Untangling Emoji Popularity through Semantic Embeddings

Wei Ai,[1] Xuan Lu,[2] Xuanzhe Liu,[2] Ning Wang,[3] Gang Huang,[2] Qiaozhu Mei[1]
[1] School of Information, University of Michigan, Ann Arbor, USA;
[2] Key Laboratory of High Confidence Software Technologies (Peking University),
MoE, PRC, Beijing, China; [3] Xinmeihutong Inc., Beijing, China
{aiwei, qmei}@umich.edu, {luxuan, xzl, hg}@pku.edu.cn, ning.wang@xinmei365.com

Abstract

Emojis have gone viral on the Internet across platforms and devices. Interwoven into our daily communications, they have become a ubiquitous new language. However, little has been done to analyze the usage of emojis at scale and in depth. Why do some emojis become especially popular while others don't? How are people using them among the words? In this work, we take the initiative to study the collective usage and behavior of emojis, and specifically, how emojis interact with their context. We base our analysis on a very large corpus collected from a popular emoji keyboard, which contains a full month of inputs from millions of users. Our analysis is empowered by a state-of-the-art machine learning tool that computes the embeddings of emojis and words in a semantic space. We find that emojis with clear semantic meanings are more likely to be adopted. While entity-related emojis are more likely to be used as alternatives to words, sentiment-related emojis often play a complementary role in a message. Overall, emojis are significantly more prevalent in a sentimental context.

1 Introduction

Good morning ☃☃☺! Wanna make some ☕ before reading our 📄? 😀

In recent years, the prevalence of emojis has been an amazing phenomenon of social innovation and appreciation. Emojis, graphic symbols carrying specific meanings, are created, quickly adopted into online conversations, supported by multiple platforms, and inducted into Unicode standards. Oxford Dictionaries even named the emoji 😂 (Face With Tears of Joy) the Word of the Year of 2015, recognizing its popularity and impact on the popular culture[1]. It won't be too surprising if they are making their ways to a Nobel Prize one day 😜😝.

Why not? Emojis have become a universal language that is used across apps, across platforms, and across cultures. They are used not only in daily communications but also as marketing tools[2], as symbols of persuasion campaigns[3], as communication channels to specific populations[4], and even as programming languages[5]. Underneath this great prosperity, however, we do observe that the popularity of emojis is highly skewed and sometimes puzzled. Indeed, we can name many emojis that are designed similarly to 😂, many of which are also associated with positive sentiments, but they are far less frequently used by the crowd (e.g. 😃😎). Is 🌑 more popular than 🌓 and 🌗? Why is one of the moons more appreciated than others? What do 🎄 and 🌴 mean and why one appears more often than the other? Simply looking at the graphics and the frequency counts, we don't have a clue.

Similar concerns about inequality have been raised by social observers. Some are worried about the increasing fragmentation of emojis which may lead to misunderstanding and misuse[6]. Others have pointed out the problems with diversity and inclusiveness in emojis, especially when they are playing increasingly important roles in social communications[7]. Recent research has reported ambiguity, misinterpretation, and cultural difference in the use of emojis (Miller et al. 2016), which may result in compromised communication experiences, social awkwardness, and even cultural offenses. All these issues point to the meanings of emojis, and they may either appear as or result in an unequal usage of emojis.

The success of 😂 and the failure of many others intrigue us to untangle the frequency counts and to understand what have attributed to the popularity of emojis. Diffusion of innovations has been well studied in the social networks literature. Important factors that affect the cascade of information (an emoji in our case) are identified, which are broadly related to who have adopted it so far and how they are interconnected (Easley and Kleinberg 2010). In this work, we take a different perspective to investigating the factors that are intrinsic to emojis, or more specifically, factors that relate directly to the semantic meanings of emojis. Indeed, the semantics of emojis are directly related to the aforemen-

[1] http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/, retrieved October 2016.

[2] http://nyti.ms/1QEC7Wt, retrieved Oct. 2016.

[3] http://nyti.ms/2aQaoZG, retrieved Oct. 2016.

[4] http://nyti.ms/29Iaspc, retrieved Oct. 2016.

[5] http://www.emojicode.org/, retrieved Oct. 2016.

[6] http://arstechnica.com/gadgets/2016/08/emoji-are-getting-ever-more-expressive-but-not-without-growing-pains/, retrieved Oct. 2016.

[7] http://motherboard.vice.com/read/the-politics-of-emoji-diversity, retrieved Oct. 2016

tioned concerns. Is the meaning of an emoji ambiguous? Can we easily describe it in words? Are the meanings of two emojis related? Do the answers to these questions affect the popularity of an emoji?

A major challenge of studying the meanings of emojis is that the emojis are officially labeled with very brief descriptions (e.g., "😀: Grinning Face"). Even if a description is informative, it may or may not be how people interpret the emoji (e.g., "🌤" officially means "sun behind" but is widely used as "Good Morning"). Inspired by the principle that "*You shall know a word by the company it keeps (Firth 1957)*," we investigate the semantics of emojis based on how they are actually used along with natural language. Recent developments in distributional representations of text have enabled us to learn the embeddings of emojis and words in the same semantic space. Our analysis is empowered by a very large corpus of online chatters collected from users of a famous emoji keyboard. Based on the semantic emoji/word embeddings learned from this data set, we present the first in-depth analysis of how the semantics of an emoji affect its popularity. We make the following contributions:

- We present the first quantitative study that correlates emoji semantics to emoji usage.
- We measure important properties of emoji semantics based on joint embeddings of emojis and words, which are learned from the largest emoji usage data to date.
- We conduct the first analysis on the relation between emojis and words - whether emojis are used as a complementarity or a supplement to the natural language.
- We identify factors that significantly affect emoji popularity, including the structural properties of an emoji in the semantic space, its complementarity to words, and the sentiment of the context.

The results of our analysis provide useful insights both to designers of emojis and to data miners who can benefit from utilizing emojis in their research.

2 Related Work

We start with introducing the background and literature related to our research. Our study is inspired by three streams of literature: nonverbal elements in communication, sentiment analysis, and information diffusion.

Emoticons and Emojis People have been long using *emoticons* to provide non-verbal cues in online communications (Walther and D'Addario 2003; Park et al. 2013). Such nonverbal cues help people better interpret the nuance of meaning and the level of emotion not captured by language elements alone (Gajadhar and Green 2005; Lo 2008). There has also been research on the sentiment of emoticons (Boia et al. 2013), which reported that the sentiment of an emoticon is in substantial agreement with the sentiment of the entire tweet.

Since the debut on Twitter and Instagram, emojis quickly expanded their territory from emotions to various objects (sports, foods, etc.). Researchers have been trying to understand the interpretation of emojis (Miller et al. 2016), and how emojis facilitate communications (Kelly and Watts 2015; Vidal, Ares, and Jaeger 2016), with a particular interest in their sentiments (Kralj et al. 2015).

A key question is to find a good representation of emojis. Some researchers use the official description on the Unicode Website (Lu et al. 2016), while others utilize the word embedding tools to represent emojis with high-dimensional vectors (Eisner et al. 2016; Barbieri, Ronzano, and Saggion 2016). Indeed, several word embedding tools have been developed to find distributed representations of words, and have shown great potentials in various tasks, such as classification and visualization (Tang et al. 2015; Mikolov et al. 2013b; 2013a).

Similar to Barbieri, Ronzano, and Saggion (2016), we also applied a state-of-art embedding model to project words and emojis onto the same high-dimensional vector space, from where we conduct extensive analysis on the popularity of emojis.

Sentiment Analysis Both emoticons and emojis have been widely used to express the emotions, which relate our work to the sentiment analysis literature. Sentiment analysis has long been a core problem of natural language processing (Pang and Lee 2008; Mei et al. 2007; Liu 2012). Although various advanced sentiment analysis techniques have been proposed, accurately identifying sentiments and emotions from free text is still very challenging.

The emergence of emoticons and emojis provide new opportunities to analyze the sentiment expressions in textual context. Many researches have attempted to model text sentiments with emoticons and emojis used in the context (Zhao et al. 2012; Kralj et al. 2015).

Instead of analyzing the predicting power of emojis on the sentiment of the message, we are curious if the sentiments of messages would predict the usage of the emojis. Since the messages in the corpus are usually short, we chose lexicon-based approaches for sentiment analysis (Wilson, Wiebe, and Hoffmann 2005; Hu and Liu 2004; Kiritchenko, Zhu, and Mohammad 2014).

Adoption and Virality Our work is also related to the literature about the propagation of messages, such as Tweets (Tan, Lee, and Pang 2014), quotes (Danescu-Niculescu-Mizil et al. 2012), memes (Simmons, Adamic, and Adar 2011), rumors (Zhao, Resnick, and Mei 2015), etc.

Researchers have discovered various factors that explain the adoption and virality of certain messages, such as social network structure (Romero, Tan, and Ugander 2013), wording (Tan, Lee, and Pang 2014), serendipity (Sun, Zhang, and Mei 2013), and other lexical characteristics (Danescu-Niculescu-Mizil et al. 2012).

The object in this study is slightly different, as the emoji is brand-new non-verbal language units, and is phenomenally adopted by Internet users. Although there have also been researches on the adoption of both emoticons and emojis (Park et al. 2013; Lu et al. 2016), these researches focus on the culture-level difference in adoption, while our work tackles directly at the popularity of individual emojis.

Most of the above work is conducted on Tweets, which cover only one aspect of users' online activities. We believe a more comprehensive data set is needed to cover more scenarios of users' online communications. In the next section, we will discuss the unique data set used in this study.

3 Data Set

The data set was originally collected by a leading input method app on Google Play, namely the Kika Emoji Keyboard (Kika). It is an emoji-oriented keyboard, which offers users easier access to typing emojis into their messages. At the time of the data collection, Kika did not offer personalized emoji recommendations, and users select emojis from a universal and fixed layout. As a security feature of Android systems, users are notified that their input may be collected when activating Kika or other third-party keyboards. With users' approval, Kika is able to collect anonymized user-input messages as well as user meta data (such as language and country information) for analysis and research purposes. In its Privacy Policy[8], Kika explicitly declares that no personal and traceable data from the user input are recorded.

We took careful procedures to protect the privacy during analyses and preserve research ethics. First, the data set is stored on a private cloud server with strict access control authorized by Kika. Second, our analysis is governed by Kika employees to ensure the compliance with its privacy policy. Third, all user identifiers are replaced with randomized hash strings, and no individual user can be identified from the data set. Moreover, the presented study is completely based on analyzing frequencies and cooccurrences of emojis and words, which does not aim to extract any named entities, relations, or other compounds from natural language. Our study has been reviewed by the University of Michigan institutional review board (HUM00124978) and exempted from ongoing IRB reviews. More details of the data set and ethic considerations can be referred to the authors' previous work (Lu et al. 2016).

As a system-wide app, Kika keyboard can be used in various apps. Therefore, the collected messages are not limited to particular apps. Such characteristics make our data set unique and especially comprehensive for studying the language usage on smartphones. Although Kika is a multi-language keyboard that supports more than 60 languages, we only focus on English-speaking users in America. Our data covers 1.03 million such users and their 1.22 billion messages in September 2015, where each message is defined as the content user typed in before pressing the "Send" button. In this work, we ignore the timestamp information and treat all messages as independent. We modified the CMU ARK Twitter Part-of-Speech Tagger (Gimpel et al. 2011) to tokenize the messages. The vocabulary size of the text corpus is 5.9 million, counting tokens that occur in at least five messages.

It is observed that 9.2% of the messages include at least one emoji, which evidences the popularity of emojis. Although Kika supports 1,281 emojis at the time of data collection, not all of them are equally popular. Indeed, the popularity of each emoji follows a power-law distribution. That is, a small portion of emojis are more frequently used, while the majority are not. Such inequality of popularity intrigues us to untangle the popularity of emojis.

[8] http://www.kika.tech/privacy/

4 Semantic Representation

To understand why an emoji is popular, we should first understand its meaning. Indeed, the Unicode Consortium provides textual description of each emoji. For example, 😂 is described as "face with tears of joy", 🌴 is described as "palm tree". Yet users don't see the description when they decide to use an emoji, so does the text really describe how users interpret it?

Instead of using the official descriptions, we decide to characterize the meaning from the context where emojis are used. And instead of *describing* an emoji, we would like to *represent* it with the words that are most similar to it.

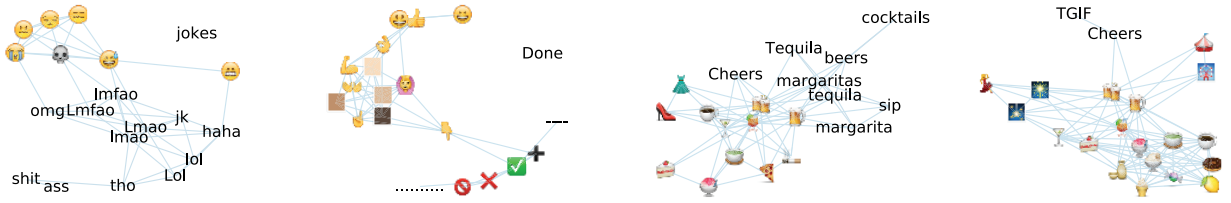
The similarity is measured semantically. That is, similar emojis are used in similar semantic contexts. To qualitatively measure the semantic similarity, and to find representing words for each emoji, we apply a network embedding algorithm that projects all language tokens on to the same high-dimensional space. In such space, words that are close to each other are semantically similar to each other.

Specifically, we choose to use LINE, a state-of-the-art embedding model that performs competitively in word analogy tasks (Tang et al. 2015). We start by constructing a co-occurrence graph from the corpus to represent the semantic structure. In the co-occurrence graph, each node represents a token (which could be a word, an emoticon, or an emoji), and the edge between the nodes represents the co-occurrence of the pair of tokens. The weight of the co-occurrence edge is defined as the logarithm of the total co-occurrences within a fixed-size window in the same message. We set the window size to be 5 as suggested in literature (Tang et al. 2015).

LINE provides two settings, which preserve first-order and second-order proximity respectively. The first-order proximity between two nodes is the local pairwise proximity, which is directly measured by the weight of the edge between the two nodes. In the co-occurrence graph, the first-order proximity measures how likely the two tokens co-occur. The second-order proximity between two nodes, however, is the similarity between their neighborhood network structures. That is, how likely the two tokens occur in similar contexts. In this work, we trained two LINE models that preserve the first and second order proximity, and obtain two sets of embeddings. We will refer to the two models as LINE-1st and LINE-2nd respectively.

The LINE embeddings enable us to find the tokens that are semantically similar to another token, as the semantic similarity can be directly computed as the euclidean distance in the embedding space. Tokens geometrically close to each other in the embedding space are also semantically similar. Therefore, the semantics of a token can be represented by its nearest neighbors in the embedding space.

Further, the nearest-neighbor relationship can be represented as a k-nearest-neighbor (kNN) graph, where the nodes are tokens, and there is an edge between two tokens



(a) 😂 (LINE-2nd) (b) ✔ (LINE-2nd) (c) 🍷 (LINE-2nd) (d) 🍷 (LINE-1st)

Figure 1: Examples of emoji egonets in the k-nearest graph.

only if one is among the k-nearest neighbors of the other. Such a graph not only preserves the list of neighbors for each emoji but also encodes the structural information in the semantic space. Since the nearest neighbors are mostly infrequent words, we filter the vocabulary and keep only those with more than 2,000 occurrences. Around 1% of the total vocabulary and 851 emojis are left after the filtering.

Figure 1 shows the egonet of some emojis in the kNN graph, where k is set to 20. The egonet is the subgraph induced by the neighbors of the emoji.

We first look at the egonet of the most popular emoji 😂 in LINE-2nd (Figure 1a), and a few observations can be made: First, many of its neighbors are popular Internet slangs, such as "lol" and "lmao". If users use 😂 and "lol" interchangeably, we may also expect 😂 to be popular. Indeed, the distance between 😂 and its closest neighbor "lol" is very small (though not directly reflected in the figure), suggesting that the meaning of 😂 is very clear. Second, the egonet is dense and there are many edges between the neighbors, indicating that the neighbors of 😂 are also close to each other. Such density suggests that 😂 is inside a semantic cluster, and has little ambiguity in its meaning.

In comparison, we look at the egonet of a less popular emoji ✔ (Figure 1b) in LINE-2nd. None of its neighbors are frequently used. We could also see that the neighbors fall into three clusters, indicating that the ✔ lies in the middle of several semantic clusters and are not close to any one in particular. Although there is a word "Done" among its neighbors, the distance between the two is quite large. Such isolated clusters and the long distance to its neighbors suggest that the meaning of ✔ is unclear and ambiguous.

We also compare the embeddings of the two settings of LINE. Figure 1c and 1d compare the egonet of another frequently used emoji 🍷 (*wine*) in LINE-1st and LINE-2nd. On one hand, we see many drinks in the neighbors in LINE-2nd, which are replaceable words with 🍷. On the other hand, the nearest words in LINE-1st, namely "Cheers" and "TGIF" (Thank God It's Friday), aren't wine or any drink, but they are indeed words that co-occur with 🍷 and are semantically closely related. Since these neighbors are frequently used, we may also expect the emoji 🍷 to be popular.

To become popular, it seems that an emoji should have clear, unambiguous meaning, and be closely related to popular words. With these insights, we extract these factors from the k-nearest graph and systematically study how they affect the emoji popularity.

5 Egonet and Emoji Popularity

The egonets in the previous section inspire us to look at how the neighbors of an emoji can affect its popularity. In this section, we will quantitatively characterize the egonet of an emoji and explore its effect on the emoji popularity. Specifically, does being similar to frequently used words increase the popularity? Does having a clear, unambiguous meaning increase the popularity? In the rest of the paper, we measure an emoji's popularity as the number of messages where it occurs.

Frequency We first look at the usage frequency of the neighbors. Intuitively, if an emoji is replacing a very popular word, we may also expect the emoji to be popular. Such popularity can be measured by the occurrences of its nearest words in LINE-2nd. On the other hand, it may also be frequently used if it goes well with popular words, which can be measured by the popularity of its nearest words in LINE-1st. Therefore, we conjecture that an emoji's popularity is positively correlated with the popularity of its neighbors in the kNN graph.

In Figure 2 and Figure 3, we plot the correlation between the max popularity of the neighbors and the popularity of the emoji. We separate different categories of tokens (emojis, words, and emoticons) and plot their popularity respectively. The plots for emoticons share similar trends with the other two categories and are omitted for the sake of space.

We can see that whatever category its neighbors are in, an emoji's popularity is always correlated with the max popularity of them, which is consistent with our conjecture.

Distance Next, we measure if the meaning of an emoji is clear. The clearness can be measured by the distance to the nearest neighbors. That is, how close an emoji is to its nearest neighbors. In LINE-2nd, a small distance indicates that the two neighbors are more replaceable to each other, while with LINE-1st, a small distance suggests that the two tend to be used together, the appearing of one usually indicates the occurrence of the other.

Figure 4 and Figure 5 show the correlation between the minimum distances to the nearest neighbors and the popularity of emojis, using both LINE-1st and LINE-2nd. We can

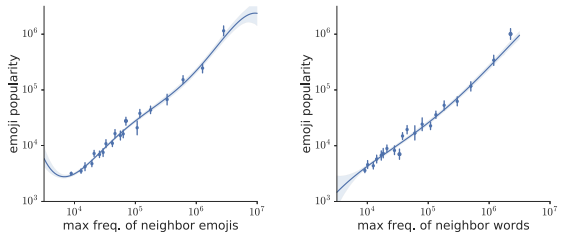


Figure 2: Correlation between max frequency of the 20 neighbors and emoji popularity in LINE-2nd

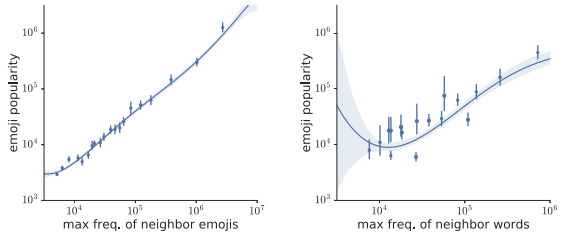


Figure 3: Correlation between max frequency of the 20 neighbors and emoji popularity in LINE-1st

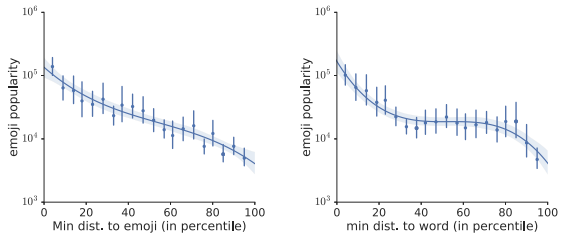


Figure 4: Correlation between distances to nearest neighbors and emoji popularity in LINE-2nd

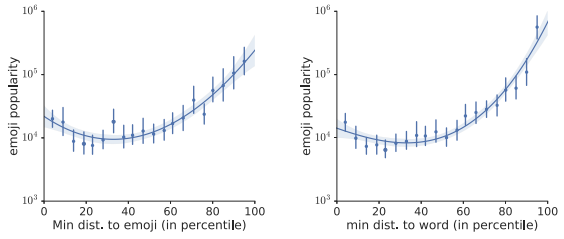
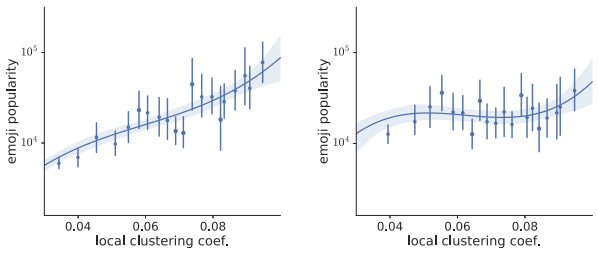


Figure 5: Correlation between distances to nearest neighbors and emoji popularity in LINE-1st

see that for the LINE-2nd, the distances are negatively correlated with the popularity. A small distance indicates that the users can easily perceive the emoji with other words, so they are more clear about the meaning of the emoji. And such emojis are more widely used.

The correlation in LINE-1st, though still significant, is positive. The closer an emoji is to another emoji or a word, the less frequently it is used. A small distance in LINE-1st



(a) LINE-1st (b) LINE-2st

Figure 6: Correlation between local clustering coefficient and emoji popularity

indicates that the emoji occurs only if its neighbor occurs. The negative correlation suggests that the users are not in favor of such emojis, and a popular emoji should be able to fit into many different contexts.

Local Clustering Coefficient The egonet of ✓ has a different structure than the others, since its neighbors are not as connected. Such emojis might be ambiguous as it lies in between different semantic clusters. Would ambiguity affect an emoji's popularity? To generalize this insight to all emojis, we need to quantify the clustering structure of the kNN graph. We could get the number of clusters for each emoji by applying clustering algorithms or community detection techniques. However, such algorithms usually need pre-determined parameters, which the numbers of clusters are very sensitive to. Instead, we choose to measure such ambiguity by *local clustering coefficient* (LCC), which is defined as the density of the egonet of a node. A high LCC can be interpreted as "my neighbor's neighbors are also my neighbors". In the kNN graph we constructed, a high local clustering coefficient indicates that the emoji is in a dense area of the semantic space, where there are highly connected token clusters, while emojis with lower LCC are more likely to be in a thinner area where tokens are scattered.

In Figure 6, we plot the local clustering coefficients for all emojis and their correlation with the popularity. As we can see, the popularity is positively correlated with local clustering coefficient in both 1st and 2nd order LINE. This suggests that users do prefer emojis that are unambiguous.

Regression Analysis The above results indicate that users prefer the emojis that are close to a popular word and have clear and unambiguous meaning. We can measure this preference with the correlation between the emoji popularity and quantitative characters of the egonet. We are curious how the correlations interact with each other. That is, does each correlation still hold given others are controlled? To answer the question, we conducted a regression analysis.

Intuitively, we want to put all the egonet statistics in LINE-1st and LINE-2nd together and run a multi-variable OLS regression. Given that there are 14 of them (min. distance to word/emoticon/emoji, max frequency of neighboring words/emoticons/emojis, and local clustering coeffi-

Table 1: Coefficients of the first 3 principle component (all features normalized to have zero mean and unit variance)

	PC1	PC2	PC3
LINE-1st			
Dist. to nearest emoji	-0.185	0.299	0.502
Dist. to nearest emoticon	-0.293	0.310	0.035
Dist. to nearest word	-0.256	0.300	0.009
Max. freq. of neighbor emoji	-0.372	-0.044	0.050
Max. freq. of neighbor emoticon	-0.318	-0.124	-0.006
Max. freq. of neighbor word	-0.260	-0.165	-0.007
Local clustering coef.	-0.151	0.289	-0.467
LINE-2nd			
Dist. to nearest emoji	0.192	0.293	0.484
Dist. to nearest emoticon	0.106	0.494	-0.029
Dist. to nearest word	0.163	0.404	0.035
Max. freq. of neighbor emoji	-0.369	-0.037	0.088
Max. freq. of neighbor emoticon	-0.363	-0.005	0.106
Max. freq. of neighbor word	-0.368	-0.011	0.048
Local clustering coef.	-0.083	0.320	-0.518
Variance explained	0.48	0.18	0.10

cient, for both LINE-1st and LINE-2nd), many of the statistics are highly correlated with each other. For example, the Pearson correlation between the maximum frequencies of neighbor emojis and words are close to 0.9. Such high correlation would result in multicollinearity in the OLS regression.

To address the collinearity, we normalize all the 14 ego statistics and perform a Principle Component Analysis (PCA). The first 3 components explained 0.77 of the variance. We report the coefficients of the first 3 principle components in Table 1.

According to the component, we can cluster the ego statistics into three groups: the distances to nearest tokens can be clustered into one group, the maximum frequencies into another. The local clustering coefficient is negatively correlated with distances to nearest emojis, but are uncorrelated with the rest, so we cluster the two LCCs into the third cluster.

Based on the observations from PCA, we selected three neighbor statistics into the regression: local clustering coefficient in 1st order LINE, distance to the nearest word in 2nd order LINE, and the maximum log-scaled frequency of nearest emojis in 2nd order LINE. The coefficients and their statistical significances are reported in Table 2. We can see that the correlations we observed earlier persist even when we put multiple statistics together.

6 Relation to Words

The correlations between the distance to neighbors and the popularity are also intriguing. The popular emojis have closer meanings to their neighbors in LINE-2nd but distantly related to their neighbors in LINE-1st. The neighbors in LINE-1st and LINE-2nd are substantially different. The neighbors in LINE-1st tend to co-occur with each other, while the neighbors in LINE-2nd have similar meanings and

Table 2: Regression Analysis. Variables in their original values without normalization (except frequency in log scale).

	Dependent variable:
	Popularity (log scale)
Local clustering coef.	1.308^{***}
(LINE-1st)	(0.224)
Dist. to nearest word	-0.004^{**}
(LINE-2nd)	(0.002)
Max. freq. of neighbor emoji	0.928^{***}
(LINE-2nd)	(0.012)
Constant	-0.262^{***}
	(0.089)
Observations	851
R^2	0.901

Note: Standard errors in parentheses
Significant at the: *** 1%, ** 5%, or * 10% level

can be used interchangeably.

Such differences suggest that there are two different relationships between emojis and words, namely complementary and supplementary. The two potential relationships correspond to complementary and supplementary goods in the economics. When the demand of a certain good increases, its complementary goods will also have an increase in demand, while its supplementary goods will be less demanded. For example, coffee and creams are complementary goods: when people drink more coffee, they consume more cream, while coffee and tea are supplementary, as people usually choose one of the two to fuel their days.

In this section, we want to see how these two types of relationships are reflected in the text corpus. Do users use emojis to replace words, or do they use emojis to complement words? Specifically, an emoji is said to be more complementary (high complementarity) if it tends to co-occur with words, and supplementary (low complementarity) if it tends to replace words. We take two alternative approaches to measuring the complementarity.

6.1 Measure by Mutual Information

The most straightforward way to measure complementarity is to measure the occurrence and co-occurrence of the emojis and their similar words (i.e. their neighbors in LINE-2nd). If an emoji is a perfect substitute for its neighboring word, they should not co-occur with each other in any messages. On the other hand, if an emoji is a perfect complement to its neighbor, we should expect them to always co-occur in messages. Admittedly, there is no perfect substitute or complement in natural language. Still, we can use the association of an emoji and its nearest word as a measure of complementarity.

Such association is measured by Pointwise Mutual Information (PMI) (Church and Hanks 1990). Denote the i-th emoji as e_i, its nearest neighbor as n_i, their occurrence probability in a message as $p(e_i)$ and $p(n_i)$, and their co-occurrence probability as $p(e_i, n_i)$. PMI can be computed

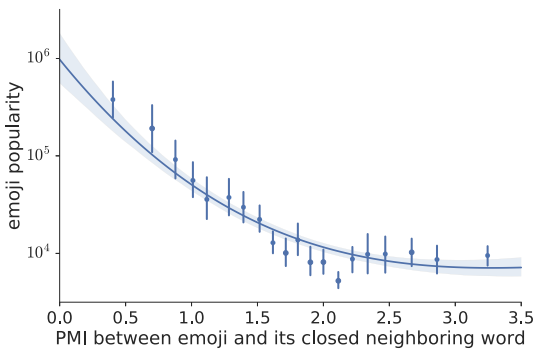


Figure 7: Correlation between the emoji-neighbor PMI and emoji popularity

as

$$\text{pmi}_i = \log \frac{p(e_i, n_i)}{p(e_i)p(n_i)}. \quad (1)$$

A positive PMI implies that the occurrence of the emoji and its nearest word is associated, indicating a complementary role of the emoji. We calculate the PMI for all emojis as well as its correlation with the emoji popularity (Figure 7).

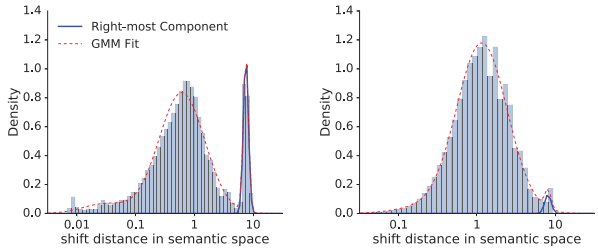
Except for a few emojis who have negative PMI with their nearest words (e.g. -0.3 for 🌚, -0.16 for 😇, and -0.02 for 😂), most emojis have positive PMI (e.g. 2.04 for 🐶), and there is a negative correlation between PMI and emoji popularity. The result suggests that most of the emojis are complementary to their similar words while popular emojis are more supplementary.

6.2 Measure by Semantic Shifting

Although straight forward, PMI measures only the complementarity between emojis and their neighbors in LINE-2nd. Yet the emojis are always used in contexts, so we should also measure the complementarity between emojis and their contexts.

If an emoji is complementary, it's meaning should be coherent with the rest of the message. For example, "That dog is sooooo cute 🐶🐶." Removing the emoji won't hurt the meaning of the message much. However, if an emoji is supplementary, it plays an important role in determining the meaning of the messages. Removing such emoji would certainly affect the meaning of the message. One such example would be "Oh that 🐶 is awesome!!".

The meaning of a message could be represented by the mean of the embedding vectors of its words. Therefore, by comparing the mean vectors of a message with and without emojis, we can calculate how much the meaning shifts by taking out the emojis. If the mean vector shifts by a large distance, we could conclude that the emoji is supplementary, while a small shift implies high complementarity of the emoji. Notice that this should be measured in LINE-2nd rather than LINE-1st since in LINE-1st, an emoji is presumably located near the words that often co-occur in the same message.



(a) dog face 🐶 (b) face with tears of joy 😂

Figure 8: Distributions of the mean vector shift distance in semantic space caused by removing certain emojis

We plot the distribution of the shifted distance caused by removing emoji 🐶 in Figure 8a, we can clearly see a bimodal distribution in the plot. The larger component on the left indicates that most shifts are small and the emoji plays a complementary role; while the right component represents the messages where the emoji plays a supplementary role and substantially determines the meaning of the message. As for comparison, we plot a similar distribution of the most popular emoji 😂 in Figure 8b. Although we could still see a bimodal distribution, the right-most component is much smaller.

Indeed, nearly all of emojis exhibit a similar bimodal distribution. On one hand, such bimodal distributions suggest that emojis can be both complementary (falling into the left component) and supplementary (right component). On the other hand, the bimodal distribution allows us to quantitatively measure the complementarity of emojis, through measuring the relative size of the two components. To measure the size of the right component, we fit the distribution with a three-component Gaussian Mixture Model (GMM) and extract the weight of the right-most component as its size. (We use an additional component to capture some artifacts in the left tail, which was brought by a precision issue during the computation.) The GMM fit for 🐶 and 😂 is already plotted in Figure 8, and their weights of the right-most components are 0.12 and 0.01 correspondingly, indicating that 😂 is more complementary.

To generalize the semantic shift measure to all emojis, we fit a GMM for each emoji and correlates the weight of the right-most component with the emoji popularity, and plot the result in Figure 9. Clearly, the complementarity is positively correlated with emoji popularity, though the fitted curve flattens as the weight goes to the long tail. Such correlation suggests that the popular emojis tend to be used to complement the meaning of the sentences. And the flattened tail is likely because that the GMM does not fit well for the emojis that don't occur a lot.

By putting Section 6.2 and Section 6.1 together, we find that the more popular emojis are more supplementary to their nearest words, but at the same time more complementary to its context.

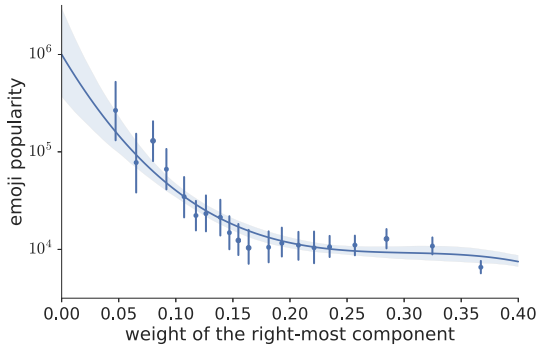


Figure 9: Correlation between the weight of the right-most component in GMM and emoji popularity

7 Sentimental Context

When we look at the spectrum of the semantic-shift complementarity, we see that many of the most complementary ones are sentimental emojis, which suggests that the sentimental emojis are more likely to be used with others and complement their context. Does that indicate that they are more likely to be used in sentimental contexts? In this section, we explore the relationship between sentimental context and the usage of emojis.

We use a lexicon-based method to annotate the sentiment of the context. Specifically, we use a standard lexicon library named LIWC[9] to generate the emotion score of each word, and we average the emotion score of the words to get the sentiment score of a message (excluding emojis). We do not distinguish positive and negative emotions as we only need to know how sentimental a message is. The sentiment scores we get are between 0 and 100, indicating the percentage of words in a message that are sentimental.

We randomly sample 100,000 messages from the corpus and calculate the sentiment scores. We want to regress the number of emojis used in a message on the sentiment score of that message. We need to be careful since one may expect that more emojis are used in longer messages. Therefore, we control the message length without emojis in the regression. We further filter the messages that are extremely long (> 30 words) or short (< 2 words) and conduct regressions on both the entire sampled messages and the filtered messages.

The regression results are summarized in Table 3. The regression on all messages suggests that the sentiment of the message seems to have no significant effect on the number of emojis used in it. However, with the messages of extreme length filtered, the effect of message length diminishes, while sentimental effect becomes prominent. The coefficient of 0.008 looks small at first sight. Yet think about in a message of median length (7 words), one extra sentimental word (which translate into an increase of 14 in sentiment score) would result in 0.11 more emoji used in that message on average.

[9]Linguistic Inquiry and Word Count. http://liwc.wpengine.com

Table 3: OLS Regression: Sentiment of Messages Predicts Emoji Usage.

	Dependent variable:	
	# Emoji in a Message	
	(1) all	(2) filtered
Message length	0.010^{***}	-0.001
(excluding emojis)	(0.002)	(0.002)
Message sentiment score	0.002	0.008^{***}
(using words only)	(0.002)	(0.001)
Constant	0.363^{***}	0.242^{***}
	(0.032)	(0.022)
Observations	100,000	90,301
R^2	0.0004	0.001

Note: Standard errors in parentheses
Significant at the: *** 1%, ** 5%, or * 10% level

Table 4: Regression Analysis on All Semantic Factors.

	Dependent variable:
	Popularity (log scale)
Local clustering coef.	1.467^{***}
(LINE-1st)	(0.223)
Dist. to nearest word	-0.005^{***}
(LINE-2nd)	(0.002)
Max. freq. of neighbor emoji	0.838^{***}
(LINE-2nd)	(0.019)
PMI complementarity	-0.038^{***}
	(0.012)
Semantic-shift complementarity	-0.065
	(0.099)
Sentiment score	0.004^{***}
	(0.001)
Constant	0.220^{*}
	(0.126)
Observations	851
R^2	0.906

Note: Standard errors in parentheses
Significant at the: *** 1%, ** 5%, or * 10% level

8 Discussion

Putting Together In the previous sections, we have found multiple factors that affect the popularity of emojis: structural properties in the semantic space, complementarity to words, and sentimental context. One may question the correlation between these factors. That is, does the correlation between each factor and the popularity still holds given the other factors are controlled?

With this question in mind, we regress the emoji popularity on all the selected structural properties and complementarity. As it is reported that sentimental emojis are more likely to be used (Kralj et al. 2015), we control the emoji's sentiment score (calculated as the average sentiment score of its neighboring words in LINE-2nd) in the regression. The result is summarized in Table 4. We can see that most

correlations are still significant even with all other factors controlled. In fact, the coefficients of the structural properties are similar to those in Table 2, indicating that their effect on the popularity is robust. The correlation between the semantic-shift-measured complementarity and emoji popularity is no longer significant, suggesting that the complementarity between emoji and contexts may be explained by the other factors. The correlation between sentiment score and emoji popularity is significantly positive, which is coherent with existing literature.

Limitations In this work, all the analyses are correlation rather than causality. Acknowledgeably, many uncontrolled variables may confound our conclusions:

First, we aggregate the messages from all users using all apps together. In fact, different users may have different preferences on emojis, as has been observed in (Lu et al. 2016). Also, people may have different language styles in different apps and different scenarios, as one would certainly use more 😘 on Kik than in emails. In the future work, we would like to have better control on users and apps to make the analyses more rigorous.

Second, our data span only one full month. Also, some external events, such as the supermoon lunar eclipse on September 27, may affect the usage of related emojis. Should a data set with longer time span be available, we would have more control of the external variations. The available emoji set remains stable throughout the period so we cannot analyze the adoption of *new* emojis as they are introduced.

Implications Despite the limitations, our work does provide implications on the development of emojis – not only the rendering and customization of existing emojis, but also designing and promoting new emojis. The emoji should have a clear and unambiguous meaning for people to perceive. It should be semantically similar to a popular word. The emoji should also be sentimental, and well fit sentimental contexts.

The correlation between sentimental context and emoji usage also indicates the emoji's potential ability in sentiment classification task. Not only can it be used as a feature in training classification models, but it can also be used to differentiate ambiguity with the nuance of its semantics.

9 Conclusion and Future Work

Whether an emoji is commonly accepted largely depends on its meaning, and more precisely how it is related to words and other emojis in the semantic space and how these words and emojis are used in the context. In this study, we have presented the first quantitative study correlating the semantics of emojis to their usage using the largest emoji usage data to date. We train embeddings of both words and emojis and construct a k-nearest neighbour graph. With the kNN graph, we are able to characterize the semantic relationship between emojis and words with structural property of the egonet. We also quantitatively measure the complementarity of emojis to words. Results suggest that the emoji popularity is affected by several factors, including its structural properties in the semantic space, its complementarity to words, and the sentiment of its context. In future, we plan to establish causal relationships between the identified factors and emoji popularity, with better controls for confounding variables and the causal inference techniques from the econometric literature.

10 Acknowledgments

This work was supported by the National Science Foundation under Grant No. IIS-1054199, an MCubed grant of the University of Michigan, the High-Tech Research and Development Program of China under Grant No.2015AA01A203 and the Natural Science Foundation of China (Grant No. 61370020, 61421091, 61528201). The authors would like to thank Ark Fangzhou Zhang from the University of Michigan for helpful discussions. Xuanzhe Liu and Qiaozhu Mei are corresponding authors of this work.

References

Barbieri, F.; Ronzano, F.; and Saggion, H. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. *LREC*.

Boia, M.; Faltings, B.; Musat, C. C.; and Pu, P. 2013. A :) is worth a thousand words: how people attach sentiment to emoticons and words in tweets. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 345–350.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: how phrasing affects memorability. In *ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*. Association for Computational Linguistics.

Easley, D., and Kleinberg, J. 2010. Networks, crowds, and markets: Reasoning about a highly connected world.

Eisner, B.; Rocktäschel, T.; Augenstein, I.; Bošnjak, M.; and Riedel, S. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, Special volume of the Philological Society. Oxford: Blackwell. 1–32.

Gajadhar, J., and Green, J. 2005. The importance of nonverbal elements in online chat. *Educause Quarterly* 28:63–64.

Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; and Smith, N. A. 2011. *Part-of-speech tagging for Twitter: annotation, features, and experiments*. Association for Computational Linguistics.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, Usa, August*, 168–177.

Kelly, R., and Watts, L. 2015. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*.

Kiritchenko, S.; Zhu, X.; and Mohammad, S. M. 2014. Sentiment analysis of short informal text. *Journal of Artificial Intelligence Research* 50:723–762.

Kralj, N. P.; Smailović, J.; Sluban, B.; and Mozetič, I. 2015. Sentiment of emojis. *Plos One* 10(12).

Liu, B. 2012. *Sentiment analysis and opinion mining*, volume 5. Morgan & Claypool Publishers.

Lo, S. K. 2008. The nonverbal communication functions of emoticons in computer-mediated communication. *CyberPsychology & Behavior*.

Lu, X.; Ai, W.; Liu, X.; Li, Q.; Wang, N.; Huang, G.; and Mei, Q. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 770–780.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007*, 171–180.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv.org*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, H.; Thebault-Spieker, J.; Chang, S.; Johnson, I.; Terveen, L.; and Hecht, B. 2016. "Blissfully happy" or "ready to fight": varying interpretations of emoji. In *Proceedings of the 10th International Conference on Weblogs and Social Media, ICWSM 2016*.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Park, J.; Barash, V.; Fink, C.; and Cha, M. 2013. Emoticon style: interpreting differences in emoticons across cultures. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013*.

Romero, D. M.; Tan, C.; and Ugander, J. 2013. On the Interplay between Social and Topical Structure. *ICWSM*.

Simmons, M. P.; Adamic, L. A.; and Adar, E. 2011. Memes Online: Extracted, Subtracted, Injected, and Recollected. *ICWSM*.

Sun, T.; Zhang, M.; and Mei, Q. 2013. Unexpected Relevance: An Empirical Study of Serendipity in Retweets. *ICWSM 2013*.

Tan, C.; Lee, L.; and Pang, B. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. *ACL* 175–185.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*.

Vidal, L.; Ares, G.; and Jaeger, S. R. 2016. Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference* 49:119–128.

Walther, J. B., and D'Addario, K. P. 2003. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 5(2):119–134.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *International Journal of Computer Applications* 7(5):347–354.

Zhao, J.; Dong, L.; Wu, J.; and Xu, K. 2012. Moodlens: an emoticon-based sentiment analysis system for Chinese tweets. In *Proceedings of the the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012*, 1528–1531.

Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. In *Proceedings of the 24th International Conference on World Wide Web*, 1395–1405.