PROJECT IDEA :

Sentiment Analysis for Code-Switched Languages

Deepti Mittal - 193050025
Ankita Singh - 19305R002

Motivation & Idea

1. Sentiment Analysis is a hot topic of research due to availability of opinions of users on social media platforms.
2. Users post comments/feedback as per their ease :
    a. *“I love this mobile”.*
    b. *“This restaurant is nice.* ***Khana lazawab hai yaha.”***
3. In the second example, a user posted his opinion by using a **code-mixed language,** in this case Hindi and English together, popularly known as HingLish
4. We wish to analyse the sentiment of users when feedback is given in a code-switched language.
5. For the purpose of our project, we will consider code-switching in Hindi and English.

Problem Statement

- Statement: Given tweets in Code Switched Language, Output the sentiment expressed. Sentiment can be positive, negative or neutral.
- Example:
  - **Input** : "*bholy bhayaa. Uffff dil jeet liya ap ne. Love you imran bhai. Mind blowing ap ki acting hai.*"
  - **Output:** Positive (1)

DataSet Statistics

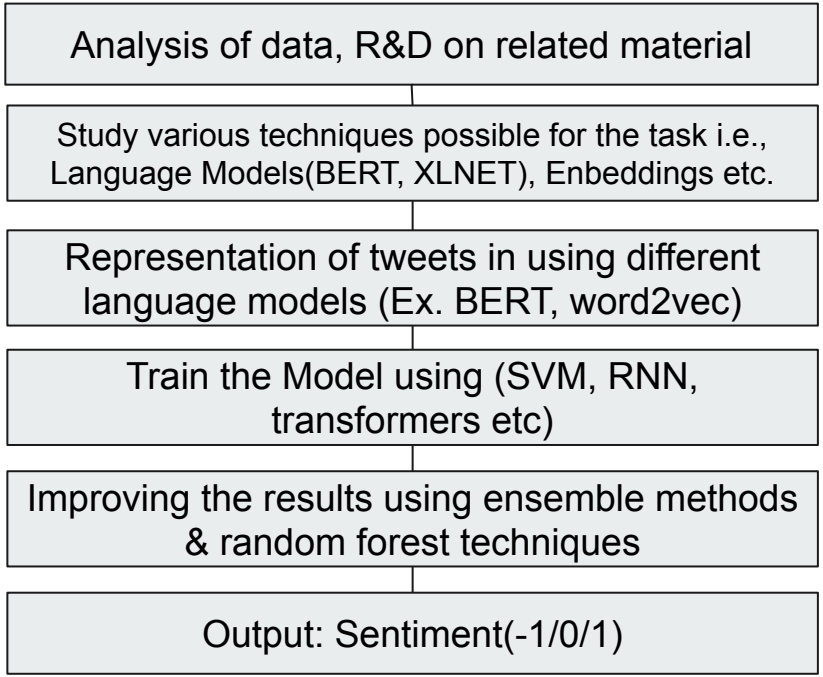
- Data is collected from tweets by using the list of tokens of hindi words released by Patra et al[2018].

Language	Split	Total	Positive	Neutral	Negative
Hinglish	Train	14,000	4,634 (33.10%)	5,264 (37.60%)	4102 (29.30%)
	Validation	3,000	982 (32.73%)	1,128 (37.60%)	890 (29.67%)
	Test	3,000	1,000 (33.33%)	1,100 (36.67%)	900 (30%)
	Total	20,000	6,616 (33.08%)	7,492 (37.46%)	5892 (29.46%)

- Example Tweets:

Hinglish	Congratulations$_{ENG}$ Sir$_{ENG}$ we$_{ENG}$ proud$_{ENG}$ of$_{ENG}$ you$_{ENG}$ ..$_{O}$ Aap$_{HIN}$ pr$_{HIN}$ pura$_{HIN}$ jakeen$_{HIN}$ hai$_{HIN}$ ..$_{O}$ aap$_{HIN}$ bohat$_{HIN}$ achaa$_{HIN}$n home$_{HIN}$ minister$_{ENG}$ Honga$_{HIN}$ ..$_{O}$)$_{O}$ (Congratulations sir we are proud of you.. We believe in you.. You will be a very good home minister..)	Positive
Hinglsih	Hostelite$_{ENG}$ k$_{ENG}$ naam$_{HIN}$ pe$_{HIN}$ dhabba$_{HIN}$ ho$_{HIN}$ tum$_{HIN}$ (you are a blot on the name of a hostelite)	Negative
Hinglish	Warm$_{ENG}$ up$_{ENG}$ match$_{ENG}$ to$_{ENG}$ theek$_{HIN}$ thaak$_{HIN}$ chal$_{HIN}$ ra$_{HIN}$ hai$_{HIN}$ (Warm up match is going fine)	Neutral

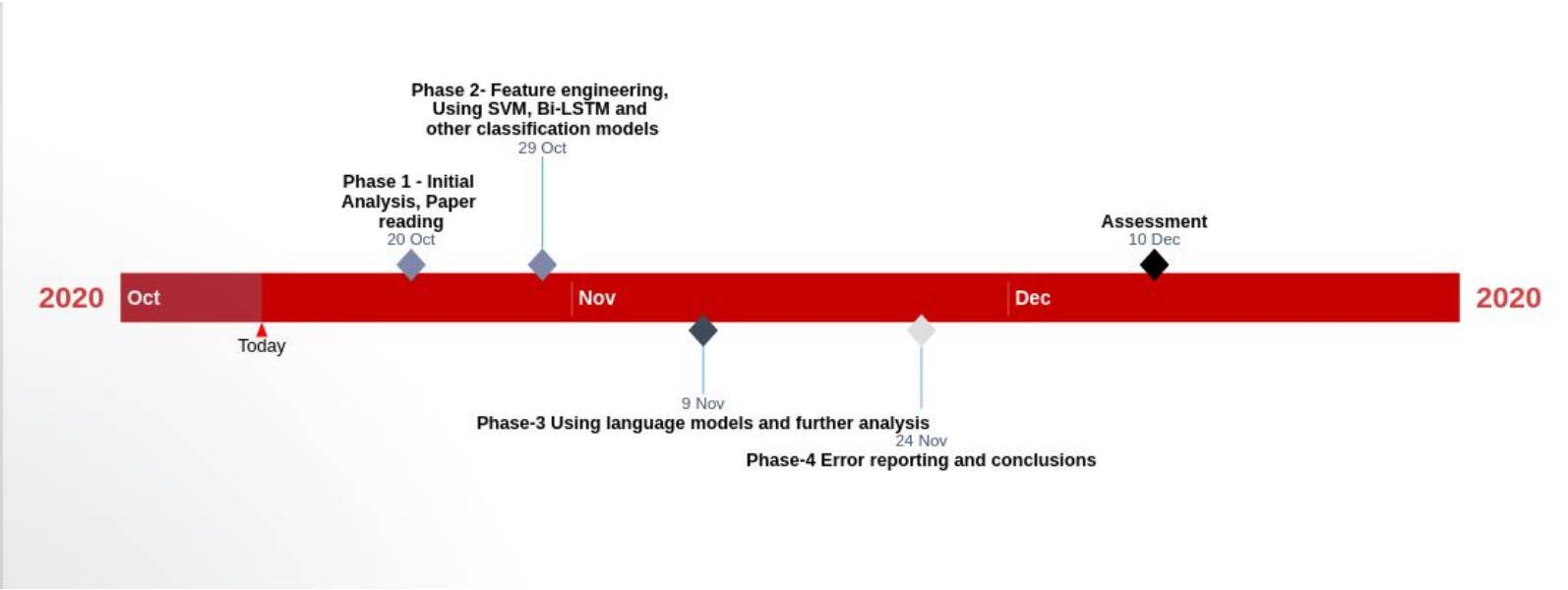
Methodology



Deliverables

1. Report the accuracy for language Identification (LId) in the given dataset.
2. Compare accuracies obtained using different models (SVM, RNN, Transformers) using different language models for the task of sentiment analysis.
3. Detailed Error Analysis for the work done.

Timelines


Phase 2- Feature engineering, Using SVM, Bi-LSTM and other classification models
29 Oct
Phase 1 - Initial Analysis, Paper reading
20 Oct
Assessment
10 Dec
2020
Oct
Nov
Dec
2020
Today
9 Nov
Phase-3 Using language models and further analysis
24 Nov
Phase-4 Error reporting and conclusions

References

1. Patwa, Parth and Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL, Srinivas and Gamb'ack, Bj'orn and Chakraborty, Tanmoy and Solorio, Thamar and Das, Amitava, ***SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets,*** Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), December , ACL 2020
2. Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. ***Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task*** @icon-2017. CoRR, abs/1803.06745.