Sentiment Analysis of Code-Mixed Social Media Text

Ashwani Kumar Jha 20305R001
Mikhil Gupta 203079016
Siri Verma Lackarsu 180050052

MOTIVATION

- ★ Sentiment Analysis is a hot topic of research due to availability of opinions of users on social media platforms.

- ★ Users post comments/feedback as per their ease :
  - ■ “This place is nice. Kya lazawab jagah hai ye.”

- ★ In the example, a user posted his opinion by using a code-mixed language, in this case Hindi and English together, popularly known as **Hinglish.**

- ★ We have analysed the sentiment of users when feedback is given in a code-switched language.

DATA DESCRIPTION

- ★ The data consists of Code-Mixed tweets containing Hindi and English words written in English script. A keyword was provided along with each token mentioning the language of the word, and a sentiment polarity is given for each tweet along with the starting word.
- ★ The tweets were classified among the Negative, Neutral or Positive sentiment polarity.
- ★ Training data contained 14000 tweets, test data and validation data contained 3000 tweets each.
- ★ Data - 169893 words tagged as Hindi and 121412 words tagged as English. On making the dictionary of their frequency distribution, it was found out that there were 26653 unique Hindi words and 26082 unique English words.

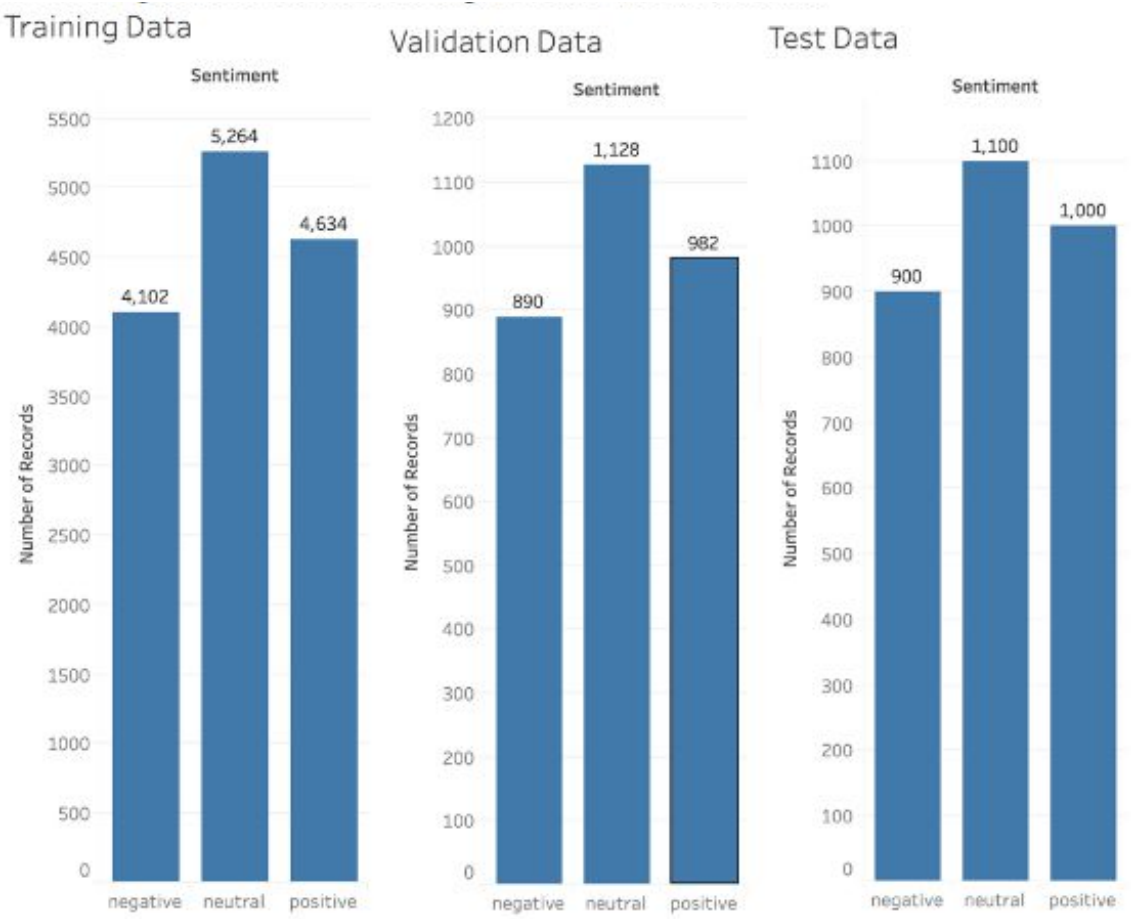

Fig. 1. Distribution of Training, Validation and Test Data among sentiment polarities.

FEATURE EXTRACTION

1. Before data cleaning, processed emoticons in tweets based on Emoji Sentiment Ranking.
2. Processed hindi slang words separately using Offensive Hindi Tweet dataset.
3. NLTK - Sentiment Intensity Analyser was applied on the data and scores have been recorded.
4. The data sentences obtained post-cleaning was transformed using CountVectorizer and tf-idf vectorizer; GloVe-twitter embedding techinque was also studied.

FEATURE SET -> Size (1 X 1007) -> (1 X 1000) Vector embedding of the data sentence; Slang Existence; Slang Rating; Positive Emoticon Rating; Negative Emoticon Rating; Neutral Emoticon Rating; Negative SIA score; Positive SIA score; Neutral SIA score.

DATA PREPARATION

1. The raw data was first consolidated into sentences from the individual words provided for each tweet.
2. @ User tags were removed, html tag symbol were removed, RT (Retweet keyword) was removed, https links were removed.
3. HTML decoding was converted to symbols and special characters and punctuation marks were removed. All the words were also lowercased.
4. English stop words were removed and English words were stemmed using the predefined PorterStemmer library of python.
5. Emoticons were converted into words using the emoji library of the python and word ‘face’ was removed as the stopword from these converted words.

TRAINING AND TESTING DATA

5 different models were created using the machine learning algorithms -

- SVM
- Logistic Regression
- Decision Trees
- XGBoost
- Random Forests (depth=20, estimator=100 parameters)

PERFORMANCE - BAG OF WORDS + FEATURES

MODEL	F1 SCORE		
	Neutral Class	**Positive Class**	**Negative Class**
SVM Classifier	0.51	0.53	0.49
XGBoost	0.49	0.52	0.34
Random Forest	0.56	0.51	0.3
Decision Tree	0.40	0.53	0.52

PERFORMANCE - GloVe twitter embeddings

MODEL	F1 SCORE		
	Neutral Class	**Positive Class**	**Negative Class**
SVM Classifier	0.57	0.55	0.58
XGBoost	0.55	0.51	0.57
Random Forest	0.54	0.50	0.51
Decision Tree	0.49	0.46	0.47
Logistic Regression	0.56	0.53	0.52

BEST RESULTS

SVM (GloVe embeddings)

```
Training Accuracy : 0.6983333333333334
Validation Accuracy : 0.57375
f1 score : 0.5779896040318199
[[484 205 235]
 [217 482  91]
 [238  37 411]]
```

Logistic regression (Bag of words)

```
Training Accuracy : 0.67124955245256
Validation Accuracy : 0.5847797062750334
f1 score : 0.5884732503173007
[[563 296 265]
 [255 638  89]
 [266  73 551]]
```

SPECIFIC OBSERVATIONS

Correct prediction:

- *“may almighty god make easi ramadan mubarak”* - positive
- “😕😕 *what rubbish* 😂😂" - Neutral
- *"ye fraud modi maha chor hai pata nahi mere desh ki janta ki basi per inh vote kar"* - Negative

Wrong prediction:

- *“ye kya bakwas hai kaam dhanda or nahi hai kya practice kro yarrrr there is no space science”* - Classified as “neutral” in dataset but actual “negative”
- *“me to saaare dialogues saath me bolti hu* 😂😂*”* - Classified as “neutral” in dataset but actual “positive”.

ERROR ANALYSIS

1. Emoticons used by people in the text could be sarcastic in context. Many tweets were seen to make positive use of negative smileys and vice versa thus not conveying the sentiment polarity of the tweets. Similar observation was also seen in case of slang words.
2. Incorrect prediction due to lack of context in incomplete sentences in the dataset.
3. Misclassifications in neutral sentences.
4. Language variance of words between Hindi and English in a tweet also added to reduced accuracy of sentiment.

HANDLING DATA QUALITY ISSUES

- 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.
- Similarly wrote regex expressions for eliminating extra characters while cleaning. Eg “rheeeeee” to “rhe”
- Smiley and slang sentiment ratings helped a lot in improving accuracies of all the models
- Tried out glove embeddings in our work. Glove-twitter embeddings gave the best accuracy with SVM.

FUTURE WORK

- Fast-text transliterated cross-lingual embeddings can be used .
- 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.
- Other languages for code switching can be also be taken up.
- More experimentation can be done on the stop words list for both English and Hindi languages.

REFERENCES

- Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets, Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), December , ACL 2020
- 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.
- 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.
- Conference Proceedings Did you offend me? Classification of Offensive Tweets in Hinglish Language Puneet Mathur, Ramit Sawhney, Meghna Ayyar, Rajiv Shah 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.


рахмат
danke
格雷西
ngiyabonga
આભાર
teşekkür ederim
баярлалаа
mersi
barka
welalin
tack
дзякуй
faafetai lava
dank je
misaotra
gracias
tapadh leat
kiitos
dankie
blagodaram
спасибо
dhanyavad
благодаря
asante
manana
bedankt
hvala
mauruuru
thank you
enkosi
mochchakkeram
djiere dieuf
dziękuję
धन्यवाद
go raibh maith agat
sobodi
dĕkuji
sukriya
kop khun krap
taiku
obrigado
mĕrsi
sagolun
ありがとう
toda
хвала
takk
dakujem
trugarez
najis tuke
terima kasih
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
дякуємо
kam sah hamnida
salamat
고맙습니다
xièxie
ευχαριστώ
eskerrik asko
ти благодарам
dankewol