Project Proposal: Multilingual Emoji Prediction

Version 1.0

Mengdan Yuan, Wen Xin

Abstract

The problem we are working on is Multilingual Emoji Prediction. We will get 500,000 English tweets which has only one emoji in each and then will analyse the contexts of them to decide which emojis should be contained in this tweets. In the end, we will compare our predictions with the real emojis.

1 Introduction

As visual icons, emojis play a pivotal role in social media messages which gives a bunch of information. Also, Twitter is one of most frequently used social media tools for people. So how does the Emoji convey information of social media message in Twitter? To solve this problem, we decide to do our project: Multilingual Emoji Prediction. 500,000 English tweets will be gotten as our dataset, every data in dataset will only have one emoji which is the one of 20 most frequent emojis. For us, we will analyse the contexts of them and them decide which emoji should be used in this context and them we will compare our prediction with the real emojis. Solving this problem is pretty meaningful. If the problem is solved, people can use this tool to predict what information could contain in these tweets and it is really useful to some foreigners and some kids who do not know much words to read the contents of tweets. Also, our methodology will be useful and other people may use our methodology to do classification tasks like this.

2 Problem Definition and Data

Given the paramount importance of visual icons for providing an additional layer of meaning to social media messages, on one hand, and the indisputable role of Twitter as one of the most important social media platforms, on the other, we decided to choose the Emoji Prediction challenge as our final project. We are going to predict, given a tweet in English, its most likely associated emoji. We will challenge our algorithms to predict emojis among a wide and heterogeneous emoji space. For simplicity purposes, we will only predict emojis for tweets with only one emoji, which is removed from the tweets for prediction, ignoring tweets including more than one emoji.(SemEval-2018 Task 2, Multilingual Emoji Prediction, 2017)

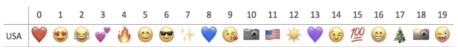

Figure 1: 20 the most frequent emojis

There are 500,000 English tweets used as the dataset for our Multilingual Emoji Prediction. We will use Twitter APIs to get these tweets which are from Oct. 2015 to Feb. 2017 and in US. Every tweet has only one of the 20 most frequent emojis in it. We will use 20 most frequent emojis as labels which are not the same in English corpora as below.And we will divide these data into three parts: trail, training and test.

3 Related Work

In the paper Short Text Classification in Twitter to Improve Information Filtering, the authors proposed an approach to address the problem of using traditional classification methods on short text data such as tweets. They proposed to use a small set of domain-specific features extracted from the authors profile and text. They used a greedy strategy to select the feature set, which generally follows the definitions of classes, extracting 8 features which consisted of one nominal (author) and seven binary features (presence of shortening of words and slangs, time-event phrases, opinioned words, emphasis on words, currency and percentage signs). This approach provided a baseline to classify new tweets online with a better ac-

curacy.(2010,Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu)

In the paper MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification, the authors proposed a new multi-class text classification approach, combining SVM and k-NN. The authors suggested that to use SVM to classify samples and to use KNN to deal with indivisible cases, which are overlapped categories borders. The result showed that the MSVM-KNN outperformed SVM and KNN in most cases. (2008, Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang)

In the paper Very Deep Convolutional Networks for Text Classification, the author proposed a new architectureVD-CNN) to process text which operates direclty at the character level and uses only small convolutions and pooling operations rather than using LSTMs, and convolutional networks which are pretty shallow.2017, Alexis Conneau, Holger Schwenk, Yann Le Cun

In this paper Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, the author had three findings (1). The inclusion of word bigram features gives consistent gains on sentiment analysis tasks (2). For short snippet sentiment tasks, NB actually does better than SVMs (3). A simple but novel SVM variant using NB log-count ratios as feature values consistently performs well across tasks and datasets.(2012, Sida Wang and Christopher D. Manning)

4 Methodology

The Emoji Prediction is essentially a multi-class classification problem. First we want to try different feature selection methods such as word2vec. But there is a pitfall because tweets are very short text data. So we need to carefully select the features. Traditional methods may fail. Then we need to try different classification methods such as Support vector machine, k-nearest neighbors, LDA, QDA, etc.

5 Evaluation and Results

For our classification models, the evaluation results are based on the following metrics: the overall precision, recall, accuracy and F1-score. We will also evaluate our result based on the precision, recall, and f1 for each class (emoji). The example table should be like the table below.

We will also evaluate our result on a confusion matrix, which looks likes the matrix below. We can intuitively get a sense of the relationship between the true labels and the predicted labels. We will compare our result to the baseline, which will be the most frequent emoji. To be clear, we will predict each tweet to be associated with the most frequent emoji, and compare this baseline precision, recall, accuracy and F1-score to our result.

Emo	P	R	F1	%
❤️	37.94	62.58	47.24	21.6
😍	28.14	30.04	29.06	9.66
😂	37.82	52.67	44.03	9.07
💕	17.35	7.75	10.72	5.21
🔥	53.76	49.22	51.39	7.43
😊	9.02	6.82	7.77	3.23
😎	21.16	14.48	17.19	3.99
✨	34.97	22.01	27.01	5.5
💙	26.34	10.78	15.3	3.1
😘	14.84	5.87	8.41	2.35
📷	30.96	47.42	37.46	2.86
🇺🇸	61.43	58.18	59.76	3.9
🌞	32.64	45.77	38.1	2.53
💜	28.57	6.64	10.78	2.23
😉	10.95	4.13	6.0	2.61
💯	29.6	20.58	24.28	2.49
😁	16.15	5.9	8.64	2.31
🎄	61.01	74.63	67.13	3.09
📸	38.7	11.83	18.12	4.83
😜	8.56	2.18	3.47	2.02

Figure 2: the evaluation of the emojis

6 Related Work

We will also evaluate our result on a confusion matrix, which looks likes the matrix below. We can intuitively get a sense of the relationship between the true labels and the predicted labels.

We will compare our result to the baseline, which will be the most frequent emoji. To be clear, we will predict each tweet to be associated with the most frequent emoji, and compare this baseline precision, recall, accuracy and F1-score to our result.

7 Discussion

8 Work Plan

The first step is to download the json of the tweets with the crawler and prepare the dataset for the emoji prediction task using the emoji extractor. So we will get the training data, The second step is to

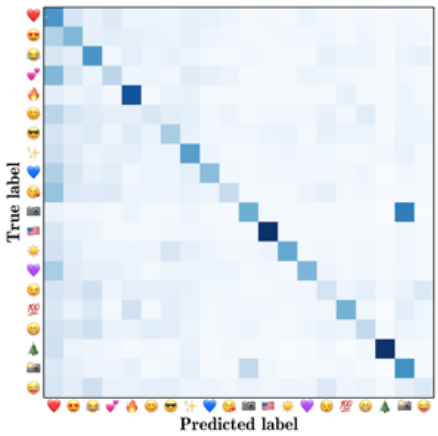

Figure 3: the evaluation of the emojis

do some research and read more papers to have a clearer view about what methods we are going to use– their advantages and disadvantages, especially papers about how to deal with the feature selection in short text data. Finally, we are going to apply all the methods and evaluate their performance on the metrics we choose.

9 Reference

SemEval-2018 Task 2, Multilingual Emoji Prediction (2017). Retrieved from https://competitions.codalab.org/competitions/17344#participate

Short Text Classification in Twitter to Improve Information Filtering,Bharath Sriram,David Fuhry, Engin Demir,Hakan Ferhatosmanoglu.*Proceeding SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*: 841-842

Sida Wang and Christopher D. Manning,2012,Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*:9094

Yuan Pingpeng ,Chen Yuqin, Jin Hai ,Huang Li. (2008). MSVM-kNN: Combining SVM and k-NN for multi-class text classification. *Proceedings - 1st IEEE International Workshop on Semantic Computing and Systems, WSCS 2008*. 133-140. 10.1109/WSCS.2008.36.

Alexis Conneau,2017,Very Deep Convolutional Networks for Text Classification:*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*1:11071116