# Hate Speech Detection using LSTM Model

**Project Documentation**

## ****Table of Contents****

1. Introduction
2. Project Overview
3. Key Features
4. Setup & Installation
5. Dependencies
6. Data Description
7. Model Architecture
8. Preprocessing Workflow
9. How to Run the Project
10. Results & Performance
11. Contributing
12. Contact Information

## ****Introduction****

The **Hate Speech Detection** project uses **LSTM (Long Short-Term Memory)** for the detection of hate speech from text data. This project leverages Natural Language Processing (NLP) methods like tokenization, preprocessing text data, and padding sequences to train a sequential LSTM model capable of classifying hate speech from tweets.

## ****Project Overview****

The goal of this project is to classify tweets into two categories:

* **Hate Speech** (Class = 0)
* **Offensive Language** (Class = 1)
* **Neither** (Class = 2)

This is achieved by:

1. Preprocessing the raw data using tokenization and cleaning.
2. Splitting data into training and testing sets.
3. Creating an LSTM-based model trained with the processed text data.
4. Evaluating the performance of the model on unseen test data.

## ****Key Features****

* Preprocessing pipeline to clean and tokenize text data.
* LSTM-based neural network model for binary classification.
* Tokenization and padding with sequence length adjustments.
* Custom stopword removal to clean text effectively.
* Data splitting and preprocessed data storage for easy retraining.

## ****Setup & Installation****

To get the project running, follow these steps:

### ****1. Clone Repository****

git clone <repository-url>

### ****2. Navigate to the Project Directory****

cd HATE\_SPEECH\_DIY

### ****3. Set Up a Virtual Environment****

python -m venv venv

source venv/bin/activate # On Linux/macOS

venv\Scripts\activate # On Windows

## ****Dependencies****

Install the required dependencies:

pip install -r requirements.txt

Dependencies include:

* pandas
* nltk
* tensorflow
* scikit-learn
* matplotlib
* Other essential libraries.

## ****Data Description****

The dataset consists of tweets labeled as either "hate speech" or "non-hate speech."

### Dataset Path

../data/hate\_speech\_data.csv

### Columns in the Data

* **tweet**: The text of the tweet.
* **class**: Binary target column.
  + 0 = Hate Speech
  + 1 = Offensive Language
  + 2 = Neither

## ****Model Architecture****

The model uses an **LSTM** architecture built with TensorFlow/Keras.

### Model Features:

* Tokenizes text data using Keras' Tokenizer.
* Pads sequences to ensure uniform input dimensions.
* LSTM architecture to classify binary text data.
* Configured hyperparameters for tokenization and sequence padding.

## ****Preprocessing Workflow****

The preprocessing pipeline is crucial for model training and includes:

1. **Text Cleaning**:
   * Removal of special characters, numbers, and unnecessary spaces.
   * Stopwords removal using the nltk stopword corpus.
2. **Tokenization**:
   * Converting raw text into numerical data using Keras' Tokenizer.
3. **Padding Sequences**:
   * Uniformly padding sequences to ensure consistent sequence lengths for model input.
4. **Train-Test Split**:
   * Random split of data into training (80%) and test (20%) sets.
5. **Saving Preprocessed Data**:
   * Processed data stored for model retraining and reproducibility.

## ****How to Run the Project****

1. **Preprocess the data**:

Run preprocessing scripts: python scripts/models/preprocess\_text.py

1. **Train the LSTM model**:

Train using preprocessed data: python scripts/models/train\_model.py

1. **Evaluate Results**:

Use evaluation scripts to analyze model performance.

## ****Results & Performance****

The model will output performance metrics like **accuracy**, **precision**, **recall**, and **confusion matrix** based on test data after training.

Expected performance results include:

* Accuracy in binary classification of hate speech tweets.

## ****Citations****

### ****Libraries & Tools****

**pandas**: [Pandas Documentation](https://pandas.pydata.org/)

* + Citation: The pandas development team. (2020). pandas-dev/pandas: Pandas.

**nltk**: [Natural Language Toolkit](https://www.nltk.org/)

* + Citation: Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python.

**TensorFlow/Keras**: [Keras Documentation](https://github.com/keras-team/keras)

* + Citation: Chollet, F. (2015). Keras: The Python Deep Learning library.

**Scikit-learn**: [Scikit-learn Documentation](https://scikit-learn.org/stable/)

* + Citation: Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python.

### ****Dataset****

* **Dataset Source**: <https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset>

### ****Model Technique****

* **LSTM Model:** Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.