EICA Team at SemEval-2018 Task 2: Semantic and Metadata-based Features for Multilingual Emoji Prediction

Yufei Xie, Qingqing Song
East China Normal University, Shanghai, P.R.China
yufeixie@ica.stc.sh.cn

Abstract

The advent of social media has brought along a novel way of communication where meaning is composed by combining short text messages and visual enhancements, the so-called emojis. We describe our system for participating in SemEval-2018 Task 2 on Multilingual Emoji Prediction. Our approach relies on combining a rich set of various types of features: semantic and metadata. The most important types turned out to be the metadata feature. In subtask 1: Emoji Prediction in English, our primary submission obtain a MAP of 16.45, Precision of 31.557, Recall of 16.771 and Accuracy of 30.992.

1 Introduction

Emojis are ideograms which are naturally combined with plain text to visually complement or condense the meaning of a message (Barbieri et al., 2017). Despite being widely used in social media, their underlying semantics have received little attention from a Natural Language Processing standpoint. Barbieri et al. (2016) compare the meaning and usage of emojis across two Spanish cities: Barcelona and Madrid. Ljubešić et al. (2017) present a set of experiments and analyses on predicting the gender of Twitter users based on languageindependent features extracted either from the text or the metadata of users tweets.

Miller et al. (2016) performed an evaluation asking human annotators the meaning of emojis, and the sentiment they evoke. People do not always have the same understanding of emojis, indeed, there seems to exist multiple interpretations of their meaning beyond their designers intent or the physical object they evoke1. Their main conclusion was that emojis can lead to misunderstandings. The ambiguity of emojis raises an interesting question in human-computer interaction: how can we teach an artificial agent to correctly interpret and recognise emojis use in spontaneous conversation? The main motivation of our research is that an artificial intelligence system that is able to predict emojis could contribute to better natural language understanding (Novak et al., 2015) and thus to different natural language processing tasks such as generating emoji-enriched social media content, enhance emotion/sentiment analysis systems, and improve retrieval of social network material.

2 Our Approach

2.1 Features

We use several semantic features and metadata features to represent the twitter.

2.1.1 Semantic Features

Semantic features represent the basic conceptual components of meaning for any lexical item (Fromkin et al., 2018). An individual semantic feature constitutes one component of a word's intension, which is the inherent sense or concept evoked (O'Grady et al., 1997).

Semantic Word Embeddings. We use semantic word embeddings obtained from Word2vec models for GoogleNews. For each twitter, we construct the centroid vector from the vectors of all words in that text.

$$centroid(w_{1..n}) = \frac{\sum_{i=1}^{n} w_i}{n} \quad (1)$$

TF-IDF. In information retrieval, tfidf or TFIDF, short for term frequencyinverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Leskovec et al., 2014). It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf-idf value increases proportionally to

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 419–422
New Orleans, Louisiana, June 5–6, 2018.

the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Nowadays, tf-idf is one of the most popular term-weighting schemes; 83% of text-based recommender systems in the domain of digital libraries use tf-idf (Beel et al., 2016).

$$tf(t,d) = 0.5 + 0.5 * \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \quad (2)$$

t represents the term, d represents the document (Luhn, 1957).

$$idf(t,d) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (3)$$

N is total number of documents in the corpus $N = |D|$, $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears. If the term is not in the corpus, this will lead to a division-by-zero (Robertson, 2004). It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

2.1.2 Metadata Features

Metadata-based features provide clues about the social aspects of the twitter. Thus, except for the semantic features described above, we also used some common sense metadata features:

Twitter containing a question mark. We think if the twitter has a question mark, it may be a question, which might indicate a negative emotions (Castillo et al., 2011).

The presence and the number of links in the twitter. We count both inbound and outbound links. Our hypothesis is that the presence of a reference to another resource is indicative of a positive emotions (Adamic and Huberman, 2000).

Twitter length. The assumption here is that longer twitter could bring more useful detail (Ogasawara, 2009).

2.2 Classifier

For each twitter, we firstly extract the features described above. Then we concatenate the extracted features in a bag of features vector and have them normalized. After the normalization, the value are mapped to interval [-1,1]. At last, we input them into the classifier. In our experiments, we use L2-regularized logistic regression classifier (Buitinck et al., 2013) and SVM classifier (Zweigenbaum and Lavergne, 2016) respectively. For the logistic regression classifier, we tune the classifier with different values of the C (cost) parameter (Aono et al., 2016), and we take the one that yield the best accuracy on 10-fold cross-validation on the training set. For the SVM classifier, we choose different kernels (Moreno et al., 2004) and achieve the best results with RBF kernel. We only show the better results of above two classifiers in the next section.

3 Experiments and Evalution

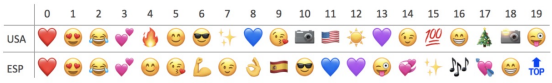
3.1 Dataset

3.1.1 Training and Evaluation Data

The data for the task will consist of 500k tweets in English and 100K tweets in Spanish (Barbieri et al., 2018). The tweets were retrieved with the Twitter APIs, from October 2015 to February 2017, and geolocalized in United States and Spain. The dataset includes tweets that contain one and only one emoji, of the 20 most frequent emojis. Data will be split into trial, training and test.

3.1.2 Label set

As labels we will use the 20 most frequent emojis of each language. They are different across the English and Spanish corpora. In the following we show the distribution of the emojis for each language (numbers refer to the percentage of occurrence of each emoji)



3.2 Evaluation Criteria

For evaluation, the classic Precision and Recall metrics over each emoji are used. The official results will be based on Macro F-score, as the fundamental idea of this task is to encourage systems to perform well overall, which would inherently mean a better sensitivity to the use of emojis in general, rather than for instance overfitting a model to do well in the three or four most common emojis of the test data. Macro F-score can be defined as simply the average of the individual label-wise F-scores. The official will also report Micro F-score for informative purposes.

3.3 Subtask 1 Result

We can see the results in Table 1. The cagri team obtains the best F1 value. The derpferd team gets

Table 1
Experimental Results on the SemEval-2018 Task 2

Team	F1	P	R	Acc
cagri(Top 1)	35.991	36.551	36.222	47.094
cbaziotis(Top2)	35.361	34.534	37.996	44.744
hgsgnlp(Top 3)	34.018	34.997	33.572	45.548
liu_man(Top 4)	33.665	39.426	33.695	47.464
lanman(Top 5)	33.354	35.168	33.108	46.296
derpferd(Top 6)	31.834	39.803	31.365	45.732
ChuhanWu(Top 7)	30.25	31.852	29.806	42.182
kennywlino(Top 8)	30.125	29.905	33.016	38.09
Shi(Top 9)	29.502	35.17	29.91	39.214
anbasile(Top 10)	29.426	30.637	29.583	40.928
EICA	16.45	31.557	16.771	30.992

the best Precision. The cbaziotis obtains the best Recall and the liu_man obtains the best Accuracy. The F1 value of our team is 16.45.

4 Conclusion

We have described our system for SemEval-2018, Task 2 Multilingual Emoji Prediction. Our approach rely on semantic and metadata-based features. Our primary submission obtain a F1 of 16.45 and accuracy of 30.992.

In future work, we plan to use our best feature combinations in a deep learning architecture, as in the Qius system (Qiu and Huang, 2015), which outperforms the other methods on two matching tasks. We also want to use information from entire threads (Joty et al., 2015) to make better predictions. How to combine them efficiently in the system is an interesting research question

5 Acknowledgments

This research was supported in part by Science and Technology Commission of Shanghai Municipality (No.16511102702).

References

Lada A Adamic and Bernardo A Huberman. 2000. Power-law distribution of the world wide web. *science*, 287(5461):2115–2115.

Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. 2016. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 142–144. ACM.

Francesco Barbieri, Luis Espinosa Anke, and Horacio Saggion. 2016. Revealing patterns of twitter emoji usage in barcelona and madrid. In *CCIA*, pages 239–244.

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 105–111. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. 2013. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Victoria Fromkin, Robert Rodman, and Nina Hyams. 2018. *An introduction to language*. Cengage Learning.

Shafiq Joty, Alberto Barrón-Cedeno, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 573–578.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of massive datasets*. Cambridge university press.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2017. Language-independent gender prediction on twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 1–6.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. Blissfully happy or ready to fight: Varying interpretations of emoji. *Proceedings of ICWSM*, 2016.

Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. 2004. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, pages 1385–1392.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Todd Ogasawara. 2009. The reason for the 160 character text message and 140 character twitter length limits. *SocialTimes. com*.

William O'Grady, Michael Dobrovolsky, and Francis Katamba. 1997. *Contemporary linguistics*. St. Martin's.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *IJCAI*, pages 1305–1311.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520.

Pierre Zweigenbaum and Thomas Lavergne. 2016. Hybrid methods for icd-10 coding of death certificates. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 96–105.