Sentiment Analysis of Hindi Tweets

Rishabh Jha
CSE
Indian Institute of Information Technology, Sri City
Andhra Pradesh
rishabh.j@iiits.in

Ashar Siddiqui
CSE
Indian Institute of Information Technology, Sri City
Andhra Pradesh
ashar.s19@iiits.in

Jainendra Prakash
CSE
Indian Institute of Information Technology, Sri City
Andhra Pradesh
jainendra.p19@iiits.in

K Avinash Reddy
CSE
Indian Institute of Information Technology, Sri City
Andhra Pradesh
avinashreddy.k19@iiits.in

***Abstract*—This electronic document states and describes our Natural Language Processing Model which we have applied to perform sentiment analysis on the Hindi Language tweets written in Devanagari Script.**

Keywords—Sentiment Analysis, Word Processing, Code Mixing, Code Switching, Time Series Model, Language Processing.

1. INTRODUCTION

The problem that we have solves in the Natural Language Processing Project is the efficient classification of the polarity or the sentiment of a given text that will be the given tweet in the Hindi Language. We have classified the given Hindi Language Tweets into three categories or classes, namely positive, negative, and neutral.

1.1 MOTIVATION

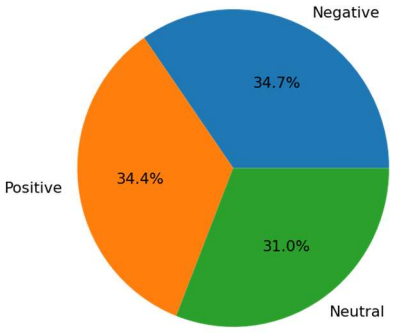
For our project we took the task of classifying tweets because these days almost everybody uses tweets to express their emotion on any given topic, be it general issues, movie reviews, or anything. We did this sentiment analysis on Hindi Language Tweets as we live in India and there are a lot of tweets which are written in Hindi Language Devanagari Script and processing these tweets are very useful for getting the feedback of any product, movie or even the general working of the government or any reviews.

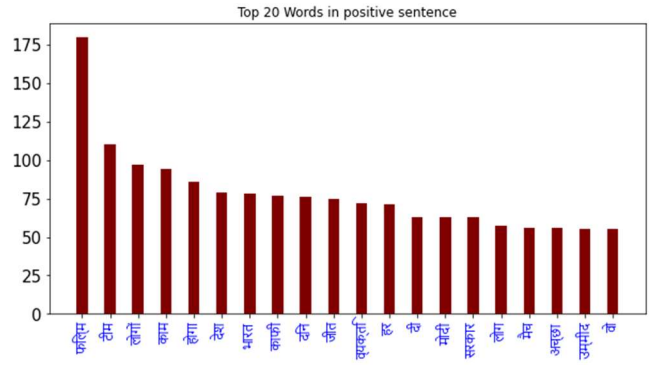
2. STATE OF THE ART / BACKGROUND

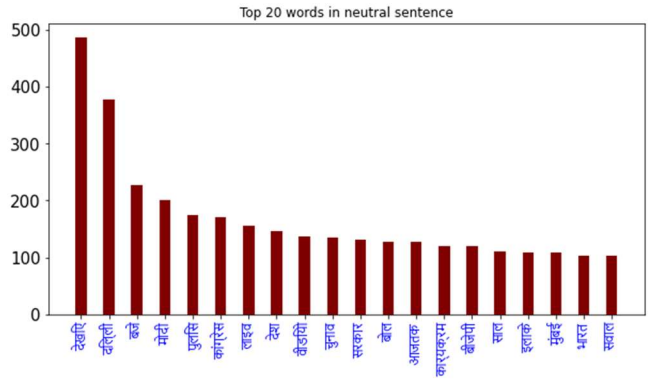
For English and Hinglish text we generally use a code – mixing technique to get the results. This code – mixing model converts the English words present into Hindi words without changing their meaning. We convert these words into Hindi Language text so that while using our word2vec model to convert the text into vectors we do not have any issues and get a better accuracy. Instead of using this model, we can also maintain an English to Hindi word dictionary pair so that we change the English words present to Hindi words and hence reduce the overall complexity of the system and get a faster working model. We can also use a Cross – Lingual Model for code mixing (XLM) but they work well on mono-lingual corpus, and they do not take any advantage of the available parallel data. Hence, we can use a Translation Language Model where we take a sequence of parallel sentences from the input corpus / transition data and randomly mask tokens from the source and all these masked tokens would in turn contribute to the given word prediction. Other than these two approaches, we can also use Word Embedding by comparing the existing Hindi Language tweet containing the Hindi word with the embedding. We also compared the three bilingual word embedding approaches, a novel approach of training skip – grams on synthetic code-mixed text on sentiment analysis and POS tagging and we got the result that CVM and CCA based embedding perform as well as the proposed word embedding technique for Hindi Language text containing some English words for the semantic and syntactic tasks respectively. Thus, there is not much need for using this existing bilingual embedding for code – mixed text processing and we can in turn use multilingual word embedding from the code – mixed tweet.

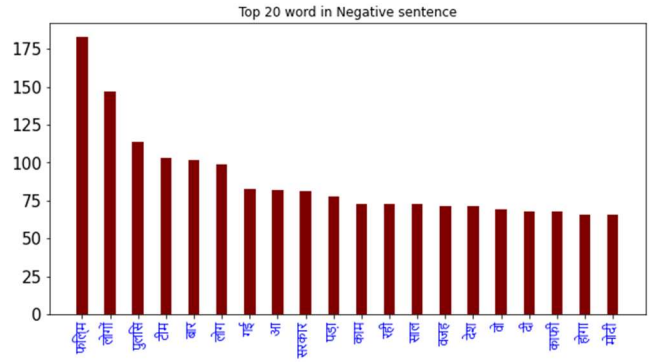
3. DATASET

We have used Snscrape, and we have collected a total of 11,209 tweets which we manually labeled into 3,850 negative tweets, 3,437 neutral tweets and 3,816 positive tweets which we used are training data / corpus for our model.









4. PROPOSED SYSTEM

In this paper, we have proposed a system which classifies Hindi Language Tweets into three categories, namely positive (0), negative (2), and neutral (1).

We have labelled these tweets and we also collected the stop words. These are the words which we would filter out from our data as they were verry common in our corpus. On this training data we did data pre-processing, which includes removing hyperlinks, hashtags, mentions and retweet tags. We also removed some English and Hinglish (Hindi + English) words are they were present out of context and would have had an adverse effect on the accuracy of our system. From our corpus we also removed some irrelevant emojis as our system is not made to distinguish emotions based on emojis and used only Hindi Language data written in Devanagari Script. To have a better data pre-processing result, we also used some data visualization as to get the most used words in positive, negative, and neutral tweets respectively and use the feedback to alter our system to work better.

In our Classification Model, we have tried using a Simple Recurrent Neural Network (RNN), a Long – Short Term Memory (LSTM) System without Word Embeddings, a LSTM with Word Embeddings, and a Bidirectional Encoder Representations from Transformers (BERT) language model. In the Simple RNN Model we have used three layers with the 'softmax' activation function and the 'adam' optimizer. The LSTM used is from Keras and TensorFlow. For using word embeddings, we have stored the word vectors in a matrix and then used a LSTM. We have generated the word vectors using the fasttext library by Facebook AI and then converted the 300-dimension word vectors to 50 and 100 dimensions for further use in word2vec and LSTM word embedding system.

5. RESULTS

We get the result that our Sentiment Analysis Model, successfully classifies the Hindi Language Tweets into three categories, namely positive (0), negative (2), and neutral (1). So, we can say that our sentiment analysis model works properly and gives the intended result. Using this sentiment analysis model for Hindi Language tweets we can get the review / sentiment for a particular group of tweets, and we can use this model for a directed approach to get the review of a particular movie, some product that is newly launched or on some new policy introduced by the government or by any organization by taking the test dataset as per the requirement. While classifying using a Simple RNN Model across 20 Epoch, we got a test accuracy of 0.3463385, using an LSTM Model without word embeddings the test accuracy was 0.7761104 and while doing sentiment analysis using an LSTM Model with word embeddings the test accuracy was found to be 0.7617047.

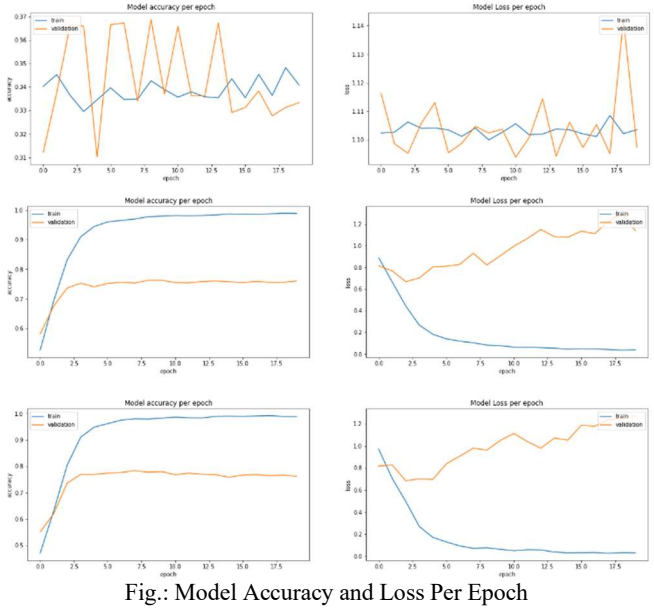


Fig.: Model Accuracy and Loss Per Epoch

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
Simple RNN	0.3482	1.0999	0.3686	1.0938	0.346338541	1.0965497493
LSTM without Word Embeddings	0.9899	0.0356	0.7627	0.6677	0.776110447	1.0134615898
LSTM with Word Embeddings	0.9916	0.0271	0.7833	0.7006	0.761704683	1.2734478712

Performance of Different Models on our Dataset

6. Conclusion and Future Work

So, in this Model, we have successfully classified the Hindi Language Tweets into three categories, namely positive (0), negative (2), and neutral (1). But in the future, we want to classify these Hindi Language Tweets on a scale of 0 – 4 where, '0' is the most positive tweet, '4' is the most negative tweet and '2' is a neutral sentiment tweet and we are currently working on it and we are collecting more data and labelling it as there is difficulty in distinguishing a tweet labeled as '0' from '1' and a tweet labeled as '3' from '4'. This changed classification would give us more insight on the sentiment of the tweet which was tweeted by the user and help in collecting the data reviews more accurately as we know that there are many tweets which are negative or positive, but they are not on the same level of negativity or positivity as each other. So, we plan to get a better working Sentiment Analysis model in the future. In order to implement this '0' to '4' bilingual classification model we also plan to implement a new code – mixing and language processing algorithm to get better results and include a better sentiment classification of emojis as we do not have any efficient bilingual sentiment analysis classification algorithm currently.

References

[1] Patwa, P., Aguilar, G., Kar, S., Pandey, S., Pykl, S., Gambäck, B., ... & Das, A. (2020, December). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 774-790).

[2] Joshi, A., Balamurali, A. R., & Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

[3] Bansal, N., Ahmad, U. Z., & Mukherjee, A. (2013). Sentiment Analysis in Hindi Text. *Indian Institute of Technology, Delhi*.

[4] Sitaram, D., Murthy, S., Ray, D., Sharma, D., & Dhar, K. (2015, July). Sentiment analysis of mixed language employing Hindi-English code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 1, pp. 271-276). IEEE

[5] Sasidhar, T. T., Premjith, B., & Soman, K. P. (2020). Emotion detection in Hinglish (Hindi+ English) code-mixed social media text. *Procedia Computer Science*, *171*, 1346-1352.

[6] Yadav, S., & Chakraborty, T. (2020). Unsupervised sentiment analysis for code-mixed data. *arXiv preprint arXiv:2001.11384*.

[7] Shanmugalingam, K., Sumathipala, S., & Premachandra, C. (2018, December). Word level language identification of code mixing text in social media using nlp. In *2018 3rd International Conference on Information Technology Research (ICITR)* (pp. 1-5). IEEE.

[8] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

[9] Snscrape.

[10] Project Melange, Microsoft. (2020).