

# Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis

Astha Mohta, Atishay Jain, Aditi Saluja, Sonika Dahiya

Computer Engineering Department

Delhi Technological University

Main Bawana Road, New Delhi, India

asthamohta@gmail.com, atishay9jain@gmail.com, saluja.aditi5@gmail.com, sonika.dahiya11@gmail.com

**Abstract**— WhatsApp is among the popular social media service with over 2 billion registered users. WhatsApp is integral in people's life as a medium of communication. They use WhatsApp to share their feelings through text messages. With WhatsApp present in over 180 countries, code-switching is common. Along with this, with the increase in usage and prevalence of emojis, emojis have become indispensable during sentiment analysis. Throughout this paper, **our approach to convert unstructured WhatsApp messages to a structured form is discussed on which various data mining techniques for sentiment analysis can be performed.** Our approach to deal with code-mixing, different emojis and the emotions they depict, and finally, perform basic analysis using this algorithm is discussed.

**Keywords**— WhatsApp; Sentiment Analysis; Emoji; Interpersonal Communication; Emotion Analysis; Code-Mixing.

## I. INTRODUCTION

WhatsApp is among the popular social media applications to connect people today, with 2 billion registered users across 180 countries and 55 billion messages sent per day [1], [2]. WhatsApp holds a key place in our life. People use WhatsApp to exchange information and also to share **intimate feelings** that are often touted as deep meaningful conversations.

In a study of students between the **ages of 18 to 23** that was conducted in Chennai, India has investigated the importance of WhatsApp among youth. Students tend to spend 8 - 16 hours a day on WhatsApp [3]. Hence, can be concluded that **WhatsApp chats are capable of depicting an individual's thought process as well as capturing their sentiments.**

Pre-processing is an important step during the **data mining** of texts. Data pre-processing should be the first step for any process involving the use of data. That is because it leads to data sets, that are cleaner, coherent, and much more manageable, a must for any business trying to get valuable insights from the collected data.

Pre-processing data consists of various steps **depending on the type of data.** In the context of the work presented in the paper, **the data obtained not only consists of textual but graphic data comprising of emojis as well.** Hence, it becomes imperative that segregation is carried out before using them for any further analysis. Furthermore, the textual data obtained is multilingual and to enforce uniformity, translation to a common language is required. The above evidence leads to the

requirement of a revised algorithm for pre-processing of these WhatsApp messages to use them in sentiment analysis. Conventional pre-processing for NLP does not account for bilingual speakers or code-switching. It does not acknowledge the importance of emojis and how they interact with the text either.

A key reason why WhatsApp is more popular than texts is that it is free. Along with this WhatsApp doesn't count words and supports multimedia [4]. Study shows that about 19.6% of tweets contain emojis and 37.6% users use emojis. Along with this, in 2015, "Face with Tears of Joy" [5] was the Word of the Year by Oxford dictionary.

Throughout the paper, **our proposed algorithm to convert this unstructured data to a more structured form to perform various sentiment analysis algorithms is discussed.** Emojis are categorized using both sentiment analyses in the following emotions: Happy, Sad, Angry, Fearful, Excited or Bored rather than which classify emotions as **either positive, neutral or negative.** For the study, people of different age groups were considered and their WhatsApp chats were analyzed (with their consent). The pre-processing on these chats are aimed to perform as well as count emojis in different categories to derive features from them to guide sentiment analysis.

In this paper, related work on pre-processing of data is first introduced used for sentimental analysis as well as the importance of emojis during texting and emotion categories for sentiment analysis in Section 2. Then presented our Data Set in Section 3. Our algorithm for pre-processing is discussed in section 4. Section 5 discusses our results and analyses on a sample of subjects. Conclusion and future work is presented in Section 6.

## II. RELATED WORKS

WhatsApp came into the market as a substitute for SMS. It is used to make voice or video calls as well as to share media. WhatsApp provides these services for free [1]. Authors of "Survey Analysis on the Usage and Impact of WhatsApp Messenger" in their study of a group of WhatsApp users of **ages between 18-50** found that about seventy-nine percent of their subjects use WhatsApp every day for at least 15-60 minutes [6]. In the study "Impact of WhatsApp on youth: A Sociological Study" 100 WhatsApp users were selected randomly, their ages lying between 18 and 30. The study

discovered that the frequency of usage for sixty-three percent users is fifty times a day, for twenty-one users are twenty times a day and sixteen percent is more than a hundred times a day. Thus can be concluded that amid the youth, WhatsApp has a huge influence on the way they communicate [7].

A study performed on the students of Abu Dhabi, eighty-five percent of women and seventy percent of men expressed that they used emojis as a substitute for facial expression on the text. Their average usage was 1 to 7 hours/day [8]. On a study on four million users of different countries, it was shown that about six million i.e., seven percent of all messages contained at least one emoji. This is a sign of the popularity of emoji among users [9].

In early research in 1982, Aravind K. Joshi describes code-switching or code-mixing as “Speakers of certain bilingual communities systematically produce utterances in which they switch from one language to another in the course of an utterance. Production and comprehension of utterances with Intra sentential code-switching is part of the linguistic competence of the speakers and hearers of these communities.” [10]. “Code Mixing: A Challenge for Language Identification in the Language of Social Media” describes that despite most social media content being in the English Language, this still makes up for only half of the content worldwide. The reason for this is that most users switch between languages mid-way [11]. This leads to the need for a kind of translation or a model for tagging multiple languages.

For our study, six emotions have been chosen and classified our emojis in these categories. When looking at **Russel's Circumplex Approach**, he states that emotions can be classified on two scales, Valence and Arousal [12]. In this model, the emotions are represented as a combination of two dimensions: Valence, which represents a degree of **pleasure - displeasure** & Arousal, which represents levels of **activation - deactivation**. For example Boredom as an emotional state with negative valence, i.e. displeasure and low arousal [13].

For the study, six emotions have been chosen to convert all states of **valence and arousal**. As seen in figure 1, happiness and excitement can be seen in the quadrant with positive arousal (or pleasure) and valence (or activation), sadness and boredom have negative arousal (or displeasure) as well as low valence (deactivation) and finally, anger and fear have high arousal (or activation) but low valence (or displeasure). Low arousal with positive valence leads to a neutral state of mind [12].

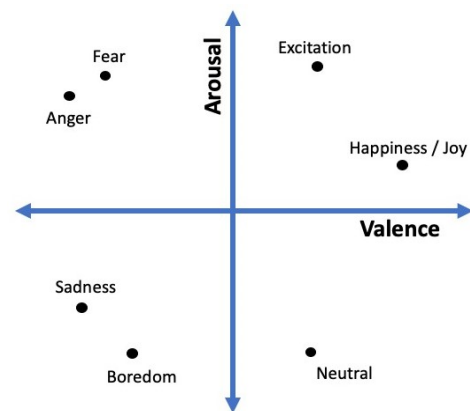


Fig. 1. Categorization of emotion based on Russell's Circumplex model

Authors of “Facebook Sentiment: Reactions and Emojis”, argue that linguistic textual messages and emojis modify one another’s meaning. The interaction of linguistic text with emojis can vary. An emoji can be a replacement for a phrase, reinforcement of a phrase, autonomously convey the emotion or attitude of the subject, provide emphasis on an already existing emotion in the message or can be used out of politeness. Given the context, **emojis can be used to detect the subject's emotions** [14]. The basic steps in any data pre-processing involve importing the libraries, data read, checking for unmitigated information, checking for categorical data, standardizing the data, PCA transformation & data splitting.

In the model prescribed in the paper “Social emotion mining techniques for Facebook posts reaction prediction”, initial pre-processing like converting to lower case, removing hashtags etc. has been done using the Stanford CoreNLP Parser which is followed by the use of a tokenizer to split posts based on spaces & after filtering the stop words [15], the list of different tokens is obtained.

In the paper “Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?“, showcases how emojis can be used to perform automatic labeling of tweets. Automatic labelling of tweets outperforms manual labeling of tweets [16].

### III. DATA SET

#### A. Data Collection and User Consent

Eighteen individuals of various qualifications and ages were considered for analysis of their WhatsApp messages for the study. WhatsApp chats are private to everyone and hence certain ethical obligations are required to ensure their privacy and fair use. Each of them was asked to sign consent forms to permit the researchers to use their conversations for analysis. Their conversations are used for research purposes only no conversation has been made public. The complete data set is formed by collecting 5 – 10 chats from each participant. To collect the data, the “Export Chat” feature is used from WhatsApp that allows you to share your chats using Email, WhatsApp, Bluetooth, etc. [17]. The user has the option of sharing including media or not. Since worked only with textual data, data was shared without media.

## B. Analysis of Subject

Around 180 chats of 18 users are collected and stored in a database. For each participant, general information such as their names, age, gender and professional background is also collected. To present the diversity of the candidates selected for the study, analyzed them on their professional backgrounds and ages in Table 1 and Table 2.

TABLE I. NUMBER OF CANDIDATES BY AGE RANGE

Age Range	Number of Candidates
15 to 25	8
25 to 35	2
35 to 45	2
45 to 55	5

Greater than 55	1
-----------------	---

TABLE II. NUMBER OF CANDIDATES BY PROFESSIONAL BACKGROUNDS

Professional Backgrounds	Number of Candidates
Engineering	5
Medicine	4
Humanities	4
Commerce	5

## IV. PROPOSED ALGORITHM

The steps of the algorithm along with a sample chat are shown in figure 2.

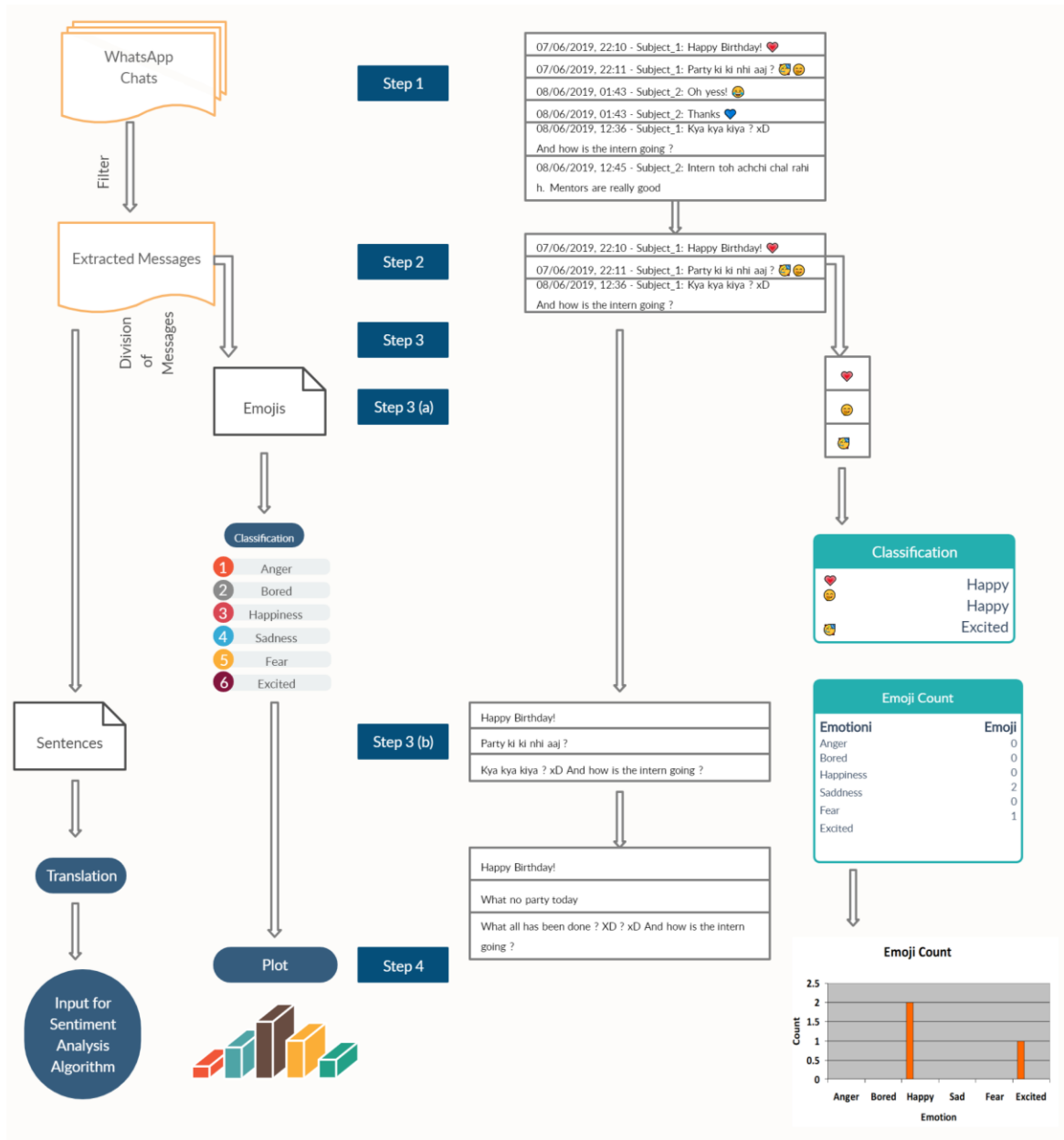


Fig. 2. Algorithm for Pre-Processing of WhatsApp Chats with example

The first step is data collection. After WhatsApp chats are taken from an individual, in the second step is to extract the message from the text file. To do so, first, the time stamp and date of the message is removed. Since a chat is between two individuals, but are only interested in analysis the sender side (our subject), the set is split into two parts using our subject name and pick the set, with messages from our subject.

For the next step, the messages are divided into two parts, one is the emoji and other is the sentence. The entire set of Unicode Emoji codes is supported along with aliases. An automatic labelling approach is used. **Unicode emoji list has been used along with the keywords associated to classify them in six categories of emojis** [18].

RegEx (Regular Expression) is a sequence of characters that represents a search pattern. RegEx can be used to check if a string contains the specified search pattern. In the work presented in the paper, the RegEx library in Python is used along with the Unicode library to figure out emoji patterns, distinguish between different emojis and classify emojis of the same pattern such as ones having the same structure but different colours as same [19]. Following is the list of Emojis classified:

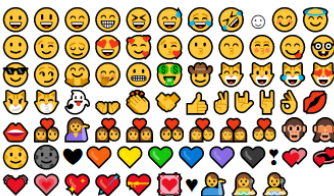





Emotion	Emojis
Happiness	
Sadness	
Fear	
Anger	
Excited	
Bored	

Fig. 3. Emojis categorized by emotions

Emojis portray emotions unambiguously. Hence an emoji indicates emotion clearly and the emotion is both domain and topic independent. [20] Quite often though emojis are used relentlessly in a single text message. For this purpose, a threshold value was chosen after much deliberation and experimentation with various values. The threshold value chosen was 3. This implies that if the same emoji is used > 3 times after one another, it might lose its weightage. To take this factor into account, if emojis of the same classification is used > 3 times, the value is divided by 3 and add the ceiling value.

Let Emoji Count<sub>i</sub> = EC<sub>i</sub>,

for i = [ Bored, Anger, Sadness, Happiness, Excited, Fear,]

$$\text{then, } f(EC_i) = \begin{cases} \lceil EC_i/3 \rceil, & EC_i > 3 \\ EC_i, & \text{otherwise} \end{cases} \quad (1)$$

Emoji count for each emotion is initialized with zero. Using the classification in figure 3 the respective counts of categories as they are encountered is increased. These emojis can be depicted in a graphical form for pictorial analysis.

Following this, the set of sentences uses GoogleTrans (python library that implemented Google Translate API). GoogleTrans is a python library using which calls to such methods as detecting and translate is made using the Google Translate Ajax API to translate the input data to a particular language. Since the data collected in the work presented in the paper is multilingual, GoogleTrans API has been used to convert all of the text to a common language (English) before analyzing the data [21]. The number of words and sentences can be counted to give a rough idea about how interactive the subject is on WhatsApp. Along with this, the ratio of Sentence to Emoji is calculated to showcase if a subject uses emojis to better express themselves. These sentences can now be used as the input for and sentiment classification model.

## V. RESULT ANALYSIS

The algorithm proposed in the last section is applied to ten chats for each of the eighteen candidates. A sample of the comparison is presented below.

Two test subjects are analyzed and compare them based on the output of the pre-processed data approach discussed in the previous section. Each chat is analyzed by considering the total number of words, sentences & emojis present. Emoji Count in an online conversation is inversely related to the degree of professionalism between the participants [22]. They are often used as a replacement for words when felt and can't express our emotions with simple words, and are instead more comfortable using emojis [23].

In a study of how often introverts and extrovert tweet (in a fixed duration of time) it was discovered that the introverts tweet approximately 14.4% more than extroverts [24]. It is also studied that introverts prefer texting as a mode of communication over the phone calls, hence they text more. Consider a comparison between two Subjects from the dataset based on emoji count, word count, sentence count and the number of emojis used per sentence.

Table 3 gives a comparison based on words, sentences, emojis (along with categories) and emoji to sentence ratio.

TABLE III. COMPARISON OF SUBJECT 1 AND SUBJECT 2

Comparison	Subject1	Subject2
<b>Sentence Analysis</b>		
Word Count	142345	151969
Sentence Count	32647	27441
Emoji: Sentence Count	0.25	0.32
<b>Emoji Analysis</b>		
Total Emoji Count	8135	8819
Anger Emoji	340	259



Bored Emoji	496	571
Happy Emoji	343	762
Sad Emoji	176	363
Fear Emoji	6159	6214
Excited Emoji	621	650

It can be observed that Subject 2 has a greater emoji count per sentence compared to subject 1. Hence according to the studies mentioned above, it can be concluded that **subject 2 mostly communicates unprofessionally and uses emojis as a method to express them better**. This matches with subject 2's own opinion who stated that they use more emojis as they feel that emojis are more fun. The sentence count also reflects that texting is a mode of communication preferred by Subject 1. It has been observed that texting is more popular among introverts compared to extroverts. In the survey filled, subject 1 identified as an introvert, whereas subject 2 identified as an extrovert.

Similarly, when applied to all of the eighteen candidates, the result was compared with if the opinion of the subject themselves to see if it matched. The results are demonstrated in table 4.

TABLE IV. ACCURACY OF HYPOTHESIS

Correctly determined if Hypothesis matched with a subject opinion on if they are introvert/extrovert	Count	Accuracy %
Matched	14	77.78%
Did not Match	4	22.22%

## VI. CONCLUSION

Data pre-processing is an important step in any machine learning algorithm. Converting raw data or unstructured data to a structured format is crucial before any further processing. Code-switching, which is quite prevalent in textual data obtained from social media poses a difficulty in processing, hence needs to be removed. The current methods do not account for this code-switching or how emojis interact with the text messages.

The algorithm explained in this paper successfully takes various **WhatsApp chats as input and separates the texts of the two participants. This is followed by separation the emojis and sentences, & translating the sentences to a given language to make it unified and generic**. The emojis are further classified for sentiment analysis using the formula proposed.

The results of the pre-processing have been used to analyze the social behavior of the subjects in context & finally classifying them as introverts or extroverts.

The data pre-processing technique described in this paper can be used for the sentiment & behavioral analysis. Blending a scientific approach for behavioral analysis with practical engineering goals of **analyzing emotions in NLP texts can inspire a better and more intelligent approach to designing systems that interact with humans such as chatbots**. Along with this psychologists can combine these approaches to enhance

their models for sociological studies as well as in mental health care. In our next step, to design such a classifier is aimed, which gives a behavioral report of a person's emotions based on the emojis as well as a text sent by them.

## REFERENCES

- [1] J. Andjelic. (2019) Retrieved from <https://fortunly.com/statistics/whatsapp-statistics#ref>
- [2] WhatsApp. Retrieved from <https://www.whatsapp.com/about/>
- [3] Ka, Jisha & Jebakumar, Dr. (2014). Whatsapp: A Trend Setter in Mobile Communication among Chennai Youth. IOSR Journal of Humanities and Social Science. 19. 01-06. 10.9790/0837-19970106.
- [4] Pinchas. (2018) Retrieved from <https://www.telemessage.com/why-is-whatsapp-more-popular-than-sms-in-europe-infographic/>
- [5] Ljubešić, Nikola & Fiser, Darja. (2016). A Global Analysis of Emoji Usage. 82-89. 10.18653/v1/W16-2610.
- [6] Kumar, Naveen & Sharma, Sudhahsh. (2017). Survey Analysis of the usage and Impact of Whatsapp Messenger. Global Journal of Enterprise Information System. 8. 52. 10.18311/gjeis/2016/15741.
- [7] Bhatt, Anshu & Arshad, Mohd. (2016). Impact of WhatsApp on youth: A Sociological Study. IRA-International Journal of Management & Social Sciences (ISSN 2455-2267). 4. 376. 10.21013/jmss.v4.n2.p7.
- [8] Kootbodien, Ammaarah & Prasad, Nunna & Ali, Muhamad. (2018). Trends and Impact of WhatsApp as a Mode of Communication among Abu Dhabi Students. Media Watch. 9. 10.15655/mw/2018/v9i2/49380.
- [9] Lu, Xuan & Ai, Wei & Liu, Xuanzhe & Li, Qian & Wang, Ning & Huang, Gang & Mei, Qiaozhu. (2016). Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. 770-780. 10.1145/2971648.2971724.
- [10] Aravind K. Joshi. (1982) Processing of sentences with intra-sentential code-switching. In Proceedings of the 9th conference on Computational linguistics - Volume 1 (COLING '82). Academia Praha, CZE, 145-150. DOI:<https://doi.org/10.3115/991813.991836>.
- [11] Barman, Utsab & Das, Amitava & Wagner, Joachim & Foster, Jennifer. (2014). Code Mixing: A Challenge for Language Identification in the Language of Social Media. 10.13140/2.1.3385.6967.
- [12] Acampora, Giovanni & Loia, Vincenzo & Vitiello, Autilia. (2011). A cognitive multi-agent system for emotion-aware ambient intelligence. IEEE SSCI 2011 - Symposium Series on Computational Intelligence - IA 2011: 2011 IEEE Symposium on Intelligent Agents. 0.1109/IA.2011.5953606.
- [13] Posner J, Russell JA, Peterson BS. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol. 2005;17(3):715-734. doi:10.1017/S0954579405050340.
- [14] Tian, Ye & Galery, Thiago & Dulcinati, Giulio & Molimpakis, Emilia & Sun, Chao. (2017). Facebook sentiment: Reactions and Emojis. 11-16. 10.18653/v1/W17-1102.
- [15] Krebs, Florian & Lubascher, Bruno & Moers, Tobias & Schaap, Pieter & Spanakis, Gerasimos. (2017). Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. 10.5220/0006656002110220.
- [16] Hussien, Wegdan & Tashtoush, Yahya & Al-Ayyoub, Mahmoud & Al-Kabi, Mohammed. (2016). Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?. 10.1109/CSIT.2016.7549459.
- [17] WhatsApp. Retrieved from <https://faq.whatsapp.com/en/android/23756533/>
- [18] Unicode (2020) Retrieved from <https://unicode.org/emoji/charts/emoji-list.html>
- [19] Regex. Retrieved from <https://regexr.com/>
- [20] Novak, P.K.; Smailović, J.; Sluban, B.; Mozetič, I. Sentiment of Emojis. PLoS ONE 2015, 10, e144296.
- [21] SuHun Han. 2018. Retrieved from <https://pyip.org/project/googletrans/>

- [22] Haji, Hadi & Nawxosh, Salam. (2019). The Use of Emoticons among University Students: A Pragmatic Study. Zanco Journal of Humanity Sciences. 23. 10.21271/zjhs.23.1.19.
- [23] Alizah K. Lowell. (2016) Retrieved from <https://www.psychologytoday.com/us/blog/contemporary-psychoanalysis-in-action/201605/why-do-we-use-emojis>
- [24] Zhou, Z., Xu, K. & Zhao, J. (2018) Extroverts tweet differently from introverts in Weibo. EPJ Data Sci. 7, 18 <https://doi.org/10.1140/epjds/s13688-018-0146-8>.