Fine-grained Emotion Prediction With Emoji and Emoticon

Haozheng Ni(hn2318), Chuqi Yang(cy2478), Somya Singhal(ss5348)

1 Abstract / Introduction

Twitter is a rich source of data for sentiment analysis, opinion mining and many other tasks. One notable feature of twitter data is the usage of emotion token such as emoji and emoticon. Intuitively such token express one's feeling regardless of the language is used, and therefore they could be very helpful in many tasks listed above. Our project aims to embed emoji and emoticon into vectors and then combine with word embedding to test whether this will help in fine-grained emotion prediction.

2 Previous Work

Sentiment analysis on twitter data has been a popular topic in recent years, and most state-of-art models uses deep learning on word embedding. One example is using gated-RNN to predict fine-grained emotions [1]. Meanwhile, there is a trend of mining emoji in text. Some researchers designed an emoji embedding based on twitter data [2] or text description of emojis [3], and they showed that combining emoji embedding could potentially improve the model performance. Some other researchers used emoji as an representative of sentiment and predicted emoji that user will use in the text. One interesting finding is the model pre-trained on such task displays better power in other domains and tasks like sarcasm detection and sentiment analysis [4].

3 Data

Since there is no available published twitter data that contain enough emojis and emotions, we will create our own twitter data by Twitter Stream API. Our plan is to get twitter data dating from 2016 and filter them. Hash-tag will be used as label.

4 Approach

After we obtain the data, our plan consists of two parts

1. Train a 300-dimensional emoticon embedding based on description or twitter data
2. Apply different models to predict fine-grained emotions, which may include
   (a) SVM
   (b) Logistic regression
   (c) KNN
   (d) Online perceptron
   (e) Decision tree
   (f) RNN/LSTM

   We will compare the results of different models can try to explain why some models are better/worse.

5 (Best Case) Impact

1. Prove the intuition that emoji and emoticon is helpful for more accurate prediction of sentiment
2. Get better understanding of model selection in sentiment analysis

6 Obstacles

1. Cleaning twitter data could be complicated
2. There may not be enough emoticon and emoji in twitter data to train an accurate embedding
3. Computation power for deep learning
4. If we want to compare with other state-of-art models, it would be hard to reproduce their models without published code/data.

References

[1] M. Abdul-Mageed and L. Ungar. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728, 2017.

[2] F. Barbieri, F. Ronzano, and H. Saggion. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *LREC*, 2016.

[3] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*, 2016.

[4] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.