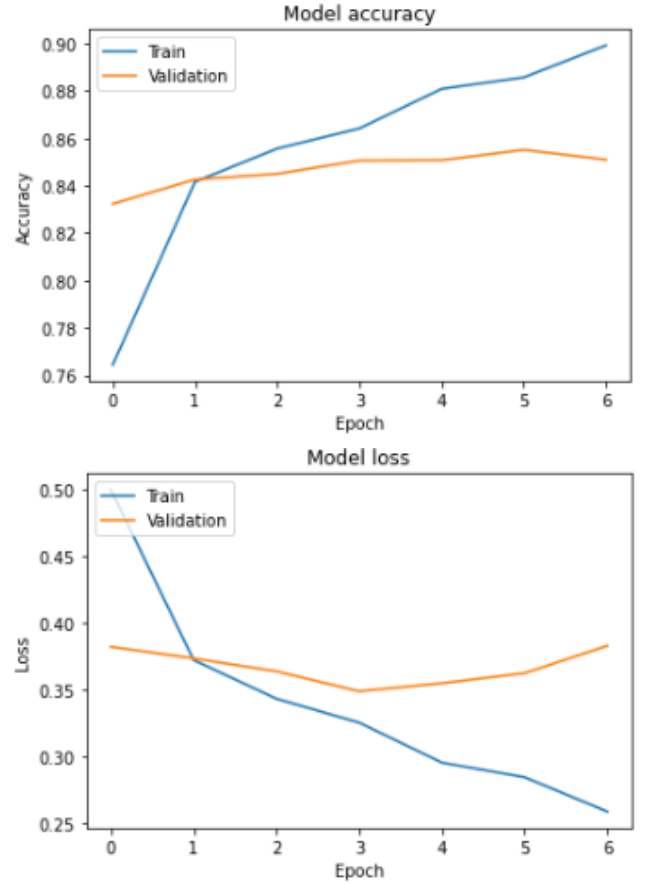
Overall, our results highlight the effectiveness of bidirectional GRU architectures for sentiment analysis on Urdu movie reviews. The combination of high accuracy, F1 score, and efficient training makes the bidirectional GRU model a promising choice for sentiment analysis tasks in languages like Urdu, where capturing nuanced sentiment nuances is essential. Below are the results of

## 1. Bidirectional LSTM



	precision	recall	f1-score	support
0.0	0.80	0.89	0.84	4564
1.0	0.90	0.82	0.85	5436
accuracy			0.85	10000
macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

## 2. Bidirectional GRU



	precision	recall	f1-score	support
0.0	0.85	0.86	0.86	4795
1.0	0.87	0.86	0.87	5205
accuracy			0.86	10000
macro avg	0.86	0.86	0.86	10000
weighted avg	0.86	0.86	0.86	10000

## VII. LIMITATIONS

While the "IMDB Dataset of 50k Movie Translated Urdu Reviews" from Kaggle provides a substantial amount of data, there are a few considerations that might impact the results and analysis:

1. **Translation Accuracy:** The dataset comprises translated Urdu movie reviews. However, the accuracy of the translations could vary, leading to potential discrepancies or loss of nuanced sentiment expressions from the original Urdu text. Inaccurate translations may introduce noise or bias into the sentiment analysis results.
2. **Limited Domain Coverage:** The dataset focuses specifically on movie reviews, which may limit the model's ability to generalize to other domains or topics. Sentiment analysis models trained primarily on movie-related content may not perform as effectively when

applied to diverse text genres or industries.

3. **Bias in Labeling:** The sentiment labels (positive and negative) assigned to the reviews may reflect subjective judgments and individual annotator biases. This could introduce inconsistencies or misinterpretations in the labeled data, affecting the model's training and evaluation.
4. **Data Imbalance:** Although the dataset is described as having an equal number of positive and negative sentiment reviews, there could be instances of data imbalance within subcategories or specific themes. Imbalanced data distributions may impact the model's ability to generalize across sentiment classes and could result in biased predictions.
5. **Limited Sentiment Nuances:** Sentiment analysis typically categorizes text into positive, negative, or neutral sentiments. However, human emotions and opinions often exhibit nuanced variations that may not be fully captured by binary sentiment classification. The model's ability to detect and interpret subtle sentiment nuances may be limited by the dataset's granularity.

#### REFERENCES

- [1] Kaggle. IMDB Dataset of 50k Movie Translated Urdu Reviews. Available online: <https://www.kaggle.com/datasets/akkefa/imdb-dataset-of-50k-movie-translated-urdu-reviews>

