Title: Boosting Performance of NLU Tasks for Indic Languages using Data Augmentation Strategies

README File for successful execution of models

In the github repository you will find 3 main folders for the languages Hindi, Kannada and Marathi. The dataset exists in the data directory inside each language folder. For understanding the data distribution of data, change directory to models/preprocessing. Preprocessing.ipynb consists of preprocessing techniques applies and EDA gives a good detail about dataset extracted as well data distribution.

To run the models, clone the repo

```
git clone https://github.com/KewlShubh/NLP_Mini_Project.git
```

You can find the models in the models directory, to change run the command,

```
cd <language>/models/
```

The models directory consists of non augmented data run models as well augmented run data models. The files name are self explanatory for eg:
`HindiBERT_RandomInsertion.ipynb` for running HindiBERT model with the Augmentation technique Random Insertion. Kindly refer to the project tree for successful execution of models:

```
Hindi
   |-- code
   |   |-- augmentation techniques
   |   |   |-- Back Translation.ipynb
   |   |   |-- Random Deletion.ipynb
   |   |   |-- Random Insertion.ipynb
   |   |   |-- Shuffle Transliteration.ipynb
   |   |   |-- Text Attack Augmenter.ipynb
   |   |-- models
   |   |   |-- Back Translation
   |   |   |   |-- DistilmBERT_BackTranslation.ipynb
   |   |   |   |-- HindiBERT_BackTranslation.ipynb
   |   |   |   |-- mBERT_BackTranslation.ipynb
   |   |   |-- Non-Augmented
   |   |   |   |-- DistilmBERT.ipynb
   |   |   |   |-- HindiBERT.ipynb
   |   |   |   |-- mBERT.ipynb
   |   |   |-- Random Deletion
   |   |   |   |-- DistilmBERT_RandomDeletion.ipynb
   |   |   |   |-- HindiBERT_RandomDeletion.ipynb
   |   |   |   |-- mBERT_RandomDeletion.ipynb
   |   |   |-- Random Insertion
   |   |   |   |-- DistilmBERT_RandomInsertion.ipynb
   |   |   |   |-- HindiBERT_RandomInsertion.ipynb
```

```
   |   |   |   |-- mBERT_RandomInsertion.ipynb
   |   |   |-- Text Attack Augmenter
   |   |   |   |-- DistilmBERT_TextAttack.ipynb
   |   |   |   |-- HindiBERT_TextAttack.ipynb
   |   |   |   |-- mBERT_TextAttack.ipynb
   |   |-- preprocessing
   |   |   |-- EDA.ipynb
   |   |   |-- Preprocessing.ipynb
   |-- data
   |   |-- BackTranslationAugData.csv
   |   |-- RandomDeletionAugData.csv
   |   |-- RandomInsertionAugData.csv
   |   |-- SynSubAugData.csv
   |   |-- clean-hindi-test.csv
   |   |-- clean-hindi-train.csv
   |   |-- gargi.ttf
   |   |-- hindi-test.csv
   |   |-- hindi-train.csv
Kannada
   |-- code
   |   |-- augmentation
   |   |   |-- Random_Deletion.ipynb
   |   |   |-- Random_Insertion.ipynb
   |   |   |-- Shuffle_Transliteration.ipynb
   |   |   |-- Synonym_Substitution.ipynb
   |   |-- models
   |   |   |-- augmented
   |   |   |   |-- GPT-2
   |   |   |   |   |-- Backtranslation_GPT_2_NLP.ipynb
   |   |   |   |   |-- RandomDeletionGPT_2_NLP.ipynb
   |   |   |   |   |-- RandomInsertion_GPT_2_NLP.ipynb
   |   |   |   |   |-- ShuffleTransliterationGPT_2_NLP.ipynb
   |   |   |   |   |-- SynonymSubstitution_GPT_2_NLP.ipynb
   |   |   |   |-- KNU-BERT
   |   |   |   |   |-- RandomDeletion_KNUBERT_NLP.ipynb
   |   |   |   |   |-- ShuffleTransliteration_KNUBERT_NLP.ipynb
   |   |   |   |   |-- backtranslation-knubert-nlp.ipynb
   |   |   |   |   |-- randominsertion-knubert-nlp.ipynb
   |   |   |   |   |-- synonymsubstitution-knubert-nlp.ipynb
   |   |   |   |-- KannadaBERT
   |   |   |   |   |-- RandomDeletion_KannadaBERT_NLP.ipynb
   |   |   |   |   |-- RandomInsertion_KannadaBERT_NLP.ipynb
   |   |   |   |   |-- ShuffleTransliteration_KannadaBERT_NLP.ipynb
   |   |   |   |   |-- backtranslation-kannadabert-nlp.ipynb
   |   |   |   |   |-- synonymsubstitution-kannadabert-nlp.ipynb
   |   |   |-- nonaugmented
   |   |   |   |-- GPT_2_NLP.ipynb
   |   |   |   |-- KNUBERT_NLP.ipynb
   |   |   |   |-- KannadaBERT_NLP.ipynb
   |   |-- preprocessing
   |   |   |-- EDA_NLP.ipynb
   |   |   |-- Preprocessing_NLP.ipynb
```

```
   |-- data
   |   |-- backtranslation_train.csv
   |   |-- clean_test.csv
   |   |-- clean_train.csv
   |   |-- clean_train.xlsx
   |   |-- fonts
   |   |   |-- Kar-Chandrashekhara-Kambara.ttf
   |   |-- randomdeletion_train.csv
   |   |-- randominsertion_train.csv
   |   |-- shuffletrasliteration_train.csv
   |   |-- synonymsubstitution_train.csv
   |   |-- test.csv
   |   |-- train.csv
LICENSE
Marathi
   |-- Data_Augmentation_Techniques
   |   |-- Back_Transilation
   |   |   |-- Augmented_Train.csv
   |   |   |-- Back_trasilation.ipynb
   |   |   |-- Test_clean.csv
   |   |   |-- Train_clean.csv
   |   |-- Random_Deletion
   |   |   |-- RandomDeletion_NLP.ipynb
   |   |   |-- randomdeletion_train.csv
   |   |-- Random_Insertion
   |   |   |-- RandomInsertionAugData.csv
   |   |   |-- Random_Insertion.ipynb
   |   |-- Text_Attack_Augmentation
   |   |   |-- Augmentation_TextAttack.ipynb
   |   |   |-- Augmented_Train.csv
   |   |   |-- Test_clean.csv
   |   |   |-- Train_clean.csv
   |-- Testing_Models
   |   |-- Distill_m_BERT_NLP.ipynb
   |   |-- M_BERT_NLP.ipynb
   |   |-- Marathi_BERT_NLP.ipynb
```