Approach

Implementation Details:

The given problem statement is a binary classification based on the multi-modal data (Text and Images)

Approach:
The problem is tackled using the following approach:

1. **Text embedding extraction:**
    - Cleaning the input text by removing the stopwords, symbols, numbers, and unwanted spaces.
    - Convert the input text into vectors using the GloVe (Global vector representation) embedding and stack LSTM network

2. **Image feature extraction**
    - CNN-based pre-trained efficientNetV2B0 model is utilized for image feature extraction

3. **Concatenate text embeddings and image features and classify using the classification layer**
    - The text embeddings and image features are concatenated and passed through the classification layer to classify the input text and image as offensive or Non-offensive
    - The concatenated Network is trained on text embeddings extracted by stacked LSTM and image features from the CNN model for the classification of the input data.

The methodology is represented in the block diagram as shown in Fig.1

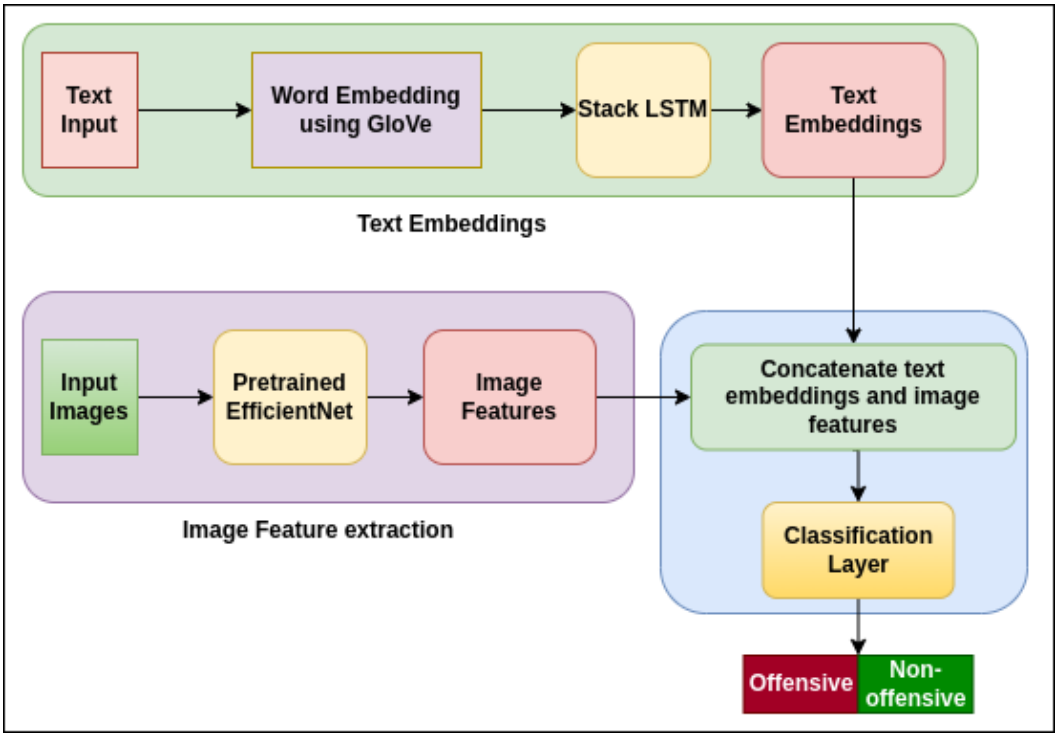

Fig.1: methodology followed for classification of text and images into Offensive or Non-offensive category

Execution of the python notebooks (.ipynb):

- Change the working directories in data analysis notebook file as per the requirements
- Install following dependencies: (requirements) [Google Colab Utilized for implementation] [Data analysis in local system]

```
PIL                 7.1.2
h5py                3.1.0
keras               2.9.0
keras_preprocessing 1.1.2
matplotlib          3.2.2
nltk                3.7
numpy               1.21.6
pandas              1.3.5
seaborn             0.11.2
session_info        1.0.0
sklearn             1.0.2
tensorflow          2.9.2
-----
-----
IPython             7.9.0
jupyter_client      6.1.12
jupyter_core        4.11.2
notebook            5.7.16
-----
Python 3.7.15
[GCC 7.5.0]
Linux-5.10.133+-x86_64-with-
Ubuntu-18.04-bionic
-----
```

Performance analysis:

Table 1: Performance Metrics

Model	Accuracy	Precision	Recall	F1-score
Combined classification Model	0.625	0.449	0.603	0.515

***Results can be varied according to the model hyperparameter settings and the data preprocessing techniques**

Fig.2: Model Training and Validation loss

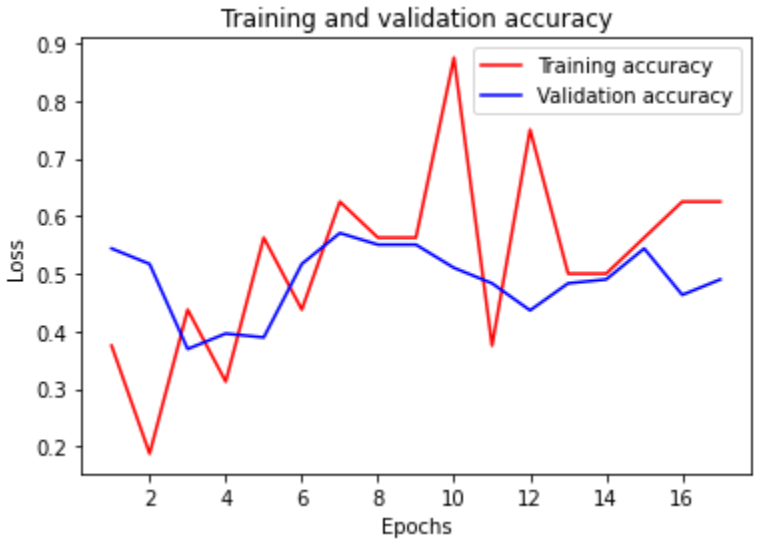

Fig.3: Model Training and Validation accuracy

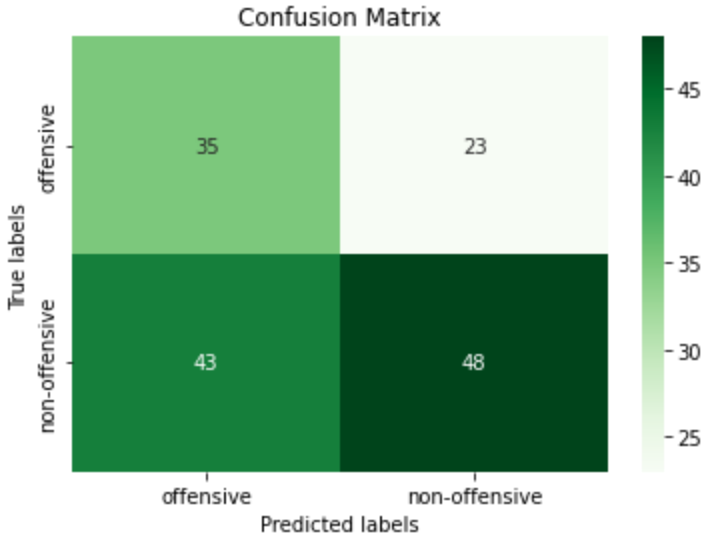

Fig.4: Confusion Matrix

References:

1. Suryawanshi, Shardul, et al. "Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text." *Proceedings of the second workshop on trolling, aggression, and cyberbullying*. 2020.