CS626 Project
Speech, Natural Language Processing and the Web

Automatic Emoji Recommendation System

12th December, 2020

TEAM MEMBERS
203059011 - Akshay Batheja
20305R002 - Shivam Ratnakant Mhaskar

Contents

1 Introduction

In 21st century when most of the communication takes place online and social media websites, emojis have emerged as a very easy and expressive way to express one's emotions in very compact way with just a single character. In most of the online communication ways on websites, like tweets on Twitter there is usually text followed by set of emojis which express the sentiment or emotion of the tweet(text). As of March 2020, the Unicode Standard has 3304 emojis one advantage of these high number of emojis is that the user gets to express himself in a variety of ways. But the disadvantage of it is that the user has to search through a lot of emojis to find the emoji he/she wants to use and this can be time consuming. Our system proposes to automatically suggest set of emojis for a given text based on the sentiment of the text, using sentiment analysis.

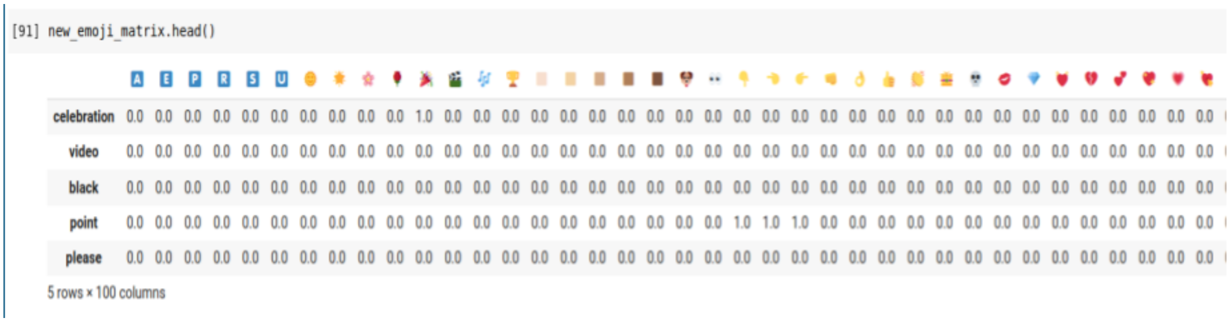
2 Dataset

The Dataset was obtained from Kaggle which consisted of sentences(tweets) with atleast 1 emoji and most of the sentences had multiple emojis in them. We selected 2 lakh sentences from the dataset. Then preprocessing was done on the dataset. The first step of preprocessing was removing all hashtags, URLs and mentions from the tweets. In the dataset most sentences had multiple occurances of the same emoji i.e. repeated emojis in the single tweet, so we removed all the repeated occurences of the emojis and kept only unique emojis in the sentence. Then we separated the emoji and text from the tweets and created a dataframe with 2 columns: text in one column and all the unique emojis in the tweet in the second column as target. Then we performed preprocessing on the text by removing stopwords, punctuations and performed lemmatization and stemming. After this we kept sentences with atleast 1 of 100 most frequently used emojis and divided the dataset with 80% sentences for train and 20% sentences for test. Our final dataset was as follows,

Train Sentences	Test Sentences	Emojis
119821	29956	100

3 Grouping

As the number of emojis is still large i.e. 100 we performed grouping of emojis using Emoji Matrix. The Emoji Matrix consists of emojis as the column labels and sentiment keywords as the index labels. In a column the values are ”1” on the rows with index label which represent the sentiment of the corresponding emoji and ”0” on the rest. A small subset of the Emoji Matrix is given below, For 100 most frequent emojis there were 195 groups/keywords that cover the entire 100 emojis. So the size of Emoji Matrix is 195x100.



```
[91] new_emoji_matrix.head()
```

	🇦	🇪	🇵	🇷	🇸	🇺	🌕	🌟	🌸	🌹	🎉	🎬	🎶	🏆	🏻	🏼	🏽	🏾	🏿	🐶	👀	👇	👈	👉	👊	👌	👍	👏	👑	💀	💋	💎	💓	💔	💕	💖	💗	💘
celebration	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
video	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
black	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
point	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
please	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 100 columns

Figure 1: Emoji Matrix(subset)

4 Models

We trained 3 models for the task of Emoji Recommendation and we tested the models in 2 approaches i.e. Without using Emoji Matrix and Using Emoji Matrix.

4.1 Logistic Regression

The Logistic Regression model was trained using Bag of Words. Without using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
Logistic Regression with BOW	15.38%	14.88%	14.72%

The accuracy of this model without using Emoji Matrix was **13.25%**

Using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
Logistic Regression with BOW	16.70%	16.70%	16.70%

The accuracy of this model without using Emoji Matrix was **16.70%**

4.2 Convolutional Neural Network

Without using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
CNN	29.17%	30.05%	28.78%

The accuracy of this model without using Emoji Matrix was **22.89%**

Using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
CNN	29.66%	29.66%	29.66%

The accuracy of this model without using Emoji Matrix was **29.66%**

4.3 Convolutional Neural Network with LSTM

Without using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
CNN-LSTM	25.61%	25.58%	18.50%

The accuracy of this model without using Emoji Matrix was **18.50%**

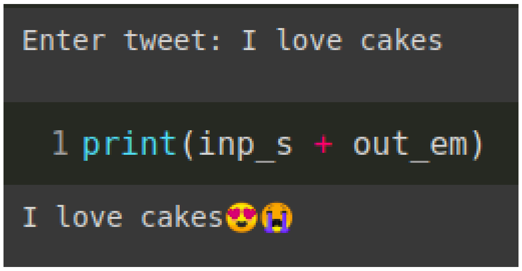
Using Emoji Matrix this model gave the following Precision, Recall and F1 score.

Model	Precision	Recall	F1-Score
CNN-LSTM	24.81%	24.81%	24.81%

The accuracy of this model without using Emoji Matrix was **24.81%**

4.4 Demonstration of CNN model predictions on sentences

The demonstration of the predictions of emojis of the CNN model on the sentence ”I love cakes” is below,



```
Enter tweet: I love cakes
```

```
print(inp_s + out_em)
```

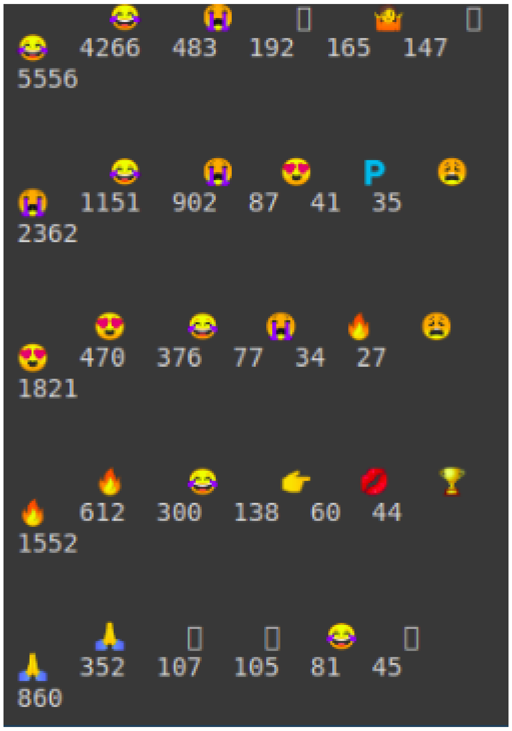
```
I love cakes😍😭
```

Figure 2: Demo

5 Error Analysis

5.1 Analysis of Predictions

For the analysis of predictions, for 5 most frequent emojis in the test sentences we displayed what top 5 emojis were predicted for them and here are the results,



```
      😂     😭     ▯     🙌     ▯
😂  4266   483   192   165   147
5556


      😂     😭     😍    🅿     😩
😭  1151   902   87   41   35
2362


     😍    😂     😭    🔥     😩
😍  470   376   77   34   27
1821


     🔥    😂    👉    💋     🏆
🔥  612   300   138   60   44
1552


     🙏    ▯     ▯    😂     ▯
🙏  352   107   105   81   45
860
```

Figure 3: Analysis of Predictions

From this we can see that for emoji 😂, the emoji 😭 was predicted very frequently because of similarity in sentiment

5.2 Emoji Matrix

As the number of emojis is high and many emojis are synonymous i.e. having similar sentimental value. Our model might predict some emoji other than the one in the test but having very close sentimental value which is also correct in the context of the sentence. If we evaluate the model by one-to-one comparison of the test target emoji and the predicted emoji then in these cases the model will be evaluated as wrong. So for this reason we use Emoji Matrix to evaluate the model. Using Emoji Matrix we find the maximal cover of both the test set of emojis and predicted set of emojis to find which sentiment keyword maximally covers the set of emojis and compare the sentiment keyword of the test set and predicted set of emojis to evaluate the model.

5.3 User Bias

In some complex sentences there are 2 or more sentimental values possible so there are 2 or more sets of emojis possible. But the user might choose to use only one set of emojis and not the other set and the model might predict the other set of emojis and not the one used by the user. In this case the model will be evaluated as wrong even if the set of emojis predicted are write in the context of the sentence. For example for the sentence, ”I like to play football and listen to music” for this sentence emojis 😍 ⚽ as well as emojis 😍 🎵 are possible but the user might choose to use only one set and not the other.

5.4 Sarcasm

Our model was not able to predict sarcasm and it only gave emojis in the literal context of the sentence and not detecting the sarcastic meaning of the sentence and giving the proper emoji for the sentence.

6 Conclusion

An automatic multiple emoji prediction system was implemented which gives multiple emojis automatically from a set of 100 emojis according to the sentimental value of the sentence and an evaluation technique was implemented which makes use of clustering with the help of Emoji Matrix.

7 References

- Do not look too far for an emoji 😃😜 : An Automatic Emoji Recommendation System Kajal Gupta (IIT Bombay); Kanchan Pal (IIT Bombay); Pushpak Bhattacharyya (IIT Patna)