CS 613 NLP Assignment 3

Jethva Utsav
17110064

I have used a neural network-based approach to build a classifier that can classify the ENG-HIN code-mixed tweets based on their sentiments.

Dataset Statistics :

	Negative	Neutral	Positive	Total
Train	4459(29.47%)	5638(37.26%)	5034(33.27%)	15131
Test	533(28.51%)	754(40.34%)	582(31.14%)	1869

From the statistics, we can infer that datasets are nearly balanced.

The tweets are given in the tokenized form. I used regex to filter out tokens. I matched if the token has only [a-zA-Z] letters. It did most of the work but some of the URL tokens like 'http', 't', and 'co' were removed using conditional regex.

In the pre-processing hashtags and mentions are kept only('#' and '@') is removed, URLs, punctuations and emojis are completely removed. The remaining tokens are either ENG or HIN.

CSNLI library is used to correct misspelled English words and transliterated roman Hindi words to the Devanagari script.

Then Google Translate API is used to convert those code-mixed [ENG+HIN(Devanagari)] to pure English tweets

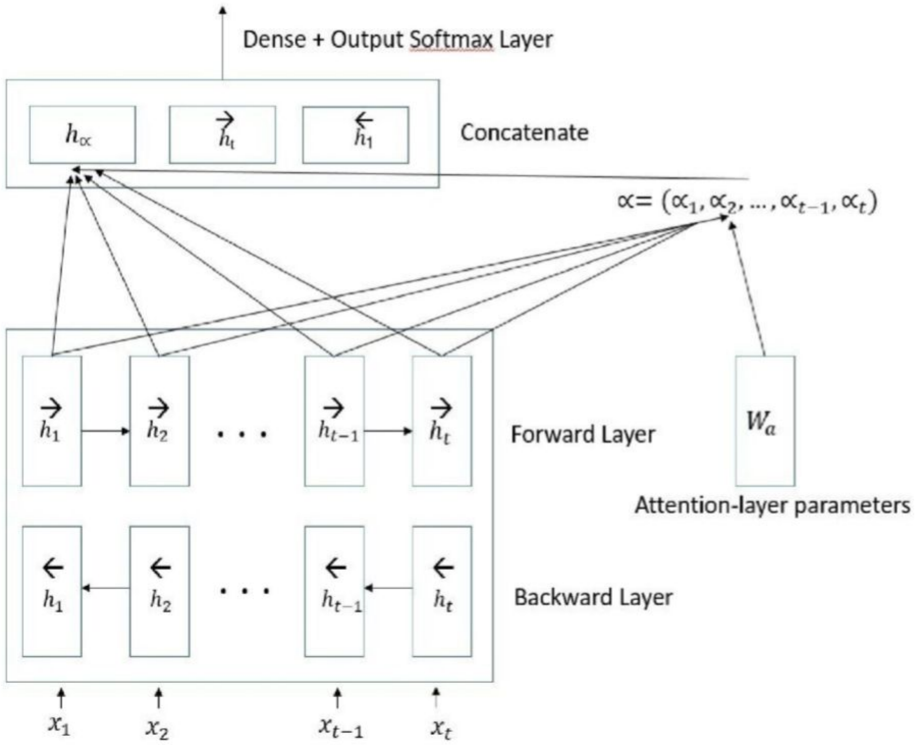
After filtering the tweets I used flair NLP library to get BERT embeddings. I created embeddings of the size 3072 dimensions (last four layers of BERT[base-cased- English pretrained model]).

Sentiments were one hot encoded to feed into the model.
Negative - 0 - [1, 0, 0]
Neutral - 1 - [0, 1, 0]
Positive - 2 - [0, 0, 1]

I used four different models to build the classifier:
Parameter and layer information along with obtained results are given the jupyter notebook for each model

1. Basic BiLSTM sequential model

2. BiLSTM with self-attention



3. BiLSTM + self-attention + GRU
4. Transformer (BERT sequence classifier)