Manchester Metropolitan at SemEval-2018 Task 2: Random Forest with an Ensemble of Features for Predicting Emoji in Tweets

Luciano Gerber
School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
l.gerber@mmu.ac.uk

Matthew Shardlow
School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
m.shardlow@mmu.ac.uk

Abstract

We present our submission to the Semeval 2018 task on emoji prediction. We used a random forest, with an ensemble of bag-of-words, sentiment and psycholinguistic features. Although we performed well on the trial dataset (attaining a macro f-score of 63.185 for English and 81.381 for Spanish), our approach did not perform as well on the test data. We describe our features and classification protocol, as well as initial experiments, concluding with a discussion of the discrepancy between our trial and test results.

1 Introduction

Written digital communication is increasingly pervaded by the use of emoji. Classic NLP systems are not well geared to handle them. Linguists are still working out how to treat them (Stark and Crawford, 2015; Danesi, 2016). Even their users may disagree on meaning (Tigwell and Flatla, 2016; Miller et al., 2016). A simple approach could be to ignore all emoji and concentrate on the words of a text, however this approach may miss valuable meaning that can be obtained by treating the emoji as semantic units.

The emoji prediction task (Barbieri et al., 2018, 2017), encourages research into the creation of text classification systems which can identify which emoji was present in a tweet. This could lead to automated suggestion systems for emoji, as well as improving the NLP communities understanding of how to deal with emoji computationally.

Emoji	top 5 words
❤️	love heart my family ve
😍	obsessed wcw heaven foodporn view
😂	lmao funny lmfao lol hilarious
💕	pink breast sanfranciscoengagement lovealwaysyje strides
🔥	lit fire mixtape heat flames
😊	802-3037 dickensfranklin dickensofachristmas bagsbycab 7171
😎	sunglasses shades cool risky coolin
✨	sparkle magical pixie magic getonshimmur
💙	royals autism bbn autismspeaks forever-royal
😘	kisses kiss princessmailyana smooches smooch
📷	: :@ bvillain shredforaliving gdlfashion
🇺🇸	merica usa ivoted imwithher election2016
☀️	sunshine sun sunny soakin beachin
💜	purple endalz purplerain alzheimer's relay
😉	mividaesunatombola multi-level silver-criketgentlemensclub azek wink
💯	facts rns realtalk salute t3t
😁	djsty cheesin braces strasberg fcpx
🎄	christmas merry christmastree tree tis
📷	opus : :@ grigsby cred
😜	martian neh silly cray jewelrydesigner

Table 1: The top 5 words according to our class occurrence features for each emoji.

2 Data Acquisition + Preprocessing

The dataset was compiled between October 2015 and May 2016 (Barbieri et al., 2018). Training, trial, and test data emerge from a 80:10:10 split based on chronological order. We followed the organisers instructions to obtain the training data, however we were only able to extract 491,486 tweets as some had been removed by their authors. We tokenised the tweets using the NLTK tweet tokeniser (Bird et al., 2009), but did not perform any further normalisation.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 491–496
New Orleans, Louisiana, June 5–6, 2018.

3 Features

3.1 Word-Class Occurrences

We created a set of features that describe which words occur with each emoji. We created a map describing how often each token occurred alongside each class. Let V be the vocabulary in terms of tokens. Let C be the number of total classes, where each class represents one emoji. We created a matrix M with size $|V| \times |C|$ such that each element $M_{i,j}$ indicates the number of times that token V_i occurs with class C_j. This allowed us to see whether one token occurred mostly in the context of one or two classes, or whether it occurred with similar frequency across all classes. This metric is similar to document frequency in information retrieval.

To further improve our metric, we applied a normalisation transformation to the rows (scaling each row by the total size of the row):

$$M'_{ij} = \frac{M_{ij}}{\sum_{k=1}^{|C|} M_{ik}}$$

This method favoured lower frequency terms (i.e., a hashtag that occurs only a few times with one emoji), so we applied a further transformation to multiply each row by the log frequency of occurrence of the token:

$$M''_{ij} = M'_{ij} \times \sum_{k=1}^{|C|} \ln M_{ik}$$

These features produced intuitive results. The top words for a few select classes are as follows (❤️: love, heart, my, family; 😎: sunglasses, shades, cool; 🎄: christmas, merry, #christmastree)

These features are at the token level, however our classification labels are at the level of the sentence. To convert these features to the sentence level, we used two strategies: average and max. We calculated the average vector as the mean of all token vectors in a tweet. We calculated the max vector by taking the highest value across all tokens for each class. This led to 40 features (20 for average and 20 for max).

3.2 Sentiment

We employed Vader (Gilbert, 2014), a lexicon- and rule-based sentiment detection system to derive a set of sentiment features. Vader fashions features, at sentence level, for positive, neutral, and negative polarities ranging from 0 to 1 and representing intensity. It also produces a combined sentiment score, with values between -1 (negative) and 1 (positive), where values in $[-0.5, 0.5]$ denote neutrality.

3.3 Psycholinguistic Features

We used the MRC psycholinguistic norms (imagery, concreteness, familiarity, meaningfulness, age of acquisition) (Coltheart, 1981) as token level features. These were averaged to give tweet level features in our classification scheme.

3.4 LIWC

We used the latest version of the Linguistic Inquiry Word Count (Tausczik and Pennebaker, 2010) system, LIWC2015, to produce a large set of features, at sentence level, concerning emotional, cognitive, and structural components derived from the texts. As shown in Table 2, our experiments with those features, arranged into different subsets, did not produce any significant improvement; therefore, we decided not to include those in our submissions.

4 Results

We performed subset analyses to determine the best feature grouping. In Table 2, we show our results for different feature sets when training on the training data and testing on the trial data.

We also optimised the number of trees in our random forest, finding 225 to be the best value for this parameter.

Table 3 shows the detailed classification report (precision, recall, F1, and support, by class), and Figure 1 displays the confusion heatmap for our best submission on the English test dataset. Our system ranked 24th, with a macro-averaged F1-score of **24.982** (n=48, median=23.919, min=2.038, max=35.991, Q1=18.278, Q3=28.410).On the Spanish challenge, our best submission (using only the average class-occurrence features) ranked 8th, with a macro-averaged F1-score of **16.338** (n=21, median=14.912, min=3.896, max=22.364, Q1=10.892, Q3=16.696) (see Table 4 and Figure 2 for detailed performance). For lack of space, we restrict our subsequent error analysis and findings

Features	Macro F1
Avg class occurrence, Vader, Topic-20, Avg MRC	0.6299
Avg class occurrence	0.6273
Max class occurrence	0.6266
Vader	0.1290
Topic-20	0.1126
Avg MRC	0.4922
LIWC	0.0425
Vader, Topic-20, Avg MRC	0.3530
Avg class occurrence, Topic-20, Avg MRC	0.6295
Avg class occurrence, Vader, Avg MRC	0.6319
Avg class occurrence, Vader, Topic-20	0.6287
Avg class occurrence, Vader	0.6301
Vader, Avg MRC	0.5211
Avg class occurrence, Avg MRC	0.6287
Max class occurrence, Avg class occurrence, Vader, Avg MRC	*0.6358
Max class occurrence, Avg class occurrence, Vader, Max MRC, Avg MRC	0.6352
Max class occurrence, Avg class occurrence, Vader, Max MRC	0.6355
Max class occurrence, Avg class occurrence, Vader, Avg MRC, LIWC	0.5400

Table 2: Analysis of different feature subsets. Score is reported as Macro F1 throughout. The best performing feature subset (which we used in our experiments) is marked with an asterix.

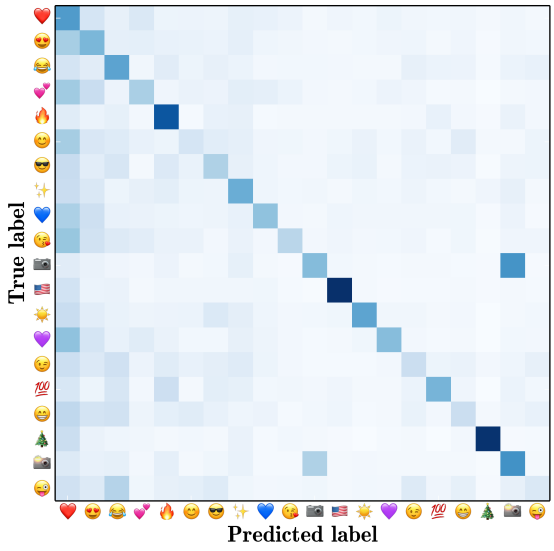

Figure 1: Confusion Heatmap For English Test Data

Emo	P	R	F1	%
❤️	35.23	62.97	45.18	21.6
😍	27.9	25.51	26.65	9.66
😂	33.0	50.64	39.96	9.07
💕	20.41	4.18	6.94	5.21
🔥	51.71	45.16	48.21	7.43
😊	10.36	5.7	7.36	3.23
😎	19.63	13.33	15.88	3.99
✨	30.49	17.06	21.88	5.5
💙	24.81	6.33	10.08	3.1
😘	17.45	4.09	6.62	2.35
📷	26.34	37.99	31.11	2.86
🇺🇸	60.64	52.8	56.45	3.9
☀️	32.76	40.47	36.21	2.53
💜	26.28	6.46	10.37	2.23
😉	13.27	5.59	7.87	2.61
💯	28.65	20.34	23.79	2.49
😁	13.45	5.2	7.5	2.31
🎄	59.81	72.43	65.52	3.09
📸	37.89	21.1	27.11	4.83
😜	8.68	3.47	4.95	2.02

Table 3: Detailed Precision, Recall, F-measure, and Support for English Test Data

to the English challenge. However, these generalise to Spanish.

The F1-score on the test data was much lower than that on the trial data (**63.185**). We hypothesise that this discrepancy might be largely due to (1) our system overfitting the training data and to (2) a test dataset whose class distribution and discriminant features differ in some measure from those of training and trial.

Figure 4 shows the class (i.e., emoji ranks) distributions on trial and test data. With respect to training (omitted here for brevity) and trial data, the shape of the distributions match al-

Emo	P	R	F1	%
❤️	32.54	48.44	38.93	21.41
😍	27.77	30.82	29.22	14.08
😂	42.11	53.77	47.23	14.99
💕	8.84	5.4	6.7	3.52
😊	11.13	11.28	11.21	5.14
😘	20.0	11.08	14.26	3.97
💪	30.93	43.32	36.09	3.07
😉	13.48	9.49	11.14	4.53
👌	11.26	9.44	10.27	1.8
🇪🇸	47.63	35.61	40.76	4.24
😎	16.26	9.73	12.18	3.39
💙	15.12	3.15	5.21	4.13
💜	3.03	0.85	1.33	2.35
😜	7.88	4.74	5.92	2.74
💞	3.51	2.15	2.67	0.93
✨	20.42	9.38	12.85	4.16
🎶	18.93	18.4	18.66	2.12
💘	1.56	0.75	1.01	1.34
😁	6.4	3.83	4.79	2.09

Table 4: Detailed Precision, Recall, F-measure, and Support for Spanish Test Data

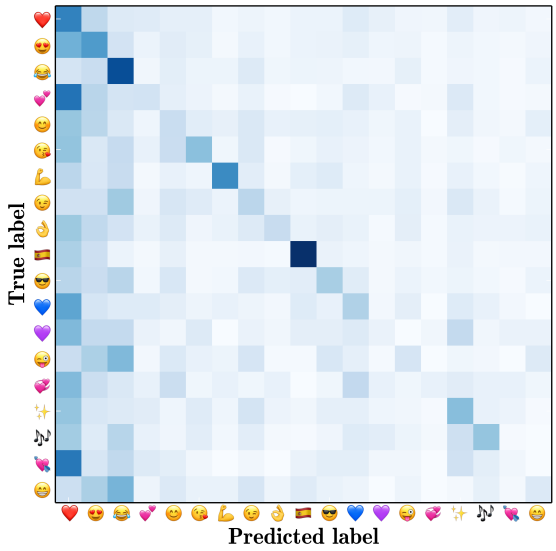

Figure 2: Confusion Heatmap For Spanish Test Data

most perfectly. Also, to a large degree, they are rank-preserving.[1] This is in contrast to the class distribution of the test data, which is not rank-preserving, particularly for those labels in the long tail (i.e., below the three most frequent).

From the classification report (Table 3) and the confusion heatmap (Figure 1) on the test data, one could infer, firstly, that our system revealed a propensity for predicting the most frequent emoji, particularly ❤️, 😍, and 😂 (accounting for about 40% of the data), which can be noticed from the consistent high values on the three left-most columns of the heatmap. Consequently, those within the surroundings of the peak of the class distribution, almost consistently, had recall significantly higher than precision.

For the majority of lower-support emoji, the system had a hard time in separating classes and quite frequently opted for higher-support ones. Secondly, it conflated classes into groups which, intuitively, could be seen as clusters of semantically-similar emoji, taking into account aspects such as emotions (e.g., joy), concepts (e.g., Christmas tree), and occasions (e.g., Christmas), to mention a few.

For instance, most of those associated with *affection*, *elation*, and other positive emotions and emotional states (e.g., 💕, 💙, 💜, 😊, 😘) presented extremely low recall and, frequently, were misclassfied as ❤️. As an example, 💕 had a recall of 4.18%, with about 64% of its tweets predicted incorrectly as ❤️.

Our system performed better at separating other seemingly distinct clusters, such as *sunny weather* (☀️,😎), *patriotism/national holidays/travelling* (🇺🇸), *occasions/special events/holidays* (🎄), *being humorous* (😉,😜), *photography* (📷,📸), to name a few. For example, 📷's recall was 37.99%, with most of its misclassified instances (18%) being assigned to 📸. Conversely, 📸's recall was 21.1%, with about 32% wrongly predicted as 📷.

5 Conclusions

We presented a system for the prediction a single emoji, out of a set of the twenty most-frequent, for Twitter datasets for (1) English and (2) Spanish. Our best model was based on a random forest (n=225) employing an ensemble of (a) max-

[1]There a few discrepancies; for example, in the trial data's class distribution, contrary to the ranking in Figure 3, (14,😉) occurs slightly more often than (13,💜).

Figure 3: Emoji Rankings For English (USA) and Spanish (ESP) (from (Barbieri et al., 2018))

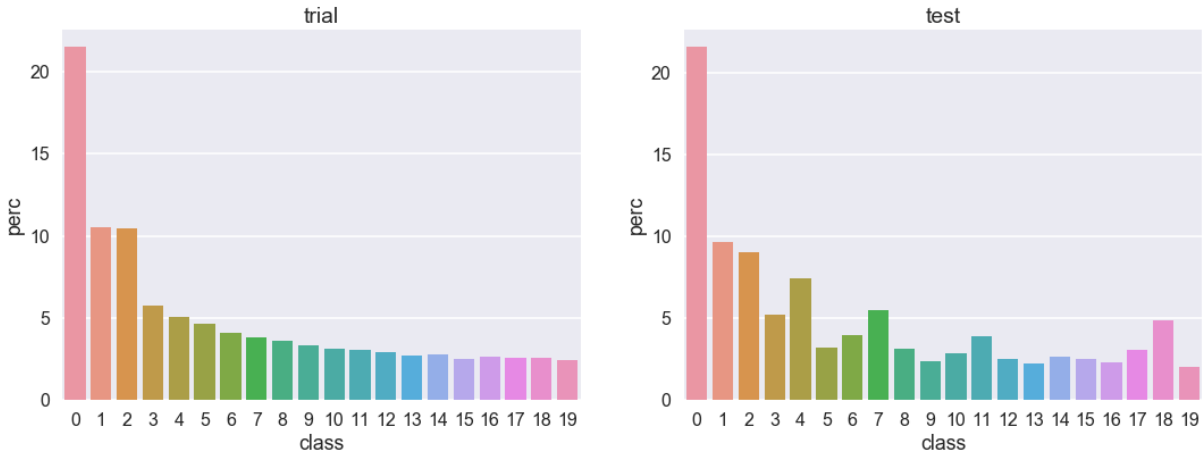

Figure 4: Class Distributions For English Trial and Test Data. The x-axis shows the classes (i.e., the emoji ranks in Figure 3), and the y-axis represents support (i.e., normalised frequencies)

and mean-aggregated normalised word-class occurrences, (b) sentiment and (c) psycho-linguistic features.

Our scores on the test data were significantly lower than those on the trial data, and we postulated that reasons for so were (1) a random forest that overfitted the training data and (2) large variance between trial and test data. It is worth investigating to which extent, and how, different periods of time explain that variance. For example, trial and test might have captured different, emerging trending topics and events; reflect drift in emoji usage; among others. It is reasonable to assume that, given the nature and the sparsity of the data, more representative samples might require much larger number of instances (say, billions of tweets) and time periods covered.

F1-scores were consistently low for all participants, which demonstrates the difficulty of the task. We are conscious that idiosyncrasies of Twitter-specific data (e.g., data sparsity, neologisms, informality, lack of grammatical structure) make it all more problematic, and some of our current research involves devising and incorporating features to address those challenges.

We believe it would be fruitful to investigate evaluation metrics that, rather than all-or-nothing (e.g., misclassification rate), reflect the semantic similarity (or distance) between labels and predicted classes.

References

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Marcel Danesi. 2016. *The semiotics of emoji: The*

rise of visual language in the age of the internet. Bloomsbury Publishing.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

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.

Luke Stark and Kate Crawford. 2015. The conservatism of emoji: Work, affect, and communication. *Social Media+ Society*, 1(2):2056305115604853.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Garreth W Tigwell and David R Flatla. 2016. Oh that's what you meant!: reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pages 859–866. ACM.