# Emoji and Self-Identity in Twitter Bios

**Jinhang Li,**[*] **Giorgos Longinos,**[*] **Steven R. Wilson** and **Walid Magdy**
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
{j.li-183,g.longinos}@sms.ed.ac.uk
steven.wilson@ed.ac.uk, wmagdy@inf.ed.ac.uk

## Abstract

Emoji are widely used to express emotions and concepts on social media, and prior work has shown that users' choice of emoji reflects the way that they wish to present themselves to the world. Emoji usage is typically studied in the context of posts made by users, and this view has provided important insights into phenomena such as emotional expression and self-representation. In addition to making posts, however, social media platforms like Twitter allow for users to provide a short bio, which is an opportunity to briefly describe their account as a whole. In this work, we focus on the use of emoji in these bio statements. We explore the ways in which users include emoji in these self-descriptions, finding different patterns than those observed around emoji usage in tweets. We examine the relationships between emoji used in bios and the content of users' tweets, showing that the topics and even the average sentiment of tweets varies for users with different emoji in their bios. Lastly, we confirm that homophily effects exist with respect to the types of emoji that are included in bios of users and their followers.

## 1 Introduction

With the rise of social media usage and online text-based communication, emoji, a simple but powerfully expressive set of visual characters (Danesi, 2016), have become a hugely popular means to express emotions, moods, and feelings over computer-mediated communication (Kelly and Watts, 2015). In the era of big data, with more and more people engaging with social media, researchers have begun to study the ways in which social media users include emoji in their posts, finding that emoji usage is associated with things like personality (Li et al., 2018), culture (Guntuku et al., 2019),

[*] Authors contributed equally.

and socio-geographical differences (Barbieri et al., 2016).

Prior work has typically focused on how people use emoji within the posts that they make online (Ljubešić and Fišer, 2016; Robertson et al., 2018), or the way that they can be used as reactions to other content (Tian et al., 2017). However, emoji are also commonly used within user's self-created profiles. In this work, we specifically examine the inclusion of emoji in Twitter bios, which are short (160 characters maximum) texts describing a Twitter account. These bios are featured prominently on a user's profile page, and given their limited length, users often use this space succinctly express the essential information about their accounts. Therefore, we expect that the choice of emoji used in these bios will have a strong connection to a user's online self-identity, or the way that they seek to portray themselves to others on a social media platform.

The goal of this paper is to give an overview of how emoji are used in Twitter bios from a computational linguistics perspective, that is, we treat emoji as a special category of tokens and make use of natural language processing methods to understand the major trends in the ways that people use emoji in their bios and what this says about both the things they tweet about and their follower network. Our results provides insights into the variety of ways in which people choose to present themselves online in their Twitter bios that may be overlooked when only considering non-emoji word tokens or only considering the ways that people use emoji in the content of tweets. More specifically, we ask, and subsequently describe the work done to answer, the following research questions:

**RQ1. How are emoji used in Twitter bios?** As a first step, we seek to characterize the ways in which users use emoji in their bios. We look at the types of emoji that most commonly used in Twitter bios,

and the position within the bios that emoji appear. We compare our findings to trends from the usage of emoji in tweets by the same set of users and note the differences.

**RQ2. What is the relationships between the emoji in a user's bio and the content that the user posts?** Next, we explore the correlations that exist between the choice of emoji to be included in a user's bio and the content that that user tweets about. We consider this from the perspectives of word-level patterns, topic usage, and overall tweet sentiment.

**RQ3. Do users and their followers use emoji in their bios in a similar way?** Last, we investigate the homophily of emoji usage with bios by studying the follower networks of our core set of users. We look at the similarities in both the absence or presence of emoji in users' bios as well as particular choices of emoji used.

## 2 Background

### 2.1 Online Self-Identity

Self-identity, or self-concept, is a collection of firm and noticeable beliefs about oneself (Sparks and Shepherd, 1992). From a general perspective, self-identity gives the answers to the question "Who am I?". Many components make up self-identity together. The self-categorization theory asserts that the self-identity consists of at least two types of self-categorization: personal identity (what makes me unique?) and social identity (which groups do I belong to?) (Guimond et al., 2006).

As social attributes are inherent, people reveal their self-identity when they communicate with others or interact with the outside world (Fisher et al., 2014). Expressing themselves is also a way for people to establish connections and bonds with the world. Therefore, social media provides a natural opportunity to study self-identity. Previous studies have shown that specific personality characteristics can be measured by analyzing linguistic behavior on social media using natural language processing techniques (Plank and Hovy, 2015). Other work analyzed the words, phrases, and topics collected from the Facebook messages, and linked these to personality traits and demographics of users (Schwartz et al., 2013). Twitter bios have been shown to be are particularly useful in discovering other aspects of self-identity such as political and religious affiliations (Rogers and Jones, 2019).

### 2.2 Self-representation in Emoji

While many studies related to online self-identity are based on the analysis of textual features, others have turned to emoji as important signals of users' identities. In one study, researchers looked at Twitter names and bios, uncovering stark differences in the emoji use of groups supporting and opposed to white nationalism (Hagen et al., 2019). Graells-Garrido et al. (2020) found that in two South American countries, different colour variations of heart emoji indicated users' opinions about abortions: tweets containing the green heart emoji '💚' were more likely to convey support of women's rights, while the blue heart emoji '💙' was more associated with stronger restrictions of abortions. In another study, researchers explored differences in emoji usage across cultures, finding that users from western countries tend to use more emoji than users from eastern countries (Guntuku et al., 2019). Although there were specific emoji that were found to be culturally specific (e.g. cooked rice '🍚'), it was suggested that many common emoji have similar meanings across cultures.

It has been shown that usage of some emoji are also correlated with aspects of identity such as personality traits (Völkel et al., 2019), and the use of skin-tone modifiers in emoji has been linked to greater feelings of self-representation online, with no evidence that the skin-tones in emoji correlated with the expression of racist views online (Robertson et al., 2018, 2020). Other work found gender stereotypes in the use of male and female emoji modifiers: male modifiers were more frequently used in emoji related to business and technology while female modifiers were used in emoji related to love and makeup more often (Barbieri and Camacho-Collados, 2018).

## 3 Data

For our study, we sampled users from Twitter who tweeted between April and July 2020. Using the Twitter streaming API, we began collecting tweets and storing all user-level information available for

| Dataset | Users | Tweets | Retweets |
|---|---|---|---|
| EmojiBio | 20,000 | 2,998,219 | 1,568,661 |
| NonEmojiBio | 2,000 | 491,646 | 247,800 |
| Followers | 7,105,521 | 425,704,661 | 169,935,436 |

Table 1: Number of users, tweets, and retweets (subset of tweets) in our datasets.

| Bios | | Tweets | |
|---|---|---|---|
| **Emoji** | **Appearances** | **Emoji** | **Appearances** |
| ❤️ | 1311 | 😂 | 115999 |
| ✨ | 1152 | 😭 | 65410 |
| 💙 | 965 | ❤️ | 48411 |
| 💜 | 720 | 😳 | 40892 |
| 💛 | 559 | 😍 | 34368 |
| 🖤 | 547 | ✨ | 33766 |
| 💕 | 543 | 🤣 | 33173 |
| 🌈 | 490 | 🥰 | 23037 |
| 💚 | 467 | 💕 | 17040 |
| ⚽ | 326 | 🔥 | 16929 |

Table 2: The most frequently used emoji in the bios and tweets of the emojiBio dataset.

| Group Name | Num. Emojis | In Bios | User ratio | Examples |
|---|---|---|---|---|
| People & Body | 2485 | 745 | 20.0% | 🧑‍🤝‍🧑🙏🧑🤱🕴️ |
| Symbols | 301 | 229 | 15.4% | 🚫➡️✝️❗🔴 |
| Objects | 299 | 219 | 15.9% | 🎙️📚🏷️🔪🛏️ |
| Flags | 275 | 215 | 16.5% | 🏁🏳️‍🌈🇺🇸🇨🇦🏴󠁧󠁢󠁳󠁣󠁴󠁿 |
| Travel & Places | 264 | 206 | 14.9% | ⛵🏯🏰🍌🌈 |
| Smileys & Emotion | 162 | 151 | **44.3%** | 😀🏠💕❤️ |
| Animals & Nature | 147 | 132 | 18.9% | 🐶🐝🦋🌹🌿 |
| Food & Drink | 131 | 117 | 5.5% | 🍑🍔🍰🍵 |
| Activities | 95 | 82 | 15.2% | ✨🎁⚽♣️❤️ |

Table 3: Emoji groups present in Unicode Emoji v13.0, number of unique emoji in the group, number of unique emoji used at least once in a bio in the userBios dataset, the percentage of users who use at least one emoji from the corresponding group in their bio, and examples of emoji from the group.

each tweet, including the bio. In order to filter out both fake or less well-established accounts, we removed all accounts that had less than 100 followers, and to remove celebrity or other widely popular accounts, we filtered out those with more than 1000 followers. From the remaining set of users, we randomly sampled 20,000 users which have at least one emoji in their bios, and collected their most recent 200 tweets, as available, labeling this dataset "emojiBio". We also collected 200 tweets each for a set of 2,000 users who did *not* use any emoji in their bio as a control group, which we label the "nonEmojiBio" dataset. Finally, the "Followers" dataset contains the user-level information and recent tweets of the followers of the users of both the emojiBio and nonEmojiBio datasets. Details about the size of the datasets are presented in Table 1.

As our dataset contains text written in many langauges, we first used the pre-trained fastText language identification model (Joulin et al., 2016a,b) to detect the language that each tweet or bio was written in. The most common languages in our datasets were English, Japanese, Spanish, and Portuguese, followed by others. After identifying the langauges, we tokenized the English-language texts using the NLTK (Loper and Bird, 2002) TweetTokenizer[1] and the texts detected as being written in other languages using the Polyglot multilingual tokenizer.[2]

## 4   Emoji Usage in Bios

First, we sought to characterize the use of emoji in users' bios, so we turn to just the emojiBio dataset.

---

[1] https://www.nltk.org/api/nltk.tokenize.html
[2] https://polyglot.readthedocs.io/

We contrast the most commonly used emoji[3] in bios and in tweets in Table 2, finding that facial expression emoji ('😂', '😭', '🤣', '😍', '😳', '🥰') are more frequently used in tweets, while different variations of heart emoji ('❤️', '💙', '💜', '🖤', '💕', '💚') are more frequently used in bios. Another emoji that is regularly used in bios is the rainbow emoji '🌈'. The sparkles emoji '✨' and the female sign emoji '♀️' (not in top 10) are frequently used in both bios and tweets. We also checked the average position of emoji within users' bios and tweets, and found that in both cases, most emoji appear at the end of the text. These emoji at the end commonly signify the overall meaning or sentiment of the text. However, we noticed that the emoji in bios are, on average, used closer to the middle of the text than emoji that are used in tweets. There is also a nontrivial number of emoji used at the *start* of texts, which happens more often in bios than in tweets. Additionally, we found that is more common for users to use a single emoji as the entire content of a bio than as the entire content of a tweet (more details in Appendix B).

Unicode Emoji 13.0 contains a total of 4,159 emoji in nine groups according to categories. We carried out analysis on emoji based on their predefined groups, and the results are shown in the Table 3. We found that the number of unique emoji in a category is directly correlated with the number of unique emoji from that group that appear in users' bios. However, after calculating the proportion of

---

[3] In their Unicode representations, some emoji with the same visual pattern are represented by different code points for historical reasons, code points can be divided into fully-qualified, minimally-qualified or unqualified (https://www.unicode.org/reports/tr51/). In this paper, we only present the qualified version of a given emoji pattern when reporting results.

Table 4: Mutual information score rank of emoji in the bios group by top 20 emoji.

| Top 20 Emojis | Mutual Information Rank | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| ❤️ | flamengo | apaixonada (in love) | god | love | kids |
| 🤍 | フォロー (follow) | たい (want) | 気軽 (feel free) | よろしく (nice to meet you) | 無言 (silent) |
| 💜 | cabj | boca | juniors | الهلال (crescent moon) | blue |
| 🖤 | bts | ot7 | army | account | fan |
| 💜 | cabj | boca | juniors | news | galatasaray |
| ❤️ | flamenguista | 推 (like) | flamengo | ucf | الاتحاد (the Union) |
| ❤️ | 大好き (like very much) | フォロー (follow) | 無言 (silent) | pop | 参戦 (participation in a war) |
| 🧡 | bi | gay | lgbt | queer | artist |
| ❤️ | pibas | scp | got7 | nct | jaehyun |
| ⚽ | jugador (player) | soccer | football | fútbol (soccer) | atleta (athlete) |
| 💙 | 円 (circle) | fire | خاص (special) | knight | vila (village) |
| 🌈 | 生活 (life) | genderfluid | ロー (low) | 定時 (on time) | 東 (east) |
| 🇺🇸 | maga | trump | american | conservative | kag |
| ❤️ | انسان (man) | caballero (gentleman) | الدعاء (pray for) | قُل (say) | الله (Allah) |
| ❤️ | sc | ig | snapchat | snap | mma |
| 🌸 | crossing | 桜 (cherry blossoms) | 垢 (account) | ライブ (live) | もっと (more) |
| 💍 | married | 元 (former) | clan | crise | motion |
| 📚 | bookstan | derecho (right) | uca | educación (education) | libros (books) |
| 🦋 | diary | vtuber | ライブ (live) | york | 木 (wood) |

Table 5: Mutual information score rank of tokens in the bios group by top 20 emoji, and translate non-English in parentheses.

the users who use at least one emoji from each group, we noticed that most users used emoji from the Smileys & Emotion group in bios, with a total of 44.3% of the 20,000 users, followed by the People & Body group with 20% of users including at least one emoji from that group. On the contrary, the number of users who used the emoji of the Food & Drink group is the least, accounting for only 5% of the total users. This suggests that users choose to represent themselves with more facial expressions, people-centric emoji, and emotions, which are connected to aspects of self-identity. We also found that many users use their bios to present their interests to others – some users use these types of emoji to express their love for certain singers or sports clubs.

Next, we examine the relationships between sets of emoji that users include in their bios. We selected the top 20 emoji used in bios and computed the mutual information between the presence of these emoji in a user's bio and the presence of any other emoji. The emoji with the highest mutual information scores are presented in Table 4.[4] . We found that high-frequency emoji also had high mutual information scores for many other emoji, such as heart emoji of various colors: '❤️', '🧡', '💜'. This indicates that these high requency emoji are not used indiscriminately, but in particular ways

---
[4]The first emoji represent a red heart, and the fifteenth emoji represent a heart suit. They are two emoji patterns with entirely different meanings and also subtle differences in the shape and color.

and have patterns in the ways that they co-occur with other emoji. Another finding is that emoji which are similar to the original emoji have high scores. This finding suggests that similar or the same types of emoji are more likely to be used together. For example, in row 10, four types of ball emoji: basketball '🏀', baseball '⚾', tennis '🎾', and American football '🏈', appear in the ten emoji that provide the most mutual information for soccer ball emoji '⚽'. People who like football may also enjoy other ball sports, and using these ball emoji in the bios at the same time indicates that they are ball sports enthusiasts (either as players or spectators). Another example is that in the 14th row, there are eight national flag emoji out of the ten emoji that have the highest mutual information with the American flag emoji '🇺🇸'. People may use multiple flags in the bios to imply their residences and national origin. Finally, we noticed that users tend to use emoji together that fit a specific context. For example, for the ring emoji '💍' in row 18, the most relevant emoji are kiss '💏', person with veil '👰', man in tuxedo '🤵', and pregnant woman '🤰'. People may use these emoji in the bios to express their relationship status, potentially indicating whether they are engaged, married, or expecting a child.

We also calculated the mutual information score of non-emoji tokens and the top 20 emoji, as shown in Table 5. Our dataset is multilingual, so the tokens obtained are also multilingual. We removed some tokens that do not capture any specific content information, such as some honorifics in Japanese. We found that the usage of emoji is related to words with similar meanings as the emoji, consistent with our previous findings that emoji with similar mean-

| | EmojiBio | | NonEmojiBio | |
|---|---|---|---|---|
| | **Bios** | **Tweets** | **Bios** | **Tweets** |
| **Average Number of Emoji** | 3.05 | 0.73 | 0 | 0.39 |
| **Average Number of Hashtags** | 0.23 | 0.06 | 0.19 | 0.08 |
| **Average Number of Words** | 8.51 | 6.75 | 9.49 | 7.74 |

Table 6: The average number of emoji, words (excluding stopwords) and hashtags in the bios and tweets of the emojiBio and nonEmojiBio datasets

ings had high mutual information. An example of this in the word-level results is in row 10 of Table 5, the tokens most related to soccer ball emoji '⚽' are words in different languages with similar meanings related to soccer and player. This finding further confirms that people prefer to use relevant emoji in a specific context. There are many other examples with similar trends, such as the rainbow flag emoji '🏳️‍🌈' in row 8 and the American flag emoji '🇺🇸' in row 14. Further, we observed that the heart emoji used in bios are more related to showing the love for celebrities or sports clubs, for example, "flamengo" (Row 1) is a sports club (shorthand name for Clube de Regatas do Flamengo), and "bts" (Row 4) is a Korean male singing group.

## 5 The Relationship between Emoji in Bios and Tweeted Content

Next, we explore the relationship between Emoji usage in bios and tweeted content. We start by comparing the overall trends in twitter usage between the sets of users with and without emoji in their bios in order to investigate whether there are notable differences in the volume of emoji, hashtags, and words (excluding emoji, hashtags, and stopwords) used by each group (Table 6).

In terms of the quantity of words and hashtags, there are no significant differences between the emojiBio and nonEmojiBio datasets. In the emojiBio dataset, we noticed that there is increased usage of emoji in bios compared to tweets (3.05 emoji in bios compared to 0.73 in tweets). The fact that the character limit for tweets is more flexible than the limit for bios makes this result even more impressive. In the nonEmojiBio dataset, the average number of emoji that appear in tweets drops to 0.39, which is roughly half the rate of emoji usage in tweets found in the emojiBio group. In terms of hashtags, there is again an increased usage in bios which is similar between the two datasets. In terms of words, users who do not have emoji in their bios tend to use a slightly higher amount of words in their bios and tweets. Specifically, the users in the

nonEmojiBio group used roughly 1 more word, on average, than their emojiBio counterparts, in both tweets and bios.

In addition to differences in the number of words, hashtags, and emoji used, we expect that aspects of a user's identity that are revealed through emoji in their bios will be reflected in measurable ways in the *content* that they choose to tweet about. We perform a case study in which we select two particular interesting emoji that were common in users' bios, and compare the content of the tweets from users who had these emoji in their bios using both topic modeling and sentiment analysis.

The emoji that we focus on for this case study are the rainbow emoji '🌈', and the American flag emoji '🇺🇸'. These emoji are both used with similar frequencies, but are rarely used together and represent distinct groups of users which we seek to understand through the lens of the twitter content that they generate. In our emojiBio dataset, the number users using these in bios are close at 324 ('🌈') and 302 ('🇺🇸'), while only two of the users use both emoji at the same time in their bios, so these two emoji can distinguish users well. These emoji also belong to different emoji subgroups within Unicode Emoji 13.0: the rainbow '🌈' belongs to the sky & weather subgroup under the Travel & Places group, and the American flag '🇺🇸' belongs to the country-flag subgroup under the Flags group.

Among the 324 users who use rainbow emoji '🌈', 155 users use English in the bios, 46 Japanese, 33 Portuguese, and 31 Spanish. For comparison, among the 302 users who use the American flag '🇺🇸', 245 use English as the language in bios, 15 Spanish, 12 Japanese, and 9 Portuguese. The tweets involved also are multilingual, but are mostly written in English. For the analyses in this section, we first translated all non-English tweets into English using the Google Translate API.[5] Considering that the topic modeling and sentiment analysis methods that we use mostly rely on bag-of-words representations of the text, issues with the grammatical accuracy of translated tweets will not have as large of an impact. After the translation, we have two sets of tweets corresponding to the two groups of users who used the emoji of '🌈' and '🇺🇸'. The number of tweets for each group are 61,239 and 58,376, respectively.

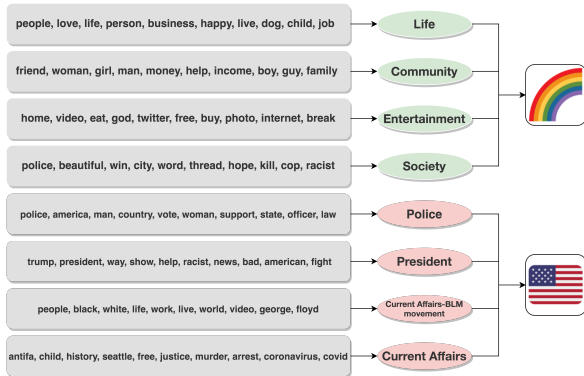We performed topic modeling using Latent Dirichlet Allocation (Blei et al., 2003) on the tweets

---

[5]https://cloud.google.com/translate

Figure 1: The most relevant tokens for topics inferred from the tweets from users who use the 'rainbow' emoji and the 'United States flag' emoji in their bios.



Figure 2: Sentiment analysis of the tweets from users who use emoji '🌈' and '🇺🇸' in bios separately.

of users who used the emoji '🌈' and '🇺🇸' in their bios. We used the coherence score provided by the gensim Python library[6] to select the number of topics. We train a separate topic model for each group of users, and select four topics for each model. In Figure 1, we visualize the process of inferring topics by zooming in on the most relevant tokens for each of the topics within the set of tweets written by each group of users. The weights between topics are unequal, decreasing from top to bottom as presented in the figure. The topics of tweets from users who use rainbow emoji '🌈' in the bios include words related to concepts like life, community, entertainment, and society. We notice some topics that contain more pleasant words, some related to gender identity, others to life and pets. The fourth topic appears to be related to issues of police brutality. However, on the whole, the tweets posted by users who use the American flag emoji '🇺🇸' in the bios are more heavy and serious. They are more concerned about topics related to police, president, and current affairs. Because of the massive surge in the #blacklivesmatter movement, caused by the death of George Floyd in the United States, broke out at the end of May 2020, and we downloaded user tweets during this time, there is a clear topic for this current affair. Besides, other current affairs discussed include Antifa and COVID, but these were part of the same topic. Comparing the two sets of different topics, we found that the different emoji included by the users in their bios are related to distinct topics, which also may reflect the self-identities of the users who used these emoji.
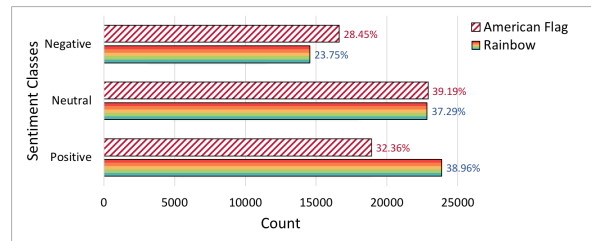
---

[6]https://radimrehurek.com/gensim/

The rainbow emoji '🌈' often represents gay pride, as well as happiness and peace in general, so the corresponding tweets also mostly reflect the love of these users for life and others. In contrast, users who use the American flag '🇺🇸' are more concerned about national politics and current affairs within the United States.

We also conducted a sentiment analysis on these two sets of tweets, using the Vader sentiment analysis tool (Hutto and Gilbert, 2014), giving the results presented in Figure 2. According to the figure, for the two datasets, the distribution of sentiment is fairly consistent overall, with more positive content than negative. While the amount of neutral sentiment in the two datasets is almost the same, the users with rainbow emoji '🌈' in their bios tweeted more positive content overall, compared the the users with the US Flag emoji '🇺🇸' in their bios. Close to 40% of the tweets from users who use rainbow emoji '🌈' in bios are positive, and less than 25% are negative. In contrast, less than 35% of tweets sent by users using the American flag '🇺🇸' in bios are positive, and close to 30% are negative.

These sentiment analysis results are mostly consistent with the results of the topic modeling. The tweets sent by users who use rainbow emoji '🌈' are more happy and light than those sent by users who use the American flag emoji '🇺🇸' in the bios. This case study suggests that users using different emoji in bios can reflect aspects of both their national identity and their personality. More specifically, this analysis shows that groups using some emoji in the bios generate more positive content than groups using other emoji.

## 6 Homophily Effects in Emoji Usage in Bios

For our final set of analyses, we explored the extent to which users and their followers use emoji in their bios in similar ways. At a very basic level, regarding the absence or presence of emoji in the

| 💚 | | ⚽ | | 🇺🇸 | | 🐶 | |
|---|---|---|---|---|---|---|---|
| 💚 | **7420** | ❤️ | 6529 | 🇺🇸 | **26199** | ❤️ | 2028 |
| ❤️ | 7361 | ⚽ | **4054** | ❤️ | 5422 | ✨ | 1943 |
| 💙 | 5509 | 💙 | 3817 | 🚫 | 2582 | 💙 | 1375 |
| 💜 | 4902 | ✨ | 2128 | ⭐ | 2421 | 🐶 | **1331** |
| ✨ | 4897 | 🖤 | 1865 | 🌊 | 2262 | 💕 | 1150 |
| 💛 | 3335 | 😍 | 1816 | ✝️ | 1933 | 💜 | 913 |
| ❤️ | 2941 | 🔴 | 1523 | ✨ | 1929 | 💛 | 828 |
| 💕 | 2880 | 🔥 | 1495 | 💙 | 1861 | ❤️ | 791 |
| 🖤 | 2042 | ❤️ | 1363 | 🙏 | 1837 | 🌊 | 782 |
| 😍 | 2042 | 💛 | 1319 | 💯 | 1568 | 💚 | 693 |

Table 7: Top 10 emoji used by followers of users with particular emoji in bios and their counts. Bold indicates the count for the same emoji that was used by the reference user. We observe that it is very common for a user and their followers to use the same kinds of emoji in their bios.



Figure 3: The distribution of the percentage of common and similar emoji appearances in the followers' bios. The lines in the graph represent the average percentage of common and similar emoji appearances.

| Common Emoji Appearances | | Similar Emoji Appearances | |
|---|---|---|---|
| Emoji | Percentage (%) | Emoji | Percentage (%) |
| 🔞 | 14.75 | 🖤 | 34.77 |
| 🚩 | 13.72 | 💕 | 34.31 |
| 💜 | 13.33 | 🙏 | 33.91 |
| ❤️ | 10.46 | 💛 | 33.81 |
| 🀄 | 10.4 | 💙 | 30.32 |

Table 8: The emoji with the highest percentage of common (exact match) and similar appearances between the users' and the followers' bios.

followers' bios, there was a considerable difference between the emojiBio and nonEmojiBio datasets. The followers of users that have emoji in their bios (emojiBio) have emoji in their bios as well 32.47% of the time. For the followers of users that do not have emoji in their bios (nonEmojiBio), this average percentage drops to 23.23%.

Next, we selected three representative emoji from the set of most frequently used emoji in the emojiBio dataset, namely, green heart emoji '💚', soccer ball emoji '⚽', and American flag emoji '🇺🇸'. Also, to eliminate bias caused by only considering high-frequency emoji, we selected the low-frequency dog face emoji '🐶' used by a total of just 157 users in our emojiBio dataset. In Table 7, we list the ten most frequently used emoji in the bios by the followers (from our Followers dataset) of the users who use these four specific emoji and mark the emoji that are the same as the users in bold text.

The green heart '💚' and the American flag '🇺🇸' are the emoji that are used most frequently by followers of users who also include these emoji. The soccer ball emoji '⚽' ranks third, and the dog face emoji '🐶' ranks fifth, only with several high-frequency emoji in front of them. There is a strong homophily relationship that indicates that the users use the same emoji with their followers in bios. Using the same emoji also reflects that emoji in the bios can reflect the users' self-identity in terms of group belonging, or their social identity. As an illustration, users using dog face emoji in bios m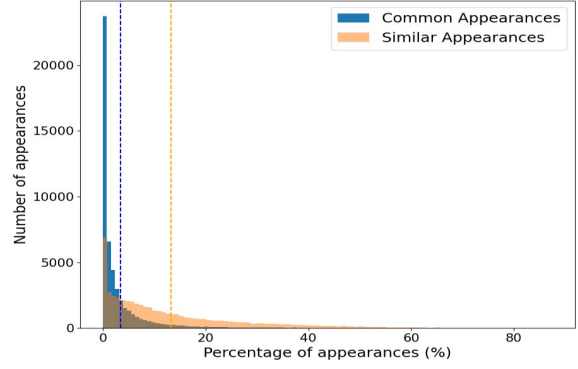ay want to signal that they are dog lovers, and they may also chose to make online connections with others who are similar, leading to many other dog lovers in their networks.

We also take a particular look at the high-frequency emoji used by followers of users who use the American flag '🇺🇸'. Prior work on emoji and American political movements on Twitter (Hagen et al., 2019) pointed out that water ("blue") wave emoji '🌊' is related to the US Democratic party, and pointed out that this emoji is frequently associated with hashtag #resist to express anti-white nationalist sentiments. We also observe the use of the red heart '❤️' and blue heart '💙' emoji, two colors are are often associated with the US republican and democratic parties, respectively. These followers may be expressing their political opinions: they use the American flag emoji along with other more specific emoji express their particular views. Lastly, we notice several emoji related to religion in this column, indicating expressions of religious as well as political affiliations.

In addition to the focused study on these four emoji, we also examined whether the emoji used in bios of Twitter users are either the same, or generally similar to those used by their followers in the

entire dataset. To assess similarity we trained our own emoji embeddings with a skip-gram model (Mikolov et al., 2013) using the tweets and bios of the emojiBio dataset, and subsequently we created a similarity lexicon of emoji based on the cosine similarity between the vectors, considering one emoji to be similar to another if it was within the top ten nearest neighbors in the learned embeddings space. We found that the average percentages for common (i.e., exact matches) and similar emoji appearances between the users of the emojiBio dataset and their followers are 3.45% and 13.30%, respectively. The respective distributions of the percentage for common and similar emoji appearances are presented in Figure 3. We used permutation tests to confirm that the difference between these two values was statistically significant, and therefore conclude that the followers of a given user seem to have a considerably high probability to use the same, or similar emoji in their bios as the users they follow. Table 8 shows the five emoji for which followers used the same, or similar emoji as the users that they follow.

## 7 Discussion

We now give answers to our original research questions based on our results:

**RQ1. How are emoji used in Twitter bios?** Our results showed that emoji are used in unique ways within users' bios on Twitter, even compared to the ways in which they are used in tweets. In general, emoji are positioned earlier in bios than in tweets, while there is a higher percentage of bios that start with an emoji compared to tweets. Also, it is more common for an emoji to be the only content of a bio than the only content of a tweet.

Moreover, facial expression emoji are the dominant type of emoji in tweets, while different variations of heart emoji are dominant in bios. Specifically, the most popular emoji in bios are from the Smileys & Emotion group, while the least frequently used emoji are from the Food & Drink group. Furthermore, we noted that the most frequently used emoji in bios have a high mutual information with other emoji that are similar to them, or from the same category (e.g. hearts, balls, flags), or related to the same concept (e.g. relationship status). In their bios, people tend to use emoji to show their support for musical groups or sports teams (or sports in general), as well as things like countries that they come from or are currently living in.

**RQ2. What is the relationships between the emoji in a user's bio and the content that the user posts?** Compared to users who do not have any emoji in their bios, users with emoji in their bios use about twice as many emoji in their tweets, on average. They also use less words in both their tweets and bios. In our case study, topic models built from the tweets of the users that use the rainbow '🌈' and the American flag '🇺🇸' emoji in their bios showed that users who have the rainbow emoji '🌈' in their bios tweet about life, community, entertainment, and society, whereas users who have the rainbow emoji '🌈' in their bios tweet about police, president, and current affairs. Also, it was shown that tweets of users that have the rainbow emoji '🌈' in bios convey a more positive sentiment on average compared to users that use the American flag '🇺🇸' in their bios. This is just one example to showcase the fact that the types of emoji that people choose to include in their bios reflect larger views, opinions, and sentiments that are expressed in the content of their tweets.

**RQ3. Do users and their followers use emoji in their bios in a similar way?**

The usage of emoji in bios also led us to some conclusions related to homophily effects in Twitter. First, our results indicate that followers of users who have emoji in their bios, are more likely to have emoji in their bios as well. We also found that users tend to use the same, or similar emoji in their bios as the users they follow. For example, followers of users with the green heart emoji in their bios also had other colored hearts in their bios, with the green heart being the most common used by the followers. These findings suggest that there are indeed similarities within user networks in the ways in which emoji are used in Twitter bios.

## 8 Conclusion

We have presented an overview of the ways in which Twitter users include emoji in their bios, and what kinds of things we can learn about those users from the particular emoji that they use. Using a range of approaches, we have shown that emoji are an important component to consider when examining the ways in which users present themselves to others in online settings like Twitter. The emoji that users choose to include reveal important aspects of their self-identities, such as the teams and musicians that they support, the activities they enjoy, their national and political identities, and show

their similarities with their followers in these same aspects. At the same time, we have only brushed the surface of the types of in-depth analyses that could be performed by consider specific sets of emoji and examining how these relate to the identities of the users who include them in their bios. This work can provide an important complementary view to other work on online-self identity that mainly focuses only on the plain text content.

# References

Francesco Barbieri and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in twitter. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 101–106.

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. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Marcel Danesi. 2016. *The semiotics of emoji: The rise of visual language in the age of the internet*. Bloomsbury Publishing.

Michael Fisher, Martin Abbott, and Kalle Lyytinen. 2014. The concept of self-identity. In *The Power of Customer Misbehavior*, pages 61–67. Springer.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Every colour you are: Stance prediction and turnaround in controversial issues. *arXiv preprint arXiv:2005.10019*.

Serge Guimond, Armand Chatard, Delphine Martinot, Richard J Crisp, and Sandrine Redersdorff. 2006. Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of personality and social psychology*, 90(2):221.

Sharath Chandra Guntuku, Mingyang Li, Louis Tay, and Lyle H Ungar. 2019. Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 226–235.

Loni Hagen, Mary Falling, Oleksandr Lisnichenko, AbdelRahim A Elmadany, Pankti Mehta, Muhammad Abdul-Mageed, Justin Costakis, and Thomas E Keller. 2019. Emoji use in twitter white nationalism communication. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 201–205.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Ryan 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*, 20.

Weijian Li, Yuxiao Chen, Tianran Hu, and Jiebo Luo. 2018. Mining the relationship between emoji usage patterns and personality. *arXiv preprint arXiv:1804.05143*.

Nikola Ljubešić and Darja Fišer. 2016. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop*, pages 82–89.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

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

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. 2018. Self-representation on twitter using emoji skin color modifiers. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. 2020. Emoji skin tone modifiers: Analyzing variation in usage on social media. *ACM Transactions on Social Computing*, 3(2):1–25.

Nick Rogers and Jason J Jones. 2019. Using twitter bios to measure changes in social identity: Are americans defining themselves more politically over time?

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013.

Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Paul Sparks and Richard Shepherd. 1992. Self-identity and the theory of planned behavior: Assesing the role of identification with" green consumerism". *Social psychology quarterly*, pages 388–399.

Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. 2017. Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 11–16.

Sarah Theres Völkel, Daniel Buschek, Jelena Pranjic, and Heinrich Hussmann. 2019. Understanding emoji interpretation through user personality and message context. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12.

# Appendix

## A   Differences across languages

A language identification analysis was conducted for the combined data of the emojiBio and nonEmojiBio datasets to identify the most frequently used languages in the tweets and bios. The analysis was conducted using the fastText language identification tool (Joulin et al., 2016b), (Joulin et al., 2016a) and the language distribution is presented in Figure 4.
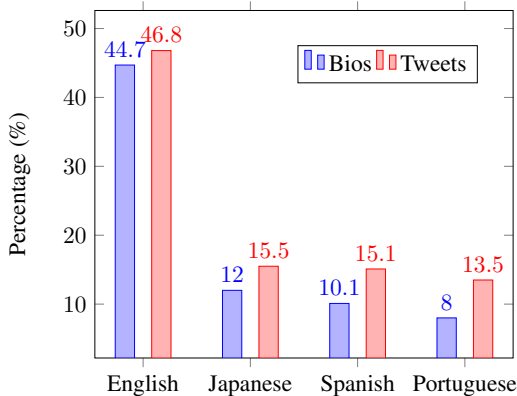


Figure 4: Language distribution in the tweets and bios of the emojiBio and nonEmojiBio datasets combined.

## B   Positioning Analysis

The positioning analysis distribution for bios and tweets is presented in Figure 5. The results of the positioning analysis indicate that emoji appear earlier in bios than in tweets. For each emoji, its positional value was calculated by computing its

distance from the first character of the text and dividing it by the overall length of the text. Therefore, emoji that were used at the beginning of the text had a positional value of 0, whereas emoji that were used at the end of the text had a positional value of 1.
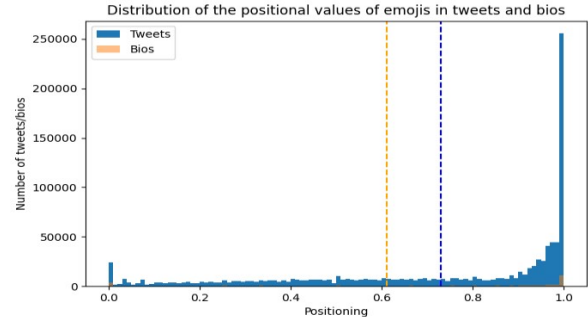


Figure 5: The distribution of the positional values of emoji in the tweets and bios of the emojiBio dataset. The vertical lines in the graphs represent the mean positional value for tweets (blue) and bios (orange).

## C   Group Analysis

In the group analysis, we divided the bios into four groups according to the language used, and we calculated the mutual information score for all emoji that appeared. Table 9 shows the 25 emoji with the highest scores in each group. We observed that in all language groups, there were multiple national flag emoji amongst the results. In most cases, those flags belong to countries where the respective language is spoken as a first or second language by a considerable portion of the population.

While emoji grouping is analyzed in Chapter 4, it is also important to consider that Unicode Emoji also provides standards for subgroups of emoji. Specifically, each emoji belongs to a group and also belongs to a subgroup under the group, which makes the classification more specific. For example, the grinning face emoji '😀' belongs to the face-smiling subgroup under the Smileys & Emotion group. Each group contains a different number of subgroups, and overall there are 98 subgroups. We counted the total number of times the emoji from each subgroup appeared in users' bios and sorted them in descending order. Table 10 demonstrates the ten most popular subgroups.

The results suggest that the most frequently used subgroup is emotion while the face-smiling subgroup also belongs to the same category (Smileys & Emotion), showing that people are commonly using emoji to express their sentiments in bios. The
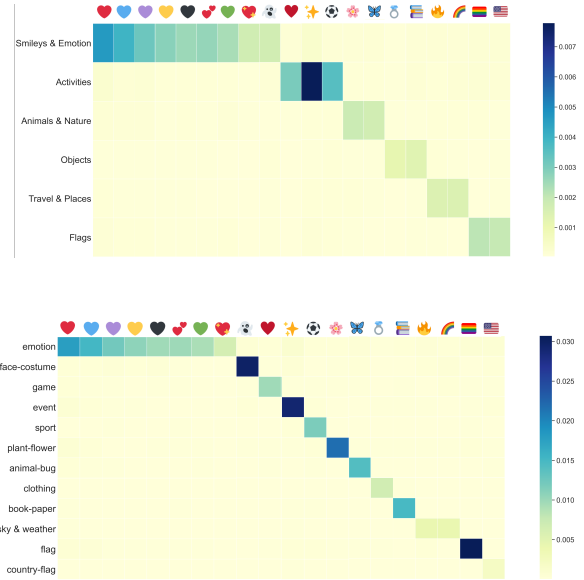
Figure 6: The average mutual information score between the 20 most frequently used emoji in bios and each group and subgroup, respectively.

Specifically, we divided the 20 most frequently used emoji into groups and subgroups, and we plotted two heat maps which illustrate the categorization of emoji, as shown in the Figure 6. While calculating the mutual information scores between a group and a specific emoji, we did not consider that emoji as part of the group, to ensure normalization. The results show that each emoji achieved a higher mutual information score with the group or subgroup in which it belongs. This suggests that emoji in bios are more commonly used with other emoji from the same group or subgroup.

## D Topic Modeling

We conducted supplementary experiments on the topic modeling analysis of Chapter 5. Specifically, we used the LDAvis tool to visualize the results, and the topic distributions are shown in Figure 7. The topic distributions visualize the weight of each topic and the connection between different topics. More precisely, the circles represent the topics, and the distance between the circle centers determine the connection between the topics. More prevalent topics are represented by larger circles.

## E Frequency Analysis

The results of the frequency analysis showed that the popularity of words and hashtags varies greatly between bios and tweets. Table 11 presents the most frequently appearing English words and hash-

second most frequently used subgroup is country-flag, which implies that users regularly use emoji in their bios to reveal their nationality or the countries where they have lived. Animal-mammal and plant-flower are also frequently used. These emoji are used to express the love of users for animals or plants, but also for decoration reasons, to make bios more attractive. Another interesting finding was that the zodiac subgroup ranks seventh. This finding shows that people like to use symbolic emoji to tell others about their zodiac, which they consider as a part of their self-identity.

To confirm that the emoji could be accurately grouped in clusters, we conducted a statistical analysis based on the results of the mutual information scores for the 20 most frequently used emoji in bios.

**Bios**

| Word | Appearances | Hashtag | Appearances |
|---|---|---|---|
| love | 645 | #bts | 33 |
| fan | 473 | #resist | 28 |
| life | 375 | #maga | 22 |
| account | 371 | #exo | 21 |
| insta | 273 | #bernie | 16 |
| instagram | 218 | #blacklivesmatter | 15 |
| flamengo | 216 | #jimin | 14 |
| god | 212 | #wwgwga | 11 |
| dm | 212 | #blm | 11 |
| follow | 198 | #mufc | 10 |

**Tweets**

| Word | Appearances | Hashtag | Appearances |
|---|---|---|---|
| like | 22843 | #peing | 868 |
| love | 16025 | #nintendoswitch | 802 |
| get | 14019 | #blacklivesmatter | 790 |
| people | 13541 | #newprofilepic | 739 |
| know | 11671 | #acnh | 656 |
| good | 10867 | #covid | 563 |
| time | 9881 | #animalcrossing | 561 |
| go | 9791 | #otgalafinal | 458 |
| lol | 9504 | #sanditon | 349 |
| got | 9241 | #psshare | 345 |

Table 11: The most frequently appearing English words and hashtags in the bios and tweets of the emojiBio dataset.

**Bios**

| Word | Appearances | Hashtag | Appearances |
|---|---|---|---|
| fan | 53 | #phish | 4 |
| love | 48 | #maga | 4 |
| account | 36 | #taehyung | 3 |
| life | 28 | #resistance | 3 |
| twitter | 22 | #kag | 3 |
| like | 22 | #ynwa | 2 |
| good | 20 | #trump | 2 |
| world | 19 | #research | 2 |
| god | 19 | #mufc | 2 |
| people | 17 | #bernie | 2 |

**Tweets**

| Word | Appearances | Hashtag | Appearances |
|---|---|---|---|
| like | 3826 | #chismesfarándulachilena | 288 |
| people | 2500 | #meigen | 182 |
| get | 2315 | #shindanmaker | 160 |
| love | 2202 | #covid | 151 |
| good | 2036 | #blacklivesmatter | 135 |
| know | 1921 | #survivor | 127 |
| time | 1656 | #nintendoswitch | 121 |
| think | 1633 | #digitalmarketing | 104 |
| go | 1625 | #lockdownhouseparty | 102 |
| see | 1526 | #bitcoin | 92 |

Table 12: The most frequently appearing English words and hashtags in the bios and tweets of the nonEmojiBio dataset.
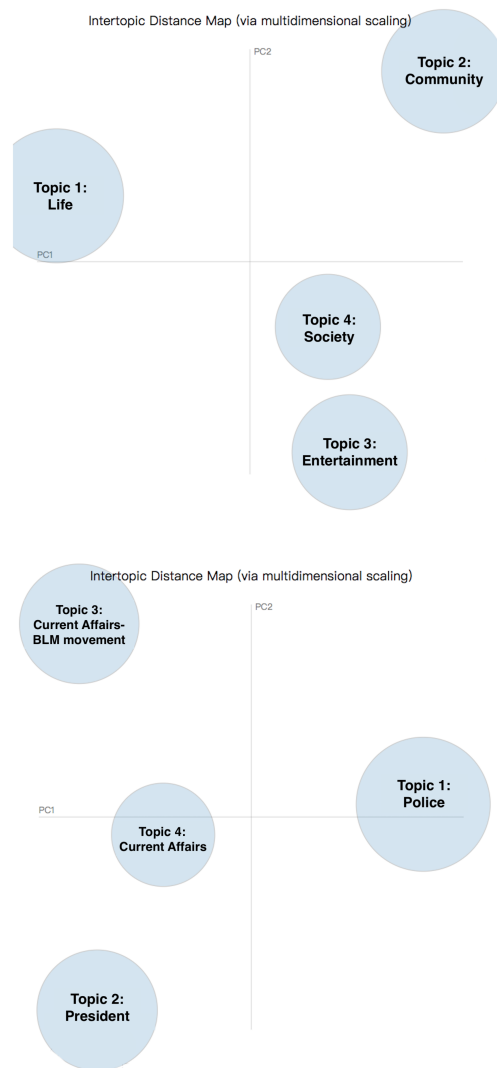


Figure 7: Topic distributions for the tweets of the users who the '🌈' and '🇺🇸' emoji in their bios.

tags in the bios and tweets of the emojiBio dataset. While the frequency analysis was conducted on multilingual data, we present only the most frequently appearing English words for consistency reasons, since there are many different English translations for words in other languages.

In terms of words, the more frequently used words in bios are nouns, in contrast with tweets, where verbs appear more frequently. The most frequently used words in bios are mostly related to the social media activity of the user (account, insta, instagram, dm, follow) and their religious or spiritual beliefs (love, life, god). On the contrary, in tweets, we can see verbs related to positive sentimental expression (like, love) or the conduction of an activity (get, go, got).

The hashtags that more frequently appear in

bios are related to music artists or bands (#bts, #exo, #jimin), presenting the user's music preferences, to political beliefs or election candidates (#resist, #maga, #bernie) and the anti-violence protest group "Black Lives Matter" (#blacklivesmatter, #blm). The hashtags related to "Black Lives Matter" are commonly found also in tweets, together with hashtags related to gaming consoles and video games (#nintendoswitch, #animalcrossing, #acnh, #psshare), TV series (#sanditon) and music competitions (#otgalafinal). Users also use hashtags to tweet about Peing - an "anonymous Q&A box" service on Twitter (#peing) and to notify others about an update of their profile picture (#newprofilepic). Additionally, users frequently use a hashtag in their tweets which is related to the COVID-19 pandemic (#covid).

Overall, the results for the words and hashtags frequency analysis per element of the nonEmojiBio dataset do not have significant differences compared to the results of the emojiBio dataset. Also, despite the decreased usage of emoji in tweets by these users, the distribution of the frequencies are very similar compared to the emojiBio dataset, since facial expression emoji are dominant again. The complete results for the frequency analysis of the nonEmojiBio dataset are presented in Table 12, but they should be interpreted with caution since the nonEmojiBio dataset is considerably smaller.