

Reading the official SEMEVAL dataset

And processing the text into columns since initially each word is separated into separate rows

(using the methods from the paper : Sentiment Analysis of Code-Mixed Social Media Text (Hinglish) Gaurav Singh)

And saving the processed sentences to texts.csv

Text

Description automatically generated

Clean Tweet function for removing the un-necessary words from the processed texts

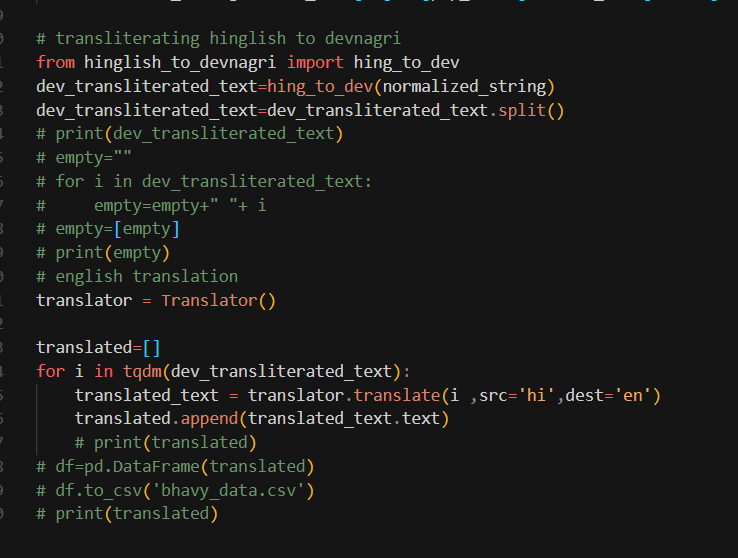
(using the methods from the paper : Sentiment Analysis of Code-Mixed Social Media Text (Hinglish) Gaurav Singh)

Text

Description automatically generated

These are done just for testing purpose

for creating words with same phoneme using pyphonetics library and checking leventise distance using Levenshtein library



Transliteration using manual letter by letter substitution method

Which is referenced from (<https://github.com/ritwikmishra/devanagari-to-roman-script-transliteration>)

But its not copied all together, so we studied their code and made necessary changes to make it for our use:

Text

Description automatically generated with medium confidence

Since they are working with converting devnagri hindi to hinglish but we wanted hinglish to devnagri

The opposite of what they have been doing so we made changes to their code to make it for our use

For converting hinglish to devnagri.

IN THE FILE hinglish\_to\_devnagri.py

But Since we didn’t get good results from this we further moved to using transformers.