PROJECT PRESENTATION :

Sentiment Analysis for Code-Switched Languages

Deepti Mittal - 193050025
Ankita Singh - 19305R002

Problem Statement

- **Statement**: Given tweets in Code Switched Language, Output the sentiment expressed. Sentiment can be positive, negative or neutral.
- **Example**:
  - **Input**: *"bholy bhayaa. Uffffff dil jeet liya ap ne. Love you imran bhai. Mind blowing ap ki acting hai."*
  - **Output:** Positive (1)
- Data collected from tweets by using the list of tokens of hindi words released by Patra et al[2018].
- 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

Data Pre-processing(½)

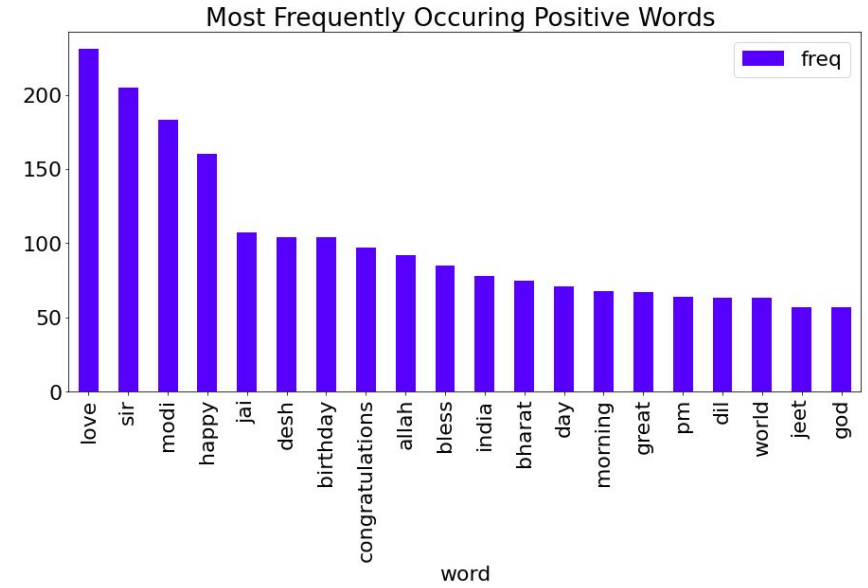
1. Data-set -> Total ***14k sentences*** (evenly distributed positive,negative and neutral)
2. Performed preliminary data processing operations like removal of unwanted links,special characters,etc and basic cleaning, expansion of short forms,stemming etc.
3. Processed emoticons separately based on ***Emoji Sentiment Ranking***. (Given in references)
4. Processed hindi slang words separately using ***Offensive hindi tweet dataset*** (Given in references)
5. Used the English/ Hindi tagged words to ***segregate chunks of hindi words*** from chunks of english words.
6. ***Translation of hindi chunks*** in total retained the meaning of the sentences accurately in most cases.
7. ***Hindi Senti-word net and nltk- SentimentIntensityAnalyzer*** used for some models
8. Train:validation:test split = 70:10:20

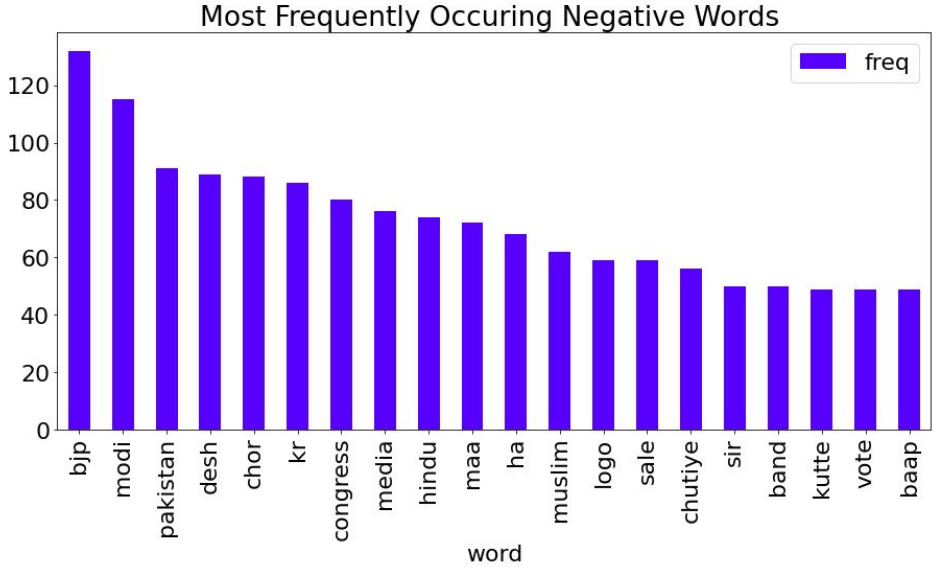
Data Pre-processing(2/2)

8. The handcrafted extra features are:
    a. Whether Slang exists or not?
    b. Chunkwise english translations
    c. Rating of slang
    d. Positive emoticon rating
    e. Negative emoticon rating
    f. Neutral emoticon rating
    g. Negative score, positive score and neutral score: calculated by using sentiment Intensity analyser

	sentence_mixed	sentence_eng	sentiment	slang_existance	slang_rating	positive_emoji	negative_emoji	neutral_emoji	neg_score	pos_score	neu_score
1	haan yaar neha 😔😔 kab karega woh post 😭 usne na sach mein photoshoot karna chahiye phir woh post karega ... URL	yes friend 😔 😔 when will he post 😭 he didn t really photoshoot then he will post ... URL	0	False	0	1896	2412	1218	0.000	0.296	0.704

Preliminary analysis





Performance using bag of words + features

Model	Positive			Negative			Neutral		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
SVC	**.62**	**.55**	**.59**	.53	.43	.48	.46	.57	.52
Random Forest	.58	.57	.57	.49	.55	.52	.48	.54	.50
XGBoost	**.58**	**.58**	**.59**	.56	.53	.54	.49	.54	.52
Logistic Regression	.51	.57	.54	.46	.48	.47	.45	.40	.42
Decision Tree	.42	.44	.43	.41	.38	.39	.41	.42	.41
Bi-LSTM	0.54	0.50	0.52	**0.62**	**0.67**	**0.64**	0.49	0.48	0.48

Performance with Glove twitter embeddings

Model	Positive			Negative			Neutral		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
SVC	**.60**	**.57**	**.58**	.56	.53	.54	.50	.54	.52
Random Forest	.58	.61	.59	.53	.58	.55	.50	.52	.51
XGBoost	.59	.59	.59	.57	.53	.55	.51	.55	.53
Logistic Regression	**.6**	**.57**	**.58**	.56	.56	.56	.5	.52	.51
Decision Tree	.50	.50	.50	.47	.44	.45	.46	.48	.47
Bi-LSTM	0.54	0.50	0.52	**0.62**	**0.67**	**0.64**	0.49	0.48	0.48
BERT	**0.63**	**0.69**	**0.66**	**0.63**	**0.62**	**0.62**	**0.53**	**0.51**	**0.52**

Error Analysis

1. Most misclassifications in neutral sentences.

rt moeshaaa a week ago we didn ’ t win the championship but we came in 5th overall and got all american ohhhh and don ’ t forget we made

Classified as “**neutral**” in dataset but “**positive**” as per our BERT implementation

2. Ambiguity in positive-neutral, negative-neutral data-set features

Example :

“ye kya bakwas hai kaam dhanda or nahi hai kya practice kro yarrrr there is no space science”

Classified as “**neutral**” in dataset but “**negative**” as per our BERT implementation

“me to saaare dialogues saath me bolti hu 😂😂 i really love that movie 😍😍”

Classified as “**neutral**” in dataset but “**positive**” as per our BERT implementation

3. Sarcasm recognition .

sahi kaha puja ji aapne aise murkh har desh me badne chahiye jo deshdrohi ko sirf sikhaye hee nahi

Classified as “**negative**” in dataset but “**neutral**” as per our BERT implementation

Error Analysis

4. Incomplete sentences in the dataset

mp home minister ke nirdesho kee rewa ke sp mahoday ne hawa nikal kar taire panchar kar diya par mp home dipartment …

Classified as “**negative**” in dataset but “**neutral**” as per our BERT implementation

5. Language variance

pehile to i btawa k ye tumko likh k kon diya hai ye hi sab jo yaha likhi ho tum ? the rest of his kilns

koch generals ky name or corruption also mentions ker monster data hy or buss sonni sonnai he chor rahy how are you wearing kangana ranaut

Words like “pahile” ,”btawa” ,”koch” and misspelled hindi words in the sentence -> translation incorrect

Classified as “**neutral**” in dataset but “**positive**” as per our BERT implementation

6. Positive use of negative smileys, and vice versa and slang words, highly negative/positive words also used in contrasting sense

“bjp brahmastra release a first brahmastra 2 ki shooting start really ? 😕😕😕 what rubbish 😂 ® “

Classified correctly when handcrafted features were used in case of SVM.

Handling issues and general observations

1. We expanded certain commonly used forms of words like *“h” to “hai”, “sb” to “sab”* and so on. Also expanded english forms like “don’t” to “do not” , I’mnt to “I am not “ etc for improving model accuracies in few cases.

2. Similarly wrote regex expressions for eliminating extra characters while cleaning.Eg “rheeeeee” to “rhe”

3. Made a function call to correct most frequently misspelled hinglish words along with english words (taken from wikipedia.dat) .

4. Smiley and slang sentiment ratings helped a lot in improving accuracies of all the models

5. Tried out senti-net for hindi- results not good because scores not given , Wiki-news pre-trained embeddings and glove embeddings in our work.Glove-twitter embeddings gave the best accuracy for Bi-lstm so we proceeded to use that.

6. Only 12.56% of total misclassifications were positive-negative or negative-positive .

7. Use of hand-crafted features like negative score etc increased accuracy for 6%(approx.) in case of SVM.

Future work

1. Review the data-set again and re-train the models based on revised data-set
2. Fast-text transliterated cross-lingual embeddings can be used .
3. We observed that even the data-set had incorrectly tagged hindi-english words , a separate project can be taken up to identify highly accurate word tagging. For eg. “are” can be both hindi and english words based on context.
4. Other languages for code switching can be also be taken up.

Contributions :

ANKITA : Data pre-processing ,cleaning,feature engineering ,Bi-LSTM, Error analysis ,tried sentiwordnet

DEEPTI : Bert implementation, Random forest,SVM , Decision Trees ,LR,Error analysis

There was no hard and fast division, as we discussed everything before implementing.

Assignment 3: CP to DP

1. **Example conversions:**

1. **Students Played football**

```
(ROOT
  (S
    (NP (NNS Students))
    (VP (VBD played)
      (NP (NN football)))))
```

```
('played', 'Students', 'nsubj')
('played', 'football', 'nobj')
('root', 'played')
```

2. Afterwards, we try implementing it with the help dataset using bi-lstm model. But it was over-fitting over PAD values.

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.
3. Construction and analysis of Emoji Sentiment Ranking is described in the following paper: P. Kralj Novak, J. Smailovic, B. Sluban, I. Mozetic, Sentiment of Emojis, PLoS ONE 10(12): e0144296, doi:10.1371/journal.pone.0144296, 2015.
4. Conference Proceedings Did you offend me? Classification of Offensive Tweets in Hinglish Language Puneet Mathur, Ramit Sawhney, Meghna Ayyar, Rajiv Shah
5. A. Das and S. Bandyopadhyay. SentiWordNet for Indian Languages, In the 8th Workshop on Asian Language Resources (ALR), COLING 2010, Pages 56-63, August, Beijing, China.