No More One Liners: Bringing Context into Emoji Recommendations

JOONGYUM KIM and TAESIK GONG, KAIST
BOGOAN KIM, Ajou University
JAEYEON PARK, Yonsei university
WOOJEONG KIM, Ajou University
EVEY HUANG, Northwestern University
KYUNGSIK HAN, Ajou University
JUHO KIM, KAIST
JEONGGIL KO, Yonsei University
SUNG-JU LEE, KAIST

As emojis are increasingly used in everyday online communication such as messaging, email, and social networks, various techniques have attempted to improve the user experience in communicating emotions and information through emojis. Emoji recommendation is one such example in which machine learning is applied to predict which emojis the user is about to select, based on the user's current input message. Although emoji suggestion helps users identify and select the right emoji among a plethora of emojis, analyzing only a single sentence for this purpose has several limitations. First, various emotions, information, and contexts that emerge in a flow of conversation could be missed by simply looking at the most recent sentence. Second, it cannot suggest emojis for emoji-only messages, where the users use only emojis without any text. To overcome these issues, we present *Reeboc* (Recommending emojis based on context), which combines machine learning and *k*-means clustering to analyze the conversation of a chat, extract different emotions or topics of the conversation, and recommend emojis that represent various contexts to the user. To evaluate the effectiveness of our proposed emoji recommendation system and understand its effects on user experience, we performed a user study with 17 participants in eight groups in a realistic mobile chat environment with three different modes: (i) a default static layout without emoji recommendations, (ii) emoji recommendation based on the current single sentence, and (iii) our emoji recommendation model that

This work was in part supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2020R1A2C1004062), Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7083534), and Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).
Authors' addresses: Joongyum Kim, T. Gong, Juho Kim, and S.-J. Lee, School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea; emails: {kjkpoi, taesik.gong, juhokim, profsj}@kaist.ac.kr; B. Kim, J. Park, W. Kim, and K. Han, Department of Computer Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16449, Republic of Korea; emails: {bokim1122, gks3284, kyungsikhan}@ajou.ac.kr; E. Huang, Computer Science and Communication departments, Northwestern University Segal Design Institute, 3.230, 2133 Sheridan Rd, Evanston, IL 60208, USA; email: eveyhuang@u.northwestern.edu; J. Ko, School of Integrated Technology, College of Engineering, Yonsei University, 85 SongdogwahakRo, YeonsuGu, Incheon 21983, Republic of Korea; emails: {jaeyeon.park, jeonggil.ko}@yonsei.ac.kr.

2469-7818/2020/04-ART9 $15.00
https://doi.org/10.1145/3373146

considers the conversation. Participants spent the least amount of time in identifying and selecting the emojis of their choice with *Reeboc* (38% faster than the baseline). They also chose emojis that were more highly ranked with *Reeboc* than with current-sentence-only recommendations. Moreover, participants appreciated emoji recommendations for emoji-only messages, which contributed to 36.2% of all sentences containing emojis.

CCS Concepts: • **Human-centered computing** → **Smartphones**; • **Computing methodologies** → *Machine learning approaches*; • **Human-centered computing** → *User studies*; *Text input*;

Additional Key Words and Phrases: Emoji recommendation, mobile applications, machine learning, user experience

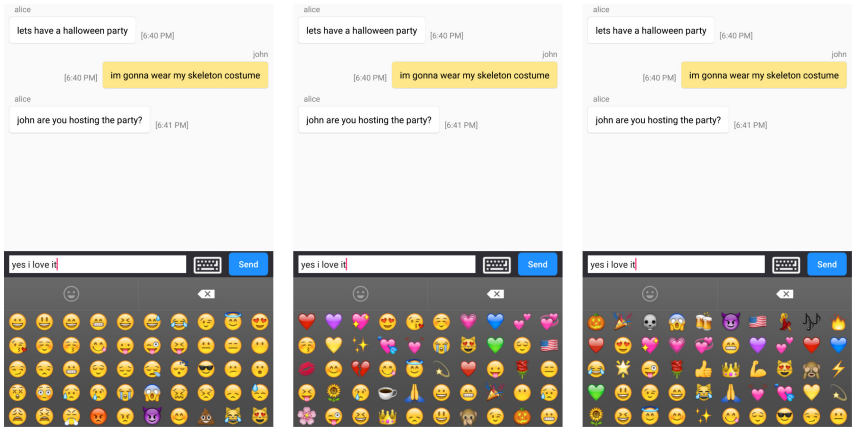
ACM Reference format:
Joongyum Kim, Taesik Gong, Bogoan Kim, JaeYeon Park, Woojeong Kim, Evey Huang, Kyungsik Han, Juho Kim, JeongGil Ko, and Sung-Ju Lee. 2020. No More One Liners: Bringing Context into Emoji Recommendations. *ACM Trans. Soc. Comput.* 3, 2, Article 9 (April 2020), 25 pages.
https://doi.org/10.1145/3373146

1 INTRODUCTION

Emojis have become a popular form of digital communication. They represent various emotions, facial expressions, objects, symbols, food and drinks, places, activities, flags, weather, animals, and even specific celebrities. People use emojis in chats, social network posts and comments, emails, and product reviews, just to name a few. It is reported that an average of 5 billion emojis are sent everyday on Facebook Messenger and 60 million on Facebook [9]. Moreover, more than half of the comments on Instagram include emojis [16]. Emoji usage is so widespread that an emoji (😂) was named the word of the year by the Oxford dictionary in 2015 [15], and there is also a World Emoji Day (July 17).

With such popularity of emojis, new emojis are often created, and as of March 2019, there are a total of 3,019 emojis in the Unicode Standard. Although users could enjoy the availability of diverse emojis, when selecting an emoji, having lots of options could lead to "the paradox of choice" [46] and delay in identifying the emoji of choice. Thus, emoji recommendation has been an active topic for both research [16, 18, 21, 50, 57] and products [26, 37, 54]. Using various machine learning and natural language processing (NLP) techniques, most emoji recommendation algorithms analyze the user's input text (e.g., the current input line of a chat) and suggest emojis. However, we believe capturing the *context* [10] of the whole conversation is important in emoji recommendation, as it is challenging to understand the context in one line of chat input. Moreover, analyzing only a single sentence leads to recommending many emojis of similar sentiment or emotion and missing various emotions or contexts expressed in conversations. Existing models have another limitation that affects user experience. As the models analyze the current input sentence, they cannot suggest emojis for "emoji-only sentences." As our user study indicates, 36.2% of the emoji-used messages were emoji-only inputs without any text. Existing recommendation models turn back to the default emoji layout for the emoji-only sentences, as they have no input text to analyze.

We propose *Reeboc* (Recommending emojis based on context), which recommends emojis based on conversation context. Instead of simply analyzing a single input sentence, we consider recent sentences exchanged in a conversation. *Reeboc* extracts various emoji usage contexts of the conversation using a long short-term memory (LSTM) network trained by conversation data captured from Twitter. Although we cannot capture all different types of context by analyzing only a part of a chat conversation, our study demonstrates the effectiveness of using beyond one chat line (e.g., five lines) in emoji recommendations. Figure 1 shows an example of emoji recommendation during mobile chat. The default keyboard (Figure 1(a)) displays the same layout whenever a user wants to send an emoji. An emoji recommendation model that considers the



(a) The default layout (b) The single-sentence model (c) The conversation model

Fig. 1. Emoji suggestion example. When an emoji suggestion system only considers a single sentence, its recommendation focuses on love-related emojis. However, when an emoji suggestion system considers conversations, Halloween and love-related emojis are recommended. The default layout used in mobile OSes is a static emoji layout without emoji recommendations.

current sentence (Figure 1(b)) shows a different layout, focusing on the the word *love* in the sentence. *Reeboc* (Figure 1(c)) instead analyzes the conversation beyond the current sentence and displays recommendations that capture various contexts appearing in the chat. Note also that as our model considers the *conversation* instead of only the current input text, we offer emoji recommendations for emoji-only sentences as well.

We evaluate the effectiveness of considering context in emoji recommendation and its effects on user experience through a user study. Results indicate that participants prefer the emoji recommendation model based on conversation than the one based only on the current sentence, as analyzing the context from the conversation provided a diverse set of emotions and information that could be missed from analyzing only the current input text. Participants were able to identify the emoji of their choice more quickly using our suggestion model. In addition, they chose emojis that were more highly ranked with *Reeboc* than with current-sentence-only recommendations. Moreover, the participants appreciated the recommendation made for emoji-only messages, which was not possible in previous recommendation models.

The contributions of the article are as follows:

- We propose *Reeboc*, an emoji recommendation system considering the conversation context. *Reeboc* recommends emojis by understanding the various contexts available through analyzing the conversation.
- We present a modeling method for analyzing the emoji usage context in a chat conversation and suggest how to recommend emojis in various emoji usage contexts.
- We evaluate *Reeboc* during a mobile group chat in a controlled setting to show its emoji suggestion effectiveness in reducing the emoji selection time and recommending relevant emojis.

This article is structured as follows. Section 2 presents related work in the domain of emoji recommendation and its social interactions. We perform a preliminary Amazon Mechanical Turk study with 230 participants in Section 3 to observe and understand the impact of analyzing conversations, rather than single sentences, for emoji suggestion, and present the design of our approach

in Section 4. The effectiveness of *Reeboc* is presented in Section 5 through a user study and its results. Finally, we discuss topics relevant to the social computing research community to foster further improvements in emoji suggestion models in Section 6 and conclude the article in Section 7.

2 RELATED WORK

We review related emoji suggestion methods from both recent research and products, and discuss the impact of emojis on social interactions.

2.1 Emoji Prediction Using Machine Learning

Characterizing and predicting emojis based on a given text is an active research topic in computational linguistics. Many prior studies apply machine learning models using NLP techniques on user-written text for emoji prediction [16, 18, 50].

Many existing NLP models in the context of social media rely on representation learning or pretrained word embeddings that can be obtained from tools such as word2vec [34] or GloVe [40]. For example, Barbieri et al. [5] trained emoji embeddings using the skip-gram method and showed increased accuracy. Eisner et al. [18] presented emoji2vec embeddings designed for emoji Unicode symbols learned from their description in the unicode emoji standard. It was shown that emoji2vec outperformed the skip-gram model on sentiment analysis tasks for a large collection of tweets.

Along with such efforts, many researchers proposed computational models using various machine learning algorithms (e.g., LSTM, bi-LSTM, autoencoder, sequence-to-sequence, other vision-related algorithms—CNN, ResNet) with vector representations of word- or character-level embeddings. Wu et al. [57] developed a CNN- and LSTM-based model that recommends emojis for plain tweet data. Thirty emojis were considered as dependent variables in their work. Felbo et al. [21] developed DeepMoji, a deep neural network model for emoji recommendation. The model uses two bi-LSTM models with an attention layer and is designed using 1.2 billion tweets containing one of 64 common emojis. Its performance is reported to outperform other models such as fasttext [27]. Zhou and Wang [63] proposed a conditional variational autoencoder (CVAE), which adds a response encoder, recognition network, and prior network (generated from the original encoder) on top of the conventional seq2seq (sequence-to-sequence) model. Their model was trained on around 650,000 tweets for predicting emojis from 64 candidates and showed greater performance than the attention-based seq2seq model. Similar to our work, Xia et al. [59] suggested that the accuracy of emoji recommendation could improve through considering conversation. They predicted only 10 emoji categories in the presence of two-person dialogue. Guibon et al. [24] showed that the sentiment analysis results of the dialogue were helpful in recommending emojis.

In addition to text-based modeling, recent research has focused on utilizing multiple modalities (e.g., images, demographics, date, time, location, or other emojis used together) instead of using only plain text, for characterizing and recommending emojis. Barbieri et al. [2] employed a multimodal approach using pictures, text, and emojis included in Instagram posts. They built a text-based model (using bi-LSTM), a visual model (using ResNet), and a combination of both for predicting the top 20 emojis relevant to the current input. Their results show that a multimodal model can show better recommendation performance compared to text-only models. Zhao et al. [61] benchmarked an emoji-Twitter dataset to observe the use of emojis on tweet content, tweet structure, and user demographics. They proposed a multitask (i.e., emoji categories and positions) multimodality (i.e., text, image, user demographics) gated recurrent unit (mmGRU) model that predicts the category and location of emoji in a user input text string. They found that users in different geographic regions have different diversity in emoji selection, and users in different regions have different understandings on the emotions that a single emoji imposes. Furthermore, the work showed that emojis in different shapes (e.g., heart shape or face shape) tend to appear at different positions in a

sentence. In another work, Barbieri et al. [4] considered three types of modalities in chat data (i.e., characters, words, and date) and used LSTM with the goal of predicting emojis from 300 candidate emojis. They studied whether the semantics of emojis change over different seasons, comparing emoji embeddings trained on a corpus of different seasons (i.e., spring, summer, autumn, and winter) and showed that some emojis are used differently depending on the time of the year.

2.2 Emoji Suggestion Products

Emoji suggestion has been a popular feature in products as well. Although there are new innovations such as Animoji/Memoji [1] on recent iPhones and ARemoji on Samsung Galaxy S9 that utilize the camera to recognize facial muscle movements to create an animated emoji, a simple and more intuitive way of suggesting (and using) emojis is using the "favorite emoji list," supported by most default mobile OS keyboards. This favorite emoji list typically includes and sorts emojis using the recent usage count. A more advanced technique used in Line [13], Google GBoard [30], TouchPal [52], Word Flow [44], and SwiftKey [31] is mapping specific emojis to a word. When a user types in a predefined word, the app suggests a relevant emoji (e.g., the word *love* maps to "❤️").

Dango [26], Swype [37], and Minuum Keyboard [54] provide a *model-based* emoji suggestion, which captures the current user's text as input and suggests appropriate emojis based on a suggestion model. However, these applications only suggest a limited number (between 5 and 13) of emojis, which restricts the user's freedom of selecting from a wider variety of options. Moreover, as they analyze only the user's current input text, they might not capture various contexts of the conversation and do not recommend emojis for emoji-only sentences.

2.3 Emoji and Social Interactions

In social computing research, user interactions with emoji usage has received great attention. Most such studies highlight the discrepancies and misinterpretations of emoji usage by senders and receivers and present design implications for mitigating such challenges. It is shown that emoji, as a simple "ubiquitous language," is used not only generally but also in a private and secret fashion, highlighting intrinsic characteristics of possessing various meanings and interpretations by individuals [56]. For example, Cha et al. [11] studied stickers, which are similar to emojis but without universal unicodes. The study identified the semantic ambiguity of the stickers on the senders and receivers' standpoints due to the misunderstanding of the context. They also suggested a function of adding an annotation to the stickers to indicate the sender's intention.

A relationship among users' demographic, cultural, and generational factors and different interpretations on emojis has been investigated. Lu et al. [32] analyzed the emoji usage patterns and interpretation differences based on country, gender, and cultural backgrounds. Zhou et al. [62] presented qualitative analysis on emoji usage in communication including nonverbal cues, competitions, compliments, and concise message delivery. In addition, they found that emoji interpretation differs by generations—for example, although most people interpret 🙂 as a smile, many young generations use it to express sarcasm or speechlessness. Li et al. [29] analyzed the relationship between personalities (i.e., openness, consciousness, extraversion, agreeableness, and neuroticism) and emoji usage patterns.

Another group of studies discovered the influence of platforms on emoji usage. Although emoji unicodes are universal, the specific design of emojis and layouts vary noticeably on each platform, which possibly causes different interpretations of emojis [36]. Pohl et al. [41] explored how to optimize emoji keyboard design using emoji similarity (i.e., Jaccard coefficient). In another work, Pohl et al. [42] proposed a new keyboard layout by allowing users to zoom into the full emoji list and select the most preferred area rather than sweeping a list of emojis.

2.4 Summary

Our literature review indicates that many prior studies have looked into ways of predicting emojis based on large-scale datasets and various text mining and machine learning methods. Although such studies offer useful insights on emoji-related research and system development, we identify several limitations, which form our motivation in performing this research.

First, regarding machine learning modeling, most prior studies used a *single sentence* for emoji classification. Datasets crawled from Twitter has been primarily used, and an emoji or the most representative emojis (if multiple emojis are in the tweet) in the given single tweet was used for designing the recommendation model. Surprisingly, less focus has been placed on utilizing multiple consecutively related tweets (or conversation sentences) for emoji prediction. Given that it is common to see conversations (through comments or replies) among Twitter users in the same conversation thread, our expectation is that considering the entire conversation would provide more contextual information for understanding the given conversation and the emojis used; thus, better emoji prediction could be possible.

Second, existing research has primarily focused on presenting model performance and real examples of model prediction using qualitative analysis. Relatively little research has investigated the user experience of interacting with emojis suggested by the proposed models, and the influence of such models on social interactions in a more realistic sense—for example, how do users interact with the suggested or reordered emojis? What other aspects should be considered together with emoji reordering? What are the conflicts that users face between suggested emojis and users' existing practice of using emojis? What human factors should we consider for better developing prediction models and user experience? These are some of the questions we aim to answer in our study.

3 CONVERSATION VS. SENTENCE

One of the primary objectives of our work is to understand the effectiveness of using a *conversation* for learning the context associated with emojis to predict accurate emojis, and in consequence improve user experience in emoji usage and interaction. Before developing a recommendation model, we perform a preliminary study to investigate whether considering conversations would be helpful in selecting emojis from an end user's perspective, compared to using a single sentence to extract a suitable emoji. For this, we design a user survey study on the Amazon Mechanical Turk (MTurk) platform.[1]

We study a user's ability in interpreting information in messages under two different configurations: (i) using the entire conversation and (ii) using only the last sentence of the same conversation (in which the emoji is used). We design such an experiment to see how a target emotion (or emoji group) might change with respect to the context. Using only the final sentence represents current emoji recommendation systems that use only the user's current text input. We are interested in which emoji(s) the users select as a response to the given conversation/sentence, and how the results would be different between the two cases. For the data used for this experiment, we collect conversations from Twitter and use the last sentence of each conversation as the single-sentence statement.

Our hypothesis for this study is that users would be able to better select the emoji(s) that accurately capture the context when provided with the entire conversation rather than a single sentence. This is intuitively valid given that the conversation would provide us with the full contextual "story" behind the text.

[1] https://www.mturk.com/.

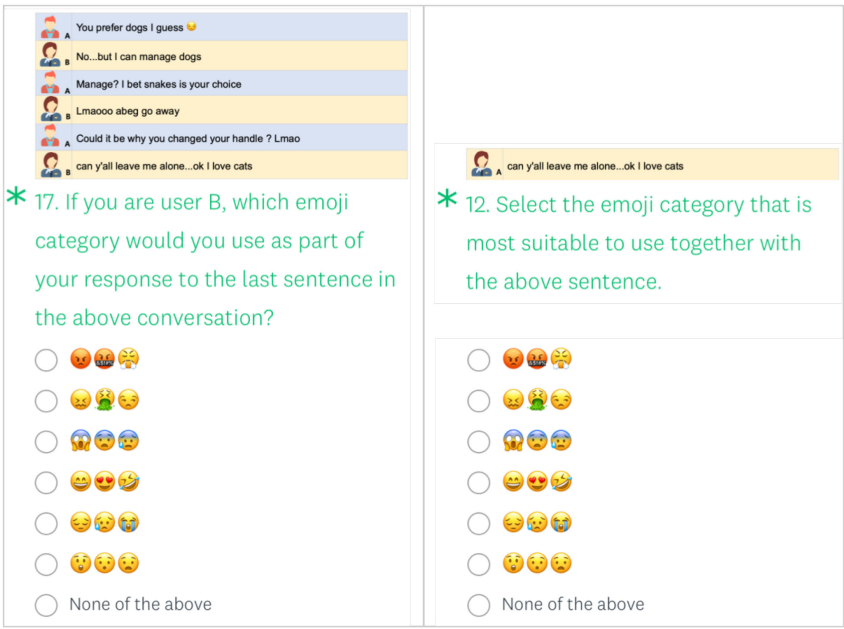


Fig. 2. Survey question examples. On the left side is an example question for the conversation case, whereas the right side shows the single-sentence scenario. Note that the question numbers of two versions are different as the order of questions are randomly presented.

3.1 Emoji Set Categorization

The purpose of this user survey is to understand whether the emotions or contexts captured from a conversation versus from a sentence are different, not the actual selection of specific emojis. We thus use diverse emojis of a similar emotion rather than a specific emoji itself as the ground truth. As the number of unique emojis exceeds 3,000, many emojis have similar meanings (e.g., 😄 and 😊 are similar, and likewise 😡 and 😤 are similar). Hence, we limit the participants' selection option by applying emoji categorization. Specifically, we categorized emojis based on Ekman's six emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) [19, 20]. For each of the six emotion categories, we present three representative emojis that we selected. This selection process involved four researchers independently performing open coding. Conflicts among the coders were resolved based on discussions. In addition to the six categories, we allowed the "none of the above" option, which totals to seven selection options to a question. Figure 2 illustrates an example of the questions (conversation and single-sentence cases) used in the survey.

3.2 Conversation Data

We extract data from Twitter for the survey questions based on the following criteria (we detail the reason for using Twitter data for our experiments in Section 4.1.1):

- We limit the number of sentences in a conversation to a maximum of 10. Although Twitter conversations can be lengthy, we noticed that limiting to 10 sentence conversations covers more than 90% of our crawled dataset. To make sure that the MTurk study participants stay focused on the questions, we limit the conversation to 10 sentences.
- The last sentence of a conversation contains an emoji. This emoji is considered as the "ground-truth" emoji of the conversation.

- At least two users (average 3, maximum 6) are participating in the conversation. We removed many conversations consisting of sentences posted by the same single user in our dataset.
- Conversations with slang, swearing, or indecent words are excluded.[2]
- Conversations with restricted topics (e.g., conversations among fans of a celebrity) are excluded.
- Conversations with image(s) or video(s) are excluded.

Based on these criteria, we provide MTurk participants with three conversations per emotion category. We categorized conversations based on which one of six emotions the ground-truth emoji belongs to.

3.3 Study Procedure

We conduct the MTurk study as a between-subjects study by preparing two independent user groups. Each group was asked to complete either a conversation-based survey (approximately 7 minutes to complete) or a single-sentence-based survey (approximately 4 minutes to complete). A conversation-based survey consists of conversations between two and six users, whereas the single-sentence-based survey consists of only the last sentence of the conversation. Each user was asked to select a proper emoji set that suits the conversation or given sentence. The purpose of these surveys was to evaluate whether the emojis representing the context captured from a conversation (with multiple sentences in the context) and a single sentence differ.

We collected 117 and 113 responses from the conversation survey and the single-sentence survey, respectively. For the analysis, we removed all incomplete responses, ending up with 90 and 93 valid responses for the conversation survey and the single-sentence survey, respectively (a total of 183 responses).

3.4 Results

Table 1 summarizes the statistical results of our MTurk survey. We focus on measuring two aspects from the responses: the number of correct responses (selection of the emoji category that matches the ground-truth emoji in the Twitter data) and the "none of the above" responses.

First, we compare the number of *correct* answers made by each group. We expect that the conversation group would make more correct answers, as they were given more contextual information on the entire conversation, whereas the single-sentence group might struggle due to lack of information. The results match our hypotheses; the conversation group gave significantly more accurate answers than the single-sentence group ($X^2 = 5.12$, $p = 0.02$). On a per-question basis, the conversation group noted significantly more correct answers on four questions (Q1: $X^2 = 2.91$, $p = 0.08$; Q3: $X^2 = 6.03$, $p = 0.01$; Q6: $X^2 = 4.79$, $p = 0.02$; and Q13: $X^2 = 9.10$, $p < 0.001$), whereas the single group scored higher on just one question (Q8: $X^2 = 9.43$, $p < .001$). Figures 3 and 4 show what these questions were.

We also examine the difference in the number of "none of the above" answers made by each group. This is quite pertinent to our expectation in the number of correct answers, because with more contextual information, people would be more clear about what they want to choose; thus, they are less likely to choose the "none of the above" option. As expected, results indicate that the conversation group selected the "none-of-the-above" answer significantly less than the single-sentence group ($X^2 = 11.53$, $p = 0.007$). On a per-question basis, the conversation

[2]Note that slangs are removed for this MTurk study, but a subset of commonly used slang words (e.g., bull****, wtf) were included in our modeling and its evaluation.

Table 1. In Our MTurk Survey, Each Question Was Given as Either a Single Sentence or in the Form of Conversations

	Correct Answers		"None of the Above" Answers	
Question	X^2	*p-Value*	X^2	*p-Value*
1	2.91	0.08	0.01	0.95
2	0.73	0.39	0.69	0.40
3	6.03	0.01	0.42	0.51
4	0.25	0.61	2.43	0.12
5	0.05	0.81	1.30	0.25
6	4.79	0.02	1.73	0.18
7	0.24	0.61	0.23	0.63
8	9.43	0.00	0.60	0.43
9	0.24	0.62	4.27	0.04
10	0.01	0.92	0.01	0.90
11	0.65	0.41	2.33	0.12
12	1.46	0.23	0.12	0.72
13	9.10	0.00	10.11	0.00
14	0.09	0.75	2.86	0.09
15	0.22	0.63	2.68	0.10
16	0.01	0.92	4.11	0.05
17	1.14	0.28	0.42	0.51
18	1.29	0.25	9.46	0.00

We use the chi-square (X^2) test because we have two user groups, and two options for the correct answers (correct or not) and for the none-of-the-above answers (none selection or not). The yellow and orange cells highlight the cases that are statistically meaningful ($p < 0.05$). Orange cells indicate the case where single-sentence-based answers showed significantly higher values, and yellow cells represent the cases where the conversation-based answers showed significantly higher values. First, the correct answers column shows the X^2 value for correctly selecting the ground-truth emoji. Results here show that conversation-based answers yield higher accuracy in more cases. The second column suggests that single sentence-based answers show more of the none of the above responses, implying greater confusion of the participants in selecting emojis.

Fig. 3. Questions 3, 13, and 18 asked in our MTurk study that resulted in significantly different answers from the two groups.

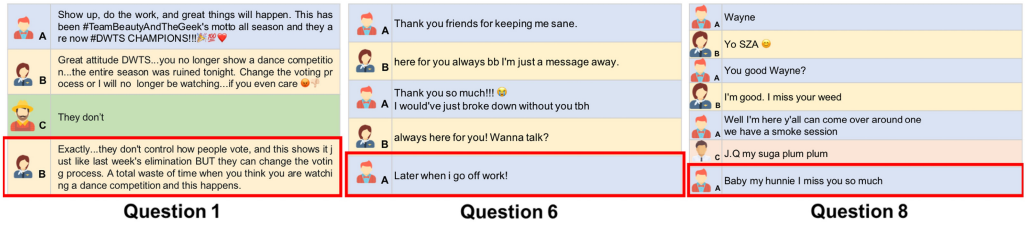


Fig. 4. Questions 1, 6, and 8 asked in our MTurk study with significantly different accuracy between single-sentence groups and conversation-based groups.

group had significantly fewer "none of the above" answers on three questions (Q13: $X^2 = 10.11$, $p < 0.001$; Q16: $X^2 = 4.11$, $p = 0.05$; and Q18: $X^2 = 9.46$, $p < 0.001$), whereas there was only one case when the single group made fewer "none-of-the-above" choices (Q9: $X^2 = 4.27$, $p = 0.04$).

In summary, our results suggest that using the entire conversation to select emoji groups showed more correct answers and less "none of the above" answers. This highlights the importance of taking contextual information (i.e., conversation) into account for developing an emoji prediction model and suggesting a set of emojis to end users for better user experience.

4 EMOJI RECOMMENDATION BASED ON CONTEXT

With the lessons from our preliminary study, we design our emoji recommendation system with the following considerations.

First, our model must consider conversations instead of using a single sentence for emoji prediction. Conversations not only include the text (and emojis) exchanged between the participants but also the identity of the person who typed the sentence (i.e., the speaker) and the sequence of a chat sentence.

Second, our model must capture various contexts of a conversation. There could be many different types of emotions, information, and sentiments expressed through conversations. The selection of emojis could be based on various contexts in an ongoing chat, and the recommendation model should provide the users with emojis that represent various contexts.

Third, our model should provide recommendations for emoji-only sentences without any text input. By analyzing the conversation as a whole, not just the current input text, our model suggests relevant emojis to the users even for emoji-only inputs. Note that in our pilot study where we analyzed users' emoji usage behavior in mobile chat, it was shown that 54% of emoji-used sentences were emoji-only sentences. Therefore, suggesting emojis for emoji-only messages could greatly impact user experiences.

Fourth, recommendations must be made in real time. From a user-experience perspective, it is important that the suggested emojis appear within the duration of the topic conversation. Therefore, the suggestion delay, which includes the delay for querying, machine learning algorithm processing, and display, should be kept minimal. Note that as the processing must be done on a mobile device (e.g., smartphone) that is not as powerful as a desktop or data center servers, the model should be lightweight.

With these design considerations in mind, we propose *Reeboc*, an emoji recommendation system considering the context of mobile chat conversations. Our system consists of (i) an emoji suggestion model that predicts emojis by analyzing a chat conversation using an LSTM model and (ii) a context analyzer that generates emoji recommendation results considering various emoji usage contexts.

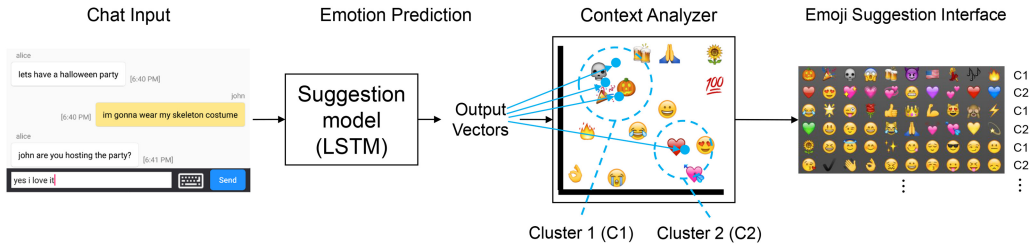


Fig. 5. *Reeboc*'s emoji suggestion process.

4.1 Suggestion Model

We present the details of the suggestion model and implementation of our system.

4.1.1 Conversation Data Collection. To train our model, we need a large amount of text data that use various emojis with the following requirements. First, the data must be a complete conversation with at least two participants exchanging messages. Second, the data must include the information on the time order of the chat sentences and the identity of the conversation participants. Third, the collective data as a whole must include conversations from various people on different topics so that we obtain a balanced dataset.

Although we identified several publicly available chat data [6, 14, 17, 22, 38, 48, 49], none satisfied all of the preceding conditions (e.g., no information on the speaker or time order of the messages). We hence crawled our training and testing data from Twitter [21]. Unlike articles or newspapers, a tweet is very similar to chat data in terms of length (e.g., relatively short texts with character limits); grammar (e.g., grammatically erroneous sentences and ill-formed text like abbreviations, acronyms, and slang); and, importantly, wide usage of emojis. To collect conversation data, we specifically targeted conversations through the reply function on Twitter that allows replying to a specific tweet to communicate between users or expressing opinions on a tweet. We only collected tweet threads that keep tweet conversation using the reply function. Our tweet reply thread data collection satisfies all of the preceding three conditions. We collected a total of 6.3 million English tweets from 1.5 million conversations by 1.8 million unique users from July to October 2018. The average character count in our crawled tweets was 84.1.

Our crawled Twitter data can be generally categorized into two types based on discussion topics. The first category is personal conversation between two and three users that are similar to mobile chat, using the reply function. The topics of such conversations include making arrangements for dinner gathering or other personal matters. The other category is discussion among multiple users (typically more than two), also using the reply function. The discussion topics in this category include news events, movies, and games, among others.

4.1.2 Data Preprocessing. Using the crawled Twitter data, we perform preprocessing to remove noise from the data. Specifically, we select and use the top 111 emojis based on the frequency of use. This covers more than 90% of the total emoji usage from the 6.3 million crawled tweets. In addition, from the list of words from crawled Twitter data, we use only the most frequently used words that cover more than 90% of the total word usage. We also ignored tweets that have two or more different emojis in a single tweet, as they potentially confuse and complicate our learning model. We limited the maximum number of sentences per conversation to 10 due to similar reasons from our preliminary study. Except for the emoji used in the last sentence of a conversation, which we use as the ground truth for the conversation, we removed the rest of the emojis from the conversation. We removed such emojis because for them to be used as inputs to modeling, they

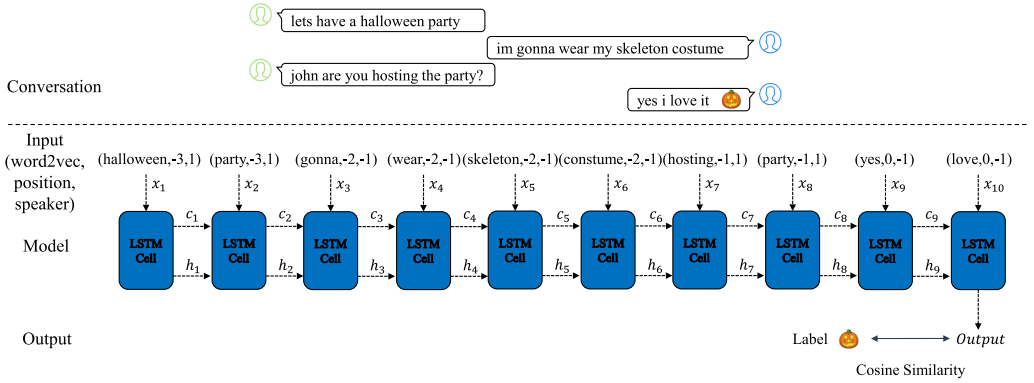


Fig. 6. System training pipeline.

need to be converted to a vector on the same vector space that was converted from words (e.g., using word2vec), but currently there is not such a scheme.

We additionally removed stop words such as "I, a, are, an, and my" from the Natural Language Toolkit (NLTK) stop word list [51]. Stop words removal is conventional in NLP because they do not add information and removing them saves time and space [7, 23, 45, 47, 60]. We only considered tweet threads with two or more participants. Finally, we ignored duplicate conversation caused by retweets and tweets collected by the same conversation. Furthermore, we removed noises from the tweets such as links, tags, hash tags, and retweet (RT) signs. With the preceding preprocessing, we reduced the crawled 1.5 million conversations down to nearly 519,000. The validation and test set are assigned based on the ratio of the number of data in each 111 emoji categories and are assigned up to 1,000 conversations per emoji category.

We have upsampled and downsampled the number of training data for all emoji categories for balancing. The "joy" (😂) emoji, the most dominant in our data (21% of total emojis used), was downsampled in half, and the remaining emoji training samples were upsampled but did not exceed the number of downsampled "joy" emojis.

4.1.3 LSTM Model for Emoji Suggestion. We use an LSTM network [25] for our emoji suggestion model. LSTM is evolved from a recurrent neural network (RNN) architecture that has been widely applied in various NLP-related issues. Unlike traditional RNNs, LSTM has a distinctive property in that it can learn from a long series of inputs; previous inputs inform the current input to consider long-term dependencies and find the best relation between the previous and current inputs. Leveraging LSTM to train our emoji selection model allows our system to consider a sequence of words across multiple sentences in a conversation for emoji suggestions.

4.1.4 Implementation. Prior to using the preprocessed tweets as LSTM training data, we convert each word and emoji to a vector form using word-to-vector [33]. In our system, we use the pretrained Natural Language Processing Group (TALN) word2vec model [5], which has mappings from a word (word-to-vector) and an emoji (emoji-to-vector) to a 300-dimensional vector, respectively. Through the word-to-vector and emoji-to-vector conversions, we represent words and emojis as 300-dimensional vectors that are kept constant throughout the training process.

The training process of the model is shown in Figure 6. In addition to the 300-dimensional word vectors, we add two more dimensions to utilize the context of the conversation: (i) the speaker information (whether the speaker of the sentence is the current user) and (ii) the relative position of the sentence with respect to the last sentence of the conversation (e.g., 0 if it is the last sentence,

and −1 if it is the second to the last sentence). Each word in the sentence is then converted into a 302-dimensional vector and is fed into the input of the LSTM model for training. The LSTM model outputs a 300-dimensional vector given the series of 302-dimensional vectors as input. In training, we maximize an objective function of the cosine similarity between the 300-dimensional output and the 300-dimensional ground-truth emoji vector. The LSTM model is trained using the Adam optimizer [28] with a learning rate of 0.001 and a mini-batch size of 256. We implemented the LSTM model using the PyTorch [39] framework.

4.2 Context Analyzer

Taking various output vectors from the LSTM model as an input, the context analyzer generates emoji recommendation results by extracting contexts from the current conversation. The context analyzer first clusters the output vectors to identify the contexts that are clearly distinct. It then makes emoji recommendation sets by merging the results of emoji vectors representing the recent conversation through clustering. This process is done in real time.

4.2.1 Selecting Input Sentences. Mobile chats could have multiple, abruptly altering topics and thus contain multiple context switches in a series of sentences. To suggest appropriate emojis according to the chat context, an emoji suggestion system should take a number of previous messages (i.e., a conversation) into account. A naive solution could be putting all of the previous sentences to the model. However, putting sentences that are not related to the current input text might result in an unwanted output. Therefore, a desired emoji suggestion system should consider a "proper" number of sentences.

To resolve such an issue and support diverse emoji suggestions, ideally the number of messages to consider should be decided dynamically while analyzing the context of the current conversation. However, this would require complex computation and latency. We thus opt for a simple mechanism where we feed the last five messages, such as $\{s_{t-4}, s_{t-3}, s_{t-2}, s_{t-1}, s_t\}$ (s_t being the current sentence), and then provide five inputs, such as $\{s_t\}, \{s_{t-1}, s_t\}, \{s_{t-2}, s_{t-1}, s_t\}, \{s_{t-3}, s_{t-2}, s_{t-1}, s_t\}$, $\{s_{t-4}, s_{t-3}, s_{t-2}, s_{t-1}, s_t\}$, into the LSTM model. We empirically decided on five samples, as it helps us capture multiple contexts in a conversation in low latency. The LSTM model then generates five different output vectors, and each would provide a meaningful emotion, possibly different from each other. Such emotions and/or information will be represented as different clusters, as we detail in the following.

4.2.2 Context Clustering. We cluster the output vectors of similar contexts through k-means clustering. The distance between the vectors is calculated as their cosine similarity. Our system finds that the output vectors of all clusters have the minimum number of clusters with a cosine similarity of at least 0.9 from the centroid while controlling the k value of k-means clustering from 1 to the number of output emoji vectors. A centroid of the cluster is the mean value of the unit vector of each cluster and represents similar emoji usage contexts of each recent conversation.

The threshold value impacts the number of clusters and thus the diversity of emoji recommendations. The higher the threshold, the more the number of clusters, with each cluster representing different contexts or emotions. However, if the threshold value is too high, similar types of emojis could be separated into different clusters. On the contrary, if the threshold value is too low, different types of emojis would be clustered together. We selected the threshold of 0.9 from our empirical analysis of Twitter data, as it gave the best performance.

Let us use the chat input in Figure 5 as an example. After the LSTM model provides four output vectors, the three vectors near the Halloween emoji vector are grouped into one cluster, and the vector near the love emoji vector is grouped into one cluster.

Note that we use clustering to put similar output vectors from LSTM into similar contexts. Although different algorithms could be used for clustering, we believe that *k*-means clustering is an effective and efficient vector quantization method for our purpose. As our scheme requires real-time processing, complex approaches such as density-based or distribution-based clustering algorithms that consume longer processing latency would not fit our system.

4.2.3 Emoji Suggestion Interface. *Reeboc* provides recommended emojis based on the context clusters. Each cluster has its emoji list that is sorted by the cosine similarity between the centroid and emoji vectors in descending order. With these emoji lists from the clusters, *Reeboc* application (the right-most image in Figure 5) displays the suggested emojis when a user clicks the emoji button. Each row of the "emoji pad" represents the sorted emoji list in each cluster (i.e., each detected context, emotion, or information). The order of the row is determined by the number of output vectors in each cluster given that having more outputs for a cluster means that the emotions from the conversation are more focused on that specific cluster. For example, if there are two clusters and one has three vectors (C1 in Figure 5) while the other has one (C2 in Figure 5), C1 is presented in the first row and C2 in the second row on the emoji pad. For the remaining rows, the remaining emojis in each clusters' emoji list are presented alternately at each row (i.e., remaining emojis of C1 in the third row, and those of C2 in the fourth row, and so on) until none of the 111 emojis remain. When there is a redundant emoji that is already presented in a previous row (therefore associated with other clusters), we skip additional presentation of the emoji to avoid duplicates.

5 USER STUDY

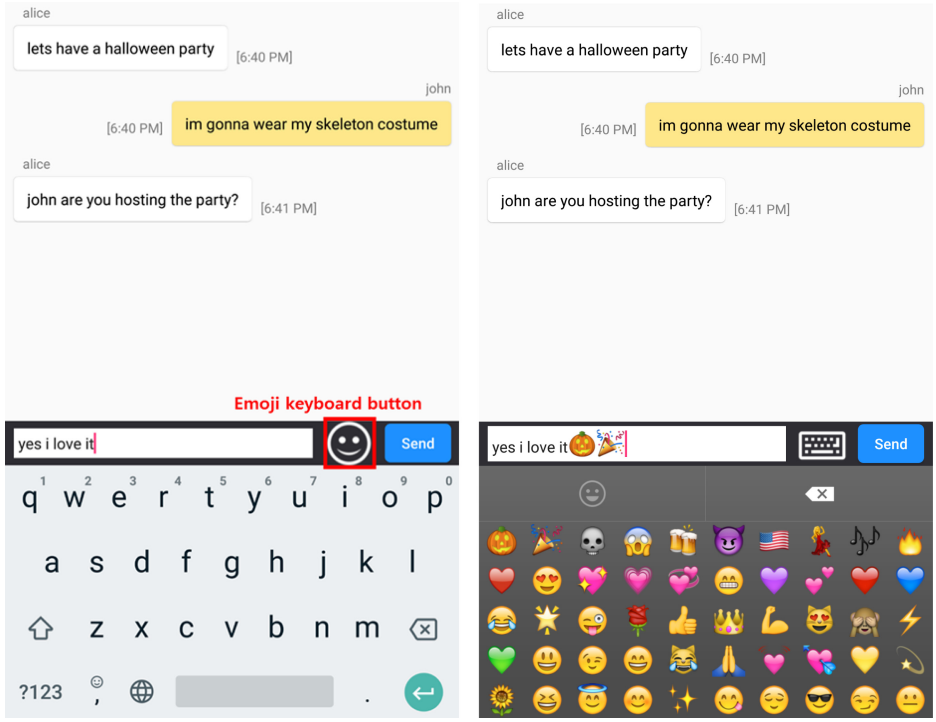
We conduct a user study to investigate how *Reeboc*, our emoji recommendation system that considers a chat conversation instead of only current chat input, affects the user's emoji selection and selection latency.

5.1 Study Design

To see the effect of using our emoji recommendation system in an actual mobile chat environment of users, we performed the following. We (i) built a prototype chat application with an emoji recommendation system for actual mobile chat environments, (ii) recruited users as a group to create a natural chat environment, and (iii) created an environment that compares among emoji modes that are the baseline (i.e., no recommendation), the current-sentence model, and the conversation model. A detailed description of our study procedure follows.

5.1.1 Prototype Application. We implemented a prototype chat application with the suggestion system for analyzing the effect of emoji suggestions on real users. Figure 7 is the prototype application that includes common mobile chat features such as chat rooms, user nicknames, text-based chat, and emoji support. We also implemented logging for all user conversations, emoji usage, and emoji selection latency to analyze the emoji usage patterns of users. The suggestion system runs on the server, and the prototype application passes the conversation to the server to get the emoji recommendation results. In our user study, emojis were recommended within 300 ms after users click the emoji keyboard button, and the average delay was 112 ms (the average network delay was 79 ms within the total delay).

We made three emoji modes for the experiment. The first is the baseline mode where the emoji layout is always the same as the usual emoji keyboard layout, without any emoji recommendation [53]. The second is the current-sentence mode that makes emoji recommendations based only on the current user input text sentence. Finally, the conversation mode (i.e., *Reeboc*) is the emoji recommendation considering the whole conversation, not just the current sentence. The second



(a) Text keyboard layout. (b) Emoji keyboard layout.

Fig. 7. The prototype application designed for the study. By default, the text keyboard layout is presented to the user (a). If the user wants to use an emoji in a conversation, the user taps on the emoji keyboard button to open the emoji keyboard layout (b). Different experimental conditions present emojis in different orders.

and third modes display recommended emojis in order of relevance (from top left to bottom right) on the emoji keyboard.

5.1.2 Participants. We recruited 17 participants by posting advertisements through campus online communities at KAIST. We used questionnaires to select participants with experiences in using emojis in English mobile chats. In the end, 11 male and 6 female students took part in the study (ages 20 to 29 years). To ensure that we create a natural mobile app chatting environment, the recruitment was done on a group basis, consisting of seven groups of two members and a group of three. The self-reported relationships among the groups were five groups of friends, one couple, and one group of coworkers. Thirteen people were native English speakers or had between 4 and 20 years of residence in English-speaking countries. The other four reported familiarity with chatting in English. Participants received $15 for 1 hour of participation.

5.1.3 Tasks and Conditions. To ensure that participants have a natural chat experience, we designed three tasks based on common topics in mobile chat scenarios: planning for a dinner get-together, planning for a weekend group activity, and sharing recent memorable experiences.

Using a within-subjects design, we instructed participants to chat about the three topics, each for 10 minutes with the three emoji modes. The task order was fixed, whereas the order of the emoji modes was randomly assigned for each task. This means that everyone in the group had the same order of emoji modes in the study. This was to avoid potential confounds from having participants in the same group use different emoji modes.

Participants used the prototype chat application that we developed for the experiment. All participants were given a mobile phone to use during the study, with the chat application installed. Three types of Android phones were used in the study: Nexus 6P, Essential Phone, and Google Pixel 2 XL, which have similar screen sizes of 5.7 to 6.0 inches.

5.1.4 Procedures. We carried out the following steps in our experiment. We first explained to participants how to use the app. After a brief tutorial, participants spent 10 minutes to become familiar with the device and keyboard. We asked them to casually chat with their group members.

Next, participants were instructed to perform three 10-minute chatting tasks within their groups. We assured them that their chat history will be anonymized and would only be used for research purposes. Each group spent 10 minutes on each of the three tasks, each with different emoji modes. Note that the participants were placed in different rooms during the mobile chat so that emotions or additional information could not be delivered using facial/verbal expressions. In addition, to prevent participants' expectation of different recommendation modes from affecting the emoji usage, we did not explain the exact difference between the three emoji modes. They were told this information after completing all three chat sessions.

Finally, we performed a semistructured one-on-one interview with each member of the group for about 20 minutes. Participants then answered a post-questionnaire about whether they noticed any differences between the three modes, which recommendation modes they preferred, and whether they were satisfied with the recommendations.

5.1.5 Measures. To measure the effectiveness of emoji recommendation and understand how our system affects participants' chatting behavior and emoji selection time, we kept track of the actual chat content along with the following chat interactions:

- Emojis recommended by the system and the order they were presented,
- Emojis participants selected and where they were presented on the keyboard,
- Emoji selection time, defined as the time between the user taps the emoji keyboard button (Figure 7 (a)) and selects the (first) emoji, and
- The type of a message (whether it is an emoji-only sentence).

If the desired emoji is recommended at the top of the emoji keyboard near the focal point, participants are likely to select the emoji in a short period of time. Measuring emoji selection time serves as a proxy for the recommendation quality. In measuring emoji selection time, we only consider the first emoji selection time. In some cases, participants might use multiple emojis consecutively, but from the second time and onward, the selection often takes much shorter because they often repeat the first emoji or choose randomly. Therefore, we only focus on the first emoji selection that directly reflects the emoji recommendation quality.

5.2 Results

Our goal of the user study was to answer the following questions through system logging and post-interviews: (1) How effective is our emoji suggestion? (2) How quickly can participants choose emojis? These questions are influenced by emoji usage behavior of the participants. We first analyzed the emoji usage propensity of participants.

In the user study, the participants used an average of 67.8 sentences per 10-minute chat task. The average ratio of sentences using emojis among all sentences was 34.1%. The mean number of characters per sentence was 16.59 (standard deviation = 12.26).

5.2.1 Effectiveness of Emoji Prediction. To evaluate the effectiveness of emoji recommendations, we analyze which emojis the participants selected during their chat. We hypothesize that with

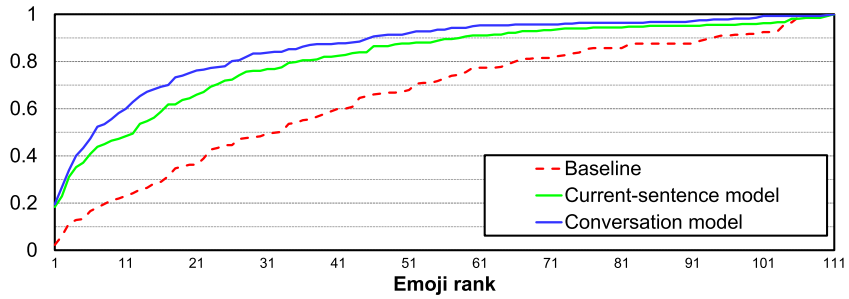


Fig. 8. CDF of used emoji "ranks." The *x*-axis represents the presentation order of the 111 emojis on the keyboard.

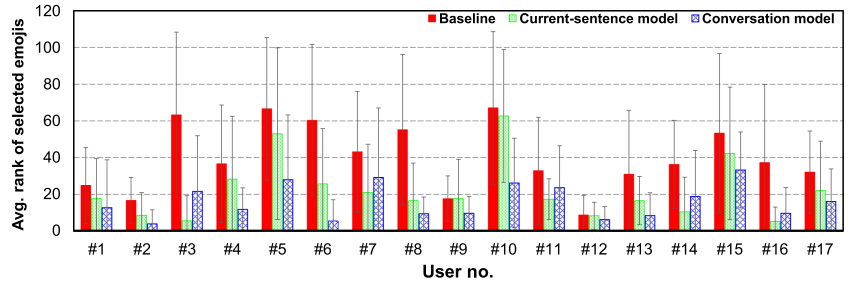


Fig. 9. Average emoji selection "ranks" for each participant in the user study. The error bars represent standard deviations.

better emoji recommendations, participants are likely to choose emojis that are presented on the top left of the emoji keyboard (as opposed to the bottom right). To measure this effect, we use the *emoji selection rank*, which is the order in which the chosen emoji is presented on the emoji keyboard. For example, an emoji with rank = 10 means that it is the 10th presented emoji in the recommendation conditions (or the 10th emoji on the baseline emoji keyboard). The closer the selection rank is to 1, the higher the emoji recommendation. We compared the emoji selection rank of the three emoji modes used in the user study. Figure 8 plots the cumulative distribution function (CDF) of emoji selection rank comparisons for the baseline, the current-sentence mode, and the conversation mode (i.e., *Reeboc*). The *x*-axis represents the rank of suggested emojis. The median rank (50% of CDF) is 32 for the baseline (i.e., 50% of emoji selections were made for emojis ranked from 1 to 32 (out of 111)), 13 for the current-sentence mode, and 7 for the conversation mode.

Overall, the two recommendation modes showed a considerably better selection rank than the baseline. In Figure 9, all participants showed improved average selection rank using the two emoji recommendation modes when compared to the baseline. Comparing among the recommendation modes, the median rank of the CDF is improved from 13 to 7 in the conversation mode over the current-sentence mode (see Figure 8). The average emoji selection rank for 12 of 17 participants was better with the conversation mode than the current-sentence mode (see Figure 9). This suggests that the conversation mode presented more effective emoji recommendations to participants than the baseline and the current-sentence mode.

Participants' comments in the post-study interview echoed the emoji selection rank differences among the three modes. Of the 17 participants, 15 said the two recommendation-based modes were much better than the baseline for emoji selection. P7 said, "Appropriate emojis were recommended,

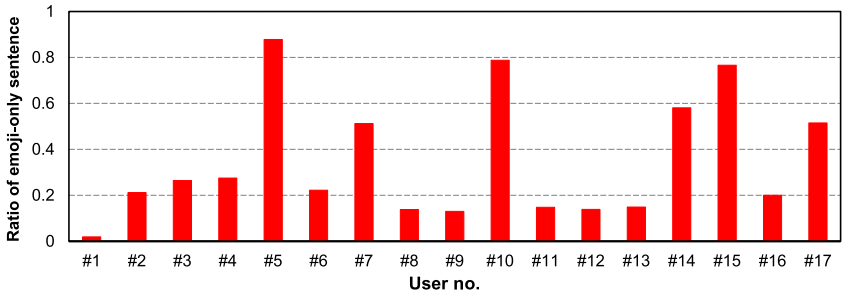


Fig. 10. The ratio of emoji-only sentence among chat sentences containing emojis for each user.

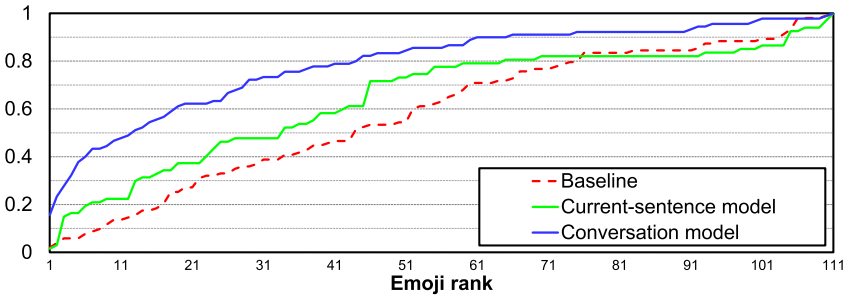


Fig. 11. CDF that displays the ranks of selected emojis, specifically for emoji-only sentences.

so I could choose my emoji from the top results and reduce the time to find the emoji I wanted to use."

Some expressed inconvenience with emoji recommendations because the constantly changing order of emojis display was confusing. These participants noted that they know by heart the static default layout of the baseline and can quickly select the desired emoji. P10 responded, "I use only certain emojis. My muscle memory remembers the locations of the emoji I use the most, so I can choose quickly. Emoji use was more inconvenient in the recommendation modes because there were no same emoji in the remembered position."

5.2.2 Emoji-Only Messages. In our user study, 36.2% of the sentences containing emojis were emoji-only sentences. The likelihood of using emoji-only sentences varied widely between participants, with a mean of 0.35 (standard deviation = 0.27) for the participants, as shown in Figure 10. The participant who had the highest ratio of emoji-only sentences among all emoji-containing sentences was 88%.

The conversation mode showed clear benefits when participants used only emojis without typing any text (i.e., emoji-only sentences). The model considering only the current sentence cannot recommend emojis for emoji-only messages, as there is no text to seed the recommendation engine. Our results suggest that emoji-only sentences were used frequently and participants reported that they primarily use emoji-only sentences to quickly react to a conversation partner's previous sentence or to immediately append emotion to their previous sentence. These motivations of using emoji-only sentences bode well for *Reeboc*, as it uses conversations (i.e., a few previous sentences) to recommend emojis. Figure 11 plots the CDF of emoji selection rank comparisons of the baseline, current-sentence mode, and conversation mode only for emoji-only sentences. The median rank (50% of CDF) was 44 for the baseline, 34 for the current-sentence mode, and 13 for the conversation mode. Intuitively, the baseline and current-sentence modes should show similar results, as

Table 2. Average Emoji Selection Latency

	Baseline	Current-Sentence Model	Conversation Model
All Emoji-Used Sentences	4,601 ms	3,262 ms	2,723 ms

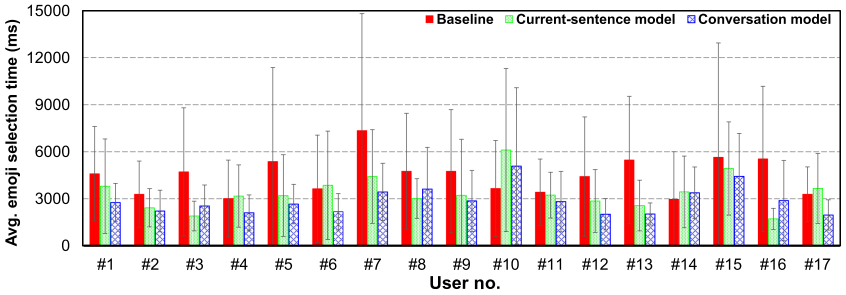


Fig. 12. The average emoji selection time for each participant in the user study. The error bars represent standard deviations.

they literally show the same layout in such cases. Nevertheless, given that the participants have built some level of trust on the suggestion of the current-sentence mode compared to the baseline (from previous chats), they tend to select the lower-rank emojis in this case.

Compared to Figure 8, the median emoji selection rank difference between the conversation mode and the current-sentence mode is larger. The difference for all emoji use was 6 median rank difference, whereas for emoji-only sentences, it was 21 median rank difference; Figure 11 shows the median rank (50% of CDF) of 34 for the current-sentence mode and 13 of the conversation mode.

This result suggests that the conversation mode presents stronger recommendations for an emoji-only sentence, as it suggests emojis based on analyzing previous sentences. P10 mentioned that the emoji-only sentence recommendation was good for reacting to the emoji used by the conversation partner: "I respond with emojis when others use emojis. In this case, I want to use a similar emoji to what others are using, but not the same emoji. I like the emoji suggestion result."

To summarize, most participants effectively selected emojis in the emoji recommendation modes over the baseline emoji layout. The conversation mode (i.e., *Reeboc*) generally showed better emoji recommendations than the current-sentence mode. Figure 8 shows the median rank (50% of CDF) of 13 for the current-sentence mode and 7 for the conversation mode. Especially for the emoji-only messages, the conversation mode showed a much better recommendation, with the median emoji selection rank being 34 for the sentence mode and 13 for the conversation mode.

5.2.3 Emoji Selection Latency. With emoji recommendations, participants selected emojis faster than when using the baseline. Table 2 presents the average time taken to select an emoji for each mode. The emoji selection latency in the conversation mode was 2.73 seconds with 1.83 seconds of standard deviation, whereas the baseline was 4.6 seconds with 4.4 seconds of standard deviation and the current-sentence mode was 3.26 seconds with 2.64 seconds of standard deviation. This suggests that the conversation model improves the latency compared to the baseline by 38% and 14% compared to the current-sentence mode.

Figure 12 shows the average emoji selection delay for the three emoji modes for each participant. The conversation model showed the best performance for 13 of 17 participants. Among the four participants whose best latency mode was not the conversation model, only one participant (P10)

had the best latency performance with the baseline layout, and the other three showed the fastest selection performance with the current-sentence mode.

Most participants selected emojis faster using a recommendation system compared to the baseline. Moreover, 14 of 17 participants selected emojis faster in the conversation mode than in the current-sentence mode. P1 said: "I think emoji recommendation of the conversation mode was better. Various and more appropriate emoji were recommended. I trusted the recommendation results and used them comfortably to pick any of the top recommendations." For P10, emoji selection time in the baseline was much shorter than the other modes, as explained in Section 5.2.1. She said that she remembers the location of commonly used emojis and that changes in the emoji order made emoji selection difficult.

5.2.4 Participants' Perception of Emoji Suggestions. Through the post-study interview, we asked the participants about their experience in chatting with emoji recommendations and whether the mode switch was recognizable.

Q1. Did participants recognize the difference between the three emoji modes? We did not provide specific details about the changes among the three sessions in the study. However, most participants naturally noted the differences among the conditions during the study. All participants responded that they easily noticed the difference between the baseline and the two emoji recommendation modes. The difference between the current-sentence mode and the conversation mode was recognized by 13 of 17 participants. Six of the 13 identified the recommendation difference in the emoji-only sentences, whereas 2 participants recognized the difference only for emoji+text sentences. The remaining 5 participants recognized the differences in both emoji+text sentences and emoji-only sentences.

Q2. Which of the three emoji modes did participants prefer? With the exception of one aware of the differences between the sentence mode and the conversation mode, participants favored the conversation mode's emoji recommendation. For those 6 participants who recognized the difference in emoji-only sentences, they all expressed that the recommendation for emoji-only sentences was better in the conversation mode. When using emoji+text, participants were even more satisfied with the recommendation results of the conversation mode of *Reeboc*. P5 responded, "The conversation mode recommends various types of emojis so it's easy to choose the emoji I want."

However, two participants found the conversation mode to be more inconvenient to use than the current-sentence mode. One such case was when using an emoji for a particular word. P16 specifically said, "When I typed 'fire,' I hoped that the fire emoji would be recommended at the top, but sometimes other types of emojis were recommended first. In the current-sentence model, it was easier to find word-emoji matches with words such as lol, soccer, and fire." This suggests that when users have a specific intent (e.g., replace a specific word with an emoji), conversational context might actually degrade the recommendation quality. However, in most cases where participants wanted to express emotion and react to previous sentences, conversational context improves the recommendation quality.

6 DISCUSSION

Based on our experiences in designing a conversation-based emoji recommendation model for mobile chat applications, we now outline several discussion points and share the current limitations of our design.

6.1 Tweets vs. Real Chat Data

Our work, and many other previous works focusing on emoji suggestion models for chat scenarios, utilizes Twitter-crawled data for model training [3, 55, 57]. The main reason behind using Twitter

data was because real personal chat data is difficult to gather at a large scale. Studies that use chat samples for model design usually focus on small-size datasets that are gathered locally. It is difficult to capture general observations with such limited (and potentially biased) data. However, Twitter data is relatively easy to crawl, and they share many similarities with chat data; the messages are short, they include conversational information, and emojis are commonly used [12, 35]. In our crawled data (1.5 million conversations), more than 0.5 million conversations included emojis. As a result, many studies, including ours, design the model trained on tweets and evaluate the model with user studies in a real chat setting.

Although tweets and real chat data are similar, they also have some noticeable differences. Based on our user study experiences, we realized that our chat study participants tend to deliver many personal and implicit messages in their chat sessions compared to tweets. Specifically, because real chat data is generated within a conversation between designated or limited users in a virtual space (i.e., chat room), there is a high level of personalization and implicit meanings abstracted in the use of emojis [62]. However, to the best of our knowledge, there is no previous work that utilizes large-scale real-world chat data for emoji suggestion model design, which we believe would be valuable future work in this domain. Although outside the scope of this article, we argue that, similar to the Crawdad Project (http://crawdad.org/) for collecting wireless network-related data for research purposes and the Stanford Network Analysis Project (http://snap.stanford.edu/) for network analysis and graph mining, a similar large-scale database where researchers and users can voluntarily share and utilize chat data can be a solution to such "training versus usage" discrepancy as a research community.

6.2 Diversifying Emoji Usage

In our user study, we noticed that approximately half of the participants used a more diverse set of emojis in their conversations with our context-aware emoji recommendation model. One participant (P16) noted that he used the laughing with tears emoji (😂) in 80% of cases when using emojis. In the post-study interview, he pointed out that the main reason for doing so was because he wanted to add an emoji to smoothen the conversation, but did not want to scroll through to select a proper emoji for the sentence he wrote. In such cases, the ability to use a more diverse set of emojis, without spending too much time on picking one, can enable participants to express more nuanced and accurate emotions.

6.3 Privacy

One participant (P1) in our user study pointed out that she would be worried about privacy issues when knowing that the chat application (or the underlying software) can analyze the context of a conversation. We see this as a valid concern from the users' perspective. Although solutions to resolve privacy-related concerns are outside the scope of this work, we believe that this is an important aspect to consider. Compared to many existing chatting applications that—although they also could read users' messages—simply deliver them, our emoji recommendation system gave the impression that it was analyzing her conversations and knew her intents accurately. From the chat UI design perspective, this brings an important question about privacy and transparency. We believe that chat applications should clearly communicate to the user what they do with conversation content and further give the user control of what gets processed by the system.

One possible design approach we can take is to allow all data processing to occur on the smartphone itself rather than sending the data to a server for emoji recommendation. Such a design choice can limit the scope of data sharing to the user's smartphone, thus reducing the risk of privacy leaks by an external server entity. Nevertheless, to do so, the computational complexity of the

emoji recommendation module needs to be optimized to fit on resource-limited mobile platforms while minimizing the impact of recommendation accuracy.

6.4 Integrating Additional Data Modalities

Our current design of *Reeboc* focuses on suggesting emojis based on the conversation as a whole and not only the input sentence of a user. We see this itself as an important step, but as in many previous works [8, 43, 58], other approaches, such as adding in data modalities, can provide additional hints to emoji suggestion as well. Having additional data such as location, time of day, or seasons along with temporal and historical context can be factors that allow for a more accurate emoji recommendation. However, on a practical perspective, the collection of such data can be difficult without proper support from chat applications or even the mobile OS itself. Such additional data would need to be carefully aligned with chat context, and the collection of such additional data might inflate the privacy issues mentioned earlier.

6.5 Limitations

Our system and evaluation have several limitations. Our study participants were limited in size (17) and were mostly university students in their 20s, and different demographic groups might exhibit different chat behaviors and reactions to emoji recommendations. In addition, most of our groups were pairs, so we cannot generalize the findings to larger group settings in which more than two people participate in a single conversation, such as group chat rooms. Understanding a conversation context in a larger group can be more challenging, as more topics might be discussed simultaneously and more users might bring in different perspectives. Although we leave addressing these limitations as future work, we believe that this work still presents a meaningful step toward this promising line of research by showing the feasibility and effectiveness of context-aware emoji recommendations.

7 CONCLUSION

Although context awareness is one of the most studied subjects in social computing, human computer interaction, and mobile computing, surprisingly little research has focused on considering context in emoji recommendations. We present *Reeboc*, which extracts various emotions and topics of a conversation and uses them to suggest emojis that represent diverse contexts. With our method, we also recommend emojis for emoji-only messages, which was impossible in existing schemes. Findings from our user study are quite encouraging. By analyzing the conversation instead of only the current input text, *Reeboc* provided users with a better experience in emoji selection and social communication. Here we we focused on emoji recommendations for mobile messaging; however, we believe that our technique could also be applied in other domains, such as emails and social networks. Although user concerns such as privacy should be seriously considered, we plan to further improve our context understanding model by considering additional modalities such as the relationships between chat partners, physical contexts (e.g., location), and cultural and historical contexts.

REFERENCES

[1] Apple. 2017. Animoji. Retrieved November 19, 2019 from https://www.apple.com/newsroom/2017/09/the-future-is-here-iphone-x/.

[2] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 679–686. DOI: https://doi.org/10.18653/v1/N18-2107

[3] Francesco Barbieri, Luís Marujo, Pradeep Karuturi, and William Brendel. 2018. Multi-task emoji learning. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media*. http://hdl.handle.net/10230/35646.
[4] Francesco Barbieri, Luis Marujo, Pradeep Karuturi, William Brendel, and Horacio Saggion. 2018. Exploring emoji usage and prediction through a temporal variation lens. arXiv:1805.00731.
[5] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this emoji mean? A vector space skip-gram model for Twitter emojis. In *Proceedings of the Language Resources and Evaluation Conference (LREC'16)*.
[6] Dave Beckett. 2004. Public IRC Chat Data. Retrieved November 19, 2019 from http://chatlogs.planetrdf.com/swig/.
[7] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. arXiv:1508.05326.
[8] Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Thomas Schievenin. 2014. Context-aware points of interest suggestion with dynamic weather data management. In *Information and Communication Technologies in Tourism*, Zheng Xiang and Iis Tussyadiah (Eds.). Springer International, Cham, Switzerland, 87–100.
[9] Jeremy Burge. 2017. 5 Billion Emojis Sent Daily on Messenger. Retrieved November 19, 2019 from https://blog.emojipedia.org/5-billion-emojis-sent-daily-on-messenger/.
[10] H. Cappelen and J. Dever. 2016. *Context and Communication*. OUP Oxford.
[11] Yoonjeong Cha, Jongwon Kim, Sangkeun Park, Mun Yong Yi, and Uichin Lee. 2018. Complex and ambiguous: Understanding sticker misinterpretations in instant messaging. *Proceedings of the ACM on Human-Computer Interactions* 2 (2018), Article 30, 22 pages. DOI: https://doi.org/10.1145/3274299
[12] H. H. Clark. 1996. *Using Language*. Cambridge University Press.
[13] LINE Corporation. 2011. LINE: Free Calls & Messages. Retrieved November 19, 2019 from https://line.me/en/.
[14] Cristian Danescu-Niculescu-Mizil. 2011. Cornell Movie–Dialogs Corpus. Retrieved November 19, 2019 from https://www.cs.cornell.edu/ cristian/Cornell_Movie-Dialogs_Corpus.html.
[15] Oxford Dictionaries. 2015. Oxford Dictionaries "Word" of the Year 2015. Retrieved November 19, 2019 from https://languages.oup.com/press/news/2019/7/5/WOTY.
[16] Thomas Dimson. 2015. Emojineering Part 1: Machine Learning for Emoji Trends. Retrieved November 19, 2019 from https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad.
[17] Eibriel. 2016. rDany Chat: Messages with a Virtual Companion. Retrieved November 19, 2019 from https://www.kaggle.com/eibriel/rdany-conversations/.
[18] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of the International Workshop on Natural Language Processing for Social Media*. 48–54. DOI: https://doi.org/10.18653/v1/W16-6208
[19] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3–4 (1992), 169–200. DOI: https://doi.org/10.1080/02699939208411068
[20] Paul Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power (Eds.). John Wiley & Sons, West Sussex, England, 45–60.
[21] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv:1708.00524.
[22] freeCodeCamp. 2017. freeCode Camp Gitter Chat, 2015-2017. Retrieved November 19, 2019 from https://www.kaggle.com/freecodecamp/all-posts-public-main-chatroom.
[23] K. V. Ghag and K. Shah. 2015. Comparative analysis of effect of stopwords removal on sentiment classification. In *Proceedings of the International Conference on Computer, Communication, and Control (IC4'15)*. 1–6. DOI: https://doi.org/10.1109/IC4.2015.7375527
[24] Gaël Guibon, Magalie Ochs, and Patrice Bellot. 2018. Emoji recommendation in private instant messages. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18)*. ACM, New York, NY, 1821–1823. DOI: https://doi.org/10.1145/3167132.3167430
[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735
[26] Whirlscape Inc. 2016. Dango: Your Emoji Assistant. Retrieved November 19, 2019 from https://getdango.com/.
[27] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv:1607.01759.
[28] D. Kinga and J. Ba Adam. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
[29] Weijian Li, Yuxiao Chen, Tianran Hu, and Jiebo Luo. 2018. Mining the relationship between emoji usage patterns and personality. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM'18)*.
[30] Google LLC. 2018. Gboard—The Google Keyboard. Retrieved November 19, 2019 from https://apps.apple.com/us/app/gboard-the-google-keyboard/id1091700242.
[31] TouchType Ltd. 2010. SwiftKey. Retrieved November 19, 2019 from https://swiftkey.com/en.

[32] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, New York, NY, 770–780. DOI: https://doi.org/10.1145/2971648.2971724

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

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. Curran Associates, 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[35] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

[36] 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. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM'16)*.

[37] Nuance. 2013. Swype Keyboard. Retrieved on 19, March 2020 from https://swype-dragon.en.aptoide.com/.

[38] NUS. 2013. NUS Chat Corpus. Retrieved on 19, March 2020 from https://www.comp.nus.edu.sg/~nlp/corpora.html.

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*.

[40] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543. DOI: https://doi.org/10.3115/v1/D14-1162

[41] Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond just text: Semantic emoji similarity modeling to support expressive communication 👫📲😃. *ACM Transactions on Computer-Human Interaction* 24, 1 (March 2017), Article 6. DOI: https://doi.org/10.1145/3039685

[42] Henning Pohl, Dennis Stanke, and Michael Rohs. 2016. EmojiZoom: Emoji entry via large overview maps 😄🔍 In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'16)*. ACM, New York, NY, 510–517. DOI: https://doi.org/10.1145/2935334.2935382

[43] Shuyao Qi, Dingming Wu, and Nikos Mamoulis. 2016. Location aware keyword query suggestion based on document proximity. *IEEE Transactions on Knowledge and Data Engineering* 28, 1 (Jan. 2016), 82–97. DOI: https://doi.org/10.1109/TKDE.2015.2465391

[44] Microsoft Research. 2016. Introduction to Bayesian Statistics. Retrieved on 19, March 2020 from https://www.microsoft.com/en-us/garage/profiles/word-flow-keyboard/.

[45] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. 810–817.

[46] Barry Schwartz. 2004. *The Paradox of Choice: Why More Is Less*. Harper Perennial.

[47] Catarina Silva and Bernardete Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks*, Vol. 3. 1661–1666. DOI: https://doi.org/10.1109/IJCNN.2003.1223656

[48] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. arXiv:1405.1438.

[49] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. 613–624. DOI: https://doi.org/10.1145/2872427.2883081

[50] Channary Tauch and Eiman Kanjo. 2016. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp'16)*. ACM, New York, NY, 1560–1565. DOI: https://doi.org/10.1145/2968219.2968549

[51] Harry Thornburg. 2018. NLTK's List of English Stopwords. Retrieved November 19, 2019 from https://gist.github.com/sebleier/554280.

[52] TouchPal. 2013. TouchPal Keyboard. Retrieved on 19 March, 2020 form https://www.touchpal.com/.

[53] Unicode. 2019. Emoji Order by Unicode Common Locale Data Repository (CLDR). Retrieved November 19, 2019 from https://unicode.org/emoji/charts/emoji-ordering.html.

[54] Whirlscape. 2015. Minuum Keyboard. Retrieved November 19, 2019 from http://minuum.com/.

[55] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. A semantics-based measure of emoji similarity. In *Proceedings of the International Conference on Web Intelligence (WI'17)*. ACM, New York, NY, 646–653. DOI: https://doi.org/10.1145/3106426.3106490

[56] Sarah Wiseman and Sandy J. J. Gould. 2018. Repurposing emoji for personalised communication: Why 🍕 means "I love you." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, NY, Article 152, 10 pages. DOI: https://doi.org/10.1145/3173574.3173726

[57] Chuhan Wu, Fangzhao Wu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Tweet emoji prediction using hierarchical model with attention. In *Proceedings of the 2018 ACM International Joint Conference and the 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp'18)*. ACM, New York, NY, 1337–1344. DOI: https://doi.org/10.1145/3267305.3274181

[58] Biao Xiang, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen, and Hang Li. 2010. Context-aware ranking in web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 451–458. DOI: https://doi.org/10.1145/1835449.1835525

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

[60] Chong Tze Yuang, Rafael E. Banchs, and Chng Eng Siong. 2012. An empirical evaluation of stop word removal in statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (EACL'12)*. 30–37. http://dl.acm.org/citation.cfm?id=2387956.2387960.

[61] Peijun Zhao, Jia Jia, Yongsheng An, Jie Liang, Lexing Xie, and Jiebo Luo. 2018. Analyzing and predicting emoji usages in social media. In *Companion Proceedings of the World Wide Web Conference 2018 (WWW'18 Companion)*. 327–334. DOI: https://doi.org/10.1145/3184558.3186344

[62] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye text, hello emoji: Mobile communication on WeChat in China. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17)*. ACM, New York, NY, 748–759. DOI: https://doi.org/10.1145/3025453.3025800

[63] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1128–1137. DOI: https://doi.org/10.18653/v1/P18-1104

Received December 2018; revised October 2019; accepted November 2019