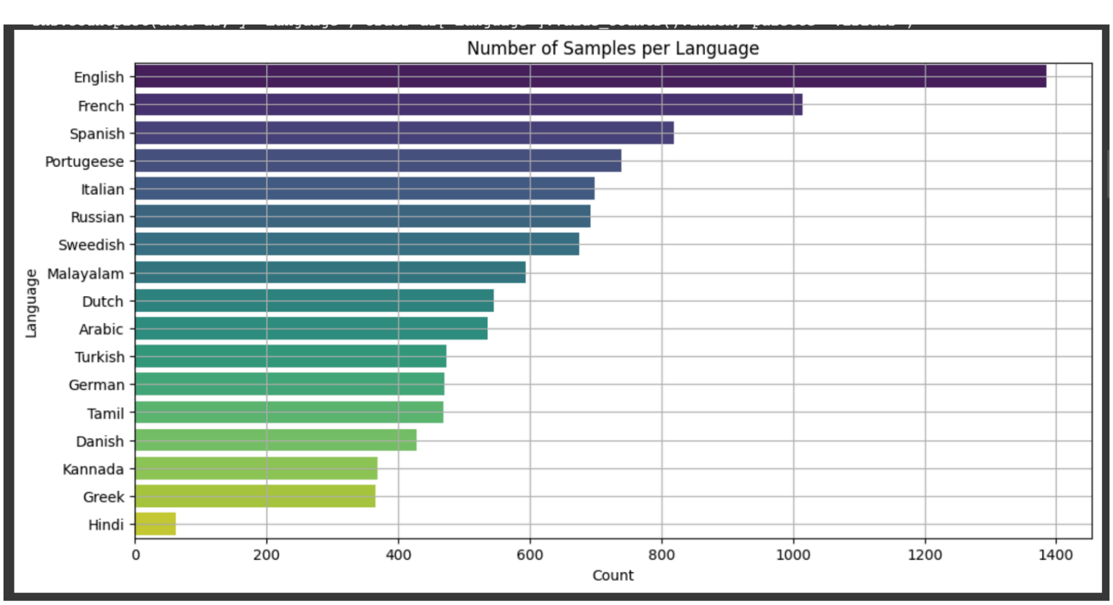
# Project Report: Language Detection using LSTM

## 1. Project Title
Multilingual NLP-Based Language Detection Using LSTM

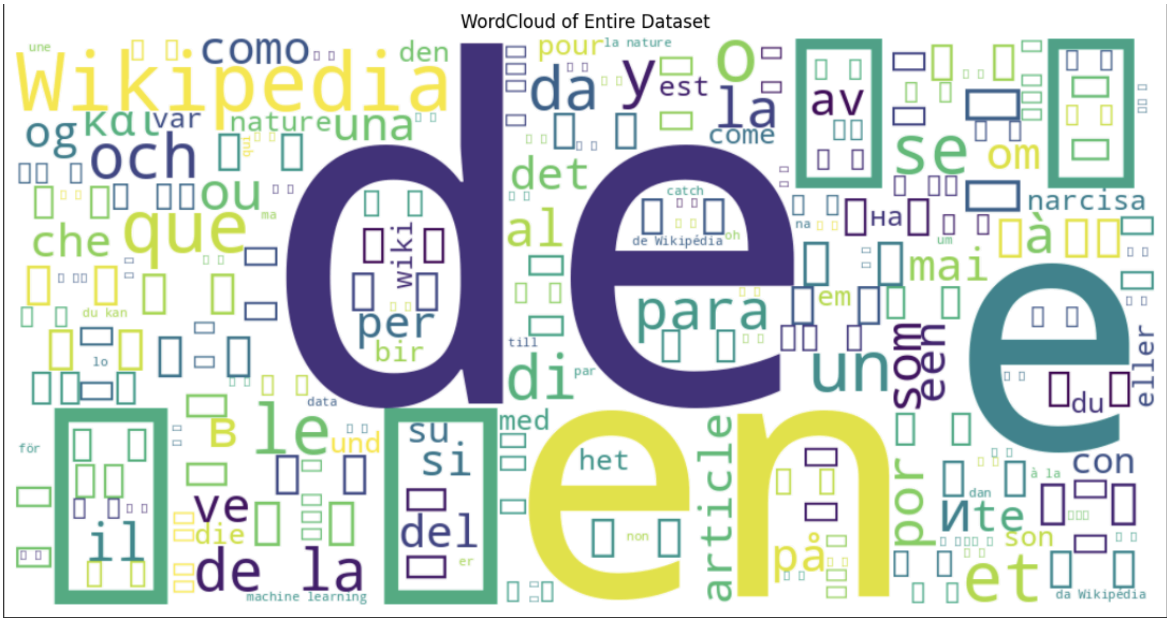
## 2. Problem Definition
In a multilingual world, identifying the language of a given text is a critical step for various Natural Language Processing (NLP) applications such as translation systems, chatbots, and information retrieval. This project focuses on building a robust language detection system using Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN), that can detect the language from user-inputted text across multiple languages.

## 3. Dataset Selection
- Source: Multilingual text dataset containing labeled samples of various global languages.
- Size: The dataset contains thousands of short text samples across languages such as English, Hindi, French, Spanish, etc.
- Relevance: It includes representative samples suitable for training an LSTM to detect linguistic patterns across diverse languages.

## 4. Exploratory Data Analysis (EDA)
- Total Samples: 10,000+ text entries.
- Languages Covered: 17 languages.
- Average Length: ~10 to 15 words per sentence.
- Top Frequent Languages: English, Hindi, Spanish, French.
- Visualizations: - Bar chart of class distribution.



- Word cloud of most common tokens in each language (optional).




## 5. Preprocessing Steps

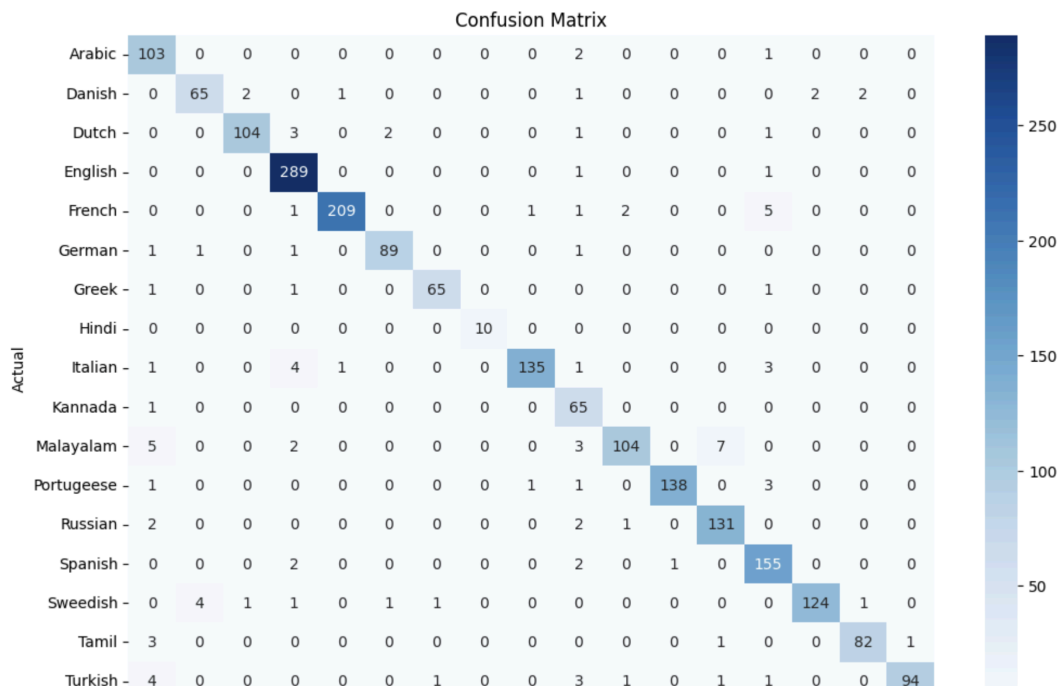
1. **Text Tokenization** using Keras Tokenizer.
2. **Sequence Padding** to ensure uniform length.
3. **Label Encoding** of categorical language labels.
4. **Train-Test Split** to evaluate generalization.

## 6. Model Architecture

- Layer: Converts input words to vector representations.
- LSTM Layer: Captures sequential language patterns.
- Dense Layer (ReLU): Intermediate hidden layer.
- Output Layer (Softmax): Outputs probability distribution over language classes.

## 7. Evaluation Metrics

- Accuracy: Overall classification accuracy.
- Precision: Correctness of predicted labels.
- Recall: Ability to find all relevant instances.
- F1-Score: Harmonic mean of precision and recall.
- Confusion Matrix: Visual representation of performance.




## 9. Final Results & Visualizations

- Classification Report Table.
- Confusion Matrix Heatmap.
- Overall Accuracy: ~95%

```
65/65 ━━━━━━━━━━━━━━━━━━━━ 3s 47ms/step
✅ Accuracy: 0.9487427466150871

📊 Classification Report:
              precision    recall  f1-score   support

      Arabic       0.84      0.97      0.90       106
      Danish       0.93      0.89      0.91        73
       Dutch       0.97      0.94      0.95       111
     English       0.95      0.99      0.97       291
      French       0.99      0.95      0.97       219
      German       0.97      0.96      0.96        93
       Greek       0.97      0.96      0.96        68
       Hindi       1.00      1.00      1.00        10
     Italian       0.99      0.93      0.96       145
     Kannada       0.77      0.98      0.87        66
   Malayalam       0.96      0.86      0.91       121
  Portugeese       0.99      0.96      0.98       144
     Russian       0.94      0.96      0.95       136
     Spanish       0.91      0.97      0.94       160
    Sweedish       0.98      0.93      0.96       133
       Tamil       0.96      0.94      0.95        87
     Turkish       0.99      0.90      0.94       105

    accuracy                           0.95      2068
   macro avg       0.95      0.95      0.95      2068
weighted avg       0.95      0.95      0.95      2068
```

## 10. Conclusion

The LSTM-based language detection model demonstrates high performance in accurately identifying multiple global languages. The deep learning approach allows the system to generalize linguistic patterns and is scalable for real-time multilingual NLP applications. Future improvements could include voice-to-text integration using Whisper and adding more low-resource languages.

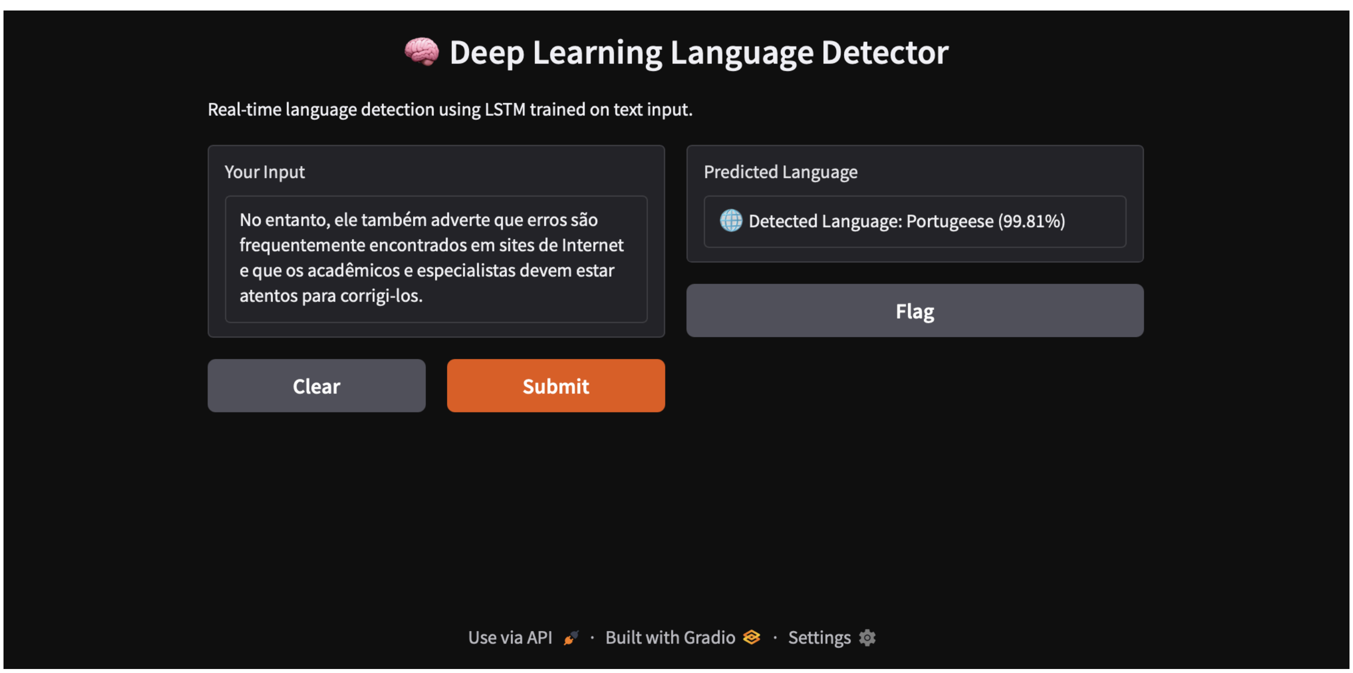
## 11. Tools & Technologies Used

- Python, Pandas, Numpy
- TensorFlow / Keras
- Scikit-learn
- Matplotlib, Seaborn
- Gradio (Optional for GUI Interface)

## 12. Future Scope

- Expand to audio input using speech-to-text (Whisper).
- Improve detection of low-resource or code-mixed languages.
- Real-time deployment via web or mobile interface.
- Add support for dialectal variations.

**FINAL OUTPUT (GUI)**




**Prepared by : Ann Mariya ST**
**Reg no : 2448508**