Learning from the Ubiquitous Language: an Empirical Analysis of Emoji Usage of Smartphone Users

Xuan Lu[1,2], **Wei Ai**[3], **Xuanzhe Liu**[1,2] *, **Qian Li**[1], **Ning Wang**[4], **Gang Huang**[1,2], **Qiaozhu Mei**[3]
[1] Key Laboratory of High Confidence Software Technologies (Peking University), MoE, Beijing, China;
[2] Beida (Binhai) Information Research, Tianjing; [3] University of Michigan, Ann Arbor, USA;
[4] Xinmeihutong Inc., Beijing, China
{luxuan, xzl, liqian515, hg}@pku.edu.cn, {aiwei, qmei}@umich.edu, ning.wang@xinmei365.com

ABSTRACT
Emojis have been widely used to simplify emotional expression and enrich user experience. As an interesting practice of ubiquitous computing, emojis are adopted by Internet users from many different countries, on many devices (particularly popular on smartphones), and in many applications. The "ubiquitous" usage of emojis enables us to study and compare user behaviors and preferences across countries and cultures. We present an analysis on how smartphone users use emojis based on a very large data set collected from a popular emoji keyboard. The data set contains a complete month of emoji usage of 3.88 million active users from 212 countries and regions. We demonstrate that the categories and frequencies of emojis used by these users provide rich signals for the identification and the understanding of cultural differences of smartphone users. Users from different countries present significantly different preferences on emojis, which complies with the well-known Hofstede's cultural dimensions model.

ACM Classification Keywords
H.1.2. User/Machine Systems: Human information processing

Author Keywords
emoji, data mining, cultural difference

INTRODUCTION
Emojis, also known as ideograms or smileys, have been widely used as complements or surrogates of plain text. First introduced as "*picture" (e)* + "*characters" (moji)* in Japanese electronic messages and Web pages, many emojis have made their way into the Unicode in recent years (722 were included in version 6.0 of the Unicode and 291 were added to version 7.0 and 8.0). Being encoded in Unicode has resulted in a rapid diffusion of emojis to many other countries and regions of the world. As of September 2015, the Unicode provides a full list of 1,281 emojis.

The wide adoption of emojis has been an interesting practice of ubiquitous computing. They are built into many different devices and applications, especially on handsets such as smartphones and tablet computers because of their compactness and liveliness. As a result, they are popularly shared by users in many different countries, from many different demographic groups, and with many different cultural backgrounds. The emoji "*Face with Tears of Joy*" was even elected as the "*Oxford Dictionaries word of 2015*,"[1] as it best represents the mood, the ethos, and the preoccupation of the world in that year. Emojis are used in daily communications, in marketing ads, in persuasion campaigns, and in many other creative ways. For example, Coco Cola once employed emojis as an entrance to their Website[2] in order to deliver the "happiness" culture that they advocate. Some companies have even created their own emojis and made them downloadable via app stores.

From a human-computer interaction perspective, emojis have significant advantages over plain text in facilitating the communications of smartphone users. The compactness of emojis reduces the effort of input; the rich semantics they convey expresses ideas and emotions more vividly; emojis do not have language barriers, making it possible to communicate among users from different countries. These advantages have attributed to the popularity of emojis all over the world, making them a "**ubiquitous language**" that bridges everyone.

The ubiquitous adoption of emojis has also created a great opportunity for researchers. Because they are shared by users worldwide, research questions that are previously restrained by language and geographical barriers can now be pursued through emojis as bridges; because they are so widely populated and frequently used, research questions that previously rely on small-scale user surveys can now be answered through analyzing large-scale behavioral data; because they are compact and conveying clear semantics, research that previously suffers from the insufficiency of natural language processing can now be facilitated using much more robust approaches.

*This work has been approved by the research ethics committees of authors' institutes (a.k.a., IRB). We protect user privacy by removing all user identifiers and textual content of user input other than emojis. Please contact the corresponding author Xuanzhe Liu (xzl@pku.edu.cn) for more information.

UbiComp '16, September 12-16, 2016, Heidelberg, Germany
 ISBN 978-1-4503-4461-6/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2971648.2971724

[1] http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji
[2] See details at http://www.emoticoke.com.

For example, one can infer how users in different countries express their emotions and how the variance in such a behavior can be explained by the difference of the cultures in these countries. Answers to such questions can help app developers better profile and categorize users, more accurately infer their status, moods, and preferences, and therefore provide personalized services and optimized user experience. These questions can be answered through analyzing how a large population of users use emojis in their daily communication, which would have been much more difficult otherwise.

Despite this great opportunity, very few studies have been done so far to systematically analyze and compare the usage of emojis, likely because of the lack of behavioral data at scale. To facilitate this type of explorations, we collected the largest emoji usage data to date, through a leading input method app on Google Play. The app " *Kika Emoji Keyboard*,"[3] or simply "*Kika*," is an emoji-oriented keyboard which has been downloaded by millions of users. A full month of input message log of 3.88 million Android users are collected, who are from 212 countries and regions.

In this paper, we present an empirical analysis based on this large-scale, cross-regional emoji usage data set. We are interested in answering the following questions: do the users from different countries have the same preferences of using emojis? Is there significant difference in their preferences of which emoji to use and which emojis to use together in certain contexts? If yes, can this difference be explained by or be correlated to the cultural difference among these countries? Answers to these questions not only provide a proof-of-concept example of how to utilize emoji usage data to approach research problems that are previously impossible, but also provide direct insights on how to personalize input services and enhance the experience of smartphone users.

To the best of our knowledge, this is the largest study of emoji usage to date. We find that users from different countries have a considerable divergence in emoji usage, and such a variance is highly correlated to the difference of cultural backgrounds, measured by the classical Hofstede culture index [8]. Our findings are valuable to the research community of mobile computing and human-computer interaction, which shed lights on both how to utilize large-scale analyses of emoji usage and how to improve user interfaces and enhance user experience.

RELATED WORK

We start with introducing the background and literature related to our research. Existing studies on emojis are very limited, probably because emojis are newly invented elements of user interfaces and there lacks emoji usage data at scale. This motivates us to collect such a data set and demonstrate the value of analyzing large-scale emoji usage.

Emotion, Emoticon, and Emoji

Emojis were originally created as a compact expression of emotions (sometimes also referred to as sentiments, moods, or attitudes) in online communications. Accurately inferring and understanding the emotions of users is critical for ubiquitous and context-aware applications. Sentiments and emotions were traditionally collected through survey-based [29], biometric-based [4], Audio/Video-based [18], and behavior-based approaches [13]. Manually labeling emotion states is inevitable in such studies. For example, a logger app is used in *MoodScope* [13] for users to report their pleasure and activeness levels four times a day. These approaches are often costly and hard to scale.

Before the emergence of emojis, emotions have been expressed in the form of natural language, Internet slangs, or emoticons (i.e., "*emotional*" icons). Sentiment analysis has long been a core problem of natural language processing [19, 16, 15]. Although various advanced sentiment analysis techniques have been proposed, accurately identifying sentiments and emotions from free text is unfortunately still very challenging, given the complexity of human language. The accuracy of the state-of-the-art, bi-class sentiment classification is widely believed to be around 80-85% . The performance is even lower when dealing with online messages (because of the nonstandard use of language, such as Internet slangs) or dealing with non-English texts (because of the insufficiency of cross-lingual language processing).

Emoticons are artificial combinations of keyboard symbols, which could contain alphanumerics, punctuations, or other characters [28] (comparing to emojis that are preloaded pictographic characters). Similar to emojis, emoticons are often used to express emotions in a compact and vivid way. Many of them are also shared across languages, and are especially popular among smartphone users due to the simplicity. For example, Boia *et al.* [3] studied emoticons in Tweets and concluded that the sentiment conveyed by an emoticon generally agrees with the sentiment of the entire Tweet. However, emoticons have considerable disadvantages. On one hand, the limited morphological variation of ASCII symbols limits the visual expressive power of emoticons, making it hard to use them to express more complex objects and semantics. Instead, the pictographic nature of emojis quickly expanded their territory from emotions to objects, topics, and ideas (e.g., food, faces, events). On the other hand, because emoticons are essentially free combinations of symbols, the nonstandard creation and use of them introduce considerable challenge to data analysis (comparing to natural language where a dictionary can usually be obtained).

Emojis have been a much more interesting, and yet unconventional practice of ubiquitous computing. The rich visual representation enables them to express arbitrarily complex objects. Emojis' compliance to the Unicode standard also guarantees that they are created and diffused in a clean and standard way. These two characteristics have made emojis both extremely useful in user interfaces and especially efficient in data analysis.

Researchers can potentially exploit the emoji usage to infer users' sentiment. Indeed, Zhao *et al.* [30] built a system called *MoodLens*, which tracks public sentiments on *Weibo*, using 95 customized emojis; Kelly and Watts [12] studied how emojis mediated close personal relationship; Vidal *et al.* [27]

[3] https://play.google.com/store/apps/details?id=com.qisiemoji.inputmethod

analyzed food-related tweets and suggested that using emojis and emoticons seem to be an easy and intuitive way to express emotions in a food context.

User Behavior and Cultural Difference

The ubiquitous adoption of emojis motivates us to compare how they are used in different countries. Indeed, existing studies have reported significant differences in behaviors, either offline or online, of users from different countries. Simply identifying different behaviors across national boundaries itself is not comprehensive enough. Explaining such differences is more important and intriguing. In sociological and psychological literature, such behavioral variations are usually explained by the difference in culture instead of nationality. A culture is a shared way of life of a group of socially interacting people, transmitted from one generation to the next via acculturation and socialization processes that distinguish one group's members from others [24, 1]. Regional grouping and culture clustering have been derived from studies of nations based on relatively similarities in history, religion [9], work-related values [7], etc. Gupta *et al*. [6] proposed GLOBE as 10 a priori clusters (i.e., South Asia, Anglo, Arab, Germanic Europe, Latin Europe, Eastern Europe, Confucian Asia, Latin America, Sub-Sahara Africa, and Nordic Europe). Ronen *et al*. [24] created a synthesized cultural clustering of countries based on similarity and dissimilarity in work-related attitudes. Based on their adjacency and cohesiveness, these clusters vary from highly cohesive Arab and Anglo clusters to the least cohesive Confucian and Far Eastern clusters. Hofstede [8] concluded six dimensions of national culture based on a research of how values in the workplace are influenced by culture conducted in IBM.

The concepts and conclusions about cultural differences in sociological literature have been borrowed by computer and information scientists to understand and model online behaviors of users at scale. Findings of such cross-cultural analyses usually provide many useful insights on better designs of behavior models, human computer interfaces, and Internet services. For example, Lim *et al.* [14] investigated how users adopt the concept of app stores, their needs of apps, and their rationale for selecting or abandoning an app; they identified cultural differences in these behaviors and compared the differences with Hofstede's culture index [8]. Reinecke and Bernstein [21] designed systems which automatically generate personalized interfaces according to users' cultural preferences; experiments revealed that such cultural adaptive systems could improve the perceived usability and aesthetics. Cultural factors have also been considered in existing research about emoticons. Tossell *et al.* [26] collected private communication data from individual users' smartphones over a 6-month period, and found that females sent more messages with emoticons, while males used a more diverse set of emoticons. Online chatters have been a new channel to infer cultural differences. Tan *et al.* [25] hypothesized that users who are somehow "connected" are more likely to hold similar opinions, and thus are likely to use similar emoticons. Park *et al.* [20] investigated the semantic, cultural, and social aspects of emoticon usage on Twitter and demonstrated that emoticons are not limited to conveying specific emotions or jokes, but further present sociocultural norms, the meanings of which can vary depending on the identity of the user. Jack *et al.* [10] suggested that culture variation may affect how people distinguish facial expressions, which may explain the fact that easterners and westerners prefer different style of emoticons [20].

Essentially, our work differs from these previous efforts as we analyze national and cultural differences from a new channel - the usage of emojis. We believe as the ubiquitous language, emojis provide a unique basis to study cultural factors. Compared to existing work on emojis and emoticons, our data set is much larger and representative. Although we directly apply concepts and theories of cultural differences (e.g., the Hofstede's culture index) from the sociological literature, we believe our findings can provide unique insights to and potentially complement these theories.

DATA COLLECTION

To facilitate the cross-cultural analysis of emoji usage, we construct a large-scale data set that represents the behavior of millions of users from hundreds of countries and regions. In this section we describe the data collection process and some important considerations.

Kika Emoji Keyboard

The data set was originally collected by the Kika Emoji keyboard (i.e., Kika), a leading Android input method app in Google Play (Figure 1). As one of the most popular third-party keyboards, it has gained millions of downloads and installations across the world, and was ranked as the top 25 most downloaded apps of Google Play in 2015. Kika supports the input of 1,281 emojis (compliant with the Unicode Standard) and more than 60 languages. Just like other popular third-party input methods, the system explicitly notifies that the user input may be collected while enabling the Kika Keyboard. With users' approval, Kika is allowed to collect the meta data, e.g., the language in use, the anonymized content of text messages (identified by "*Send*" action), and the country information (optional at user registration), for research purposes. In particular, Kika explicitly declares in its Privacy Policy that no personal and traceable data from the user input are recorded.[4]

User Privacy and Ethical Consideration

Undoubtedly, preserving user privacy is a critical concern of any input method app, including Kika. To preserve user privacy, we adopted serious procedures before analyzing the data set. First, we removed all textual contents and extracted only the usage of emojis. Second, the data set is stored on a private, HIPPA-compliant cloud server, with strict access authorized by Kika. Third, our analysis pipeline is entirely governed by Kika employees to ensure the compliance with the public privacy policy stated by Kika. Finally and the most significantly, the user IDs are replaced with randomized strings before storage. In other words, one cannot identify individual users with information from the data set.

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

In particular, our analysis is approved by the Research Ethical Committee of the institutes of the authors. The ethical considerations have been carefully addressed during the entire life cycle of this study.

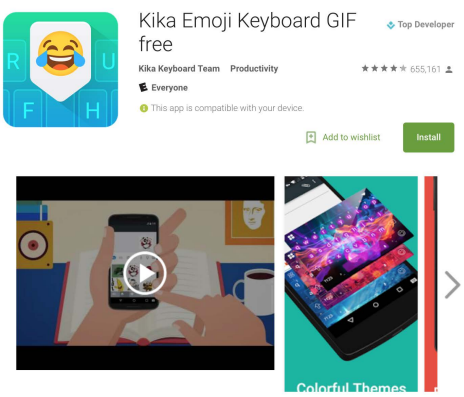

Figure 1. The Kika Emoji Keyboard

Data Set Description

In this study, we use the information including the anonymized user identifier (device ID replaced with random string), the country where the user comes from, and the messages typed by users (with timestamps and emojis and no other information). We associate the messages with user identifiers and aggregate users from the same country.

Finally, we constructed a data set that covers 3.88 million active users from 212 countries and regions and their 427 million messages from September 1 to September 30, 2015, each containing at least one emoji. All users involved in this data set were active ones who used Kika to send at least one message during the period. We further plot the world-wide distribution of active users in Figure 2. Each country is colored according to the number of active users in that country. The deeper the color, the more users in this country use Kika.

All emojis supported in Kika are compliant with the Unicode Standard. We therefore use the name and annotations of the full emoji list on the Unicode Consortium Website[5] in the following analysis.

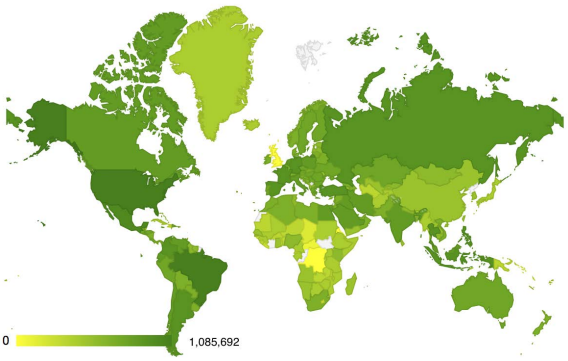

Figure 2. The distribution of active users across the world. The depth of color corresponds to the number of users using Kika.

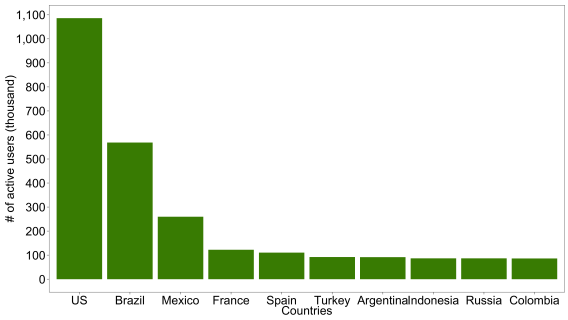

Figure 3. Top 10 countries with the most active users

Note that this data set is not only the largest to date, but also more representative than most similar ones. First, Kika supports 60 languages and is directly downloadable from Google play, hence the distribution of its users does not have a strong geographical or language bias. Second, because Kika is an input method, the data it collects are not limited to particular applications (compared to studies using Twitter or Weibo). These benign characteristics make our data set unique and especially comprehensive for testing hypotheses about smartphone users. We continue with a descriptive analysis of the data set.

DESCRIPTIVE ANALYSIS

In this section, we present a descriptive analysis of the emoji usage data set: the distribution of emoji users and the popularity of emoji usage.

User Distribution

We first report the demographic distributions of users covered by our data set. As shown in Figure 2, the monthly active users spread across 212 countries. The top 10 countries ranked by their active users are US, Brazil, Mexico, France, Spain, Turkey, Argentina, Indonesia, Russia, and Colombia, as shown in Figure 3. In particular, US, Brazil, and Mexico have the dominant number of users compared to others. Over 1 million active users in the US constitute nearly 1/3 among all active users in our data set.

Some demographic information, such as gender, age, religion, and relationship status, can help us better understand how users use emojis. Kika made a survey of the age and gender distribution of its users, as shown in Figure 4. More specifically, 68.3% of the users are female while the other 31.7% are male. Most of them are young people, as 74.3% are under 25 years old.

Arguably, these demographic information may be confounding factors of the correlation between cultural difference and the emoji usage. Yet we assume that they do not contribute to the cross-cultural differences in emoji use.

[5]http://unicode.org/emoji/charts/full-emoji-list.html

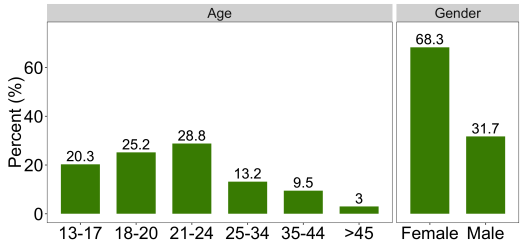

Figure 4. Gender and age distribution of Kika users

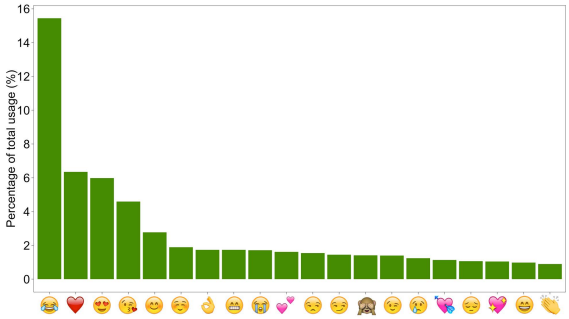

Figure 5. Top 20 most used emojis

Popular Emojis

We then investigate how often users use emojis in their text messages. It is observed that 7.1% of the 6.06 billion messages include at least one emoji. Such a result can evidence the popularity of emojis. However, not all emojis are popular. The popularity of every single emoji, computed by the number of messages in which it occurs, typically follows a power-law distribution. In other words, a small portion of emojis are more frequently used while the majority of emojis are not. More specifically, we find that 119 out of the 1,281 emojis can constitute around 90% of the usage.

The top 20 "*emojis-of-the-month*" are illustrated in Figure 5. The emoji 😂, known as *face with tears of joy*, is the most frequently used one, followed by ❤️ and 😍. We can find that all of the 20 emojis fall into the categories of *face*, *heart*, and *hand*. Such an observation indicates that expressions and body signals play an important role in expressing ideas when using emojis. The face 😂 is overwhelmingly popular, comprising 15.4% of the total emoji usage, which is consistent with the report from Oxford Dictionaries.

COUNTRY DIFFERENCE IN EMOJI USAGE

Although emojis can be ubiquitously used as a simple language, we are interested in whether users from different countries have similar or different behaviors in using emojis. To this end, we study the distributions of emojis in messages from two aspects. On one hand, we investigate how often a single emoji is used. On the other hand, we explore which emojis are likely to be frequently co-used together, i.e., the co-occurrence of emojis.

Difference in the Frequently Used Emojis

Since the frequency of emoji usage typically follows a power-law distribution (i.e., a small portion of emojis contribute to a large portion of the usage), we first look at the most frequently used emojis in different countries. Due to page limit, in Table 1, we summarize only the top 10 most frequently used emojis in the top 10 countries ranked by their users.[6] We also calculate the proportion of messages that contain emojis.

Preliminary results show that variance of emoji usage does exist among countries. We can find that users from France are more likely to use emojis, i.e., 19.8% of messages involve at least one emoji, a significantly higher ratio than the users from other 9 countries. It is interesting to see that the top 10 emojis in France are also quite different from those used by other countries, i.e., they are more likely to use emojis related to *hearts*, while users from other countries prefer emojis related to *faces*.

Table 1. Emoji usage in text messages

	% emoji-msg	Top emoji
US	9.2	😂 😘 😍 😊 ❤️ 😭 😉 ☺️ 😁 💯
Brazil	5.1	😂 😍 ❤️ 👏 😘 👌 😒 😌 😏 💕
Mexico	7.9	😂 ❤️ 😍 👌 😘 ☺️ 🙈 😭 🙈 😁
France	**19.8**	❤️ 😂 💘 💖 😘 😍 💕 💞 💜 💗
Spain	3.4	😂 😍 😘 ❤️ 💕 😊 👌 🙈 😏 💞
Turkey	5.8	😂 😄 😊 😘 ☺️ 😃 😁 😍 ❤️ 😀
Argentina	3.1	😍 ❤️ 😂 😘 👌 😊 💕 😭 😏 ☺️
Indonesia	3.2	😂 😘 😁 😊 😍 😄 😢 😅 😭 ❤️
Russia	10.9	😂 😄 ❤️ 😘 😊 😍 ☺️ 😃 💋 😏
Colombia	3.7	😍 😂 ❤️ 😘 🙈 👌 ☺️ 😊 😔 😒

Difference of Co-Used Emojis

We have already demonstrated the difference in using the popular emojis. Yet we are also interested in the difference at the long tail. Characterizing such difference is not straightforward, since the distributions of the long-tail emojis are small and trivial. Therefore, we group similar emojis together and quantitatively characterize the difference between groups. More specifically, we group the emojis in two different ways, one using the annotation tags associated with each emoji, and the other using the co-occurrences.

Grouping by Annotation Tags

The Unicode Consortium provides the official annotations for each emoji. These annotations are represented as simple tags, thus a natural way for grouping is to put emojis with the same tag together. Since there can be several annotations associated with each emoji, we allow the groups to overlap.

There are 1,283 unique annotation tags provided by the Unicode Consortium. For each tag, we aggregate the occurrences of all emojis belonging to it as the tag's popularity score. We filter the tags and keep those whose popularity scores are larger than 2 million. We also filter countries and keep those with at least 2 million emoji occurrences. Finally, we have 141 tags and 38 countries.

For each of the 38 countries, we normalize the popularity scores by the emoji usage to make them comparable. We also compute the normalized popularity scores using aggregated

[6]We will list the frequencies of top-10 frequently used emojis for all 212 countries along with this paper's publication.

data from all countries. Then, we examine if a country could use certain tags significantly more or less than average. We conduct a two-side *z-test* to compare the popularity score of each tag and country with that of the aggregated score. We assign a +1(or -1) to the country-tag pair if the country uses that tag more (or less) than average at 0.05 significance level, and 0 otherwise. In this way, we calculate a 141-dimensional vector $\{+1, 0, -1\}^{141}$ for each country.

With these vectors, we calculate the similarity score between every two countries as the inner product of the corresponding vectors. The larger the similarity score is, the more similar pattern the two countries share in using emojis. We then perform a hierarchical clustering by the similarity scores, and show the result in Figure 7. The dendrogram corresponds to the hierarchical structure while the brightness of each cell indicates the similarity between the two countries.

As shown in the hierarchy tree, the 38 countries are clustered into 2 big groups: the 27 countries in the upper-right corner are categorized into the first group, 11 of which are developed countries (i.e., US, Canada, Germany, Italy, France, Belgium, Portugal, UK, Netherlands, Czech Republic, and Australia); the remaining 11 countries in the lower-left corner are categorized into the second group, where only Spain is developed country. This result indicates the potential correlation between the development levels of the countries and the emoji groups that users in this country prefer.

The clustering result can be explained by several factors, such as geographical closeness, language, and history. 10 countries (i.e., Germany, Austria, Italy, France, Belgium, Greece, Portugal, UK, Romania, and Netherlands) from Europe are closely clustered in the first group. In the second group, most countries are in South America: Ecuador, Colombia, Peru, Chile, Mexico, Costa Rica, Brazil, Argentina, Venezuela, and Dominican Republic. All of these countries used to be colonies of Spain, except Brazil. It seems that the religion and language brought by Spanish successfully create a singular and cohesive Latin American cluster [6]. Brazil, which used to be a colony of Portugal and uses Portuguese as the official language, is also categorized into the Latin American cluster. Such a result indicates that the regional factors have more effect on Brazilian expression of topic.

The country pairs at the bottom level of the clustering dendrogram share the most similarity of using emojis in different annotation groups. Some of these pairs are geographically close, e.g., Canada and US. However, it is surprising for some pairs to be clustered, e.g., Russia is in the east Europe while India is in the south Asia. Actually, the languages used in Russia and India, Russian and Sanskrit, belong to the *satem* group of the Indo-European family of languages. In addition, the distance between these two languages are closer than that between Sanskrit and other Indo-European languages, and the speakers of Russian and Sanskrit have lived close together during some period of history [23].

Grouping by Co-Occurrence

The annotation tags are predefined by the Unicode. However, users can interpret the emojis' similarities variously. We then leverage the user-generated data to cluster emojis that are frequently used together. We use the *point-wise mutual information (PMI)* [5] to measure the co-occurrence of every two emojis. A larger PMI indicates the two emojis are more likely to occur together.

We then use the PMI to build emoji graph. In such a graph, each node represents an emoji, and we connect each emoji to five emojis that have the largest PMI with it. In Figure 6, we use Gephi[7] tool to plot the graph with a force-based layout [11]. We can observe some significant clusters, such as flags, food, faces, travel, animals, and clocks.

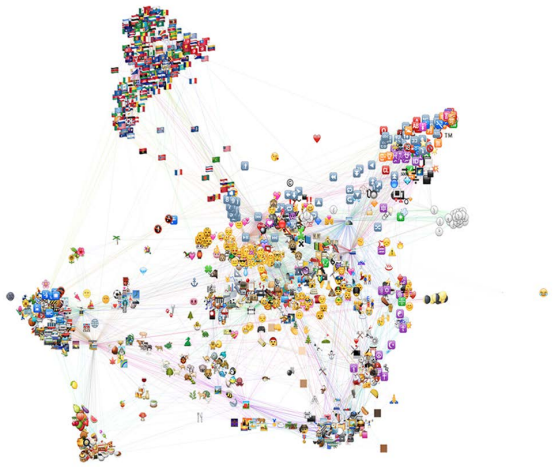

Figure 6. Network of emoji co-occurrence

We then perform community detection using the classic Fast Unfolding algorithm [2] and obtain 69 non-overlapping clusters. The algorithm splits the graph as follows: the nodes within the same cluster have more connections (larger PMI) with each other, while the nodes from different clusters have fewer connections (smaller PMI).

In addition, we build such emoji graphs for the top 10 countries that have the most active users. Some common clusters are observed, but the size and cohesion of clusters vary a lot among countries. For example, emojis related to *face* form a relatively significant cluster in Argentina, while they are mixed with situations such as *office* and *entertainment* in Mexico. Based on such an intuitive finding, we aim to investigate the country differences of dependency on these 69 clusters, i.e., how different countries differ in using emojis from different clusters.

With the community detection results (i.e., the 69 non-overlapping clusters), we calculate the correlations between countries on their usage of different clusters, and perform a similar hierarchical clustering using the correlations as similarities (Figure 8).

Using the co-occurrence of emojis, the countries are clustered into two groups, complying with the GLOBE clustering [6]. The first group includes countries from South America (Ecuador, Peru, Dominican Republic, Costa Rica, Colombia,

[7] https://gephi.org/

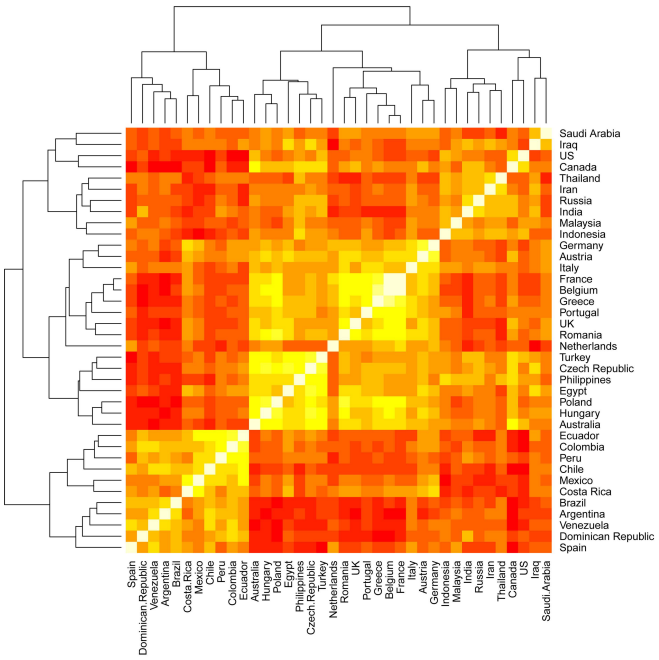

Figure 7. Similarity score of using emojis with different annotation tags

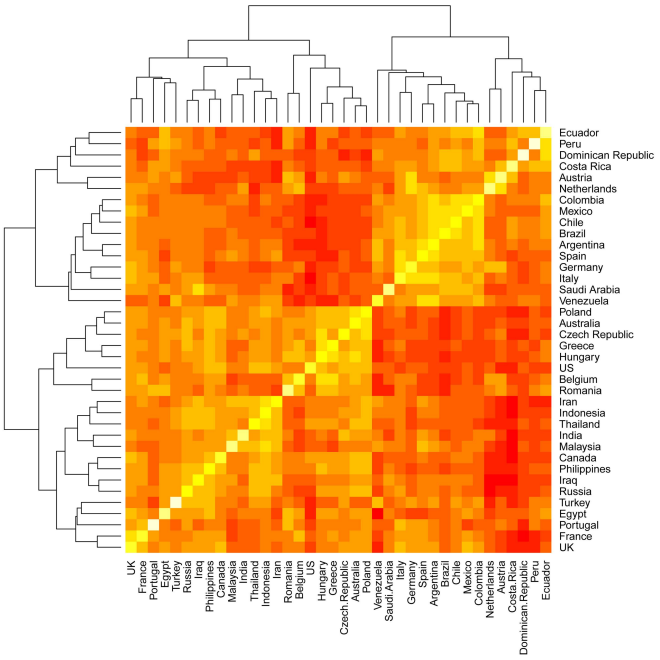

Figure 8. Correlation of using emojis in different clusters

Mexico, Chile, Brazil, Argentina, and Venezuela), Germanic Europe (Austria, Netherlands, and Germany), and Saudi Arabia. The countries in the second group are mainly from Southern Asia (Iran, Indonesia, Thailand, India, Malaysia), Anglo Cultures (Australia, US, Canada, and UK), Eastern Europe (Poland, Greece, Hungary, and Russia), and Arab Cultures (Turkey, Egypt).

In these two big groups, some sub-clusters are compact and clear (such as Southern Asia), but some are quite scattered (such as Anglo Cultures). In addition, countries belonging to Latin Europe are split into various groups, i.e., Italy and Spain are in the first group, while Portugal and France are in the second group. Such results provide a new perspective to exploit the relationships of the GLOBE clusters.

EMOJI USAGE AND CULTURE INDEX

From previous sections, we have observed that users from different countries can have obviously various patterns of using emojis. Not only do they have different go-to emojis, but they also have different preferences towards different groups of emojis. By examining the similarity matrix, we find that countries sharing the similar emoji usage patterns are more likely to share common language or geo-region. However, we do witness that sometimes the similarity goes beyond language and geological closeness.

The commonalities in users' preferences can be due to deeply-rooted factors in culture background. Previous literature [20] revealed that vertical style emoticons such as ^_^ and T_T are more popular among users with oriental culture, while horizontal emoticons like :) and :D (expressions based on the mouth shape) are more popular in western people. Formally, the vertical emoticons depict expressions based on the eye shape, while horizontal emoticons depict expressions based on the mouth shape. The style preferences are well aligned with the differences how easterners and westerners decode facial expression signals [10].

Inspired by this study on the precursor of emojis, we then examine how the culture difference can lead to the difference in emoji usage. More specifically, since emojis are invented to facilitate emotion representation, we aim to associate the sentiment of emojis with the Hofstede culture index [8].

Hofstede Culture Index

Culture can be loosely based on shared values, and researchers have made substantial efforts to find a set of tangible indicators of culture. Hofstede described the differences in national culture with a six-dimension model [8]: *power distance, collectivism versus individualism, masculinity versus femininity, uncertainty avoidance, long-term versus short-term orientation*, and *indulgence versus restraint*. These dimensions have been widely used in cross-cultural studies [22]. Definitions of the six dimensions are quoted as follows.

Power distance. This dimension describes the extent to which the less powerful members of *institutions* and *organizations* within a country expect and accept that power is distributed unequally. *Institutions* are the basic elements of society, such as the family, the school, and the community. *Organizations* are the places where people work.

Collectivism versus individualism. *Individualism* pertains to societies in which the ties between individuals are loose: everyone is expected to look after him- or herself and his or her immediate family. *Collectivism* as its opposite pertains to societies where people from birth onward are integrated

into strong, cohesive in-groups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty.

Masculinity versus femininity. The emotional gender roles are clearly distinct in masculine societies: men are supposed to be assertive, tough, and focused on material success, whereas women are supposed to be more modest, tender, and concerned with the quality of life. But such roles overlap in feminine societies: both men and women are supposed to be modest, tender, and concerned with the quality of life.

Uncertainty avoidance. This indicator describes the extent to which the members of a culture feel threatened by ambiguous or unknown situations. This feeling is, among other manifestations, expressed through nervous stress and in a need for predictability: a need for written and unwritten rules.

Long-term versus short-term orientation. The *long-term orientation* stands for the fostering of virtues oriented toward future rewards–in particular, perseverance and thrift. Its opposite pole, *short-term orientation*, stands for the fostering of virtues related to the past and present–in particular, respect for tradition, preservation of "face," and fulfilling social obligations.

Indulgence versus restraint. The dimension *indulgence* stands for a tendency to allow relatively free gratification of basic and natural human desires related to enjoying life and having fun. Its opposite pole, *restraint*, reflects a conviction that such gratification needs to be curbed and regulated by strict social norms.

The preceding six dimensions are all quantitative and measured by an index, i.e., Power Distance Index (PDI), Individualism Index (IDV), Masculinity Index (MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation Index (LTO), and Indulgence Versus Restraint Index (IVR). Hofstede estimated the scores for a number of countries and areas.[8] In this study, we have 102 countries and areas that are covered by both Hofstede model and our data set. All of the 102 countries and areas have scores of PDI, IDV, MAS, and UAI. However, only 86 countries have LTO, and 81 have IVR. In our analysis regarding each culture index, we use only the countries with corresponding scores.

Extracting Emoji Sentiment

After obtaining the culture indexes for each country, we can naturally examine the correlation between the culture index and the usage of specific emojis. However, selecting certain emojis for study may not be representative. Even using emojis from a single tag may still not be representative enough. Recalling Figure 5, the most commonly used emojis all express certain emotion. We thus choose to look at the emojis that convey user emotions and see how the usage of these emotional emojis reflect the culture background. We choose to use emojis that convey different sentiments.

[8]http://geert-hofstede.com

Table 2. Classification of emojis with the emotion semantic

	Condition		# of emojis
POS	$S_{posemo}>S_{negemo}$		141
MIX	$S_{posemo}=S_{negemo}>0$		4
NEG	$S_{posemo}<S_{negemo}$		54
ANX	$S_{posemo}<S_{negemo}$ & Max(S_{anx}, S_{ang}, S_{sad}) >0 &	Max(S_{anx}, S_{ang}, S_{sad})=S_{anx}	6
ANG		Max(S_{anx}, S_{ang}, S_{sad})=S_{ang}	11
SAD		Max(S_{anx}, S_{ang}, S_{sad})=S_{sad}	11

Finding such emojis in an objective way is not easy, since everyone can interpret emojis in his/her own way. Therefore, we make use of the official annotations again, which provide textual descriptions that "*translate*" emojis back into words. Instead of manually labeling each emoji, we take the annotations of emojis and employ a text analysis tool, named LIWC (Linguistic Inquiry and Word Count)[9] to calculate the sentiment and gender score for every single emoji. LIWC includes the main text analysis module along with a group of built-in dictionaries. Basically, LIWC reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Then, LIWC ranks the text and assigns the emotion score such as *positive*, *negative*, and so on. Due to page limit, more details of LIWC can be found on its Website.

After applying LIWC, we obtain 199 emojis with sentiment scores. The rest emojis are discarded since their annotations do not imply sentiment by LIWC. Emojis with positive scores higher/equal/lower than (to) negative ones are categorized as Positive (**POS**), Mixed (**MIX**), and Negative (**NEG**), respectively. For the negative emojis, we further compare their scores of Anxiety (**ANX**), Anger (**ANG**), and Sadness (**SAD**). Negative emojis whose anxiety scores are the highest among the three negative scores are categorized as "anxious" emojis, so do "angry" and "sad" emojis. The categorization of the emoji sentiment is summarized in Table 2.

Correlating Culture Index with Emoji Sentiment

For each category, we calculate the proportion of emojis used in each country that falls into that category. Then, we measure its Pearson's correlation ρ with the country's Hofstede culture index. Table 3 summarizes the Pearson's correlation of all culture indexes and all sentiment categories.

It is observed that some significant correlations exist. We explain them as follows.

- Users from strong power-distance countries are more likely to express negative emotion through emojis of NEG (ρ=.287, p-value=.003), especially of SAD (ρ=.365, p-value=.000). For example, users from countries such as Malaysia, Iraq, Saudi Arabia, Mexico, Indonesia, and Ecuador (high PDI) are more likely to use emojis of NEG and SAD than countries such as Czech Republic, Spain, Italy, Hungary, Netherlands, Australia, UK (low PDI).

- Users from strong-individualism index countries are more likely to express positive emotion through emojis of POS (ρ=.241, p-value=.015), but less likely to express negative emotion through emojis of NEG (ρ=-.459, p-value=.000).

[9]http://liwc.wpengine.com

Table 3. Pearson's correlation ρ

Index	# of countries	POS	MIX	NEG	ANX	ANG	SAD
Power Distance (PDI)	102	-0.059	-0.076	**0.287***	0.094	0.186	**0.365***
Individualism (IDV)	102	**0.241***	0.194	**-0.459***	**-0.200***	**-0.347***	**-0.449***
Masculinity (MAS)	102	-0.012	-0.187	0.161	0.057	0.066	0.174
Uncertainty Avoidance (UAI)	102	**-0.211***	0.165	0.050	-0.082	**0.208***	0.055
Long-Term Orientation (LTO)	86	**0.413***	0.181	**-0.474***	-0.025	**-0.281***	**-0.379***
Indulgence Versus Restraint (IVR)	81	**-0.455***	-0.042	**0.224***	-0.198	-0.218	0.123

The 3 particular types of negative emotions, anxiety (ρ=-.200, p-value=.044), anger (ρ=-.347, p-value=.000), and sadness (ρ=-.449, p-value=.000) are also less likely to be expressed through emojis by these users. For example, users from Australia, Hungary, France, and Czech Republic (high IDV) are more likely to use positive emojis of POS and less likely to use negative emojis of NEG, ANX, ANG, and SAD than users from Jamaica, Iraq, Mexico, Chile, Thailand, El Salvador, Peru, and Colombia (low IDV).

- Users from high uncertainty-avoidance countries are less likely to express positive emotion (ρ=-.221, p-value=.034). They are more likely to express emotion of anger through emojis of ANG (ρ=.208, p-value=.036), but not for the entire set of negative emojis of NEG. For example, users from Portugal, El Salvador, Peru, Chile, Argentina, Costa Rica, Iraq, Mexico, Israel, Saudi Arabia, and Colombia (high UAI) are less likely to use emojis of POS, and more likely to use emojis of ANG than users in Australia, Indonesia, US, and UK (low UAI).
- Users from strong long-term orientation index countries are more likely to express positive emotion through emojis of POS (ρ=.413, p-value=.000), and less likely to express negative emotion through emojis of NEG (ρ=-.474, p-value=.000), ANG (ρ=-.281, p-value=.009) and SAD (ρ=-.379, p-value=.000). For example, users from Bulgaria, France, Hungary, Ukraine, Romania (high LTO) are less likely to use negative emojis than users in Israel, Saudi Arabia, Thailand, Chile, Uruguay, Iraq, Peru, Mexico, Argentina, El Salvador, Venezuela, Iran, and Colombia (low LTO).
- Users from strong-indulgence index countries are less likely to express positive emotion through emojis in POS (ρ=-.455, p-value=.000), but more likely to express negative emotion through emojis in NEG (ρ=.224, p-value=.044). For example, users from Venezuela, Mexico, Colombia, Chile, Argentina, Saudi Arabia (high IVR) are less likely to use emojis in POS and more likely to use emojis in NEG, in comparison to Turkey, France, Czech Republic, Poland, Russia, and Romania (low IVR).

Among the six indexes, IDV seems to be an indicator of most of the 6 emoji sets. Expressing happiness is encouraged while expressing sadness is discouraged in individualism-oriented societies, with quite significant correlation between the individualism and the used emojis.

Other three indexes (i.e., UAI, LTO, IVR) also explain both the positive and negative sides of the results. People in high uncertainty-avoidance societies have more tendency of higher stress and anxiety, and thus express less positive emotion. However, the correlation between the UAI and the usage of ANX emojis is not quite significant, which contradicts to the characteristic of high uncertainty-avoidance countries. Instead, people from high uncertainty-avoidance countries tend to express anger through emojis. People from high long-term-orientation countries focus more on the long-term plans and goals, and thus they tend to perform more positively. People from societies with high IVR tend to constraint less of negative expression.

For the PDI that demonstrates only one side of the emotion, people from strong power-distance countries tend to express more negative emotion, especially sadness.

Consequently, this analysis suggests that country differences in emoji usage are quite significant, although not entirely well correspond to the Hofstede's culture index. Nevertheless, the derived knowledge demonstrates that emoji usage can be a useful signal to distinguish users with different culture background.

DISCUSSION

Based on the results reported above, we can confirm that the usage of emojis presents significantly different patterns across countries, which to certain extent comply with the culture backgrounds of the countries. In this section, we discuss the limitations of our empirical results, and try to derive some implications and insights from our results.

Threats and Limitations

One major threat of this study is that the covered users are those who use Kika keyboard. Indeed, most popular smartphone manufacturers support emojis in their built-in input methods. Yet the Kika keyboard is designed to optimize the input experience of emojis, and thus is more attractive to younger group of users, leading to a potential selection bias.

Besides, although emojis have been encoded as standard Unicode, and the emojis covered by Kika and other input methods are almost identical, the rendering of the same emoji is not exactly the same on different platforms. The difference in rendering may lead to different interpretations of the exact same emoji, as pointed out in [17].

Another limitation is that the time of our data set spans only one month. Some emerging events such as natural disasters in certain countries can possibly lead to unrepresentative user moods and behaviors, and can affect the usage of emojis. For example, the occurrence of some emojis can temporarily burst. We do not capture such usage patterns in this paper. In our future work, we plan to conduct time series analysis of emoji

usage and compare the results of short-term and long-term study. It would be interesting to explore whether emojis can be leveraged to predict public opinions and sentiment of a country.

Finally, although our analysis demonstrates significant correlations between emoji usage and culture indexes, it is far from enough to establishing any casual conclusion, as many confounding factors may also have an effect on the emoji usage, such as age and gender. To establish more rigorous conclusions, an in-depth statistical analysis needs to be done should such information be available.

Certain future work would make our conclusions more comprehensive: in this study, we correlate culture index with only sentiment polarities, while we can further validate whether the results can be generalized to other metrics such as age, economic, gender, and so on. We could also analyze the diverse patterns of emoji usage in different apps, at different diurnal slots, etc.

Implications

The study in this paper has demonstrated that emojis can be a signal to tell the difference between users from different countries even without any textual information. We then discuss some implications following our previous observation.

Optimizing user experience for input methods. The first and most intuitive implication is to improve the user experience of input methods not limited to Kika. Though lots of input methods support emojis, the specific optimizations for emojis have not been well addressed. The UI layout of emojis is rather fixed. Some input methods can suggest the most popular emojis to make users fast locate, but the suggestion is not optimized for users from different countries. As we reported in this paper, users from different countries can have quite various preferences to use emojis. For example, users from France prefer using *heart*-related emojis. Therefore, the rank of emojis shown on the input methods' UI should be country-aware to users. Similarly, from the community detection of frequently co-used emojis, input methods can be capable of suggesting more relevant "*next-to-use*" emojis to users. Contextual information (i.e., texts, apps, location, and time) could also be leveraged to provide better context-aware user experiences

Understanding user preferences. This paper leverages the country information that the users optionally provide in Kika, and reveals the various usage patterns of emoji among countries. We validate that the usage of emoji can comply with the classic culture difference model. In some cases, it is observed that emoji usage can capture the culture differences of users who come from different countries but speak the same native language. For example, users from Brazil have quite similar emoji usage with those who come from other countries in South America, but perform quite differently from the users of Portugal even if they speak the same language. In other words, emojis can be complementary to NLP techniques when text is sparse. It is then possible to understand user preference through such an ubiquitous language. For example, smartphone users tend to use more emojis other than type in plain texts when they commit reviews for food, movie, and so on. In such scenarios, the understanding of user preferences can be more accurate by synthesizing emoji usage with other contextual information, enabling developers to customize country-aware and personalized user experiences or place accurate in-app advertisements.

CONCLUSION AND FUTURE WORK

In this paper, we have presented an empirical study of how people use emojis, an emerging ubiquitous language for expressing emotions, topics, and ideas. We conducted our study based on a unique and large data set collected through a popular input method app. The data set consists of over 400 million emoji-contained messages generated by more than three million users from 212 countries and regions. We demonstrated considerable diversity of emoji usage among users from different countries, and linked this diversity to a classical culture index model. Based on our observations, we have presented some implications and suggestions to improve the quality of user experiences for input methods, understand user preferences, etc. To the best of our knowledge, we have made the first large-scale analysis of emoji usage.

Currently, we are working on integrating the usage patterns of emojis in Kika and improving the user interface by customizing the personalized list of suggested emojis for users from different countries or with different languages. Another interesting future direction is to study whether emojis are really consistent with the sentiments presented in texts. We would be surprised if they are not, but any gap between the "ubiquitous language" and the natural languages would be intriguing.

ACKNOWLEDGMENTS

This work was supported by the High-Tech Research and Development Program of China under Grant No. 2015AA01A202, the Natural Science Foundation of China (Grant No. 61370020, 61421091, 61528201). Qiaozhu Mei's and Wei Ai's work was supported in part by the National Science Foundation under Grant No. IIS-1054199 and an MCubed grant of the University of Michigan. The authors would like to appreciate Kaidong Wu from Peking University for the efforts on data pre-processing.

REFERENCES

1. J. W. Berry and Ype H. Poortinga. 2006. Cross-cultural theory and methodology. *Families across cultures. A 30-nation psychological study* (2006), 51–71.
2. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Statistical Mechanics: Theory and Experiment* 30, 2 (2008), 155–168.
3. M. Boia, B. Faltings, C. C. Musat, and P. Pu. 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.
4. Erin A. Carroll, Mary Czerwinski, Asta Roseway, Ashish Kapoor, Paul Johns, Kael Rowan, and Monica M. C.

Schraefel. 2013. Food and mood: just-in-time support for emotional eating. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*. 252–257.

5. Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16, 1 (1990), 22–29.
6. Vipin Gupta, Paul J. Hanges, and Peter Dorfman. 2002. Cultural clusters: methodology and findings. *Journal of World Business* 37, 1 (2002), 11–15.
7. Geert Hofstede. 1997. *Cultures and organizations: software of the mind*.
8. Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and organizations: Software of the mind (3rd Edition)*.
9. Samuel P Huntington. 1993. The clash of civilizations? *Foreign affairs* 72, 3 (1993), 22–49.
10. Rachael E Jack, Caroline Blais, Christoph Scheepers, Philippe G Schyns, and Roberto Caldara. 2009. Cultural confusions show that facial expressions are not universal. *Current Biology* 19, 18 (2009), 1543–1548.
11. M Jacomy, T Venturini, S Heymann, and M Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One* (2014).
12. R Kelly and Leon Watts. 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* (2015).
13. Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual International conference on Mobile systems, applications, and services, MobiSys 2013*. 389–402.
14. Soo Ling Lim, Peter J. Bentley, Natalie Kanakam, Fuyuki Ishikawa, and Shinichi Honiden. 2015. Investigating country differences in mobile app user behavior and challenges for software engineering. *IEEE Transactions on Software Engineering* 41, 1 (2015), 40–64.
15. Bing Liu. 2012. *Sentiment analysis and opinion mining*. Vol. 5. Morgan & Claypool Publishers.
16. Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007*. 171–180.
17. 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 Conference on Weblogs and Social Media, ICWSM 2016*.
18. Claudia Orellana-Rodriguez, Ernesto Diaz-Aviles, and Wolfgang Nejdl. 2013. Mining emotions in short films: user comments or crowdsourcing?. In *Proceedings of the the 22nd International World Wide Web Conference, WWW 2013, Companion Volume*. 69–70.
19. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1-2 (2008), 1–135.
20. Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: interpreting differences in emoticons across cultures. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013*.
21. Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 2 (2011), 8–29.
22. Katharina Reinecke, Sonja Schenkel, and Abraham Bernstein. 2010. Modeling a user's culture. *Chicago* (2010).
23. Weer Rajendra Rishi. 1982. *India & Russia: linguistic & cultural affinity*. Roma Publications.
24. Simcha Ronen and Oded Shenkar. 2013. Mapping world cultures: cluster formation, sources and implications. *Journal of International Business Studies* 44, 9 (2013), 867–897.
25. Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011*. 1397–1405.
26. Chad Tossell, Philip T. Kortum, Clayton Shepard, Laura H. Barg-Walkow, Ahmad Rahmati, and Lin Zhong. 2012. A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior* 28, 2 (2012), 659–663.
27. Leticia Vidal, Gastón Ares, and Sara R Jaeger. 2016. Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference* 49 (2016), 119–128.
28. Joseph B. Walther and Kyle P. D'Addario. 2003. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 5, 2 (2003), 119–134.
29. Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015*. 295–306.
30. Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 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.