MARK MODELS INDEX

INITIAL THINGS TO REMEMBER :-

1) Entire project's backend now has 3 Base Models prepared with every Base Model having different accuracy. Base Model 3 is latest and is currently in use.

2) Mark 13, Mark 14 and Mark 15 are untested codes and are made in efforts to make new base models.

3) This document has last been updated on 04/11/2024 Monday.

TESTED MARK MODELS INDEX :-

1) Mark 1 :- Code Description :- This is Base Model 1. It only Works for English Language. It uses BERT(bert-base-uncased) and Pytorch.
Datasets Used :- Hasoc_English_Dataset_2019

2) Mark 2 :- Code Description :- Base Model 1 is used for this. It only Works for English Language. It uses BERT(bert-large-uncased) and Pytorch.
Datasets Used :- Hasoc_English_Dataset_2019

3) Mark 3 :- Code Description :- Base Model 1 is used for this. This is first model prepared for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch.
Datasets Used :- hasoc_hinglish_train_2021 + hasoc_hinglish_train_2022

4) Mark 4 :- Code Description :- This is Base Model 2. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process Text as Input.
Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

5) Mark 5 :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process Youtube Video Comments.
Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

6) Mark 6 :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process Audio as Input.
Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

7) Mark 7 :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process Image as Input.
Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

8) Mark 8_GIF :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process GIF as Input.

Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

Mark 8_Video :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process Video as Input.

Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

9) Mark 9 :- Code Description :- Base Model 2 is used for this. It Works for Hinglish Language. It is the Combinational Model of Mark 4,5,6,7,8 together. It uses BERT(bert-base-multilingual-uncased) and Pytorch. This Model has ability to process following input types - Text, Audios, Images, GIFs, Videos and YouTube Comments.

Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

10) Mark 10 :- Code Description :- This is Base Model 3. It Works for Hinglish Language. It uses BERT(bert-base-multilingual-uncased) and Tensorflow and Binary Classifier. This Model has ability to process Text as Input.

Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 1

11) Mark 11 :- Code Description :- Base Model 3 is used for this. It Works for Hinglish Language. It is the Combinational Model of Mark 10,5,6,7,8 together. It uses BERT(bert-base-multilingual-uncased) and Tensorflow and Binary Classifier. This Model has ability to process following input types - Text, Audios, Images, GIFs, Videos and YouTube Comments.

Datasets Used :- Github_Dataset 1_Hinglish_Tanmay_Version 2

12) Mark 12 :- Code Description :- Mark 11 Code has been updated and 2 most important changes have been made :-

1) New feature of converting Video to Audio and then passing that newly created Audio file in the system for classification has been added.

2) Emoticons Angle Version 1 has been added successfully. An array of hateful Emoticons has been added in the code itself. If the user input has any of emoticon present in this array then the entered text will be marked as Hate.

Datasets Used :-

1) Saved Model 1 :- Github_Dataset 1_Hinglish_Tanmay_Version 2

2) Saved Model 2 :- Hasoc_English_Dataset_2019 + Hasoc_Hinglish_Train_Dataset_2021 + Hasoc_Hinglish_Train_Dataset_2022 + Panchal Sir Dataset + Suyash English HSD + Yash Hinglish HSD

3) Saved Model 3 :- Hinglish Hate Speech Detection + Constraint_Dataset_Panchal_Sir_Transformed

----

Mark 12 Saved Model 3 is the last successfully tested model and its results have been used in making the Research Paper that has been submitted to the IEEE.
Mark 12 Saved Model 3 is the code that has gave the maximum accuracy until now and thus this was used in the research paper submitted to IEEE.
Mark 12 Saved Model 3 also marked the end of this project and all future models made have remained untested till date.

All models prepared after "Mark 12 Saved Model 3" are not in any relation with their adjacent models and have been developed with any prior research.

UNTESTED MARK MODELS INDEX :-

13) Mark 13 :- Experimented with different training configurations and adjustments that show a shift towards optimizing performance.
    1) Mark 13_PyTorch_Submark 1 :- Includes language detection to filter unknown languages and uses 6 training epochs without mixed precision.
       Mark 13_PyTorch_Submark 2 :- Excludes language detection, increases training to 8 epochs with frequent evaluation and logging every 250 steps, and uses mixed precision for faster training.
    2) Mark 13_Tensorflow_Submark 1 :- Implements a TensorFlow BERT model with an 8-epoch training loop, calculating average training loss per epoch and printing classification reports and accuracy.
       Mark 13_Tensorflow_Submark 2 :- This code is completely identical to Mark 13_Tensorflow_Submark 1 Code.
    Mark 13 Codes are completely untested and might contain errors.

14) Mark 14 :- This is the first time that this project moves away from BERT and performs an Expansion in Model Diversity. Various Data Augmentation Techniques have also been introduced and an ensemble of traditional ML models alongside transformer models has been implemented.
    1) Mark 14_Trial 1 :- Utilizes a single TFRoberta model with synonym-based data augmentation and a custom training loop.
    2) Mark 14_Trial 1 :- Implements multiple transformer models (Roberta, DistilBERT, XLNet), adds contextual BERT-based augmentation, and evaluates using an ensemble of traditional ML classifiers with TF-IDF features.
    Mark 14 Codes are completely untested and might contain errors.

15) Mark 15 :- A special hyperparameter tuning code is made for each of these novel approaches. Initially the hyperparameters need to be tuned and then the main code has to be adjusted with the tuned parameters to train the model.
    1) Mark 15_Novel 1 :- A PyTorch-based BERT + CRF model that combines BERT embeddings with TF-IDF features, leveraging synonym replacement for data augmentation and Ray Tune for hyperparameter optimization.

2) Mark 15 Novel 2 :- A TensorFlow ensemble model with BERT and TF-IDF features, enhanced by WordNet-based synonym replacement, focal loss for class imbalance, and Ray Tune for optimized training configurations.

3) Mark 15 Novel 3 :- A TensorFlow ensemble model integrating BERT embeddings with TF-IDF features, using back-translation for data augmentation, focal loss for handling imbalanced data, and Ray Tune for efficient hyperparameter tuning.

Mark 15 Codes are completely untested and might contain errors.
