

Measuring Emoji Relatedness with PMI Matrices

Graham Adachi-Kriege

McGill University

graham.adachi-kriege@mail.mcgill.ca

Benjamin Taubenblatt

McGill University

benjamin.taubenblatt@mail.mcgill.ca

Abstract

The increased use of social media platforms and electronic messaging has resulted in the increased use of emojis — ideograms used in electronic messages or web pages — to convey meaning. We show that word embeddings attained from traditional distributional models can be used to explore the semantic relatedness between different emojis. Furthermore, we show that natural human intuition about the similarity between emojis corresponds with our quantitative relatedness results.

1 Introduction

Social media has become an integral part of today's society. Naturally, platforms such as Twitter and Facebook have become important vessels for daily communication. As a result, the use of emojis — ideograms used in electronic messages or web pages — to convey meaning has greatly increased.

Distributional semantics is the study of different theories and methodologies for quantifying and categorizing semantic relationships between linguistic items based on their distributional properties. The basic idea can be illustrated in what's called the *Distributional Hypothesis*: words with similar distributional properties tend to have similar meanings (Sahlgren 2008). Thus, by collecting large samples of language data, one can use distributional properties of linguistic items to gain insights into how people use language to convey meaning.

In this paper, we explore the hypothesis that quantitative emoji similarity should correspond with human intuition about emoji relatedness. We use modern techniques in distributional semantics such as pointwise mutual information (PMI) to evaluate

this hypothesis and compare these results with human labelled emoji sentiments.

By using pointwise mutual information (PMI), we create word embeddings which allow us to measure the correlation between emoji-to-natural-language and emoji-to-emoji pairs. We then extract the most important relationships through truncated Singular Value Decomposition (SVD) which both factorizes the PMI matrix and reduces the dimensionality by projecting the data to a lower dimensional subspace. Calculating the cosine similarity between emoji word embeddings gives us a measure of relatedness between emoji pairs which we use to test linguistic hypotheses about emoji semantic relatedness.

We find that the conclusions drawn from natural linguistic insights about the distributional context of emojis also holds for our quantitative conclusions. For example, 🤪 is most correlated with 🌲, 🧑‍🎄, and 🎁.

Finally, by using Spearman rank correlation coefficient, we compare the quantitative relatedness rankings computed with cosine similarity of emoji pairs and quantitative similarity scores of emoji pairs labelled by humans. The results were questionable, with a weak positive correlation between human relatedness judgment rankings and the quantitative relatedness rankings, although a closer look at the rankings provides a number of explanations for the difference in the lists.

2 Related Research

There has already been research done on the topic of emoji word embeddings, and there are well-developed methods for word embeddings as a whole. Levy et al. (2015) explain various methods for optimizing word embeddings which we will use in

this paper, including techniques for Point Mutual Information and Singular Value Decomposition for term-context matrices. The work we found on word embeddings for emojis that we found involves using pretrained Google News word2vec word embeddings to train 300-dimensional embeddings for the emojis (Eisner et al., 2016) (Wijeratne et al. 2017). While the emoji embeddings are evaluated with sentiment analysis tasks in these papers, we will use the secondary method for evaluating our emoji embeddings by borrowing the EmoSim508 used in Wijeratne et al.

3 Methodology

The frequency of target to context word pairs is traditionally modelled using a term-context matrix, or a matrix with target word row labels and context word column labels. Entries in the term-context matrix show the number of times a target word t_i appears in the context of a context word c_j where t_i is the target word at row i and c_j the context word at column j . We constructed a term-context matrix using emojis and context words present in over eleven-thousand tweets from the popular social media platform Twitter and used in the emoji2vec dataset (The dataset can be found here: <https://github.com/uclmr/emoji2vec>). We used the unique emojis our target words and the unique emojis and natural language words as our context words.

The term-context matrix in itself doesn't convey continuous information about the distribution of words with respect to the corpus. It only describes the counts of target words in relation to context words. In this paper, we are primarily concerned with word-embeddings—or distributed vector representations of words—and how we can use these continuous representations to explore semantic relationships.

Therefore, we use the term-context matrix to construct a different *sparse matrix*—or a matrix filled with many zeros—which represents the distribution of target words and context words in relation to each other. We use pointwise mutual information (PMI)—or a probabilistic measure of association—between target words and context

words in order to construct the PMI sparse matrix. PMI computes a quantitative score of relatedness between target and context words under independence assumptions. Positive values imply that the two values co-occur *more* frequently than would be expected under independence assumptions while negative values imply that the two values co-occur *less* frequently than would be expected under independence assumptions. A value of zero indicates that the two inputs are statistically independent.

The PMI matrix M contains $|targets| \times |contexts|$ entries. Each entry M_{ij} where

$$\begin{aligned} i &\in [0, n), n = |targets| \\ j &\in [0, m), m = |contexts| \end{aligned}$$

in the PMI matrix M is

$$PMI(i, j) = \log \left(\frac{\hat{P}(i, j)}{\hat{P}(i) \hat{P}(j)} \right)$$

Where

$$\hat{P}(i) = \frac{count(i)}{\sum_i count(i)}$$

$$\hat{P}(j) = \frac{count(j)}{\sum_j count(j)}$$

$$\hat{P}(i, j) = \frac{count(i, j)}{\sum_{i, j} count(i, j)}$$

It is important to note that for target and context pairs never observed in the corpus,

$$PMI(i, j) = \log(0) = -\infty$$

For those cases, we use

$$PPMI(i, j) = \max(PMI(i, j), 0)$$

$\hat{P}(i)$ is the probability of target emoji i occurring, $\hat{P}(j)$ is the probability of context word or context emoji j occurring, and $\hat{P}(i, j)$ is the probability of both target emoji i and context word or context emoji j occurring together in the same context.

We found that rare emojis, such as 📰 (newspaper), and rare context words, such as names, had the highest PMI scores. This makes sense as target-context pairs which almost exclusively occur together must be highly correlated. This obscured the results of our PMI matrix and made the more interesting and intuitive results appear with lower PMI.

To fix this problem, we replaced target and context words which appeared less than five times with an *UNK* term. We found that removing targets and context words with less than five occurrences improved performance significantly without losing too much interesting information but this hyperparameter should be tuned for individual datasets. It is also important to note that this process can also significantly reduce the target and context word size.

Using context distribution smoothing also significantly improved performance.

$$PMI_{\alpha}(i,j) = \log \left(\frac{\hat{P}(i,j)}{\hat{P}(i) \hat{P}_{\alpha}(j)} \right)$$

Where

$$\hat{P}_{\alpha}(j) = \frac{count(j)^{\alpha}}{\sum_j count(j)^{\alpha}}$$

(Levy et al., 2015)

By increasing the probability of sampling a rare context word or context emoji, context distribution smoothing reduces PMI's bias towards rare words.

This is because $\hat{P}_{\alpha}(j) > \hat{P}(j)$ when j is an infrequent context word. Consequently, this reduces the PMI score of the target word i co-occurring with the rare context word or context emoji j . We used a value of $\alpha = 0.75$ which significantly improved performance (Levy et al., 2015).

After creating the smoothed PMI matrix, we used Singular Value Decomposition (SVD) to factorize the PMI matrix. We used truncated SVD which reduces the dimensionality of the resulting matrices by projecting the data to a lower dimensional subspace. Reducing the dimensionality of the PMI matrix has the effect of improved

computational efficiency and better generalization (Levy et al., 2015).

SVD factorizes a matrix M into the product of three matrices $U \cdot \Sigma \cdot V^T$, where U and V are orthonormal and Σ is a diagonal matrix of eigenvalues in decreasing order. Using truncated SVD, we can choose to keep only the top k elements of Σ . We then obtain

$$M_k = U_k \cdot \Sigma_k \cdot V_k^T$$

Truncated SVD compresses the matrix M while minimizing reconstruction loss. In this way we preserve the most important information while at the same time reducing the dimensionality and reducing the degree of noise in the data.

Another way to interpret truncated SVD is finding the principal components of the data and projecting them down to a lower dimensional subspace spanned by these components.

After performing truncated SVD, it is common in NLP to take the rows of

$$W^{SVD} = U_k \cdot \Sigma_k$$

as our new word representations. We performed truncated SVD with $k = 150$. We started with $d = 365$ distinct emojis (dimension of the rows of M , the PMI matrix). We found that a $k = 150$ produced the best results, but this hyperparameter should be tuned to individual datasets. We also weighted the eigenvalue matrix Σ_k with the exponent $p = 0.5$. It is not theoretically clear why this weighting scheme is better for semantic tasks, however, researchers have found that setting $p = 0.5$ does work much better empirically (Levy et al., 2015).

Our final new word representations was

$$W^{SVD} = U_{k=150} \cdot (\Sigma_{k=150})^{p=0.5}$$

The matrix resulting from performing truncated SVD on the PMI matrix is the final emoji word embedding matrix where each row corresponds to the word embedding for an emoji, and each emoji word embedding captures the “meaning” of its respective emoji from the data.

We used cosine similarity to measure the relatedness between emoji word embeddings. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space and is the cosine of the angle between the two vectors. Therefore, any emojis that tend to share context words or context emojis will have high cosine similarity. Cosine similarity is defined as

$$SIM(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

By normalizing the rows of W^{SVD} to unit length (L_2 normalization), the dot product operation becomes equivalent to cosine similarity (Levy, Goldberg, Dagan 2015).

We measured the cosine similarity between every emoji pair in order to find the most and least similar pairs of emojis. Ranking these pairs by their similarity scores allows us to visualize the semantic information captured by the word embeddings and can show some more intuitive patterns, however, we ultimately want a more empirical evaluation metric for analyzing the embeddings. For this, we borrowed the method used in Wijeratne et al. (2017) by using the EmoSim508 dataset, a list of 508 unique pairs of emojis each with a relatedness rating averaged from ten human judgments on a scale of 0 to 4, as a gold standard dataset (The dataset can be found here: <http://emojinet.knoesis.org/emosim508.php>).



Using this dataset, we took all pairs shared between EmoSim508 and the PMI matrix (of which some emojis in EmoSim508 did not appear) to get two lists of 427 pairs, one sorted by the cosine similarity ranked of the PMI matrix and the other ranked by the human judgment scores. We then compared the two lists using Spearman’s rank correlation coefficient (Spearman’s ρ) as a measure of the PMI matrix’s correlation with human judgments on emoji relatedness. Finally, for a more qualitative analysis, we used t-distributed Stochastic Neighbor Embedding (t-SNE) class from `sklearn.manifold` to visualize emoji similarity by reducing the emoji word embeddings to 2 dimensions.

4 Results











We found that many of the most related emojis corresponded directly with our own intuition, however, often times the least similar emojis did not have as clear a reason as to why they were dissimilar. Some of the more interesting findings are summarized in Table 1 in the appendix but we have outlined a few examples here. We would like to note that, every target emoji had the highest cosine similarity with itself and we have excluded the identical emoji from the table. This result is most likely because naturally, identical emojis have the same contexts. However, this could also be the result of the online trend to use streams of identical emojis—e.g. 🍌🍌🍌.

For the target emoji 👨👩👧👦 (family) had the highest cosine similarity with 🤰 (bride), 🚗 (car), 🏠 (house), 🎓 (graduation), and 👶 (baby). It makes sense that when one settles down and starts a family, they usually have already graduated from high school or college. They are typically married and usually buy a car or house, and may also have a baby. However, for the target emoji 👨👩👧👦 (family), the emojis with the lowest cosine similarity were 🆘 (S.O.S), 🦊 (fist), 📱 (smart phone), 🌸 (bouquet), and 🔒 (lock). It is hard to see meaningful relationships between these emojis and the target emoji 👨👩👧👦 (family). However, this could make sense as these emojis may be dissimilar not necessarily because they are antonyms of each other—e.g. 😊 and 😞—but because they have little to no semantic relatedness. Among other results were that emojis which referenced alcoholic beverages or sweets had high cosine similarities with other distinct alcoholic beverages and sweets respectively.

Through our analysis of emoji relatedness, we also uncovered advanced semantic relationships which provided insight into potential cultural norms and idioms. For example, the target emoji 💍 (wedding ring) had a surprisingly high cosine similarity with 🔑 (lock and key). This high similarity could be a consequence of the English idiom “*under lock and key*”, as in “I keep my grandmother’s secret recipe *under lock and key* so that no one can steal it”, or “Her jewelry was *under lock and key* at the vault”. Another interpretation of this relatedness could be the English idiom, “*the old ball and chain*”, or a 20th century slang term meaning

wife.  (wedding ring) and  (lock and key) could have a high cosine similarity because of this idiom, the allusion being the presumption that a man's wife held him back from doing things he really wanted to do.

The spearmanr rating of the list of 427 pairs of emojis that overlap between the set of EmoSim508 pairs and the pair combinations of the PMI matrix was 0.0518 along with a P-value of 0.281, meaning there is little correlation between the human rankings and the emoji embeddings with some confidence. This suggests that the emoji word embeddings are a poor representation of human semantic understanding of emojis.

Aside from issues with a small dataset, the ranking lists explain some of the differences between the emojis, since the judgments humans have about what emojis they believe are the most related may not be the emojis that they use in co-occurrence in practice. For example, the third highest cosine similarity among the shared pairs was (, ), which was one of the least related pairs of emojis according to human judgments. The (, ) pair of emojis is sometimes used as a sexual gesture on Twitter and would result in a higher co-occurrence for seemingly unrelated symbols. This comes down to a semantic issue itself on what relatedness is. Word embeddings learn word meaning from distributional information from a corpus. Without using any explicit logical information about the words, they learn relatedness through patterns in human behavior within in the text. Humans have a higher level understanding of word meaning that involves associations built from experience (like the word embeddings) along with a logical information like synonymy-antonymy and hypernymy-hyponymy. What may get lost in the human version of word meaning are expressions that make more sense in when used in combination in the context of social media, like   (enthusiasm),   (sneaking), or   (yet another sexual gesture).

The t-SNE plots (Plot 1.2 and subregions of Plot 1.2 in Plots 2-4) show some predictable patterns in emoji word embedding similarities.

Animal-related emojis are the most secluded group, likely a result of long strings of animal emojis used in tweets. Moon and alcohol-related emojis each had their own respective clusters as well, a sign of a niche yet common group of co-occurring group of emojis suggestive of social media culture. The bottom-most cluster in the graph (Plot 4) is not as intuitive and does not have as clear of an explanation as the other clusters.

5 Conclusion

In this study, we show how in certain circumstances, human intuition about emoji similarity align, while in other situations they diverge from each other. We also highlight the importance of analyzing non-typical language data for semantic relationships, especially in an age where electronic messaging and social media platforms are so prominent.

Limitations of this work include the source of our data as we only drew emojis from the social media platform Twitter. More accurate similarity representations could be achieved using larger and more diverse datasets.

This work could be further extended by incorporating prior belief into the pointwise mutual information (PMI) calculations by using MAP estimates instead of MLE estimates of the true distribution.

Seemingly minor techniques in distributional semantics can often have a strong impact on the success of word embeddings. This study shows the success of classical techniques in distributional semantics in determining relatedness on non-typical language data. By showing how these techniques can be applied to emojis, we highlight their generalizability and timelessness.

This study also emphasizes the need for increased research into semantic relationships in non-typical language which has been popularized through the increased use of social media platforms and electronic messaging.

6 References

- Eisner, B., Rocktäschel, T., Augenstein, I., Bosnjak, M., & Riedel, S. (2016). Emoji2vec: Learning Emoji Representations from their Description. *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. doi:10.18653/v1/w16-6208
- Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. In *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.
- Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20, 33-53.
- Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2017). A semantics-based measure of emoji similarity. *Proceedings of the International Conference on Web Intelligence - WI 17*. doi:10.1145/3106426.3106490

A Appendices





































































































































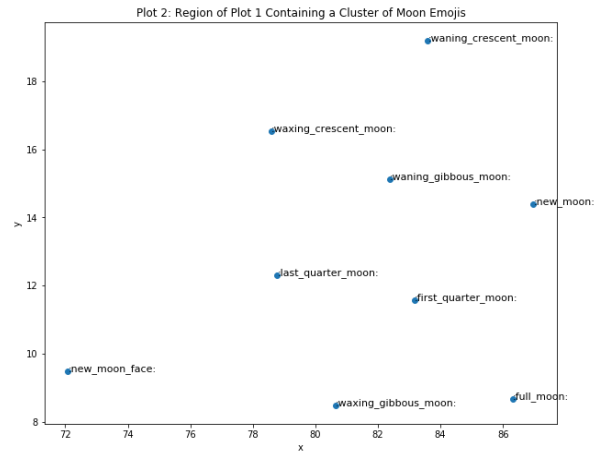
<i>Emoji Target</i>	<i>Top 5 Greatest Cosine Similarities</i>	<i>Bottom 5 Least Cosine Similarities</i>
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 
	 ,  ,  ,  , 	 ,  ,  ,  , 

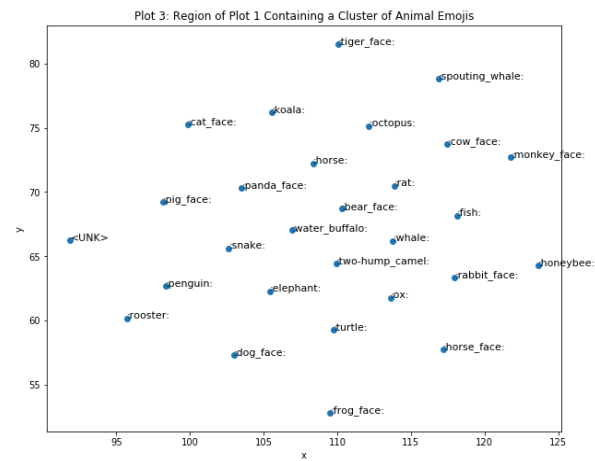
Table 1: The 5 greatest and least related target emojis and context emojis respectively, based on cosine similarity.

Plot 1.1: t-SNE Visualization of the PMI matrix.

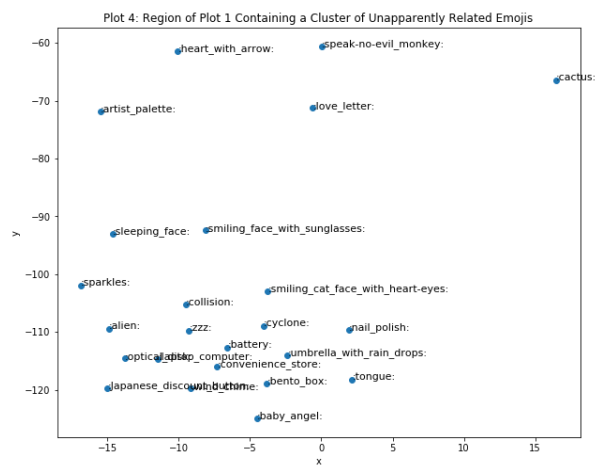
Plot 1.1: t-SNE Visualization of the PMI matrix with labels.



Plot 2: Region of Plot 1 containing moon-related emojis.



Plot 3: Region of Plot 1 containing animal-related emojis.



Plot 4: Region of Plot 1 containing cluster of emojis with an unclear relation