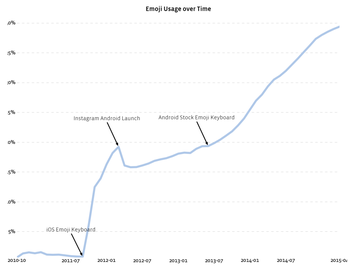
EMOJI SUGGESTER

Paras Gupta | Reuben Devanaesan | Shubh Lavti | Tarun Sharma

ABSTRACT

Emojis, a condensed form of language that can expresses emotions, have become an inescapable part of our online conversations. The emoji suggestion problem aims at predicting the emojis that are associated with a particular text. Models based on distant supervision that employ Bi-LSTMs and Transformer Networks are currently state-of-the-art for predicting emojis from a given text. On these lines, we implement a model that predicts and inserts emoji. Our model gives comparable performance to the state of the art baselines in the prediction task and also succeeds at placing the emoji at the intended position in the text.



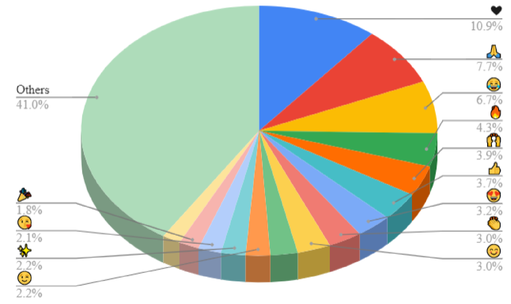
DATASETS

The dataset for the model was constructed out of the Celebrity Profiling Corpus using Twitter API. It scraped tweets from only celebrities having a verified account on Twitter and at least a Wikipedia Page, ensuring that only quality tweets were scraped.

Around 4.5 million tweets were scraped and preprocessed. Preprocessing involved:

- Cleaning the tweet by removing mentions, hashtags, punctuations, hyperlinks and escape characters.
- Choosing tweets that had at least five words and one emoji in them.
- Separating the text from the emoji and assigning each emoji a different label.
- Creating tags for each word in the text.

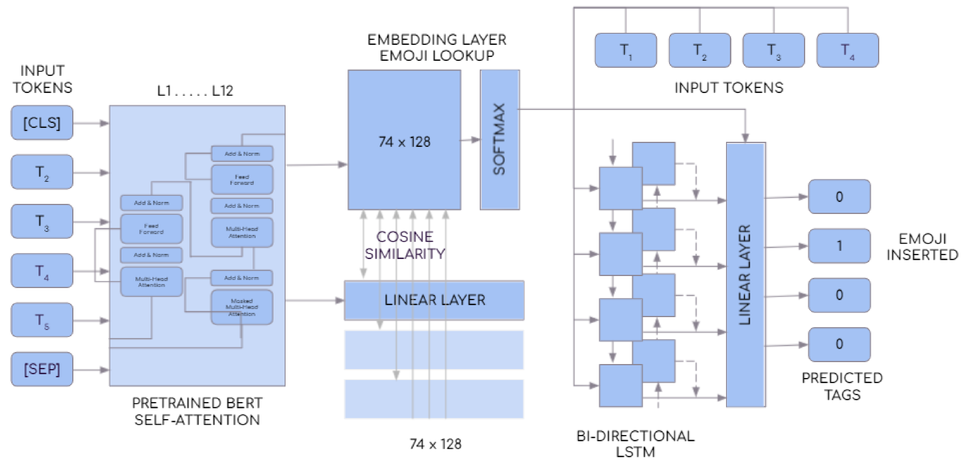
The dataset has five features: Emoji, Label, Cleaned Text, Tags and Original Tweet. Each emoji has a label (number) associated with it. Tags are sequences of numbers that indicate whether each word in the cleaned text is associated with its corresponding emoji or not. The dataset has 111334 entries and 74 emojis. These 74 emojis were manually selected by us based on popularity. Any other emojis in the tweet were ignored.



PROPOSED ARCHITECTURE

The proposed model architecture consists of a pre-trained Transformer Model (BERT) cascaded with a Multi-Layer Bi-Directional LSTM. The BERT Network is fine-tuned to our dataset by the addition of a Linear Layer and an Embedding Layer. The dense layer generates the representation of the input sentence The Embedding Layer acts as a lookup table with each of its row representing one of the emojis in the considered emoji set. The model is trained to maximize the cosine similarity between the sentence representation and its corresponding label from the lookup table.

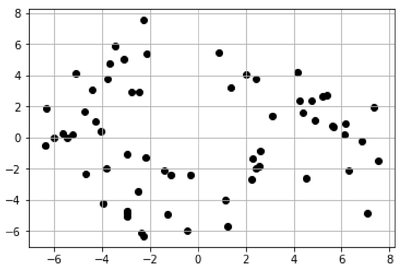
The Multi-Layer Bi-directional LSTM is used for the emoji insertion task. This network makes use of the pre-trained GloVe embeddings to get the encodings for the different tokens in the input sentence and then process them in the multi-layer Bi-LSTM. The final labelling tags are predicted using a Linear layer which then inserts the predicted emoji in the input sentence.



EVALUATION & RESULTS

	Proposed Model	Baseline Model (BERT)
Accuracy	17.17%	17.87%
Top 5 Acc	35.19%	40.92%
Top 10 Acc.	50.75%	56.23%
F-1 Score (Top 10)	39.48%	49.2%

EVALUATION TABLE



EMOJI EMBEDDING SPACE (USING DIMENSIONALITY REDUCTION)

- I got promoted to senior developer today.
==> I got promoted to senior developer today 🤞

- I am feeling sick today.
==> I am feeling sick 😩 today.

- I love my mom's cooking.
==> I love my mom's 💔 cooking.

SAMPLE EXAMPLES OF EMOJI PREDICTION

CHALLENGES

- The accuracy of our model highly depended on the availability of quality data.
- The dataset had noise due to input errors, random usage, data imbalance, and the exact input text having multiple emojis associated with it. This posed a risk to the project's success at some point in time.
- The similarity approach used required high computation power and time.

Input	I have no sense of direction, unbelievable I get lost in the school library.
Positive	I have no sense of direction 😔, unbelievable I get lost in the school library 😂.
Negative	I have no sense of direction 😔, unbelievable I get lost in the school library 😂.

CONCLUSION & FUTURE SCOPE

We consider this model a success, as it is successful at suggesting and inserting reasonable emojis at the intended position in the given text. The performance scores of our model is also considerable to the state of the art baseline models. It can also be appropriately used for downstream prediction tasks such as sentiment analysis (positive or negative) and sarcasm detection.

It also gives an high accuracy score of up to 97.6% in predicting the placement of emoji given an input sequence of texts.
Multiple consecutive emojis are challenging, and we plan to incorporate our current sequence tagging structure with text-editing methods. The impact of mixed-sentiment tweets on the use of emojis would be investigated.

REFERENCES

1. Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Emoji Prediction: Extensions and Benchmarking.
2. Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm.
3. Lin, Fuqiang, Yiping Song, Xingkong Ma, Erxue Min and Bo Liu. "Sentiment-Aware Emoji Insertion Via Sequence Tagging."
4. Barbieri, Francesco Espinosa-Anke, Luis Camacho-Collados, José Schockaert, Steven Saggion, Horacio. (2018). Interpretable Emoji Prediction via Label-Wise Attention LSTMs.

CS 613 Natural Language Processing

ACKNOWLEDGEMENT

Special thanks to Prof. Mayank Singh, and our mentor, Shivam Sahni for their constant help and support throughout the duration of this project.

KEYWORDS

Emoji Suggester, Transformer Networks, BERT, Neural Networks, Bidirectional Long Short-Term Memory (BiLSTM), Sequence Tagging