Analysis of Emoji Use in Response to News Videos

Angela Zhang (az2542@columbia.edu)
Supervisor: Dr. John R. Kender
Columbia University

ABSTRACT

As emojis have gained widespread popularity over the last decade, many linguists and researchers have opened discussion about how their usage in digital communication can be analyzed and studied. Furthermore, the question of how emojis may take on different meanings when utilized across different cultures has also been brought to the forefront of this conversation. Can analysis of emoji usage across social media mediums from different cultures tell us anything significant about these cultures? In this paper, an analysis of US-based Tiktok accounts and emoji usage in response to COVID-19 related content will be performed. We also preliminarily observe the Chinese version of Tiktok, known as Douyin, and how usage of emojis on this alternate platform might be able to help us understand cultural differences between the United States and China.

1. INTRODUCTION

1.1 News on Social Media

In 2019, the social media platform known as TikTok experienced a massive increase in popularity, particularly amongst the younger generation. Owned by the Chinese tech company ByteDance, TikTok is a short video platform which initially gained popularity for dancing and lip syncing videos [1]. Continuing its popularity boom into 2020 and over the COVID-19 pandemic, TikTok began to see a vast diversification in content and the demographic of users that were joining the platform to create content. Established news outlets and broadcasting networks such as CBS, NBC, The Washington Post, and CNN all have verified (confirmed affiliation with the companies) profiles on TikTok and post a variety of content tailored for the short video medium. The integration of news reporting on social media presents an interesting topic of discussion regarding the adaptation of traditional approaches to journalism in order to appeal to younger generations on such platforms. These companies must balance informing audiences with creating content that will succeed on the underlying algorithms that ultimately decide which pieces of content will be shown to which users [2].

1.2 Tiktok and Douyin

In order to set up a cross cultural analysis, it is important to select platforms that are as similar as possible to one another. Differences in how a platform looks or is presented can cause differences in the way users may interact with content. This was an additional factor in the decision to focus on TikTok, which is actually the international version of ByteDance's Douyin app. Douyin, originally released in 2016 to the Chinese market, also utilizes the short video format [3]. Some of China's most prominent news outlets and broadcasting channels also have verified accounts on the platform. In order to better understand cultural differences between the US and China, I have decided to focus on analyzing content from both Tiktok and Douyin regarding the COVID-19 pandemic, a topic that is certainly widely reported on in both countries. It is important to note that TikTok has a global audience and is not strictly limited to the US. For this reason, I will be focusing on TikToks accounts managed by US-based news outlets and broadcasting services.

1.3 Emojis

Paralleling the widespread growth of social media is the increase in the use of emojis- pictorial icons that can be used alongside natural language in typed text. First introduced in 1997 on Japanese mobile phones, emojis have become a worldwide phenomenon since being introduced as a part of the Unicode standard in the early 2010s [4]. Growing usage of emojis has sparked many conversations about what we can ascertain about a group of people based on how they incorporate emojis into their online interactions. Previous work by Luvena Huo on the comparison of English and Chinese natural language comments under news videos related to the COVID-19 pandemic has shown probable differences in how these two cultures responded to the ongoing health crisis [5]. Because emojis are so widely used on social media, I have decided to analyze how emoji use tends to differ between US Tiktok users and Chinese Douyin users and what this might mean regarding cultural differences between the two countries.

Something important to note is that many popular Chinese social media platforms, including Weibo, WeChat, and Douyin incorporate a mix of unicode standard emojis as well as certain platform specific emojis denoted by custom shortcode descriptions [6]. Figure 2 illustrates one such specialized emoji.

2. RELATED WORK

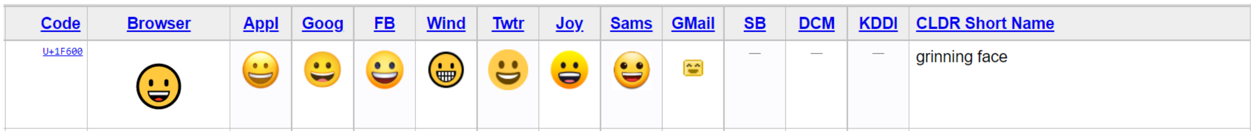
For emoji analysis across cultures, Guntuku et al. conducted a study to compare emoji usage between the East (China and Japan) and West (United States, United Kingdom, Canda) [7]. We use many of the algorithms and methodologies detailed in the Guntuku paper for our own analysis. One thing that was not mentioned in the Guntuku paper was the aforementioned platform specific emoji sets used widely on Chinese social media apps, which introduces an entirely new element of a different culture and also poses challenges with regard to data analysis.

In terms of topic exploration, we look to previous works introduced by Luvena Huo. Huo compares video comments across the US and China regarding COVID-19, and this project aims to reveal what can be found when shifting the focus to social media news [5].

3. BACKGROUND

3.1 Emoji2Vec

To assist in creating vector representations for emojis, we look to Emoji2Vec, a set of pre-trained emoji vector embeddings created from the shortcode descriptions of a given emoji [8]. For example, in the Full Emoji list which can be viewed on the Unicode website, each emoji is associated with both a unicode encoding as well as a short, natural language based description [9]. Given that unicode is not readily human readable, the natural language description is one way to help formulate an idea of what a certain encoding is meant to convey or depict.



Code	Browser	Appl	Goog	FB	Wind	Twtr	Joy	Sams	GMail	SB	DCM	KDDI	CLDR Short Name
U+1F600	😀	😀	😀	😀	😀	😀	😀	😀	😀	—	—	—	grinning face

Figure 1: A snippet taken from the emoji list on the Unicode website. Each row contains the unicode for a specific emoji as well as a natural language short description. Also shown are the different art interpretations when the emoji is used on certain platforms and web browsers [9].

Prior to formulating emoji vector representations using these short descriptions, many prior works on this subject focused on using a skip-gram model over a large set of emojis and the Tweets or other social media posts they appeared with to derive the context of the emojis. However, Eisner et al. found that Emoji2Vec outperformed these methods on a Twitter sentiment analysis task. Furthermore, Emoji2Vec holds the advantage of needing much less data to create an embedding. It also is not dependent on the input data to create a complete set of embeddings [8].

However, the key limitation of Emoji2Vec is that it will only fetch embeddings when queried by a unicode standard emoji. As mentioned before, many Chinese platforms use a different form of emoticon, which I did not find to be easily accessible using US services. For example, using the Chrome inspect tool on TikTok reveals that emojis are displayed as a unicode symbol, but on the Douyin site, many of the specialized emojis are displayed as image files instead. This means that it is not yet possible to use emoji2vec to generate embeddings for specialized Chinese emojis.

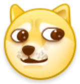


Figure 2: The above 'doge' emoji is and example of one of the platform specific emojis that is not available in the US [6]

However, this is not a huge setback as Chinese emojis also have known natural language descriptions. On Douyin, these can be accessed by using the inspect tool on a Chrome browser. There are also lists of compiled emoticons and their corresponding short descriptions which can be utilized in the computation of an emoji's vector representation. Because Emoji2vec's underlying representation utilizes Google's word2vec embeddings to form embeddings for an emoji, we can manually form our own embeddings as long as we have access to the natural language description [8]. While we do use the emoji2vec embeddings for our analysis of US Tiktok videos, it is suggested to shift to the manual method once the project expands to incorporate analysis of Douyin videos as well.

3.2 Category Association

In their paper detailing emoji use between the East and West, Gunutuku et al. performed an analysis based on emoji association using LIWC (Linguistic Inquiry and Word Count) categories. This dictionary, which exists for both English and Mandarin, consists of several categories such as 'money', 'family', 'health', 'posemo', 'negemo' (positive and negative emotion) as well as word associations which have been psychologically validated [7]. While the LIWC lexicon is commonly used and widely validated, this project utilizes a much smaller, topic-specific dataset. In order to avoid random association with categories that aren't directly related to the topic of COVID-19, I decided to move forward with forming category associations with the NRC Emotion Lexicon. The NRC lexicon, developed by Mohammad and Turney for the National Research Council of Canada, similarly associates words with certain categories, but instead forms the categories based on the set of basic emotions as proposed by Robert Plutchik: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy [11]. The NRC Lexicon additionally recognizes the categories of positive and negative emotion, making a total of 10 categories [10].

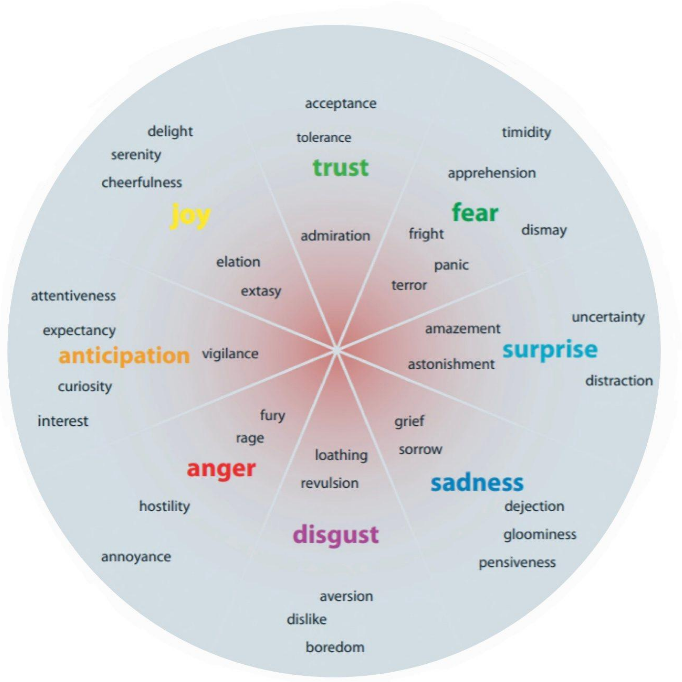


Figure 3: Plutchik's 8 categories of basic emotions. Opposite emotions are placed across from each other on the wheel [11]. Figure taken from http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf

4. METHODOLOGY

4.1 Initial Exploration - Platform Comparison and Video Selection

Although Douyin and Tiktok are very similar, there are a few differences that were discovered upon an initial exploration of both the user interface as well as some underlying demographics. Having an account on one platform does not mean that you have an account on the other- in fact, Douyin is generally not available for mobile download in US locations. Both platforms also support web access.

The top 10 most followed accounts on TikTok are mostly owned by 'social media personalities', individuals most known for the content that they post on social media platforms [11]. Almost half of the top 10 began their careers on TikTok and did not have any significant following prior. In contrast, the top 10 most followed accounts on Douyin consists primarily of news broadcast channels as well as well established actors and actresses whose careers in the public eye predate Douyin. TikTok's most followed account is owned by Charli D'Amelio, a 17 year old social media personality, while Douyin's most followed account is owned by People's Daily, the most prominent newspaper group in China.

Additionally, Douyin recently implemented a time limit of 40 minutes of usage a day between 6am and 10pm for users aged 14 and under to try and prevent internet addiction among the younger users [13]. Tiktok currently has no such ban.

On average, news based accounts on Douyin had many more followers than those on Tiktok.

Because of the many videos on both Tiktok and Douyin regarding the pandemic, I decided to narrow the search to target news about the COVID vaccine specifically. One video from each platform was selected for an initial comparison. From Douyin, a video from the CCTV Quick Take account was selected, while a video from CBS This Morning was chosen from Tiktok [14] [15]. Each video showcases a reporter relaying news of the authorization and distribution of the COVID-19 vaccination for children. The Tiktok video showcases the reporter for a much shorter period of time, opting to show a string of seemingly generic photos of vaccines and schools in the latter half of the video.

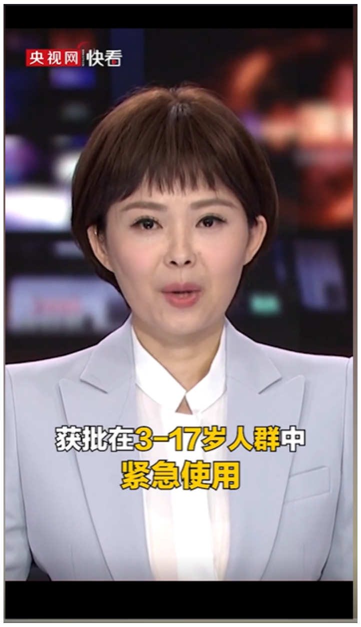


Figure 4: A screenshot from a Douyin post regarding COVID vaccinations for children aged 3-17. This was taken from the CCTV Quick Take Douyin account [15]

Figure 5: Two screenshots taken from a Tiktok post regarding COVID vaccinations for children. These were taken from the CBS Mornings account [14].

4.2 Data Collection

To create the dataset with which we would use to extract the categorical emoji associations, I utilized the comments under each video. As there is not currently any readily available software to scrape Douyin comments from the web UI and manual collection of Douyin comments, emojis, and emoji shortcode description was not possible in the time frame, the focus of this project lies with the analysis of Tiktok and US emoji use.

To scrape Tiktok comments, a preexisting software was used to extract 150 comments under each video collected [16]. Due to software limitation, a comment was only added to the final dataset if it was made from a public account. A public account refers to an account whose content is able to be accessed by anybody on the app. In contrast, a private account is only viewable by followers of that account. Each comment did not necessarily contain the use of an emoji. From the three videos selected from Tiktok, a total of 249 comments were successfully extracted and added to the corpus.

4.3 Pre-processing

Utilizing nltk, stopword removal was performed on the corpus to take out commonly used words that would contribute little meaning to finding a vector representation for each NRC Lexicon category. Common punctuation was also removed before tokenization using nltk's work_tokenize function.

We also extract all emojis used in the compiled comments and store them in a separate list for use in later steps. The 'emoji' Python library was utilized for emoji extraction. The emoji list for our corpus can be found below:

['🤔', '😅', '😏', '😳', '🤣', '😂', '☠️', '👏🏽', '😁', '😬', '🤬', '💪🏼', '💜', '☀️', '🤗', '🙏🏾', '🙏', '🥰', '😐', '😩', '🥺', '🥴', '🇺🇸', '❤️', '🙌', '🙄', '👏', '👍', '🤷🏽', '💯']

4.4 Creating Category Vectors

Following the methodologies outlined in Guntuku et al. for creating "category vectors" for each LIWC category, I created similar vectors for each of the 10 NRC categories [7]. First, each word was sorted into its corresponding NRC category. There are multiple possible categories for a single word, so it is possible to have repeat words between categories. An example category-to-word-list association from our corpus is shown below:

```
'fear': {'afraid','bad','crazy','deadly','destroyed','die','disappear','dying',
'emergency','fear','feeling','fever','forced','god',
'government','hell','hide','horrible','hurt','injection','kidnap','kill',
'manipulation','missing','nervous','pain','paralyzed','problem','risk',
'shooting','shot','syringe','volunteer','wan','watch','weight','worse'}
```

A word2vec model was trained on the full corpus to create the embeddings for each individual word. Because emoji2vec returns 300-dimensional vectors, the specified size for the model was also 300 to ensure compatibility between the category vectors and emoji vectors.

For each category i in the NRC emotion lexicon, $i \in \{anger,\ fear,\ joy,\ ...\}$ the sum of the word2vec embeddings for each word in i is taken. This sum is then divided by the number of words in the category to create an average vector for category i. This vector is known as $\vec{C_i}$.

4.5 Emoji Embedding

For proof of concept with regard to creating emoji associations with the NRC categories, we use Emoji2Vec to derive the embeddings for all of the emojis in our corpus. To expand this analysis to include Chinese specialized emojis, manually creating the emoji embeddings is recommended. For both countries, an emoji embedding would be created by summing the word2vec vectors for each word found in that emoji's short description.

Because we are directly utilizing the pretrained emoji2vec model, we can directly query the model using an emoji, j, from our emoji list. This will return a 300 dimensional vector representation for our emoji, denoted $\vec{e}_j$.

Because the pretrained emoji2vec did not contain embeddings for all of the emojis found in our corpus, the following emojis were excluded from further analysis.

☠️🤬🥰🥺🥴🤷🏽‍♂️

Figure 6: The list of emojis what were not captured by emoji2vec

4.6 Cosine Similarity and Category Association Formation

Now that we have $\vec{C}_i$ for each NRC emotion category and $\vec{e}_j$ for each emoji in our corpus, we can utilize scipy to compute the cosine similarity between each category and each emoji to determine the more heavily associated emojis and categories. After the similarity rankings were determined, the top three most associated emojis with each category were extracted. The result from our corpus is displayed below.

```
{'fear': ['🤔', '😅', '😏'], 'anger': ['🤗', '🙌', '😏'], 'anticipation': ['🇺🇸', '👏🏽', '💪🏼'], 'trust': ['😩',
'👏🏽', '😬'], 'surprise': ['💪🏼', '🙏🏾', '❤️'], 'positive': ['😏', '🤣', '👏🏽'], 'negative': ['💜', '😩', '🙏'],
'sadness': ['💜', '😩', '❤️'], 'disgust': ['🙌', '🙏', '🤗'], 'joy': ['🇺🇸', '😁', '🤗']}
```

5 RESULTS AND FUTURE WORK

The result from our emoji to emotion category at first glance doesn't seem quite intuitive. One reason for this is because the dataset we collected was very small, and our emoji list was not very extensive and we can see many emojis repeatedly being included in some categories.

There has also been discussion regarding bias within the NRC Emotion Lexicon, which certainly would affect category association. There has been work to correct these biases in the lexicon, and future work regarding this project can utilize the new lexicon [16].

Another interesting aspect to consider is the possible intention of sarcasm when using emojis. Given that many of the comments under the US Tiktok videos were seemingly negative which has led to many seemingly positive sentiment emojis being associated with the fear, anger, and sadness categories. The misalignment of emojis and emotion categories could also be a good indicator that TikTok users in the US don't always use emojis in the most straightforward and intuitive way. A good example of this is the 😭 emoji, which has a short description of "loudly crying face". One might expect this to be associated with the feelings of sadness; however among the younger demographic on Tiktok, this emoji has shifted to represent more of an exasperated laughter. Slight malalignments between an emoji's shortcode description and the way it is actually used on social media could cause discrepancies in category association.

To further continue this work, the proposed next step would be to perform this analysis for US Tiktok on an expanded dataset with a more diverse set of emojis. Expanding this work to include analysis of Douyin emojis would require a slight adjustment in the way the emoji embeddings are calculated because Chinese platform-specific emojis are not encoded in the same way as the unicode standard emojis. Furthermore, by calculating emoji embeddings this way, we can be guaranteed to have an embedding for each emoji in the corpus. The emojis that were not recognized by emoji2vec in our analysis did not have a chance of being included in the category association.

6 CONCLUSION

To summarize our findings, we were able to adapt Guntuku's method for emoji analysis across cultures to our study, which focuses on the topics of COVID-19 as reported by US-based news Tiktok accounts. Our initial findings map the top three emojis which hold the most similarity with each of the categories defined by the NRC Emotion Lexicon. While there is certainly more work to be done to complete the cross culture analysis between the US and China, our initial findings and result is helpful to form the foundational understanding of how to create category vectors and emoji embeddings.

References:

[1] Kaye DBV, Chen X, Zeng J. The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. Mobile Media & Communication. 2021;9(2):229-253. doi:10.1177/2050157920952120

[2] Vázquez-Herrero J, Negreira-Rey M-C, López-García X. Let's dance the news! How the news media are adapting to the logic of TikTok. Journalism. October 2020. doi:10.1177/1464884920969092

[3] "TikTok, WeChat and the growing digital divide between the US and China". *TechCrunch*. Archived from the original on 11 January 2021. https://social.techcrunch.com/2020/09/22/tiktok-wechat-and-the-growing-digital-divide-between-the-u-s-and-china/

[4] "Standard Emoji keyboard arrives to iOS 5, here's how to enable it". *9to5Mac*. June 8, 2011. https://9to5mac.com/2011/06/08/standard-emoji-keyboard-arrives-to-ios-5-heres-how-to-enable-it/

[5] Huo, Luvena. Cross-cultural Differences in Responses to News Videos. http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/21y_Luvena.pdf

[6] "Sina Weibo". *Emojipedia.* https://emojipedia.org/sina-weibo/

[7] Guntuku, Sharath Chandra, Mingyang Li, Louis Tay and Lyle H. Ungar. Studying Cultural Differences in Emoji Usage across the East and the West. *ICWSM* (2019).

[8] Eisner, Ben, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak and Sebastian Riedel. emoji2vec: Learning Emoji Representations from their Description. *ArXiv* abs/1609.08359 (2016): n. pag.

[9] "Full Emoji List, v14.0" Unicode. https://unicode.org/emoji/charts/full-emoji-list.html

[10] Saif M Mohammad and Peter D Turney. 2013b. Nrc emotion lexicon. National Research Council, Canada, 2.

[11] Plutchik, Robert. (1980). Emotion: A Psychoevolutionary Synthesis. Harper & Row, Publishers.

[12] "Top 50 Most Followed TikTok accounts in 2021 | Insiflow". *Social Media Marketing & Management Dashboard*. https://insiflow.com/tiktok/top

[13] "Chinese version of Tiktok limits use of app by those under 14". *Reuters.* https://www.reuters.com/technology/chinese-version-tiktok-limits-use-app-by-those-under-14-2021-09-18

[14] CBS Mornings [@cbsmornings]. (2021, September 20). *Pfizer says its COVID-19 vaccine is safe and effective in kids ages 5 -11. #news #coronavirus #pfizer #covid19 #parentsoftiktok #kids* [Video]. TikTok. https://vm.tiktok.com/ZM8NmMnVa/

[15] 央视网快看 [@kuaikancctv]. (2021, July 17). 重磅！国药集团新冠病毒灭活疫苗获批在*3-17*岁人群中紧急使用 [Video]. Douyin. https://v.douyin.com/dWwvrSv/

[16] Export Social Media Comments. https://exportcomments.com/

[17] Zad, Samira & Jimenez, Joshuan & Finlayson, Mark. (2021). Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon. 102-113. 10.18653/v1/2021.woah-1.11.