PickleTeam! at SemEval-2018 Task 2: English and Spanish Emoji Prediction from Tweets

Daphne Groot
University of Groningen
d.groot.2@student.rug.nl

Rémon Kruizinga
University of Groningen
r.kruizinga.2@student.rug.nl

Hennie Veldthuis
University of Groningen
h.veldthuis@student.rug.nl

Simon de Wit
University of Groningen
s.de.wit.2@student.rug.nl

Hessel Haagsma
University of Groningen
hessel.haagsma@rug.nl

Abstract

We present a system for emoji prediction on English and Spanish tweets, prepared for the SemEval-2018 task on Multilingual Emoji Prediction. We compared the performance of an SVM, LSTM and an ensemble of these two. We found the SVM performed best on our development set with an accuracy of 61.3% for English and 83% for Spanish. The features used for the SVM are lowercased word n-grams in the range of 1 to 20, tokenised by a TweetTokenizer and stripped of stop words. On the test set, our model achieved an accuracy of 34% on English, with a slightly lower score of 29.7% accuracy on Spanish.

1 Introduction

The way people communicate with each other has changed since the rise of social media. Many people use visual icons, so-called emojis, to complement their social media messages. Emojis are frequently used on online platforms like Twitter, Facebook, Instagram and WhatsApp. The wide use of emojis in social media means that processing these emojis can be relevant for NLP applications dealing with social media data.

Social media text has been studied in the field of author profiling, but only recently the interest in the research on emojis started growing. Author profiling is used in different fields such as marketing, forensics, psychological research and medical diagnosis. Author profiling focuses on stylometric features, and since this new popular way of expressing meaning by using emojis has become mainstream, its important to research if and how this data can be used in addition to the textual data. It could be possible that emojis reveal a great deal about the author's gender, location, age or other characteristics.

We describe our approach to SemEval-2018 Task 2 on Multilingual Emoji Prediction (Barbieri et al., 2018) in this paper. We will discuss the features, the machine learning methods we used and analyse the performance of our best method.

2 Related Work

Author profiling tasks are focusing more and more on social media. Oftentimes, the data that is provided is data obtained from social media platforms Rangel et al. (2017). However, research on emojis is more scarce. Some research on emojis is done by Barbieri et al. (2017). They investigated the relation between words and emojis, and found that neural models outperform baseline bag-of-words models as well as humans when predicting which emojis are used in tweets.

Xie et al. (2016) researched automatic emoji recommendation using neural networks. Emojis can express more delicate feelings beyond plain text, and suggesting valid emojis to users of messaging systems can enhance user experience. They approached this problem with neural networks, and they found an Hierarchical-LSTM system significantly outperformed all other LSTM approaches.

Zhao and Zeng (2017) also looked at emoji prediction. The task described in this paper is very similar to the SemEval task. They achieved an accuracy of 40% using a CNN. As features they used the Twitter GloVe embeddings[1]. Since they worked with a noisy dataset they constructed themselves and we are provided with a

[1] http://nlp.stanford.edu/projects/glove/

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 454–458
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

clean dataset, a similar approach might yield high scores.

Author profiling on tweets is not new. At PAN 2017 (Rangel et al., 2017), Basile et al. (2017) were able to achieve a score of 82% on gender prediction of English tweets. They approached the task with an SVM using combinations of character and tf-idf word n-grams. This yields good results for predicting gender, and can provide a good basis for an emoji prediction system.

In the light of this task, sentiment analysis might be helpful. The sentiment of a tweet might point the classifier in the right direction. Mohammad et al. (2013), Han et al. (2013) and Da Silva et al. (2014) all looked into sentiment classification of tweets using machine learning algorithms. Da Silva et al. (2014) achieved an accuracy of 84.85% on predicting sentiment on a Tweet dataset using an ensemble where SVM, Random Forest and Multinomial Naive Bayes were combined using majority voting. It might be fruitful to try some features and methods used in these papers to see if sentiment can be a distinctive feature for emoji prediction. Unfortunately, we did not manage to experiment with these features.

3 Data

The dataset used for this task was provided by the organizers of the SemEval task, and is derived from Twitter and only includes Spanish and English tweets from respectively Spain and the United States. An overview of the emojis in the dataset is shown in Tables 1 and 2.

4 Method

For the task of emoji prediction, we explored a neural network approach and an SVM approach. We established a basic machine learning model per approach and improved on these models for both Spanish and English development dataset. With this approach, we aim to develop a robust model that is able to predict the emojis for both the Spanish and the English dataset accurately.

Architectures we tried for the neural network approach ranged from a simplistic sequential model with a few hidden layers to a stacked LSTM model with word embeddings.

The highest results for our neural network approach were achieved by a sequential neural network model. Our first layer was a 200-dimensional embedding layer, using the GloVe Twitter embeddings (Pennington et al., 2014). Secondly, we used an LSTM layer. After the LSTM layer, our model included a Dropout of 0.2 (Srivastava et al., 2014). The output layer was a dense layer with the sigmoid activation function. Our model used a categorical cross-entropy loss and was optimized by the Adam optimizer (Kingma and Ba, 2014). We used zero masking, 20 epochs and a batch size of 128. Other parameters were left to the Keras defaults.

By establishing a basic SVM system, we tried to improve the system with divergent features. Our basic model consisted of word and character n-gram features. Improvements on this model were applied by using different kinds of preprocessing, tokenization, stemming and POS-tagging methods. We tried tokenization with the NLTK Word Tokenizer and the NLTK Tweet Tokenizer For stemming, we tried the Porter and Snowball stemmer, also from NLTK. Both of them did slightly decrease the accuracy of our system. The POS tagger we tried was NLTK's default POS tagger.

After trying several setups for both systems on the development dataset, we concluded that our SVM approach was the most accurate for both the English and Spanish tweets.

For our best SVM system, we found that some special characters and punctuation had to be removed. Besides, we replaced the Twitter URLs with the placeholder 'URL' and we substituted '. . . ', which was a reference to Instagram, with the placeholder 'INSTAGRAM'. Lastly, we applied a method to reduce each character sequence to a maximum sequence of three characters. E.g., if a user uses the word 'wooooooooooow', we normalize it to 'wooow', so the textual input to the system is less sparse.

The SVM system which yielded the best results on the development set, used the NLTK Tweet tokenizer and merely one feature, namely a tf-idf word vectorizer with a word n-gram range of (1,15), no lowercasing, removing of English stopwords for both the English and Spanish dataset (unconventional, but improved the scores) and a minimum document frequency of one. Our model was trained with sklearn's SGDClassifier[2] with a hinge loss and a maximum number of 50 iterations. All other parameters were left to the sklearn defaults.

[2] http://scikit-learn.org/stable/modules/generated /sklearn.linear_model.SGDClassifier.html

Emoji	❤️	😍	😂	💕	🔥	😊	😎	✨	💙	😘
SVM	45.35	26.58	40.65	8.22	46.93	7.98	14.14	23.79	11.64	6.14
Ensemble	43.67	25.82	38.82	7.45	39.92	7.67	13.79	21.04	9.79	5.53
Emoji	📷	🇺🇸	☀️	💜	😉	💯	😁	🎄	📸	😜
SVM	17.85	56.48	34.86	9.50	3.61	19.51	6.13	60.27	15.77	1.70
Ensemble	17.15	40.26	30.88	7.49	3.55	15.77	5.70	48.60	13.34	1.63

Table 1: Macro F1-score per emoji on test-set for English.

Emoji	❤️	😍	😂	💕	🔥	😊	😎	✨	💙	😘
Spanish	38.79	29.12	49.89	4.25	10.32	17.34	32.01	7.85	12.96	41.68
Ensemble	38.57	28.18	47.06	4.61	10.29	16.40	27.08	6.34	13.65	41.93
Emoji	📷	🇺🇸	☀️	💜	😉	💯	😁	🎄	📸	
Spanish	10.64	3.56	0.71	2.33	1.56	12.84	20.58	3.30	1.52	
Ensemble	9.48	3.95	0.71	2.34	2.72	12.22	16.19	3.02	1.49	

Table 2: Macro F1-score per emoji on test-set for Spanish.

In addition to the SVM and LSTM, we tried an ensemble approach that combined both. Our assumption was that both systems performed slightly better or worse in different aspects. By combining our best SVM and LSTM, we tried to achieve a higher accuracy. When the LSTM system is 95% certain about a label prediction, our ensemble system takes this label as the predicted label. When the LSTM is less certain, the ensemble system takes the label predicted by our SVM system as the predicted label. This threshold was chosen after a short trial of different thresholds, where the 0.95 provided the best results. Yet, it turned out that combining both systems yielded a slightly worse accuracy than our best SVM system alone.

5 Results

The baseline results, obtained by always predicting the most frequent label from the training set, are presented in Tables 4 and 5

The results obtained on the development set are presented in Table 3, with the highest scores, i.e. those achieved by the best systems, are printed in bold.

The results of the final SVM model that we submitted on the test set are presented in Tables 4 and 5, for English and Spanish, respectively. The scores on individual classes (==emojis) are presented in Tables 1 and 2.

Our final system achieves a macro F1-score of 22.86% for English and 15.86%.In order to provide additional insights into the system's performance, the confusion matrices for English and Spanish on the test set, are presented in Figure 1 and Figure 2.

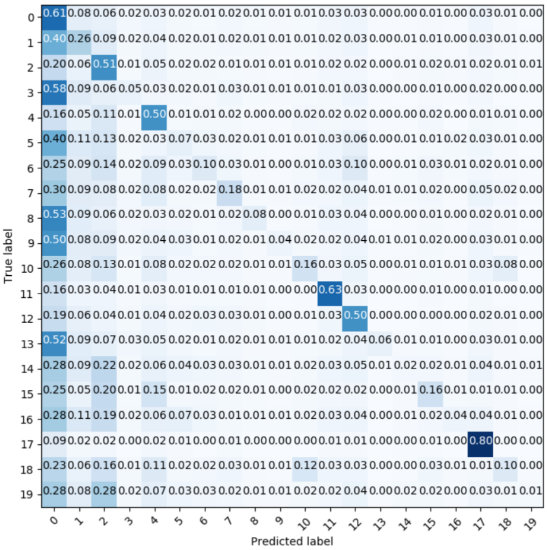

Figure 1: Confusion matrix for predicted and gold labels on the English test set.

Model	Setup	F1 (EN)	F1 (ES)
SVM	Word 1-grams	0.443	0.652
	Word 1- to 3-grams + Character 3- to 5-grams	0.519	0.790
	Word and PoS-tag 1- to 3-grams + Character 3- to 5-grams	0.556	0.804
	Word 1- to 10-grams	0.583	0.821
	Word 1- to 15-grams, no punctuation	**0.613**	**0.830**
	Word 1- to 15-grams, no punctuation + tweet length	0.525	–
LSTM	Dropout of 0.2 before LSTM, 6 epochs	0.427	–
	Dropout of 0.2 before LSTM, 20 epochs	0.529	–
	No Dropout before LSTM, 10 epochs	0.525	0.728
	No Dropout before LSTM, 20 epochs	**0.553**	**0.790**

Table 3: Macro F1-score for various system setups on the development sets for English and Spanish.

	F1	Precision	Recall	Acc.
Baseline	1.78	1.08	5.00	21.60
SVM	22.86	26.17	24.37	34.09
Ensemble	19.89	21.97	20.89	30.57

Table 4: Macro-averaged F1, Precision & Recall and Accuracy for English on the test-set.

	F1	Precision	Recall	Acc.
Baseline	1.86	1.13	5.26	21.41
SVM	15.86	17.57	16.76	29.70
Ensemble	15.06	16.38	15.90	28.17

Table 5: Macro-averaged F1, Precision & Recall and Accuracy for Spanish on the test-set.

6 Discussion & Conclusion

In the confusion matrices, the diagonal lines of correct predictions can be seen. However, as also reported in the paper of Zhao and Zeng (2017), there is also a bias towards predicting the most frequent emojis. For the English tweets, the Christmas tree emoji was predicted most accurately. This is understandable, since this is an emoji that is mostly used in very distinct circumstances. For emojis 3, 8, 9 and 13 this is not the case. They were often incorrectly predicted as emoji 0 (a red heart), which is explainable by the fact that all these emojis relate to love and hearts. For the Spanish tweets, the same issues can be seen with similar emojis.

In this paper, we explored two approaches (an LSTM and an SVM) and a combination of both for predicting emojis of English and Spanish Tweets. Ultimately, the SVM classifier achieved the highest results: F1-score of 22.86 for English and 15.86 for Spanish. Compared the other participating groups, the results were in the mid-range. These results showed that our system ranks 26^{th} out of 49 for English and 10^{th} out of 22 for Spanish. The results on the test set were lower than what we achieved on the development set. This is possibly due to the fact that there seemed to be an overlap between the training set and the development set. This would cause the classifier to be able to make more correct predictions, because it has seen the exact same tweets before.

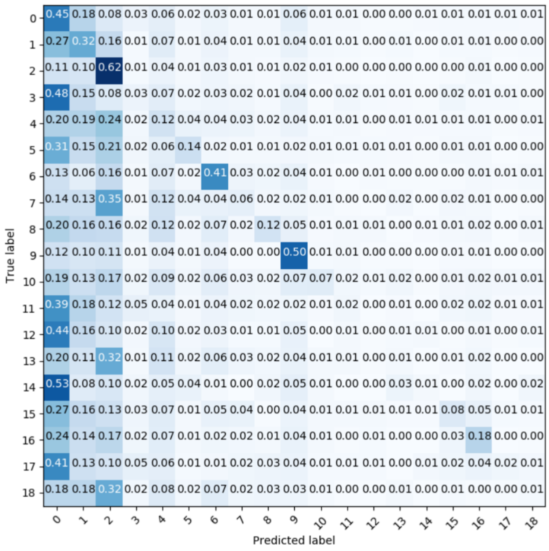

Figure 2: Confusion matrix for predicted and gold labels on the Spanish test set.

References

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Lapata*

M, Blunsom P, Koller A, editors. 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017 Apr 3-7; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 105-11. ACL (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.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model. *arXiv preprint arXiv:1707.03764*.

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.

Qi Han, Junfei Guo, and Hinrich Schütze. 2013. Codex: Combining an SVM classifier and character n-gram language models for sentiment analysis on Twitter text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 520–524.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ruobing Xie, Zhiyuan Liu, Rui Yan, and Maosong Sun. 2016. Neural emoji recommendation in dialogue systems. *arXiv preprint arXiv:1612.04609*.

Luda Zhao and Connie Zeng. 2017. Using neural networks to predict emoji usage from Twitter data.