Bilingual Lemmatizer

For اُردو & پنجابی

Team

Malaika Basharat
20021519-004

Muhammad Umar
20021519-139

Supervisor: Dr. Zafar Mehmood Khattak

UNIVERSITY OF GUJRAT
A WORLD CLASS UNIVERSITY

CONTENTS

Introduction

1. Lemmatization is the process of reducing words into their root forms.
2. For example, the words “running”, “ran” and “runs” would all be lemmatized to the word “run”.
3. Lemmatization can be a difficult task as Urdu and Punjabi have complex morphology, but it is essential for a variety of NLP applications like
   - Information retrieval
   - Machine translation
   - Text classification
   - Sentiment analysis.

Words	Lemmas
کھانا	کھا
کتا باں	کتا ب
پہاڑیاں	پہاڑی

Problem and Proposed Solution

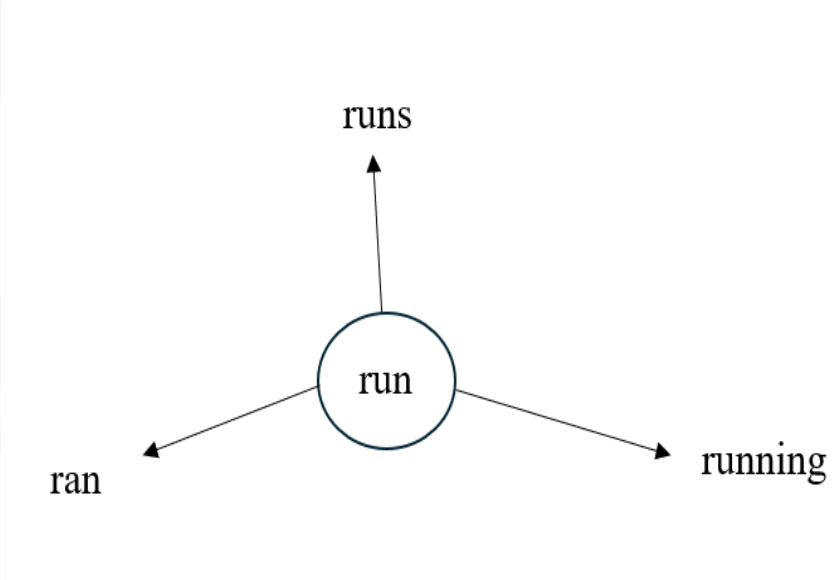
Problem Statement:

- Currently, there are no public Lemmatizers available for Urdu and Punjabi.

Proposed Solution

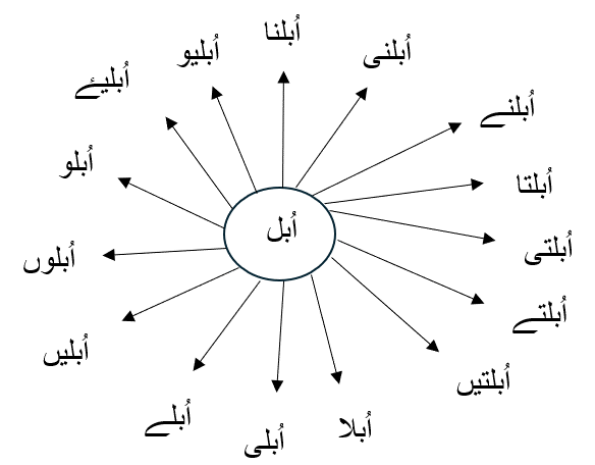
- Our solution is to develop a Deep Learning-based Bilingual Lemmatizer for Urdu and Punjabi.
- It will fill a major gap in the NLP toolkit
- Perform wider range of NLP tasks with Urdu and Punjabi text
- This lemmatizer will be more precise and efficient than earlier lemmatizers.

Morphology of Urdu and Punjabi languages as compared to other languages

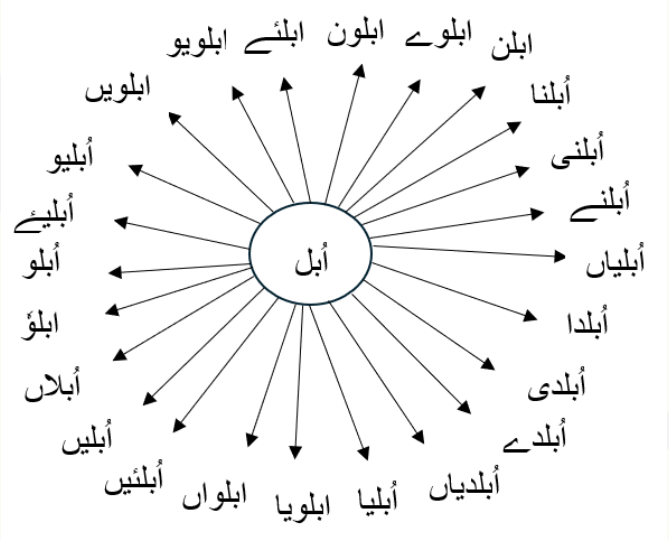


English forms

Urdu forms



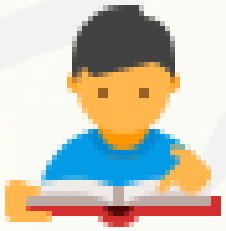
Punjabi forms



Targeted Audience

NLP Researchers

NLP researchers can use for language-specific studies, cross-linguistic research, and evaluation of NLP models.



Students

Students can enhance their educational experience and language learning journey.

Developers

Developers seeking to integrate language processing tools into their applications.

Language Enthusiast

Persons who are highly interested in learning a language.

Goals and Objectives

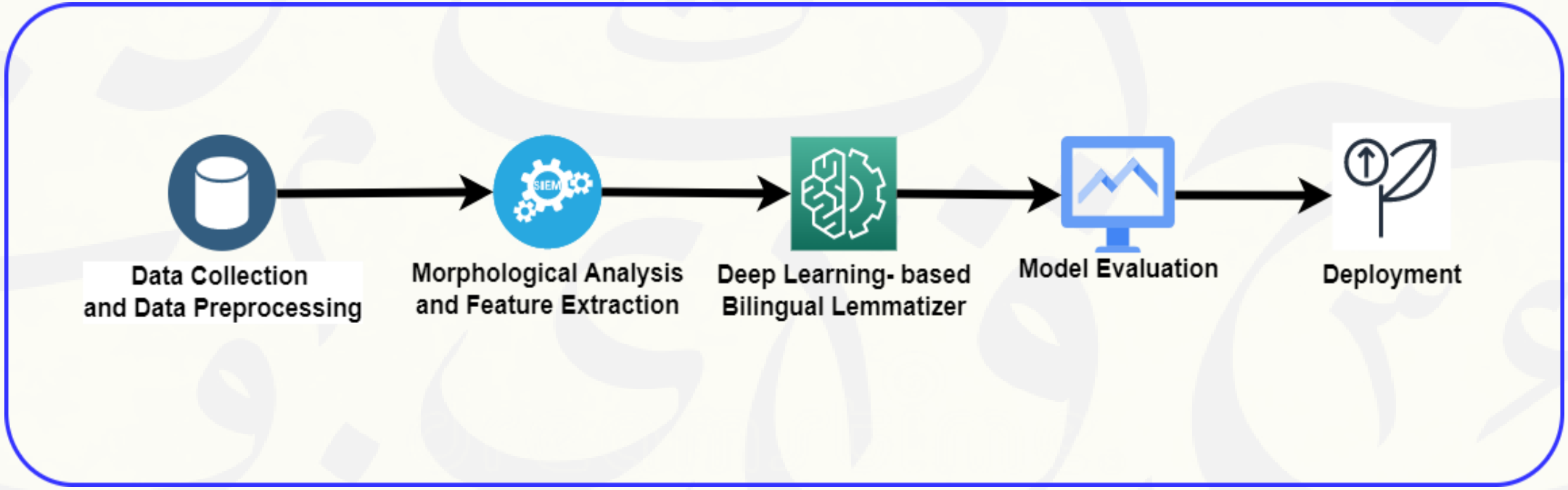
Goals

1. Develop high quality lemmatizer for Urdu and Punjabi language.

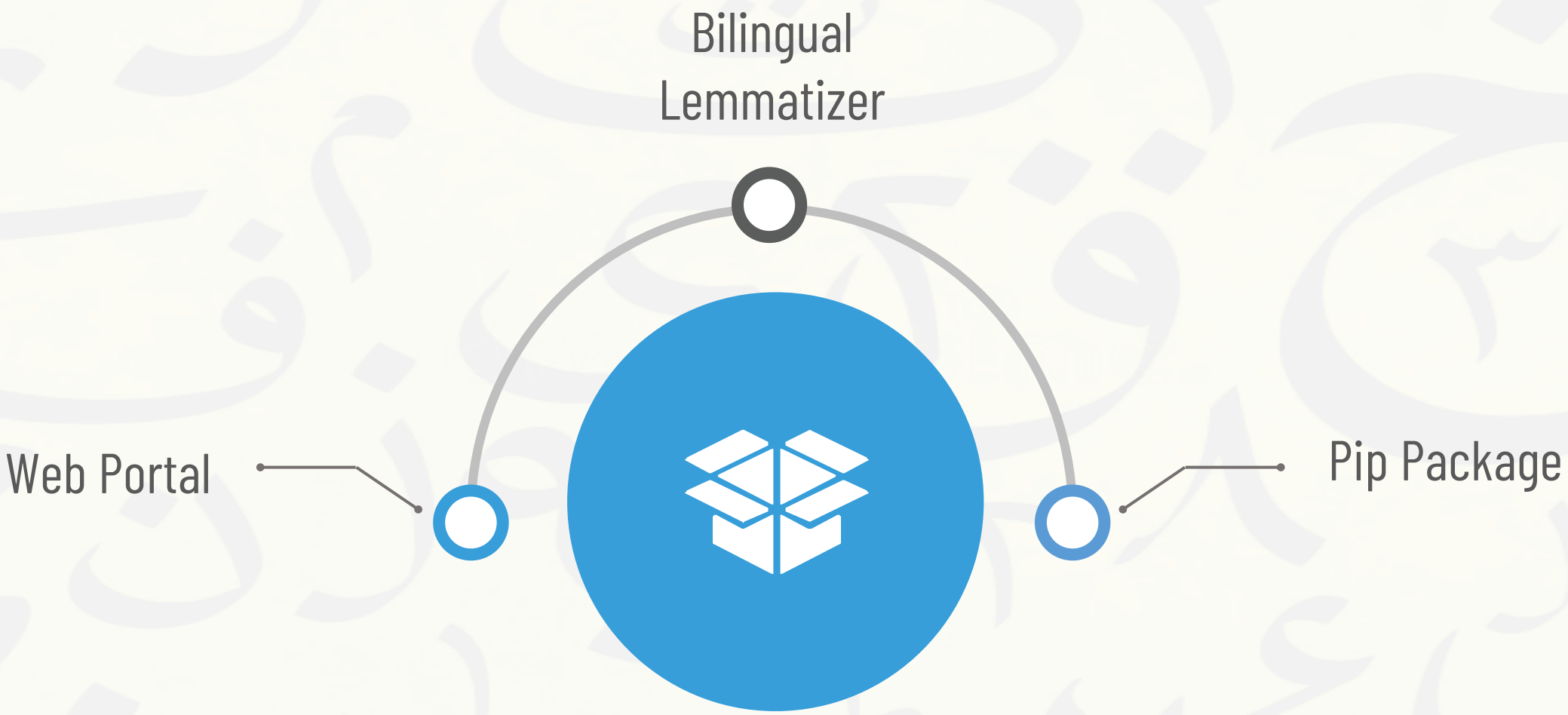
Objectives

1. Collect and preprocess a large corpus of Urdu and Punjabi text.
2. Extract relevant features from the tokenized text for effective lemma identification.
3. Develop a bilingual lemmatizer using deep learning models and Python libraries.
4. Deploy the lemmatizer as a pip package and web portal.

Architecture



Modules



Working of Lemmatizer

Text → Lemmatizer → Lemmatized Text

أبلتا

أبل

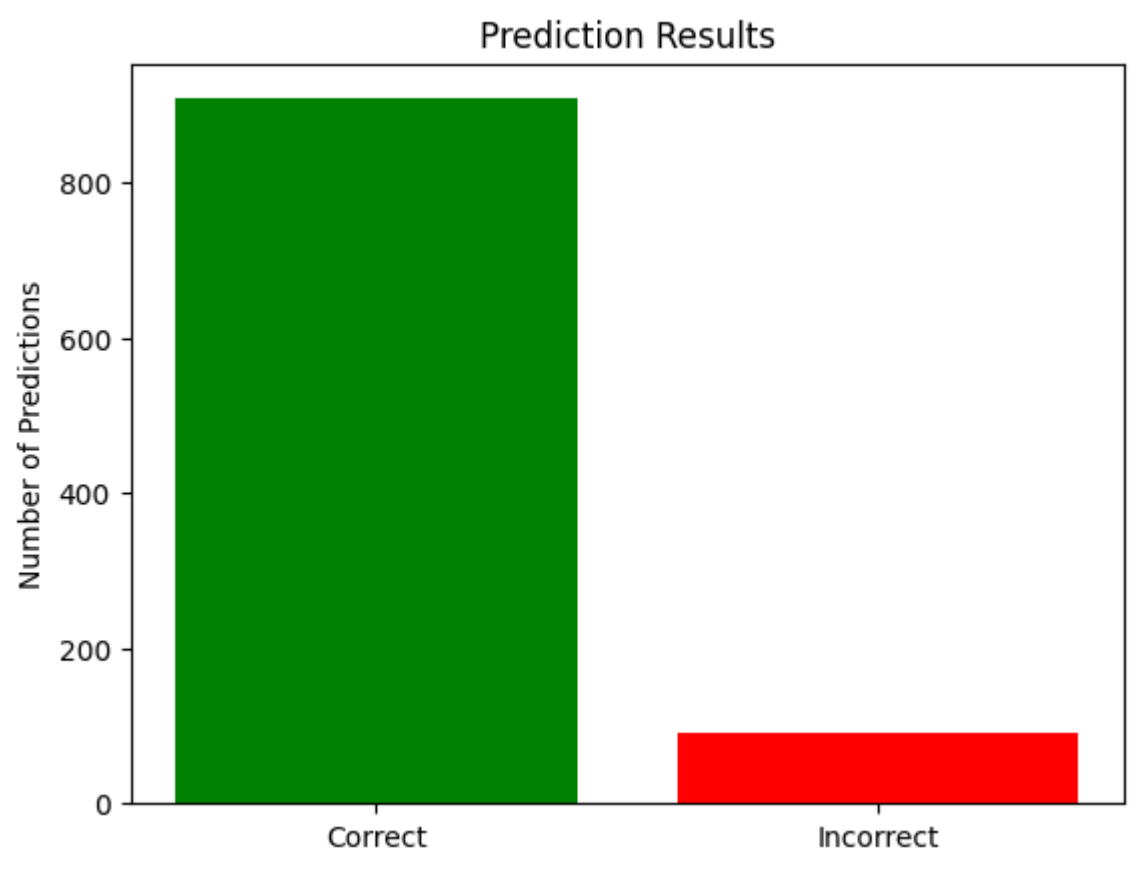
Results

Below table shows the overall results of both Urdu and Punjabi dataset. The accuracy, precision, recall and F-score.

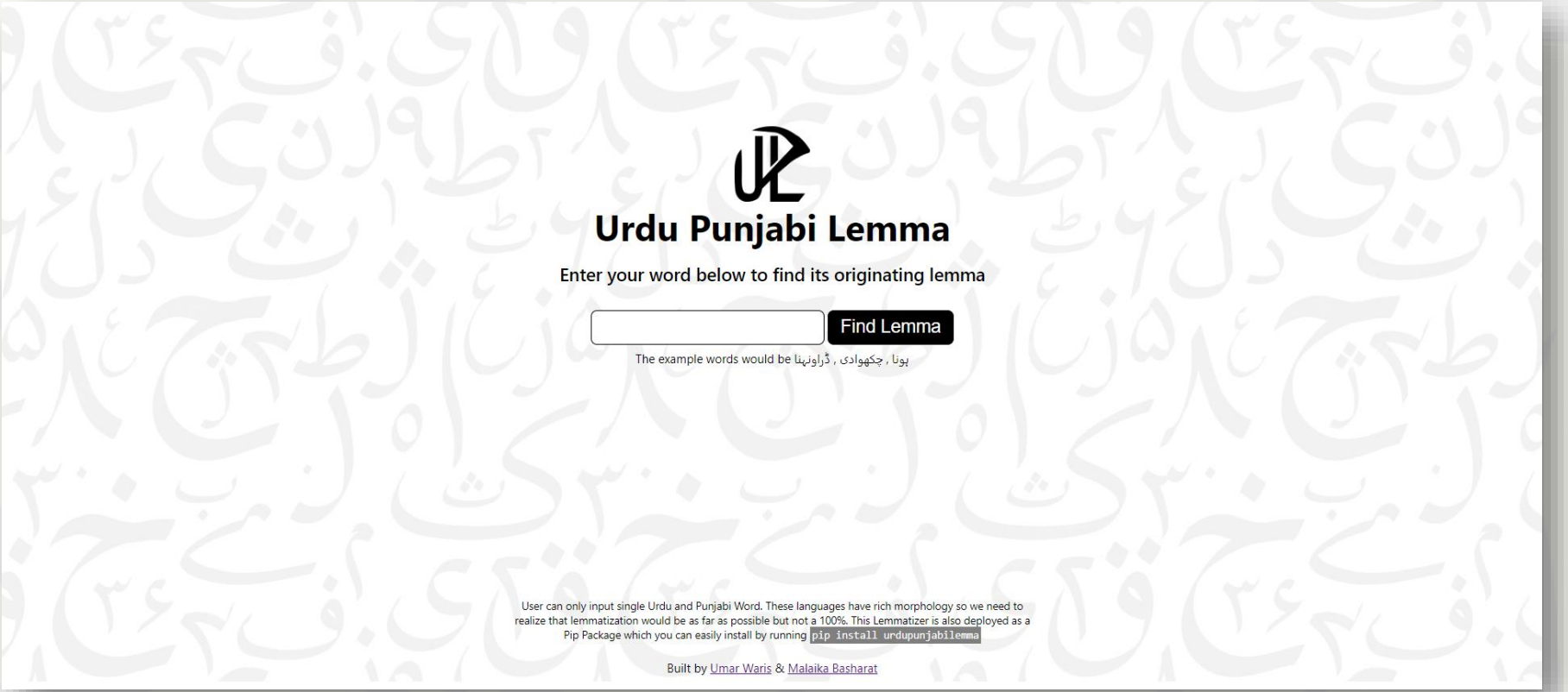
Models	Accuracy	Precision	Recall	F-score
RNN	96.4%	96.9%	92.6%	94.2%
LSTM	97.1%	94.9%	91.4%	92.8%
BiLSTM	97.8%	97.1%	93.6.%	95.1%
GRU	96.3%	94.9%	90.6%	92.3%

Results

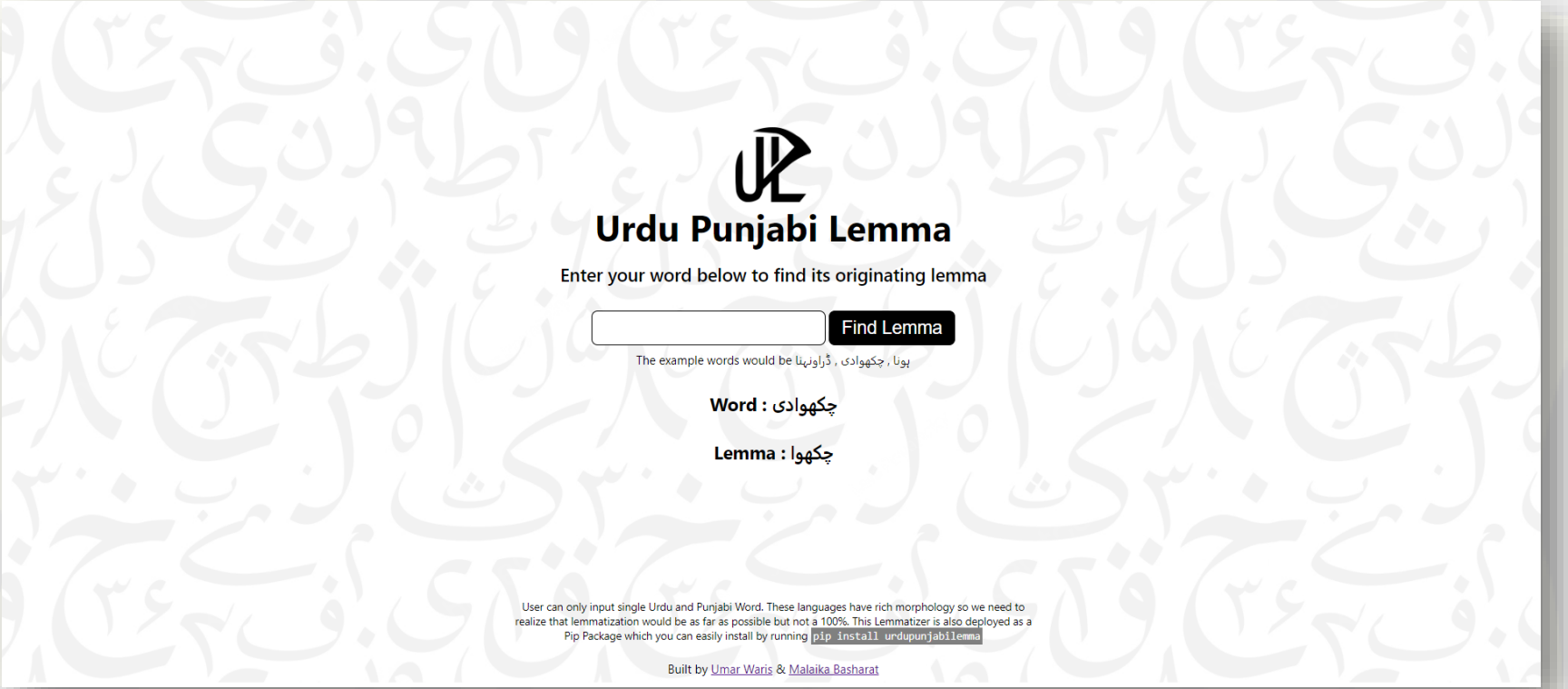
- Total Input Words: **1000**
- Correct Prediction: **908**
- Incorrect Prediction: **92**



Deployment as Web Portal



Testing of Web Portal



Deployment as Pip Package

Package could be easily installed by using command: pip install urdupuj

```
PS C:\Users\PMLS> pip install urdupunjabilemma
Collecting urdupunjabilemma
  Downloading urdupunjabilemma-0.2-py3-none-any.whl.metadata (596 bytes)
Requirement already satisfied: numpy in c:\users\pmls\appdata\local\programs\python\
a) (1.26.2)
Requirement already satisfied: tensorflow in c:\users\pmls\appdata\roaming\python\py
16.1)
Requirement already satisfied: MarkupSafe>=2.1.1 in c:\users\pmls\appdata\local\pr
kzeug>=1.0.1->tensorboard<2.17,>=2.16->tensorflow-intel==2.16.1->tensorflow->urdup
Downloading urdupunjabilemma-0.2-py3-none-any.whl (875 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 876.0/876.0 kB 147.0 kB/s eta 0:00:00
Installing collected packages: urdupunjabilemma
Successfully installed urdupunjabilemma-0.2
```

Testing of Pip Package

```
import urdupunjabilemma as upl
lemma=upl.lemmatize("کرنا")
print(lemma)
```

```
1/1 [==============================] - 0s 31ms/step
کر
```

```
import urdupunjabilemma as upl
lemma=upl.lemmatize("کتاباں")
print(lemma)
```

```
1/1 [==============================] - 0s 31ms/step
کتاب
```

THANKS