EmojiPred

Emoji Prediction on the Fly!

See Website

See Code

Team Members:- Varun Khurana(2019124)

Jahnvi Kumari(2019469)

Harsh Kumar(2019472)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- Emojis are extensively used by both English and Non-English speakers.
- Very few Emoji recommendation keyboards
- No such keyboard for Indian languages.

Literature Review

Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm

- ➢ Their dataset had 1.2 billion tweets.
- ➢ Used transfer learning technique chain-thaw from DeepMoji
- ➢ Used BiLSTM and attention.

Emoji Recommendation in Private Instant Messages

- ➢ MultiLabel-Random Forest algorithm on real private instant message corpus
- ➢ Predicted upto 169 emojis.
- ➢ Modelled it as a multi-label classification approach, each emoji being a possible label.

Preprocessing

Dataset Description

- Kaggle dataset
- For English, 70,000 anonymised tweets containing 20 unique emojis. For Bengali and Hindi, we have 60000 tweets. We have separate 20000 tweets for Telegu.
- The percentages of Hindi, Bengali and Telugu tweets used were as shown.

❤️	😊	😍	😏	😐	😠	😢	🤔	😡
9.54	9.51	9.42	8.58	7.72	7.50	7.22	6.78	6.73
😭	💕	😉	😁	💙	😘	😶	😜	💜
4.92	4.32	3.82	2.90	2.81	2.56	2.33	1.87	1.46

Table 2: Percentage of emojis in Hindi tweets

😊	❤️	😍	😐	😠	😏	😘	😢	🤔
15.65	13.99	10.46	10.22	8.66	8.59	7.44	5.80	5.70
😶	😭	😡	😁	😜	😉	💕	💙	💜
4.64	3.66	1.32	1.08	1.04	0.67	0.52	0.49	0.07

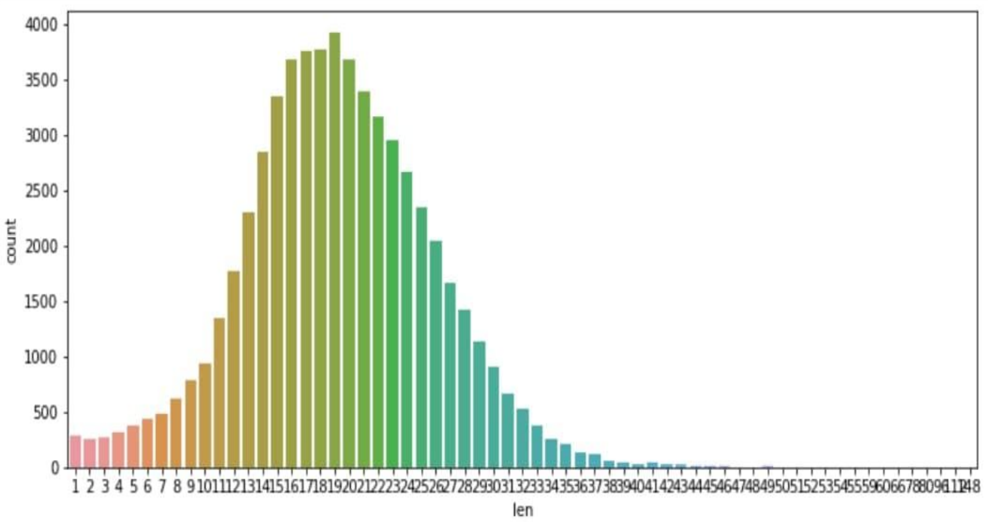
Table 3: Percentage of emojis in Bengali tweets

😊	😠	😍	❤️	😐	🤔	😏	😢	😭
22.67	16.55	13.38	10.77	6.68	5.67	5.05	4.19	3.91
😡	😜	😁	😉	💕	😶	😘	💜	💙
3.01	2.29	1.57	1.36	0.94	0.84	0.54	0.31	0.25

Table 4: Percentage of emojis in Telugu tweets

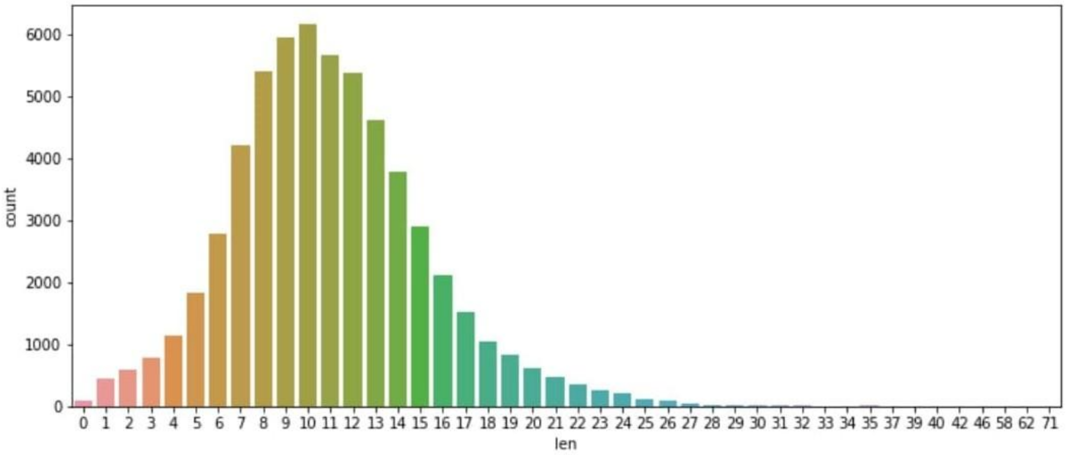
Dataset Preprocessing

1. Data Cleaning (removed punctuation, hashtags etc.)
2. Data Augmentation (using Back Translation)

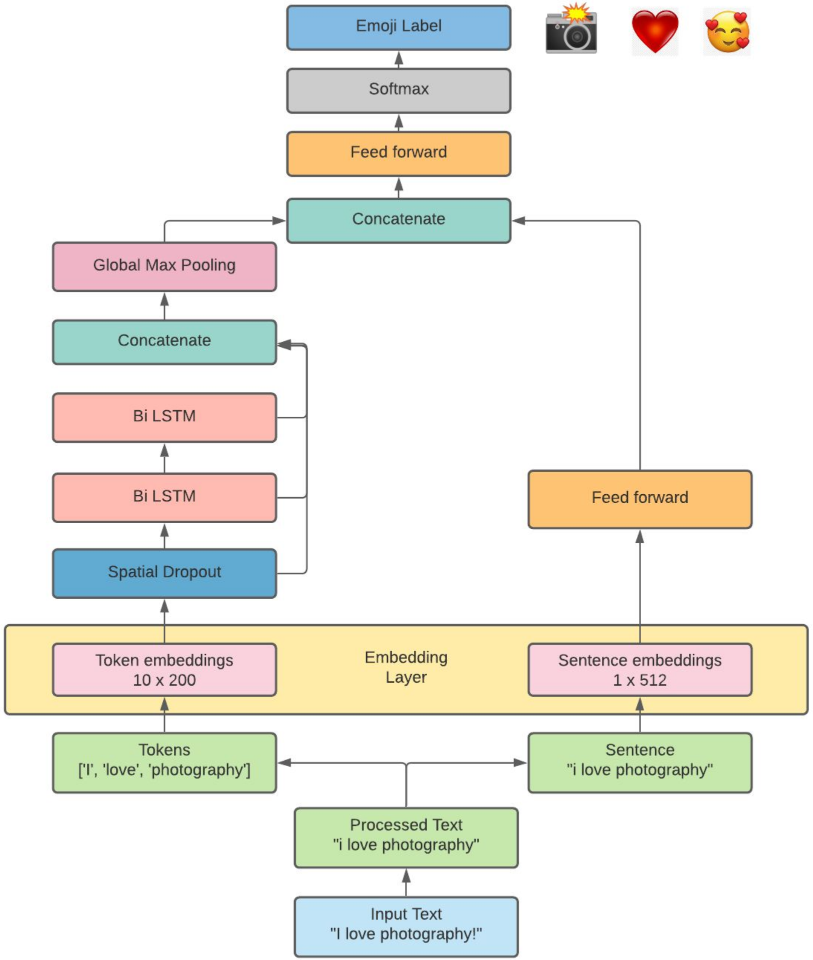


<= Before Processing

After Processing =>



Methodology/ Model


Emoji Label
Softmax
Feed forward
Concatenate
Global Max Pooling
Concatenate
Bi LSTM
Bi LSTM
Spatial Dropout
Feed forward
Token embeddings
10 x 200
Embedding
Layer
Sentence embeddings
1 x 512
Tokens
['I', 'love', 'photography']
Sentence
"i love photography"
Processed Text
"i love photography"
Input Text
"I love photography!"

Embedding

- Token-level and Sentence Embedding.
- Used Google Translate API for processing Hindi, Bengali and Telugu.
- GloVe for Token-level embedding pre-trained on 1.2 billion

Model Architecture

EmojiPred implements two parallel pipelines- one at the token level and the other at the sentence level. These are integrated together at a later stage to yield the final predictions.

Token level Pipeline: Word-level embedding is used followed by a spatial dropout layer helping in regularisation and overfitting. The output is fed to two BiLSTM layers implementing skip-connections. The result is then passed to Global Max Pooling stage to reduce the output from 3 dimensions to 2 dimensions.

Metrics

- Accuracy
- Precision
- Recall
- F1-Score

Result & Analysis

Baselines

Machine Learning based models:

- MultinomialNB
- Decision Trees based

Deep Learning based models:

- Multi-layer Perceptron
- Causal Convolution

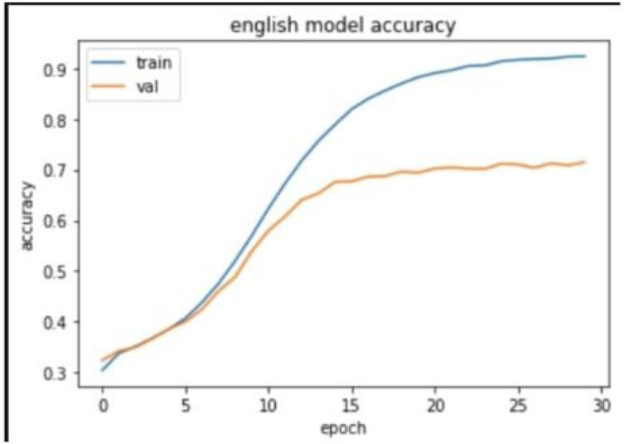
Comparison of EmojiPred against Baselines

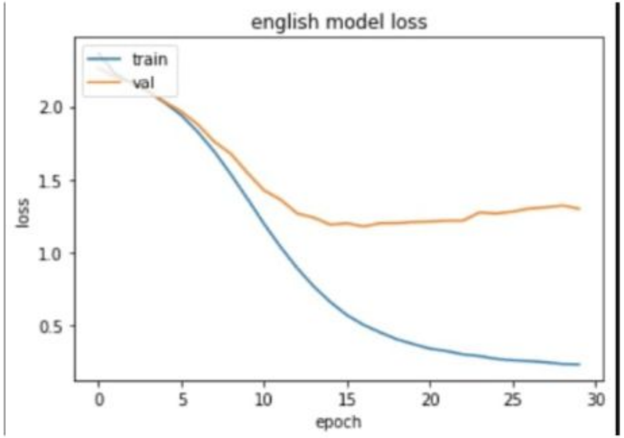
Model	Accuracy
Multinomial Naive Bayes	0.24
SVM	0.40
Multilayer Perceptron	0.57
Causal Convolutions	0.66
EmojiPred	**0.72**

Results on various languages

Language	Accuracy	Top-5 Accuracy	Precision	Recall	F1-score
English	0.72	0.77	0.69	0.66	0.67
Hindi	0.17	0.59	0.17	0.17	0.14
Bengali	0.23	0.68	0.13	0.23	0.14
Telugu	0.42	0.76	0.43	0.42	0.41

Accuracy and Loss Plots for Emoji Pred





Experiments

Ablation Studies

Model	Accuracy
EmojiPred	**0.72**
EmojiPred – USE	0.69
EmojiPred – Spatial Dropouts	0.64
EmojiPred – BiLSTM	0.44

Table 3: Importance of components of EmojiPred towards overall performance.

Conclusion

- EmojiPred is a well-performing multi-recommendation model.
- It used very less data for training and can be made better by training it on large datasets.

Contributions

Ideation: All

Data scraping and preprocessing: Harsh

Baseline Models: Jahnvi

EmojiPred Model Pipeline: Varun

Machine Translation: Varun

Model Deployment: Harsh

Report and ppt: Jahnvi

Thank You