Sentiment Analysis Project Report

By Tanishq Selot, 3rd year (IIT Indore) for TrueFoundry

Please find the GitHub Repo Link here.

Model/Metric	Training Accuracy (90:10 train:test split)	Validation Accuracy	Testing Accuracy
LSTM + Word2Vec (Main)	0.9815	0.9647	0.9687
LSTM + GloVe	0.9481	-	0.4880
BERT	0.9966	0.9831	-

Steps:

1. **Data Reader Class & Text Pre-processing:**
    1. Cleaning tags/mentions
    2. Cleaning non-alphanumeric characters
    3. Cleaning URLs
    4. Cleaning Punctuations
    5. Cleaning repeating characters
    6. Cleaning numbers
    7. Stemming

    Resampling because of class imbalance in the favour of "Negative" class.

2. **Tokenization** – Converting sentences into list of strings

3. **Word Embedding** – Representing words in a form which can be understood by Deep Learning Models. Used Word2Vec because it performed better than GloVe and the most_similar() method gave reasonable results even for a small vocabulary. Fitting the model on all text. Later, using it for the embedding layer of the DL model.

4. **Train-test split & Padding** – Making the length of all the input vectors same.

5. **Model class** – Building, compiling, summarising, and fitting the model. A Spatial Dropout of 40% avoided overfitting and LSTM was used to retain the memory throughout longer texts. Later, I fine-tuned BERT as well but couldn't download the model from Colab due to its large size and abrupt runtime disconnection. Will train in local machine from next time.

6. **Model Evaluation**

7. **Inference class & testing** a random example. Finally deployed the model using FastAPI & documented using Swagger.