**Emojis Prediction for English Tweets**

(Jubair Ibn Malik Rifat, Shabbab Algamdi and Mohammad Al Olaimat)

Github Link: <https://github.com/jubairmalik/Emoji_Prediction.git>

1. **Motivation**

Communication in social media using short text messages has become very popular in our daily life. Using of emojis is common although people have different writing styles of these text messages. Emojis electronically express people emotions based on ideograms and smiles [1]. The usage of emojis increases quickly and changes our communication ways because these emojis are very short and easy to send.

Dealing with emojis is considered crucial task when we try to model social media communications. However, prediction of an emoji which is associated with a piece of text in an effective way might be helpful to improve several Natural Language Processing (NLP) tasks including information retrieval, suggestion of emojis while writing text messages and sharing pictures online [2]. In addition, modeling the semantics of emojis can improve sentiment analysis, emotion recognition and irony detection. Based on the importance of modeling social media communication, we propose a model to predict emojis for English text that are retrieved from Twitter.

1. **Significance**

Recurrent Neural Networks (RNNs) are used to model the text classification. However, the using of RNNs may lead to vanishing or exploding problem which make the training of RNN difficult. To overcome this problem, Bidirectional Long Short-Term Memory (Bi-LSTM) is used by [1, 2] which is a variant of the RNN. In our project, we are proposing to use Bidirectional Encoder Representations from Transformers (BERT) which is based on transformer model architecture [4] instead of LSTMs. We are planning to use BERT-Base-Uncased model which has 12-layers, 768-hidden, 12-attention-heads, 110M parameters. We will compare our results with [1, 2] results and show the improvements.

1. **Objective**

Our proposed task is to build a system which will predict emojis from English tweets. The input of that system will be a text message (tweet) and the output will be the corresponding emoji related to the content of the text. For example,

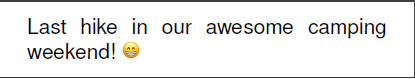


Fig 1: Example of tweet with an emoji at the end [1].

From figure 1, we can say that the text part “Last hike in our awesome camping weekend!” is the input and the emoji part should be the output of our system.

1. **Features**

In this project, we are going to use transformers which are considered the state of the art in NLP domain. The dataset consists of a list of tweets associated with a given emoji (label). The dataset includes tweets that contain one and only one emoji, of the 20 most frequent emojis. We split the data in training, tuning and testing data. The tweets were retrieved with the Twitter. As for the training and tuning data, the tweets were gathered from October 2015 to February 2017, and the test data is gathered from May 2017 to Jan 2018. The dataset can be found here [3].

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Training** | **Tuning** | **Testing** |
| **Number of Tweets** | 45000 | 5000 | 50000 |

Table 1: Number of training, tuning and testing tweets

1. **References**
2. Liu, Man. "Emonlp at semeval-2018 task 2: English emoji prediction with gradient boosting regression tree method and bidirectional lstm." In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 390-394. 2018.
3. Baziotis, Christos, Nikos Athanasiou, Georgios Paraskevopoulos, Nikolaos Ellinas, Athanasia Kolovou, and Alexandros Potamianos. "Ntua-slp at semeval-2018 task 2: Predicting emojis using rnns with context-aware attention." *arXiv preprint arXiv:1804.06657* (2018).
4. <https://github.com/cardiffnlp/tweeteval/tree/main/datasets/emoji>
5. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html