

**NLP Spring 2025**

**Assignment 3**



**Section:** 8A

**Name:** Abdullah Fayyaz

**RollNumber:** 21L-5208

**Submitted to:** Mamoona Akbar

National University Of Computer and Emerging Sciences

Department of Computer Science

Lahore, Pakistan

# Question # 1:

## Table:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| XLMBERT | 0.5388 | 0.6667 | 0.0348 | 0.0661 |
| mBERT | 0.6490 | 0.8372 | 0.3130 | 0.4557 |
| BILSTM | 0.6245 | 0.5852 | 0.6870 | 0.6320 |
| GRU | 0.4694 | 0.4694 | 1.0000 | 0.6389 |
| LSTM | 0.6163 | 0.5669 | 0.7739 | 0.6544 |
| RNN | 0.5469 | 0.5156 | 0.5739 | 0.5432 |

### **Best Model: LSTM**

## **F1-Score** is the primary metric for binary classification when dealing with imbalanced classes or when balancing Precision and Recall is critical.

## **LSTM** achieved the highest **F1-score (0.6544)**, which indicates the best overall performance among the tested models.

## While **mBERT** had the highest **Precision (0.8372)** and **GRU** had the highest **Recall (1.0000)**, their overall F1-scores were lower than LSTM's.

# Question # 2:

## Table:

| **Embedding** | **F1-Score** | **Accuracy** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| **Word2Vec** | 0.6269 | 0.5918 | 0.5490 | 0.7304 |
| **GloVe** | 0.6167 | 0.6245 | 0.5920 | 0.6435 |
| **FastText** | **0.6434** | **0.6245** | **0.5804** | **0.7217** |
| **ELMo** | 0.5630 | 0.5184 | 0.4903 | 0.6609 |

### **Best Embedding: FastText**

* **FastText** achieved the highest **F1-score (0.6434)**, matching the best **accuracy (0.6245)**.
* Its balanced precision and recall indicate robust classification performance across positive and negative sentiments.
* Therefore, **FastText is the most effective embedding for the LSTM model on this dataset.**

# Question # 3:

## Table:

| **Model** | **BLEU Score** |
| --- | --- |
| RNN Model | 0.0018 |
| BiRNN Model | 0.0538 |
| LSTM Model | 0.0573 |
| Transformer | **0.5158** |

### **Best Model: Transformer**

* The **Transformer model** achieved a **BLEU score of 0.5158**, dramatically outperforming all RNN-based models.
* This demonstrates the **power of self-attention and multi-head attention mechanisms** in handling the complexities of sequence-to-sequence translation, especially for morphologically rich languages like Urdu.
* RNN-based models, including LSTM and BiRNN, showed limited performance, highlighting their comparative weakness in capturing long-range dependencies without attention mechanisms.

# Question # 4:

## Table:

| **Model** | **BLEU Score** |
| --- | --- |
| RNN (Random) | 0.0016 |
| RNN (GloVe) | **0.0019** |

### **Best Model: RNN (GloVe)**

* **RNN (GloVe)** achieved a **BLEU score of 0.0019**, outperforming **RNN (Random)** with a score of **0.0016**.
* **GloVe embeddings** provide richer, pre-trained word representations, improving model accuracy.
* The model with GloVe embeddings converges **faster** and performs better due to prior semantic knowledge.
* **RNN (Random)** struggles with learning relationships from scratch, leading to slower performance.