#TeamINF at SemEval-2018 Task 2: Emoji Prediction in Tweets

Alison P. Ribeiro
Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brasil
alisonrib17@gmail.com

Nádia F. F. da Silva
Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brasil
nadia@inf.ufg.br

Abstract

In this paper, we describe a methodology to predict emoji in tweets. Our approach is based on the classic bag-of-words model in conjunction with word embeddings. The used classification algorithm was Logistic Regression. This architecture was used and evaluated in the context of the SemEval 2018 challenge (task 2, subtask 1).

1 Introduction

Over the years, technology has significantly changed the way people communicate. It was changed especially due to social media like Twitter[1], Facebook[2], WhatsApp[3], among others. Such media provide users with the ability to express their opinions/emotions not only with words, but through images, the so-called *emojis*.

However, within the context of the sentiment analysis, little research has been dedicated to explore the semantics of *emoji* (Barbieri et al., 2016), thus becoming an interesting challenge to investigate.

Understanding the meaning of *emojis* in relation to their context of use is important for indexing multimedia information, retrieval, or content extraction systems. In addition, *emoji* can complement the meaning of a message, that is, an *emoji* can determine the feeling of a text, however, such emotive figures may become fragile in the ironic/sarcastic context.

In this paper, we developed a methodology to predict *emoji* in tweets, especially our method is based on the bag-of-words model in conjunction with word embeddings (GloVe[4] pre-trained) and n-grams[5], applying a classification algorithm. This configuration was employed and evaluated in the SemEval 2018 challenge (task 2, subtask 1), in which the goal is to predict the *emoji* of a tweet (Barbieri et al., 2018).

This work is organized as follows: section 2 explains some related works, section 3 describes the data set, section 4 addresses the methodology applied in the task, section 5 presents the results, and finally section 6 final considerations as well as future work.

2 Related Works

Emojis can express diverse types of contents in a visual way, adapting to the informal style of communication in social networks. The meaning expressed by emoticons has been explored to allow or improve various tasks related to the sentiment analysis, as in (Hogenboom et al., 2013, 2015).

Emojis can also be used to label excerpts of texts where they occur, thus making it possible to construct sentiment lexical. In this context, in (Go et al., 2009) and (Castellucci et al., 2015) use a distant supervision over the emotionally marked textual contents to form a sentiment classifier and construct a lexicon of polarity. While Novak et al. 2015 constructed lexicons and drew a map of sentiments of the 751 most used *emoji*.

In the work of Barbieri et al. 2017, the authors investigated the relationship between words and *emojis*, studying the new task of predicting which *emoji* are evoked by text-based tweet messages. The authors trained several models based on Long Memory Short-Term networks (LSTMs).

In (Barbieri et al., 2016) the authors explore the meaning and use of *emojis* in four languages: American English, British English, Peninsular Spanish and Italian. By performing several experiments the researchers were able to compare

[1] https://twitter.com/
[2] https://www.facebook.com
[3] https://www.whatsapp.com/
[4] https://nlp.stanford.edu/projects/glove/
[5] terms composed by n words.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 415–418
New Orleans, Louisiana, June 5–6, 2018.

how the semantics of *emoji* vary according to the languages. In a first experiment, they investigated whether the meaning of a single *emoji* is preserved in all variations of language. In the second experiment, they compared the general semantic models of the 150 most frequent *emoji* in all languages. In this study it was possible to find out that the general semantics of the most frequent *emoji* is similiar.

Finally, given the context of the challenge of Semeval 2018 (task 2, subtask 1), we propose a model capable of predicting *emoji* corresponding to the tweets.

3 Dataset and Task

Dataset. The data for the task consists of 500k tweets in English for training, 50k for trial and 50k for test. The tweets were retrieved with the Twitter APIs, from October 2015 to February 2017, and geolocalized in United States. The dataset includes tweets that contain one and only one *emoji*, of the 20 most frequent *emojis*. The amount of tweets for dataset can be seen in Figure 1.

Task details. Because of the importance of visual icons with the ability to provide additional meaning for social messaging and Twitter's key role as one of the most important communication platforms, the Semeval 2018 team invites participants to predict the *emoji* associated with a tweet in English (Barbieri et al., 2018).

Emojis	Train	Trial	Test
❤	105663	10760	10798
😍	51015	5279	4830
😂	50028	5241	4534
💕	26852	2885	2605
🔥	24316	2517	3716
😊	22957	2317	1613
😎	20982	2049	1996
✨	18043	1894	2749
💙	16695	1796	1549
😘	15861	1671	1175
📷	15870	1544	1432
🇺🇸	15067	1528	1949
🌞	13617	1462	1265
💜	12712	1346	1114
😉	13255	1377	1306
💯	13180	1249	1244
😁	12873	1306	1153
🎄	12621	1279	1545
📸	13065	1286	2417
😜	12106	1214	1010

Figure 1: Number of labels per classes.

4 Methodology

The methodology applied in this task consists of two phases, one based on the bag-of-words model and another based on the word embeddings (GloVe) model, in the end both are concatenated, as shown in Figure 2.

4.1 Preprocessing

This step consists in eliminating noises and terms that have no semantic significance in the sentiment prediction. For this, we perform the removal of links, removal of numbers, removal of special characters, removal of *stop words* (words with low discriminative power, for example, "is", "that" etc.). The standardization of tweets in lowercase was also applied, and finally, *stemming*. The purpose of stemming is to reduce words to their radical, for example, the word "*belivies*" will be transformed into "*believ*" (Perkins, 2014).

4.2 Bag-of-words

We apply bag-of-words as baseline, since it has been successfully employed in various classification tasks (Da Silva et al., 2014; Barbieri et al., 2017; Pak and Paroubek, 2010; Kouloumpis et al., 2011; Socher et al., 2013). We represent each message with a vector of tokens, selected using term frequency-inverse document frequency (TF-IDF) with quadrigrams, and min_df = 1, $max\_features$ = 3500, and $ngram\_range$ = (1,4). In the Logistic Regression it was considered C = 10.0, while in the Support Vector Machine and Random Forest the hyperparameters were used by default.

4.3 Word embeddings

Word Embeddings (Bengio et al., 2003) is a supervised statistical language model trained using deep neural networks. The purpose of this model is to predict the next word, given the previous context in the sentence, so similar words tend to be always close. The vector presentation of words was a great advance in relation to the strategies based on bag-of-words. For the proposed task we apply the GloVe model (with 200 dimensions) by (Pennington et al., 2014), GloVe is based on a counting model, in which the vectors are derived from an array of co-occurrences used to extract statistical information about the corpus. With this model an array was generated through the simple arithmetic mean of the word vectors.

4.3.1 Challenges

Because of the need for high computational power to perform the task and the high dimensionality of the table, both in terms of number of attributes and number of rows, only a sampling of 10% of training data was used, this sampling reflects the distribution of real classes.

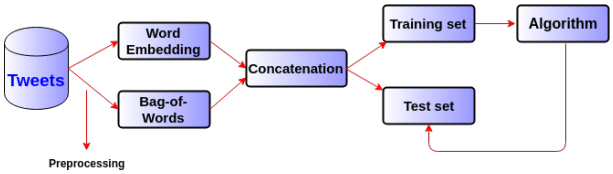

Figure 2: Model used in competition.

5 Results

In this section, we report the obtained results by our model according to the metric evaluation of the challenge, macro f1, precision and recall, accuracy, and f1 for all the *emojis* (Barbieri et al., 2018). Results are reported for five diverse configurations: (i) the system based on word embeddings and baf-of-words with Logistic Regression (LR); (ii) the system based on word embeddings and baf-of-words with Support Vector Machine (SVM); (iii) the bag-of-words system with Logistic Regression (LR); (iv) the bag-of-words system with Support Vector Machine (SVM); and (v) the bag-of-words system with Random Forest (RF). In Table 1 we show model's performances and in Figure 3 we present the predicted score for one of the 20 *emojis*.

Model	F1	P	R	Acc
WE+BoW-LR	**21.497**	**26.208**	**20.843**	**31.588**
WE+BoW-SVM	21.023	27.034	21.403	32.570
BoW-LR	20.351	24.923	19.824	30.830
BoW-SVM	20.194	26.659	20.518	31.966
BoW-RF	15.793	19.890	15.310	25.842

Table 1: Result Semeval-2018.

The obtained results on the testing data indicate that word embedding together with bag-of-word produces the best F1, on the other hand the three configurations represented only by bag-of-word obtained their results close to the central work model (Word Embedding + Bag-of- Words). It is important to remember that only 10% of training data was used, such choice directly influenced the final result.

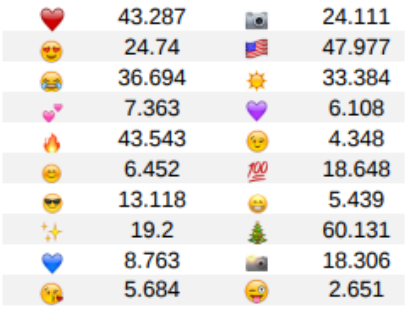

Figure 3: F1 per classes.

6 Conclusion

In this paper, we propose several configurations based on word embeddings and bag-of-words for the Semeval 2018 task 2, subtask 1. As base classifiers we use Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) to predict *emojis* in tweets. Our best model got F1 of 21.497.

As future works we intend to explore the semantics of *emojis* more, as well as apply new word embeddings templates, such as Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2016) and Doc2Vec (Le and Mikolov, 2014) with more computational resources.

References

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? *arXiv preprint arXiv:1702.07285*.

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.

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 531–535. ACM.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015. Acquiring a large scale polarity lexicon through unsupervised distributional methods. In

International Conference on Applications of Natural Language to Information Systems, pages 73–86. Springer.

Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40.

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jacob Perkins. 2014. *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.