Folder Structure Overview

All relevant files and folder links are organised in the Google Document titled **“Links and File Structure”**. This document contains direct access to the necessary resources.

Folder: n-gram

Purpose: Contains all essential code, data, and model files for the n-gram-based language detection model.
Link: n-gram Folder on Google Drive

Contents:

- `clf.joblib` – Serialised classification model.
- `n-gram.ipynb` – Jupyter Notebook implementing the n-gram approach.
- `Ultimate_100_data.zip` – Compressed dataset used for training and evaluation.
- `vectorizer.joblib` – Serialised vectorizer used for feature extraction.

Folder: Fast_Text

Purpose: Contains all necessary code, data, and models for FastText-based language detection.
Link: Fast_Text Folder on Google Drive

Subfolder: 30_language

Description: FastText model trained on 30 languages.
Contents:

- `fast_text_30Lang.ipynb` – Jupyter Notebook for training and evaluation.
- `lang_detect_model.bin` – Trained FastText model binary.
- `mini-Ultimate_100_data_30lang.zip` – Compressed dataset for the 30-language model.
- `train.txt` – Training data formatted for FastText.

Subfolder: 200_language

Description: FastText model trained on 200 languages.
Contents:

- `fast_text_200Lang.ipynb` – Jupyter Notebook for training and evaluation.
- `lang_detect_model.bin` – Trained FastText model binary.
- `Ultimate_100_data_FastText.zip` – Compressed dataset for the 200-language model.
- `train.txt` – Training data formatted for FastText.
