Sentiment Analysis

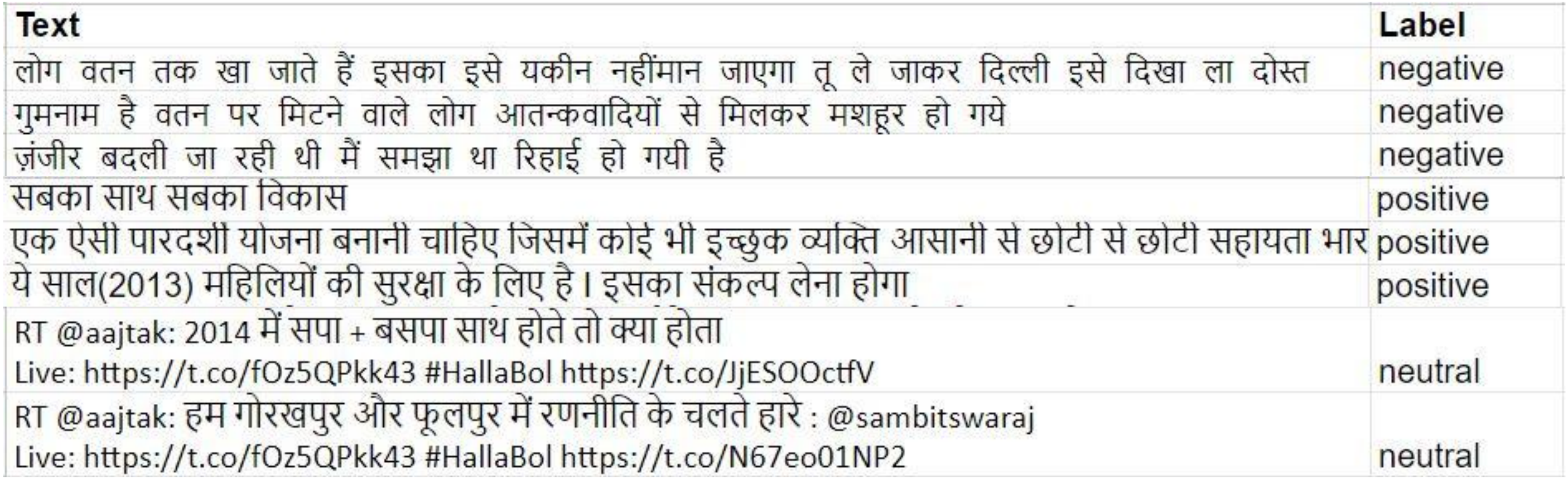
Analyzing hindi tweets into positive, negative, and neutral classes.

Group - 10

Quick Overview

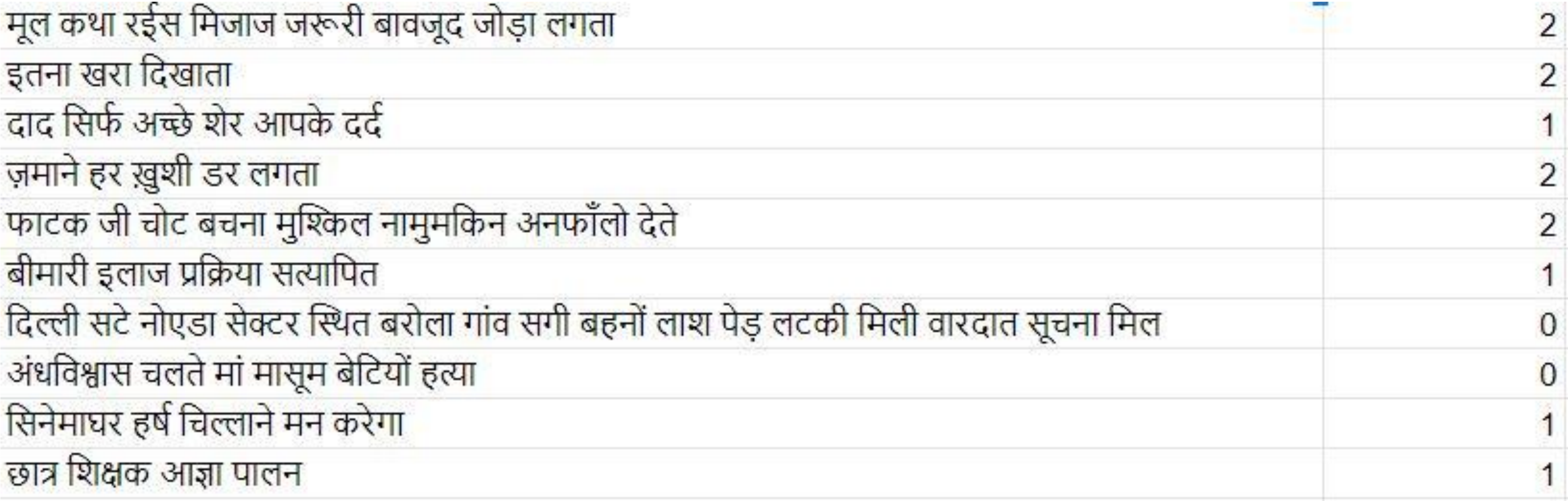
The problem that we hope to solve in this project is the efficient classification of the polarity or sentiment of a given text that will be the tweet in the Hindi language. We would classify the tweet in three categories, namely positive(1), negative(2) and neutral(0). We have chosen to use tweets as they are frequently used these days to express the tweeter’s emotion on any given subject.

Sample Data(Before Preprocessing)



Text	Label
लोग वतन तक खा जाते हैं इसका इसे यकीन नहींमान जाएगा तू ले जाकर दिल्ली इसे दिखा ला दोस्त	negative
गुमनाम है वतन पर मिटने वाले लोग आतन्कवादियों से मिलकर मशहूर हो गये	negative
ज़ंजीर बदली जा रही थी मैं समझा था रिहाई हो गयी है	negative
सबका साथ सबका विकास	positive
एक ऐसी पारदर्शी योजना बनानी चाहिए जिसमें कोइ भी इच्छुक व्यक्ति आसानी से छोटी से छोटी सहायता भार	positive
ये साल(2013) महिलियों की सुरक्षा के लिए है। इसका संकल्प लेना होगा	positive
RT @aajtak: 2014 में सपा + बसपा साथ होते तो क्या होता Live: https://t.co/fOz5QPkk43 #HallaBol https://t.co/JjESOOctfV	neutral
RT @aajtak: हम गोरखपुर और फूलपुर में रणनीति के चलते हारे : @sambitswaraj Live: https://t.co/fOz5QPkk43 #HallaBol https://t.co/N67eo01NP2	neutral

Processed Data



मूल कथा रईस मिजाज जरूरी बावजूद जोड़ा लगता	2
इतना खरा दिखाता	2
दाद सिर्फ अच्छे शेर आपके दर्द	1
ज़माने हर ख़ुशी डर लगता	2
फाटक जी चोट बचना मुश्किल नामुमकिन अनफाँलो देते	2
बीमारी इलाज प्रक्रिया सत्यापित	1
दिल्ली सटे नोएडा सेक्टर स्थित बरोला गांव सगी बहनों लाश पेड़ लटकी मिली वारदात सूचना मिल	0
अंधविश्वास चलते मां मासूम बेटियों हत्या	0
सिनेमाघर हर्ष चिल्लाने मन करेगा	1
छात्र शिक्षक आज्ञा पालन	1

Data Processing steps

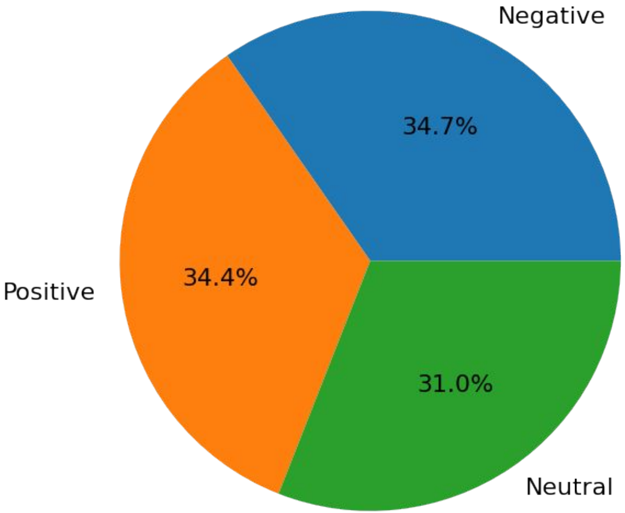
- Removing Hyperlinks, Hashtags, Mentions etc.
- Removing emojis.
- Removing hinglish words and some english word in between hindi tweets.
- Generating stopwords list and removing stopwords from text corpus.

Data

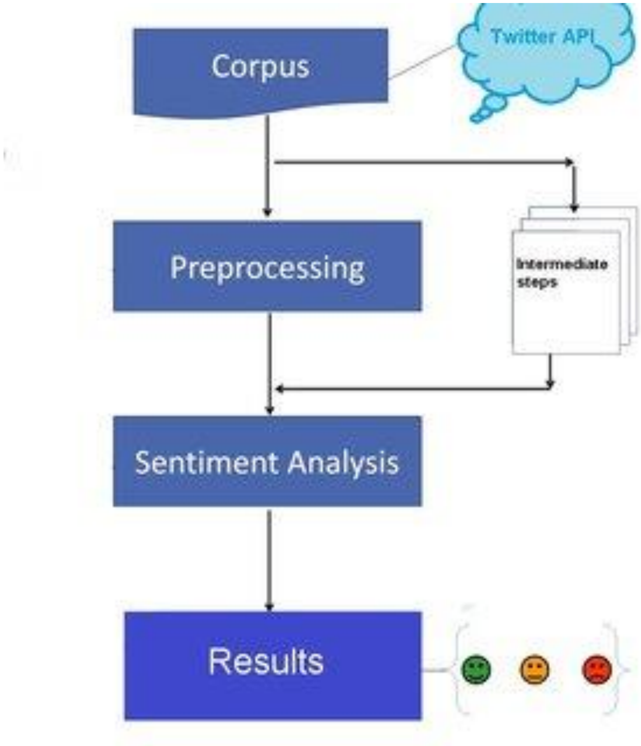
We have a total of 11,209 tweet which he have pre-processed and removed any tweet which was totally in English or Hinglish(Hindi written in English Text). **We collected and annotated the dataset manually.**

Of these 11,209 tweets we have:

- 3850 negative tweets
- 3437 neutral tweets.
- 3816 positive tweets.



Workflow


Corpus
Twitter API
Preprocessing
Intermediate steps
Sentiment Analysis
Results

Code Mixing

Approach 1 : (English to Hindi Translation) [word ref link]

- Generally, we make a model which converts the English words into Hindi words by which

  example: अच्छा - GOOD

- With the help of the model, we convert the English words in our tweets to Hindi so that meaning doesn’t change, and we can easily go with the Hindi words flow so that we don’t get any issue with word2vec model.

- OR we can also maintain a dictionary of words with English - Hindi words pair so that we can change the English words to Hindi words and move ahead with model. This approach will reduce the overall complexity of our design.

Code Mixing

Approach 2 : (Cross lingual model for code mixing(XLM))

- The CLM and MLM tasks work well on monolingual corpora, however, they do not take advantage of the available parallel translation data. Hence, the authors propose a Translation Language Modeling objective wherein we take a sequence of parallel sentences from the translation data and **randomly mask tokens from the source as well as from the target sentence**. For example, in the figure above, we have masked words from English as well as from the French sentence. **All the words in the sequence contribute to the prediction of a given masked word**, hence establishing a cross-lingual mapping among the tokens.

- *In **XLM** work, we consider cross-lingual language model pretraining with either CLM, MLM, or MLM used in combination with TLM.*

Code Mixing

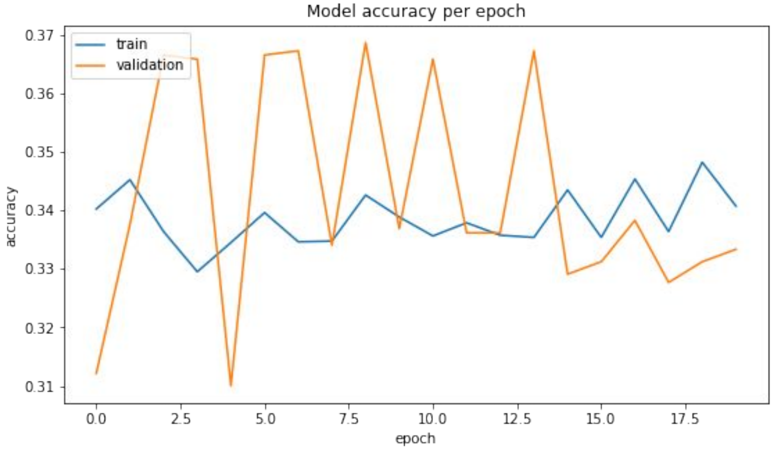
Approach 3 : (Word embedding for code mixing) [1]

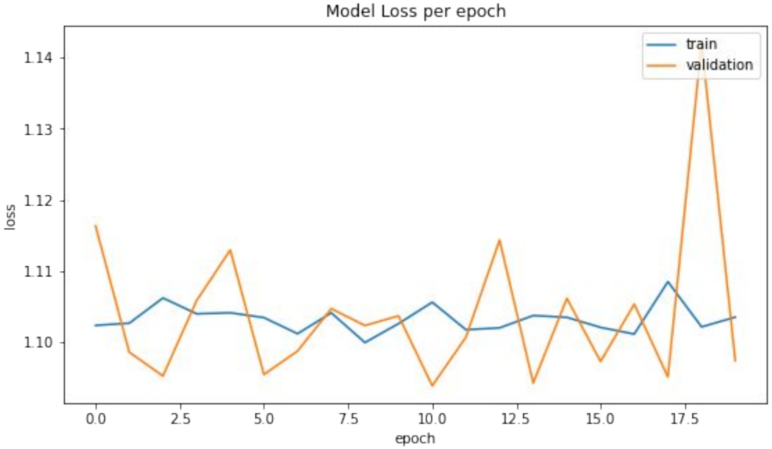
- We compare three existing bilingual word embedding approaches, and a novel approach of training skip-grams on synthetic code-mixed text generated through linguistic models of code-mixing, on two tasks - sentiment analysis and POS tagging for code-mixed text. Our results show that while CVM and CCA based embeddings perform as well as the proposed embedding technique on semantic and syntactic tasks respectively, the proposed approach provides the best performance for both tasks overall. Thus, this study demonstrates that existing bilingual embedding techniques are not ideal for code-mixed text processing and there is a need for learning multilingual word embedding from the code-mixed text.

Classification Models Used

Simple Recurrent Neural Network

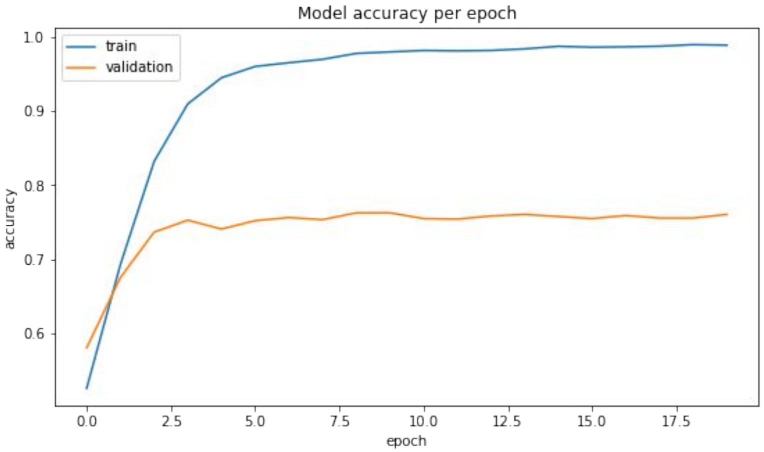
- The first classification model which we have used is a simple RNN model to get the sentiment classification. The RNN model has three layers and the optimizer used is ‘adam’ optimizer and dropout is also used to prevent overfitting.

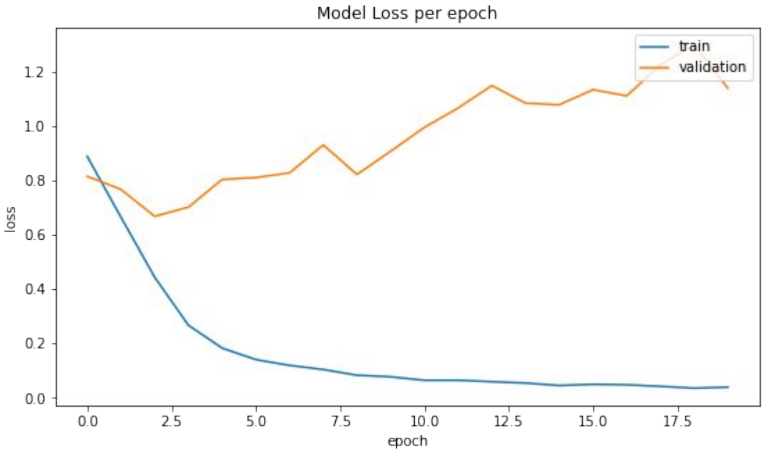




LSTM without Word Embeddings

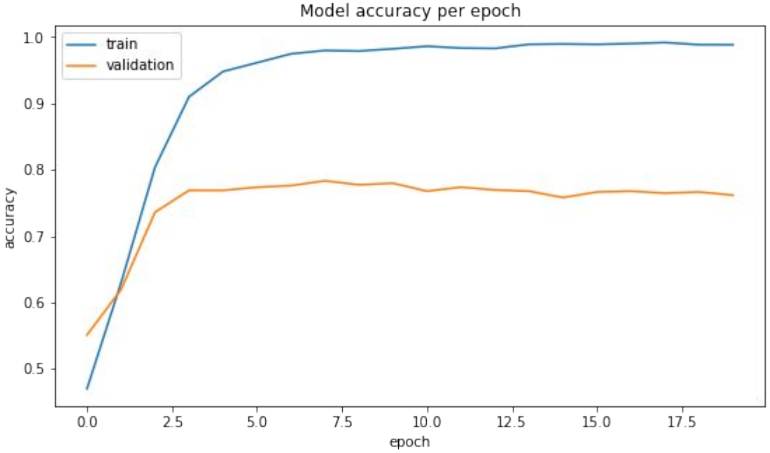
- The second classification model which we have used is a LSTM which used one hot vector in the embedding layer to represent a word but we did not use word embeddings like word2vec for this model.

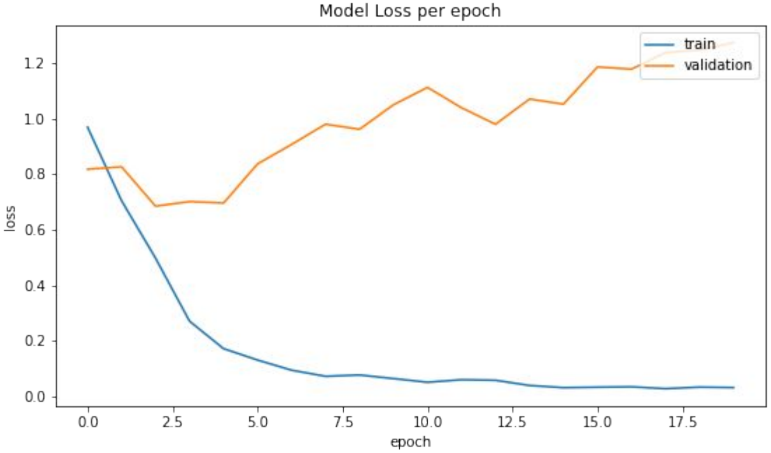




LSTM with Word Embeddings

- The third classification model which we have used is a LSTM which uses a word2vec word embedding for Hindi Language Text.
- We used pre – trained word embeddings provided by fasttext [1].





Work to be done

- We plan to make a better classifier by dividing the Hindi Language Tweets written in Devanagari Script into 5 classes i.e. from 0 – 4 with ‘0’ being the most positive tweet, ‘4’ being the most negative tweet and ‘2’ being a neutral tweet. This classification would give us more insight on the sentiment as we know that there are many tweets which will be positive or negative but they wouldn’t be on the same level of positivity or negativity as each other.

- To achieve this result, we are working on collecting and labelling more data and we also plan to implement a code –mixing and language processing model to get better results.

Software Components

- Python
- Tensorflow
- Natural Language Toolkit
- Colab
- Word2Vec
- Snscrape
- Twitter API
- Fasttext (pre - trained word embeddings)
- Keras
- TensorFlow

Summary

So far we have worked on data collection, data preprocessing and we have visualize the words which occur in each class i.e. neutral (0), positive (1) and negative (2). We have also implemented three different classification models namely – a Simple RNN model, a LSTM system and a LSTM which uses word embeddings. For better results in future we are planning to work on code mixing models. So that we can also work with multilingual corpus. We did literature survey on code mixing and code switching. As well as looked into microsoft Project ‘Project melange’ for better understanding.

References

- Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, Amitava Das. Overview of Sentiment Analysis of Code-Mixed Tweets , 2020[1]
- Naman Bansal, Umair Z Ahmad, Amitabha Mukherjee. Sentiment Analysis in Hindi , 2013[2]
- Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya, A Fall - Back Strategy for Sentiment Analysis in Hindi : a Case Study, 2010[3]
- Aanusha Ghosh, Indranil Dutta. Real - Time Sentiment Analysis of Hindi Tweets, 2014
- Snscrape [4]

References for Code Mixing

- D. Sitaram, S. Murthy, D. Ray, D. Sharma and K. Dhar, "Sentiment analysis of mixed language employing Hindi-English code switching," 2015 International Conference on Machine Learning and Cybernetics (ICMLC), 2015, pp. 271-276, Doi: 10.1109/ICMLC.2015.7340934.

- T Tulasi Sasidhar, Premjith B, Soman K P, “Emotion Detection in Hinglish(Hindi + English) Code-Mixed Social Media Text”, Procedia Computer Science, Volume 171, 2020, Pages 1346-1352, ISSN 1877-0509, [1]

- Yadav, Siddharth, and Tanmoy Chakraborty. "Unsupervised sentiment analysis for code-mixed data." *arXiv preprint arXiv:2001.11384* (2020).

- K. Shanmugalingam, S. Sumathipala and C. Premachandra, "Word Level Language Identification of Code Mixing Text in Social Media using NLP," 2018 3rd International Conference on Information Technology Research (ICITR), 2018, pp. 1-5, Doi: 10.1109/ICITR.2018.8736127.

- Project Mélange [Microsoft] [GLUECoS GitHub]