The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

A Quantitative Analysis of Comparison of Emoji Sentiment: Taiwan Mandarin Users and English Users

Fang-Yu Chang
Graduate Institute of Linguistics,
National Taiwan University
R07142005@ntu.edu.tw

Abstract

Emojis have become essential components in our digital communication. Emojis, especially smiley face emojis and heart emojis, are considered the ones conveying more emotions. In this paper, two functions of emoji usages are discussed across two languages, Taiwanese Mandarin and English. The first function discussed here is sentiment enhancement and the other is sentiment modification. Multilingual language model is adopted for seeing the probability distribution of the text sentiment, and relative entropy is used to quantify the degree of changes. The results support the previous research that emojis are more frequently-used in positive contexts, smileys tend to be used for expressing emotions and prove the language-independent nature of emojis.

Keywords: emoji, sentiment enhancement, sentiment modification

1 Introduction

With the considerable growth of social media, emojis have become increasingly popular and widely used across the world. During the last few years, emojis change the way we communicate online. They allow us to interact with each other more clearly when we struggle to express our emotions through pure texts. It is an explicit acknowledgment that emojis are now part of how we express our emotions, intents and feelings.

Aside from emotion expression, emojis are believed to enhance and modify the sentiment of a text. A sentence may convey an emotion, but its emotion would be strengthened after the addition of emojis, which is called sentiment enhancement.

On the other hand, if the emotion of the sentence is weakened or altered to the opposite emotion, that is called sentiment modification. Moreover, research shows that emojis used to convey emotions are mostly smileys and hearts, which belong to the Smiley & Emotion group in the Unicode emoji categories.

However, though emojis' non-verbal nature suggests that they are universal across cultures, their usages may change from language to language and culture to culture. Over the past few years, great concern has arisen in the research of cross-cultural or cross-language emoji usage. Some of them discuss which kinds of emoji patterns are the same or different across cultures. Some studies build the emoji sentiment lexicon with the data from different languages. But little was done on the degree of how each emoji can enhance or modify the text across languages.

This paper aims to compare the degree of sentiment enhancement and sentiment modification of emojis which are used by Taiwanese Mandarin users and English users with the help of a multilingual language model. Recently, large pre-trained neural models such as BERT have achieved great success in NLP, motivating more and more research to investigate what aspects of language they are able to learn from unlabeled data.

2 Background and Related Work

Many people believe emojis are like the older emoticons, which provide a visual representation using punctuation marks. Emojis and emoticons function as the non-verbal cues (paralanguages) in face-to-face communication, which are believed to convey emotions more effectively and efficiently than the words themselves are saying. In fact, the reason punctuation marks came into existence was to complement emotional engagement in written texts. (Evans, 2017) Moreover, Guibon et al. (2016)

propose that emojis not only add an emotion to a sentence, but also enhance and modify the emotion of a sentence. They also state that emojis are ambiguous and unreliable without context and emojis are often placed at the end of sentences to express emotions.

Further works are done to support that emojis and emoticons result in higher sentiment and have higher contribution in overall sentiment score. Davidov (2010) uses KNN-like strategy to show that punctuations, words and pattern features (including emoticon tags) can improve the quality of sentiment classification tasks. Agarwal (2011) suggest that specific features like emoticons and hashtags also add marginal value to the sentiment classifier. Hogenboom (2013) puts forward that sentiment classifiers are more accurate when they train on emoticons. Ayvaz and Shiha (2017) collects positive and negative data to analyze the influence of emojis in sentiment analysis, they find that emojis not only increase sentiment score in both polarities, but more frequently used to show positive opinions. Tian's study on Facebook data across four different countries (2017) proves that emojis and texts can update the meaning of each other, suggesting there is a correlation between emojis and linguistic contexts, the author also states that sarcasm, irony and politeness can be interpreted by analyzing emojis.

However, previous works mainly take emojis as features to better the performance on sentiment analysis. This approach would not take the impact of emojis on the texts into account, since an emoji has different influences on different contexts. We can know a smiley has a positive impact on texts, but it might be difficult to obtain how much degree of the impact of a smiley on two unrelated texts. Along with the development of the attention network, Lou et al. (2020) first use attention mechanisms to train emoji and text embeddings simultaneously on a Bi-LSTM model. Conneau et al. (2020) present a transformer-based multilingual pre-trained on texts in 100 languages.

The most widely used genre among emoji is facial expressions, Gao (2020) states that they are keys to convey emotions. When people look at a smiley face online, the same parts of the brain are activated like they look at a real human face. However, facial expressions are not universal signals. The interpretation of emotions and attitude is strongly influenced by different cultural backgrounds. (Jack et al., 2009) According to the study, overt emotional demonstration is the norm in Western cultures, while subtle emotional demonstration is the norm in Eastern cultures. Researchers also suggests that these differences extend to the use of emojis. (Gao and VanderLaan, 2020)

From another perspective, there are researches exploring the meanings and usages across cultures and languages. Barbieri et al. (2016) adopt various experiments to compare the usage of emojis across four Western languages. They observe that the frequently used emojis share similar semantic usages across these four languages, supporting that emojis are language independent. On the other hand, they find that the usages of particular emojis differ due to the cultural influences.

3 Methods

Data collection: Taiwanese Mandarin users data is from Dcard and Instagram, since they are both popular among young people in Taiwan. Dcard is the largest anonymous social media platform in Taiwan with over eighteen million unique visitors per month, and there have been over ten million Taiwanese Instagram users until 2022. On the other hand, English users' data is from Twitter and Instagram. According to the statistics[1], both are on the list of the top five social media platforms in the US (Instagram and Twitter are more closely related to microblog/social media platforms on the list). Dcard data were collected from the public Dcard API, Instagram data were crawled from the Instagram-scraper, and Twitter data were collected from the Twitter API using tweepy [2] Python package2. All data were randomly collected from October 2021 to July 2022 with no repeat. The texts in Dcard articles, Instagram posts and tweets were splitted into sentences, and only sentences with one emoji remained. Since the number of Taiwanese Mandarin sentences (23646 sentences) exceeds the number of English ones (17876 sentences), the Taiwanese Mandarin data were randomly selected from the original data in order to make the two dataset have equal amounts.

[1] https://www.oberlo.com/statistics/most-popular-socialmedia-in-the-us

[2] https://docs.tweepy.org/en/stable/api.html

Therefore, both dataset contain 17876 sentences respectively, with one emoji in each sentence.

Data pre-processing: To clean the data, irrelevant and redundant information like hashtags (#happy), URLs, user tags(@username) and spams were deleted. A sentence is made to a sentence pair, one with the emoji and one without emojis.

Multilingual language model: The language model adopted here is XLM-T, which is trained on XLM-R (Conneau et al., 2020), and then finetuned for various monolingual and multilingual applications. Emoji plays an important role in this model, which is applicable to explore the impact of the emojis on texts. XLM-T and associated data is released at https://github.com/cardiffnlp/xlm-t. I use the NLP pipeline in huggingface[3]. The output of each sentence contains 3 labels (positive, neutral, negative) with three scores being probability distribution.

Measuring frequency: The quantity of each emoji is divided by the sum of total quantities of emojis for two languages.

Measuring frequency of Unicode categories: The quantity of each emoji is divided by the sum of total quantities of emojis in each Unicode emoji category for two languages.

Measuring the degree of sentiment enhancement and sentiment modification: Relative entropy or Kullback-Leibler Divergence (Kullback and Leibler, 1951) is a method of comparing probability distributions over the same variables. Higher values of the divergence mean less similarity between the distributions. It can be used to quantify the change between sentence pairs. To measure the degree of sentiment enhancement and sentiment modification, all sentence pairs in two languages are grouped into four categories. For positive sentiment enhancement, both sentences in the same pair must be labeled with "positive", while both sentences must be labeled with "negative" in the negative sentiment enhancement category. On the other hand, for positive sentiment modification, the sentence without emoji in a pair is labeled with "negative", and the sentence with the emoji is labeled with "positive". For negative sentiment modification, the sentence without emoji in a pair is labeled with "positive", and the sentence with the emoji is labeled with "negative". Applied to the comparison of sentence pairs in four categories, KLD gives us an indication of the degree of sentiment difference between two languages as well as the features that are primarily associated with a difference. In addition, the Spearman correlation coefficient (SCC) of the emojis' degree (those appearing in both languages) in four categories are measured. The SCC is abbreviated as "r s".

Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Taiwan															
Ratio	6.33	3.94	3.6	2.35	2.32	2.2	2.11	2.1	1.64	1.56	1.44	1.27	1.24	0.97	0.9
US															
Ratio	6.76	5.03	3.8	3.49	2.59	2.01	1.94	1.44	1.25	1.21	1.05	0.98	0.92	0.92	0.83

Figure 1: Top 15 frequent emoji in Taiwanese Mandarin and English, their rank order correlation is 0.767.

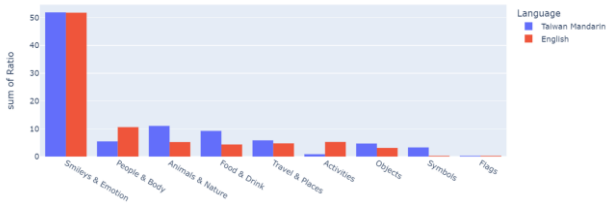

Figure 2: Frequency of emojis grouped by Unicode categories.

Category	Spearman Correlation Coefficient
Smileys & Emotions	0.84
People & Body	0.73
Animals & Nature	0.42
Foods & Drink	0.61
Travel & Place	0.46
Acitity	0.76
Objects	0.38
Symbols	0.63
Flags	0.32

Figure 3: The correlation of the frequency of emoji usage in each category across Taiwanese Mandarin and English.

4 Results and Discussion

Frequency of emoji usages: Figure 1 shows 20 most frequently seen in both languages. Across two languages, Spearman correlation coefficient (SCC) is 0.767, indicating that two groups of different language users favor similar types of emoji. "Face with tears of joy" emoji has high ranks in two languages. Figure 2 and Figure 3 show the

[3] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-basesentiment

frequency of emoji categories and their SCC values. The values range from 0.32 to 0.74, depending upon categories. Not surprisingly, the frequency of emojis in "Smiley and Emotions" exceeds other categories greatly in both languages. There are relatively low correspondences in "Animal & Nature", "Travel & Place", "Objects" and "Flags". To drill down the details, the data shows that Taiwanese Mandarin users use more animals while English users use more plants. In the "Activity", the highly frequent emojis are related to birthday or celebration, which is quite different from the previous research (Guntuku et al., 2019). The correlation of emojis in "Activity" category in their research across the east and the west users is low.

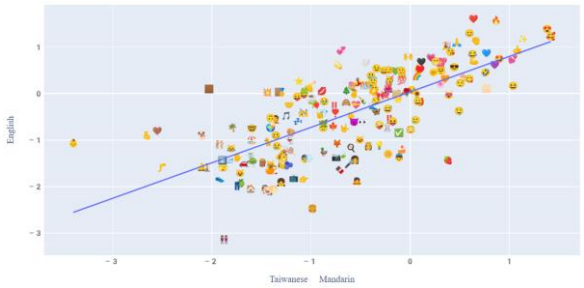

Figure 4: The scatter plot of emojis used by Taiwanese Mandarin users and English users in positive enhancement. ($r_s = 0.604$)

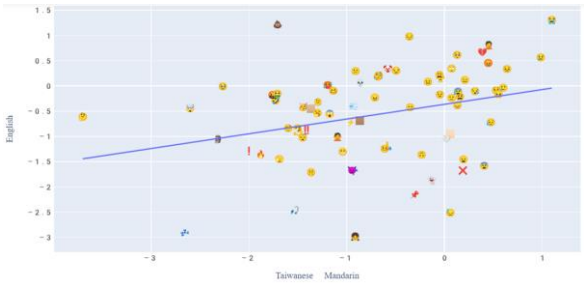

Figure 5: The scatter plot of emojis used by Taiwanese Mandarin users and English users in negative enhancement. ($r_s = 0.547$)

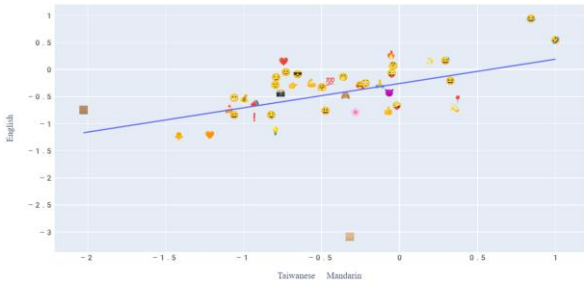

Figure 6: The scatter plot of emojis used by Taiwanese Mandarin users and English users in positive modification. ($r_s = 0.751$)

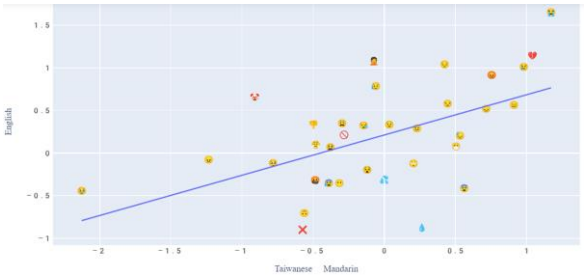

Figure 7: The scatter plot of emojis used by Taiwanese Mandarin users and English users in negative modification. ($r_s = 0.771$)

Sentiment Enhancement: Figure 4 and Figure 5 show the scatter plots of emojis for sentiment enhancement in both languages. In Figure 4, 192 types of emojis are used for positive enhancement, the amount is more than emojis used for negative enhancement (only 74 types). Moreover, smileys and hearts highly increase positive feelings in both languages, whereas emojis in other categories also enhance the positive sentiment. Therefore, emojis are frequently used in positive feelings and convey positive feelings in general. In Figure 5, most of the emojis for negative enhancement are classified into negative emojis (sad faces and angry faces), which are considered to increase negative feelings. And most of the emoji types for negative enhancement belong to the "Smiley and Emotion". The percentage of emoji types in "Smiley & Emotion" category is 22.3% for positive enhancement and 72.3% for negative enhancement. And the total number of smileys in positive enhancement and negative enhancement across two languages are over 99%.

Sentiment Modification: Figure 6 and Figure 7 show the scatter plots of sentiment modification in both languages, the usages for this purpose across two languages have relatively high correlation, implying that the effects of emojis might surpass the texts in both languages. Compared with sentiment enhancement, the emoji types for sentiment modification are relatively low with 42 types for positive modification and 34 types for negative modification. The percentage of emoji in the "Smiley & Emotion" category is 47.6% for positive modification and 79.4% for negative modification. Similar to negative sentiment enhancement, the percentage of smileys faces and hearts emojis are higher than the positive ones. The total number of smileys in positive modification and negative modification across two languages are over 99%.

5 Conclusion

In this research the sentiment degree of emojis used by Taiwanese Mandarin users and English users is compared. For the similar part, the usages of smileys and hearts support the agreement that emojis can be used universally. While there is no significance in the difference in two languages due to data amount. These are only preliminary results, more extensive analyses of the function of emojis are planned to run further.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM '11)*. Association for Computational Linguistics, USA, pages 30–37.

Subashini Annamalai and Sobihatun Abdul Salam. 2017. Undergraduates' Interpretation on WhatsApp Smiley Emoji. *Jurnal Komunikasi, Malaysian Journal of Communication.* 33(4):79-103. https://doi.org/https://doi.org/10.17576/JKMJC-2017-3304-06.

Serkan Ayvaz and Mohammed O. Shiha. 2017. The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering*. 9. 360-369. https://doi.org/10.17706/IJCEE.2017.9.1.360-369.

Francesco Barbieri, Luis Espinosa Anke and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. LREC 2022. https://doi.org/10.47550/arXiv.2104.12250.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. Language Resources and Evaluation conference, LREC, Portoroz, Slovenia, May 2016.

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How Cosmopolitan Are Emojis?: Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics. *MM '16: Proceedings of the 24th ACM International Conference on Multimedia*, Oct 2016, pages 531–535. https://doi.org/https://doi.org/10.1145/2964274.2967277.

Owen Churches, Mike Nicholls, Myra Thiessen, Mark Kohler, and Hannah Keage. 2014. Emoticons in mind: an event-related potential study. *Social neuroscience*, 9(2), pages 196–202. https://doi.org/10.1070/17470919.2013.773737.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7440–7451. https://doi.org/10.47550/arXiv.1911.02116.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China. Coling 2010 Organizing Committee.

Vyvyan Evans. 2017. The Emoji Code: The Linguistics Behind Smiley Faces and Scaredy Cats. Picador.

Boting Gao and Doug P VanderLaan. 2020. Cultural Influences on Perceptions of Emotions Depicted in Emojis. *Cyberpsychology, Behavior, and Social Networking*, 23(7):567-570. https://doi.org/10.1079/cyber.2020.0024.

Gaël Guibon, Magalie Ochs, and Patrice Bellot. From Emojis to Sentiment Analysis. 2016. *WACAI 2016, Lab-STICC; ENIB; LITIS*, Jun 2016, Brest, France.

Sharath C. Guntuku, Mingyang Li, Louise Tay, and Lyle H. Ungar. 2019. Studying Cultural Differences in Emoji Usage across the East and the West. *Proceeding of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019*).

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13)*. Association for Computing Machinery, New York, NY, USA, 703–710. https://doi.org/10.1145/2470362.2470497.

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*, Sep 2009, 19:1543-1547. https://doi.org/10.1016/j.cub.2009.07.051

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–76.

Lou Yinxia, Yue Zhang, Fei Li, Tao Qian and Donghong Ji. 2020. Emoji-Based Sentiment Analysis Using Attention Networks. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 19, pages 1-13. https://doi.org/10.1145/3379035.

Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. 2017. Facebook sentiment: Reactions and Emojis. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 11-16, Valencia, Spain. Association for Computational Linguistics.