Emojis Usage in Social Media by Demographics

Stylianos Nicoletti

4th Year Project Report
BSc Computer Science
School of Informatics
University of Edinburgh

2019

Abstract

Twitter is a social networking service which allows users to create short messages to be posted and shared with public online. This kind of messages are called “Tweets”. Tweets can be in different forms, with a variety of content, including text, emoji, videos, pictures, audio etc. Twitter has become extremely popular type of blogging, especially on the mobile web.

Emoji are pictorial representations of facial expressions, foods, sports, buildings or even animals. Emoji are an advancement of the text-based, picture-like emoticons that users were used in e-mails and early SMS text messages. They are supported by almost all latest platforms, such as iOS, Android, Windows and macOs. By using emoji in tweets can help us identify the mood or state of mind that a person may have been when posting. Emoji are more engaging than simple text, and can express a lot of things visually without using extra words.

This project involves analysing the usage of emoji across Twitter by gathering millions of public tweets. We would like to know if emoji are used differently by different persona. A quantitative analysis study on the differences of emoji usage on Twitter by gender and race is presented. Our experiments use the timelines of a set of up to 40,000 Twitter users from four different locations that are manually labelled for gender and race. Our observations show that there are clear differences in the emoji used by males and females, and also among different racial groups. To demonstrate the significance of these differences, we built gender and race classifiers that are trained solely on the emoji used by each user. Our classifiers achieved 78% and 80% accuracy on the gender and race respectively, which confirms the significant difference in the emoji used by each demographic.

Acknowledgements

First of all, I would like to express my very great appreciation to my supervisor Dr Walid Magdy for his helpful and effective suggestions during the process of this research work. The successful outcome of this research is mostly due to the help Walid has provided me and I can not thank him enough.

Next, I would like to thank Alexander Robertson for providing the Twitter data, emoji extraction tool and generally for all of his assistance.

I would also like to thank my family and friends for motivating me through all the long nights I spent working in the computer labs.

Table of Contents

Chapter 1

Introduction

Emoji are characters used in electronic messages and social network posts. They can exist in various forms such as facial expressions, animals, plants, buildings and common objects.

Our work contributes the first in-depth analysis of the connection between emoji and two demographic features, gender and race. We apply a quantitative analysis for the difference in emoji usage between different gender and race users. Our results show that there are differences in emoji usage between male and female Twitter users, and between Asian, Black, Hispanic and White users. To further explore the importance of these dissimilarities in usage, we examined using emoji distribution of users as the only features for training gender and race classifiers. Our hypothesis is that the distribution of emoji used in a user's timeline might be a sufficient indication to detect the gender and the race of this user.

All analysis and experimentation in our study are based on a set of 40,000 manually annotated Twitter user accounts, where around 20,000 of them were both gender and race annotated. The results of our classification show that these dissimilarities in emoji usage can be successfully used by classifiers to automatically detect the gender or race of Twitter users. Using *only* the emoji found in a users posts, the gender and race of a user can be predicted with accuracy of 78.2% and 80% respectively. These classifier results highlight the significant variation in emoji usage between different demographics, which allows prediction of gender and race using emoji *only*; something that was not attempted in past researches. As a result, we prepared a paper to be submitted to the 11th International Conference on Social Informatics [soc19], (see Appendix A).

The rest of the study is organised as follows: Section 1 provides a background and history of the research topic. Section 2 gives a background on the related work. Section 3 describes the methodology used for our data collection, preparation, analysis and experimentation. Section 4 presents the analysis results of tweet and emoji distribution, emoji usage by gender and race as well as percentage usage of emoji in different locations of the world. Section 5 reports the results of making use of emoji as features for gender and/or race detection. Finally, section 6 concludes the work and provides possible future research.

1.1 Background

Social networking services are fast growing networks, allowing users to share their ideas, photos, videos, links and posts to the rest of the world. Social networking services vary in structure and features they provide. They can run on many devices such as tablets, smart phones, desktops and laptops. Some examples of these services are Facebook, Google+, Twitter and Linked-In. In this paper we are interested in Twitter.

Twitter, is one of the rapid expanding networks. As a user of Twitter, you can follow other people you are interested, they can follow you back, and exchange messages with them. If your followers find your post interesting they can re-post it mentioning you ("Re-Tweet"). If their followers are interested in your "Tweet", they can follow you back.

Emoji (icons used to represent emotions, ideas, or objects) became a formally recognised component of the Unicode Standard in 2010 [DE14] and have a penetrating influence in computer-mediated communication, such as Twitter. The use of the word "emoji" has increased hugely. An evidence to this is that in 2015, 😂, officially called the "Face with Tears of Joy" emoji, became Oxford Dictionaries' "Word" of the Year [Dic15]. Apart from that, starting from the 20th of May 2015, Domino's customers are able to order pizza via Twitter with a quick "Pizza" 🍕 emoji post at the brand's Twitter account [Mon15].

The increasing usage of emoji has received a great attention in social media and there is a growing body of research into their linguistic and sociological properties since most natural language processing techniques are unsuccessful due to spelling, grammatical and punctuation mistakes. Several studies have recently analysed the usage of these emoji by users [KMW17], including semantic meaning [NSSM15], sentiment effect [MKTS$^+$17], representation of user identity [RMG18], and usage by location [LF16]. The importance of these studies consists better understanding of online human communication, which has wide range of applications in social linguistics and social science in general.

1.2 History

The birth of social media began in 1997 with the introduction of Six Degrees. Six Degrees allowed it's users to create a profile, add friends and post in forums with other users. Six Degrees lasted only until 2001. In 2002, Friendster was the next big social media site, using features from Six Degrees, connecting users by common interests.

A year after that, LinkedIn took a different point of view, using social media in a business manner. LinkedIn continuous to be one of the thriving social media sites as of today with around 590 million users [Asl18]. In 2003, MySpace was introduced, with common characteristics of today's social media sites. MySpace not only allowed to create user profiles and add friends, but also add music to others' page or even instant message them.

In 2004, Facebook was created by Mark Zuckerberg and other university students at Harvard, which was originally designed to be used only for college students. Facebook is a social networking site that makes it easy for you to connect and share with family and friends online. Facebook introduced the action to "Like" other users content, such as photos or posts. As of 2006, Facebook has become the most successful social media site and continues to add new features, such as the possibility to stream live videos.

YouTube was introduced in 2005, a video sharing social media site, allowing users to upload their own videos to public. A lot of users earn money from sharing their recordings, having corporate sponsors who pay for product placement in their videos, known as "YouTubers".

Twitter began as an idea in 2006. Originally envisioned as an SMS-based communications platform, and referred as "twttr". The first tweet was posted in March 2006 [Dor06], and from then the company experienced rapid initial growth. Statistics from October 2018 showed that Twitter had 326 million monthly active users [Gro18].

From then, other social media sites were introduced such as Instagram, Google+ and Snapchat. Table 1.1 shows a complete time-line of major social network sites. Today, social media are supported by thousands of platforms, all providing the same, but slightly different intentions.

Table 1.1: Time-line of major social network sites.

Year	Site Name
1997	Six Degree AOL
1998	Google
2002	Friendster
2003	LinkedIn MySpace
2004	Facebook
2005	YouTube
2006	Twitter
2007	Tumblr
2010	Instagram Snapchat Google+
2012	Vine

Emoticons from the late 1990s are pictorial representation of facial expressions. An example of emoticon is **:-)** , the "Smiley Face" emoticon. On the other hand, emoji offer a wider spectrum of concepts such as foods, sports, buildings, animals etc. An equivalent emoji to the emoticon above can be 🙂, the "Slightly Smiling Face" emoji.

Emoji were invented in 1999 by Shigetaka Kurita and were intended for users in Japan. The first emoji set was very simple and only 12 by 12 pixels. The idea came when image-based text messages could soften the problem of the limited 250 characters of

"i-mode" mobile internet service of that time [Lee18]. Emoji are used outside of Japan since 2010, when they were adopted into the worldwide character-encoding standard Unicode.

Today, emoji are available in almost all messaging apps and platforms, including Twitter. However, different platforms have distinct emoji styles, emoji can translate across platforms, thanks to Unicode character-encoding. Twitter has introduced "Twemoji" [twe14], which is their emoji graphics together with a JavaScript library to handle them. Today, Twemoji offers support for 2,841 emoji for all devices. Table 1.2 shows an example conversion of five emoji from Unicode to Twitter style.

Table 1.2: Unicode to Twitter emoji style.

Unicode	Twitter Emoji	Description
U+1F609		Winking Face
U+1F622		Crying Face
U+1F349		Watermelon
U+1F42C		Dolphin
U+1F45F		Athletic Shoe

Chapter 2

Related Work

Previous work has broadly examined the range of semantics and pragmatics functions satisfied by emoji. Some researches focused on ranking emoji as positive or negative by analysing people's attitudes and opinions from the sentiment of the tweets in which emoji occur [NSSM15]. A similar work, analysed how emoji definitions vary for rendering across five different platforms (Google, Apple, Microsoft, Samsung and LG) since emoji look different by each platform, identifying the variance of interpretation in terms of sentiment [MTSC$^+$16]. Effort has been put on defining a system that will maintain emotional intent by replacing the senders emoji with one that the receiver clarifies as most similar, trying to reduce significance of individuals and platform differences in emoji understanding with and without textual content [TF16, MKTS$^+$17, MLK$^+$18].

In addition, research across three virtual platforms (email, text message and a social networking site) has revealed different themes and sub-themes for emoji usage: "aiding personal expression", "establishing emotional tone" as well as "to lighten the mood" [KWM16]. In similar fashion, research focused on people willingness in using positive, negative, neutral and non-facial emoji for non linguistic signals in communication; reporting differences between emoji types in terms of intended uses [HGaTNL17]. Apart from that, examination is done on how emoji used in highly personalised and purposefully secretive ways between close friends, family or partners to express personal sentiments [WG18]. Another research examined the way that Japanese teen mobile users use emoji, revealing that "emoji allow teens to manage communication climate" and "construct and express their aesthetic selves" [Sug15].

Several works trained models for generating emoji based on input images [CMS15], as well as for predicting which emoji are recalled by text-based tweet messages [BBS17]. Investigating the relation between words and emoji; and capturing the emphasised semantics of emoji. In a similar way, models were trained to predict emoji in Instagram posts, containing both pictures and text [BBRS18]. A model was also build for emoji similarity, by making use of semantic data collected from Twitter [PDR17].

Moreover, embeddings were trained for all Unicode emoji, based on their description that can be used in natural language processing applications and outperform skip-gram

models [ERA+16]. In a similar manner, analysis on emoji semantic meaning was done by computing the embeddings of emoji and words based on a large corpus containing a collection of inputs from a popular emoji keyboard [ALL+17]. Relative work has extended distant supervision through emoji prediction on tweets dataset, achieving better performance over previous approaches [FMS+17].

Research has also explored broad patterns of emoji usage based on skin colour. This includes analysing the use of the skin tone emoji geographically based on Twitter data, showing that skin tone emoji by country have close similarity to the skin tone of country's population [Coa18]. Besides that, researches have explored the connection between emoji usage and ethnic identity. By using skin tone modifiers on emoji on Twitter, this study revealed that users with darker-skinned profile photos use darker-skinned emoji more often than users with lighter-skinned profile photos; and that most skin tone emoji usages match the colour of user's profile picture [RMG18]. Research has analysed the use of emoji analogous to both skin tone and gender using Twitter data, showing that patterns associated to skin colour and gender seem to be reflected on the use of emoji modifiers. Additionally, a study has revealed that female modifiers tend to be semantically close to emoji related to "love" and "makeup", where male modifiers appear closer to emoji related to "technology" and "business" [BCC18].

A research based on emoji smart-phone users' data confirms that there is a noticeable difference in emoji usage by males and females; suggesting that these difference are acceptable by machine learning algorithms [CLS+17]. Furthermore, a research based on 86,702 Facebook users showed that emoticon usage counts were predicted primarily by age and gender, explaining 16% of the variance, whereas user's personality level explained less than 2% of the variance, suggesting that emoticons usage may support predicting user's demographic [OKP+17]. Equivalently, another study analysing emoticons in WhatsApp communities suggests that a user's gender has a great impact in defining how emoticons are included in conversations for relational intents [PS19]. Besides that, a study identified a change in emoticon usage when moving from same-gender to mixed-gender newsgroups [Wol00].

Significantly more consideration has been on demographic prediction. Researches in fields related to social science, linguistics and psychology are interested in massive amounts of data generated by online social media. Nevertheless, in order to take advantage of this kind of data, it is generally crucial to determine social variables based on demographic information. As an example, a sociolinguist (a person who studies the social and cultural factors that influence linguistic communication) that explores the use of a particular construction online, wants to know at least one of race, age, sex, ethnicity or social class of users in the collected data. Without knowing one of these demographic information, it is challenging for them to do anything except from simply counting examination of facts. Demographic information is rarely made available for social media users by most platforms. However, platforms like Facebook encourages their user to provide their personal information. On the other hand, platforms such as Twitter and Instagram allow users to provide limited amount of personal information. As a result, multiple methods have been developed by researchers for deducing the required variables from the data that is actually available.

The most comprehensive overview of these methods could be found in [CGN17], were a range of facial recognition to classic supervised and unsupervised machine learning methods focuses on detecting age, sex and race/ethnicity from text extracted from a user's online presence. This is by far the most advanced common resource and there is a large body of work on precisely which source of text provides the strongest signal for which demographic. These include clustering of gender behaviour and linguistic style in user's tweets [BES14] as well as gender prediction from Twitter users' profile information. They also include predictions of user's demographic in micro-blogs, by taking into consideration their video analogous behaviour [WXMX16]. Besides that, identification of Twitter users' age, occupation and social class using text found in their profiles [SMBW15], plus political orientation and ethnicity identification using users' behaviour, linguistic content and network structure of users' profiles in Twitter [PP11].

While text is arguably the most useful resource available for determining demographics of Twitter users, metadata can be extracted to improve age classifiers accuracy, such as birthday announcement, language and selected emoji in tweets [MLKCR17]. However, there have to date been no attempts to predict demographics using *only* emoji. In this study we analyse only emoji in users' feed, to demonstrate the significance usage differences of them by users' gender (male, female) and race (Black, Asian, White, Hispanic). Resulting to notable accuracy of 78% and 80% of users' gender and race identification respectively using emoji alone.

Chapter 3

Methodology

In this section the methodology used during the research is presented. This section explains different approaches on how data was collected and processed as part of both statistical analysis and machine learning applied to it. It also includes methods used for results visualisation, difficulties faced and actions taken to overcome them.

Python[1] programming language was used for almost all tasks of the research since it comes with a huge amount of inbuilt libraries, including machine learning libraries. Python is widely used in scientific researches because it is easy to experiment and prototype algorithms. In addition, Jupyter notebook [2] (document containing both computer code and rich text elements) was used making our code human-readable (containing both the analysis and the results) and executable. Git version control was also used for management of changes in our code where most of our implementation can be found [3].

3.1 Data Collection

In general, the most important factor of a research is the data collected. This is because evaluation and experimentation is based on that data. At the beginning of our research Tweepy [4] was used to capture Twitter data to be analysed. Tweepy is a Python library used for accessing the Twitter API . However, the appropriate API authorisation was required to access Twitter stream. Thus, a Twitter developer account application needed to be approved by Twitter after informing them the purpose of our application. The email informing Twitter the purpose of our study, together with Twitter's confirmation email can be seen in Appendix B. After Twitter's confirmation, we were able to use the API credentials to stream public tweets. Part of the streaming code can be seen in Appendix C.

[1] https://www.python.org/

[2] https://jupyter.org/

[3] https://github.com/stylianosnicoletti/Emojis-usage-in-social-media-by-demographics

[4] https://tweepy.readthedocs.io/en/3.7.0

We let the stream run for 3 weeks on a server provided by the University of Edinburgh since the process of capturing tweets required twenty-four hours service to store an enormous amount of tweets. In Figure 3.1 the structure of a simple tweet captured can be seen. Apart from the actual tweet text and user's name, restricted information of the corresponding user that posted the tweet could be captured, including "friends_count", "language" and "profile_image_url_https". However, no demographic information (race, age, sex, ethnicity or social class) was available for users in our captured tweets.


```
"created_at":"Thu Apr 06 15:24:15 +0000 2017",
"id": 850006245121695744,
"id_str": "850006245121695744",
"text": "Today we're sharing our vision",
"user": {
   "id": 2244994945,
   "id_str": "2244994945",
   "name": "TwitterDev",
   "screen_name": "TwitterDev",
   "location": "Internet", ++
   "url": "https://dev.twitter.com/",
   "description": "Your source for Twitter news",
   "verified": true,
   "followers_count": 477684,
   "friends_count": 1524,
   "listed_count": 1184,
   "favourites_count": 2151,
   "statuses_count": 3121,
   "created_at": "Sat Dec 14 04:35:55 +0000 2013",
   "utc_offset": null,
   "time_zone": null,
   "geo_enabled": true,
   "lang": "en",
   "profile_image_url_https": "https://pbs.twimg.com/"
 }
```


Figure 3.1: Streamed tweet structure.

Subsequently, by having each user's profile image address, we attempted to predict Twitter's users gender and age from their profile pictures. We have used Open-CV[5] (Open source computer vision library) trained classifiers for detecting faces in profile pictures as well as age and gender classifiers trained on IMDB celebrities and Wikipedia face images [RTG15][RTG16]. The classification accuracy wasn't convincing, therefore we needed another way to collect adequate data and proceed with the research. Identifying the gender and age from users' profile picture was an extremely challenging task, requiring a lot of time and it was not part of our research scope.

Thankfully, gender and race annotated data was provided by Alexander Robertson, a research postgraduate student at the University of Edinburgh. Data consists of over 18 million tweets collected from 40,000 Twitter accounts posted from March to October 2018.

Twitter Streaming API (1% sample) was used for the data provided to us, sampling 3.4 million tweets made by 2.6 million unique users on the 14th of March, 2018. The location field was inspected in the collection, if self-provided by users, and a random

[5]https://opencv.org/

sample of 20,000 users from three different location was collected. Locations focused was London, Johannesburg and New York City. The reason for choosing these three location is their demographic composition, having different ratios of ethnic groups, giving variations for the experimentation to be applied. For example, Johannesburg has generally Black population [Sta11], London mostly White [Off11]. However, New York City is balanced in ethnic groups and a large amount of residents could be identified as Hispanic or Latinx, besides their racial identity [Uni10]. Additionally to these three locations, a forth group was selected of 20,000 users that were sampled randomly, regardless to their location, making sure than none of these users exists in other groups.

Effort was made to retrieve the public profile pictures for all 80,000 users collected, to be used for annotating the gender and race of the user. For some users, retrieving their profile pictures was not possible since many had removed their accounts, set their profiles to private or been banned by Twitter. Therefore the amount of users for each group was randomly reduced to 10,000 users. Table 3.1 shows in more detail the final number of tweets collected from users by each of the four groups.

Table 3.1: Number of tweets collected in final dataset from 4 different locations.

Location	Number of Users	Number of Tweets
Random	10.000	11.326.306
Johannesburg	10.000	2.631.549
London	10.000	2.308.977
New York City	10.000	2.524.608

Then, human annotators were hired using Figure Eight[6] to validate user photos as well as check the gender and race of valid photos. Two choices were provided for gender: {male, female}, and four choices for ethnicity: {White, Black, Asian, Hispanic}. Annotators classified a photo as "invalid" if it contained multiple or no faces, if the photo was not fully coloured, or if the photo subject's face was obscured. Each profile photo was annotated by three annotators. Only Twitter users where at least two annotators agreed on both photo validity and gender/race were used in this study.

Out of the 40,000 user accounts, only 19,382 accounts had profile pictures that were agreed on by at least two annotators on the gender and race of the user. Statistics on the annotated accounts are shown in Table 3.2. As shown, males and females distribution are almost even, while the distribution of race have the majority as Black and White, while Asian and Hispanic are much less.

Table 3.2: Number of users with valid photos, per gender and race.

	White	Black	Asian	Hispanic	Total
Female	5,244	4,203	509	249	10,20
Male	4,704	3,936	782	205	9,627
Total	9,948	8,139	1,291	454	**19,382**

[6]http://www.figure-eight.com

Later, the most recent 3,200 tweets were retrieved for each user and re-tweets were excluded. Users who had since set their profile to private, or had been deleted, were removed from the final dataset. The total number of tweets collected for the 40,000 accounts to be used in our analysis are over 18 million tweets. Data provided could also be stored in local machine requiring only 4.3 GB of space, compared to the initial heavy stream approach were everything had to be stored and accessed from a server machine.

3.2 Data Preparation

Data was provided in CSV (Comma-Separated Values) format, separated for users and tweets for all four locations. In Figure 3.2 you can see in more detail the fields of both tweets and users CSVs. In users data some of the important attributes for our study are: "ethnicity", "gender" and "user_id", where for the tweets data, important attributes are: "user_id" and "text", containing the actual tweet's text. In order to make use of the data, joining users with their posts by "user_id" field was required.

Users Data	Tweets Data
_unit_id	user_id
_golden	user_name
_unit_state	user_screen_name
_trusted_judgments	user_statuses_count
_last_judgment_at	tweet_id
ethnicity	hashtags
ethnicity:confidence	is_quote_status
gender	text
gender:confidence	created_at
reason_not_valid	source
reason_not_valid:confidence	in_reply_to_screen_name
valid_photo	
valid_photo:confidence	
which_icon_most_closely_matches_the_persons_skin_tone	
which_icon_most_closely_matches_the_persons_skin_tone:confidence	
orig__golden	
ethnicity_gold	
ethnicity_gold_reason	
gender_gold	
gender_gold_reason	
image_url	
reason_not_valid_gold	
reason_not_valid_gold_reason	
user_id	
valid_photo_gold	
valid_photo_gold_reason	
which_icon_most_closely_matches_the_persons_skin_tone_gold	
which_icon_most_closely_matches_the_persons_skin_tone_gold_reason	

Figure 3.2: Fields of tweets and users data.

Tweet and user CSV files from all four locations were loaded into Pandas[7] dataframes (two dimensional labeled data structures aligned in a tabular fashion of rows and columns) and from there joined on similar attribute "user_id". Once joined, we extracted the dataframes into 4 different CSV files, one for each location (Random Worlwide, New York City, London, Johannesburg).

Since we were interested only in tweets containing emoji, Emoji-extractor[8] (Python

[7] https://pandas.pydata.org/

[8] https://github.com/alexanderrobertson/emoji-extractor

package that counts the emoji in a string and returns the emoji and their counts) was used to count emoji in tweets. Figure 3.3 shows a simple use of Emoji-extractor, printing the counts of emoji in a sample tweet. As you can see, frequencies of repeated emoji are also taken into account, since a single tweet can consist an emoji multiple times.

```
sample_tweet = "This 🤔 is 🔥🔥 a 💗 sample 🤣🤣 tweet 🤣"
emoji_frequencies = emoji_extractor.count_emoji(sample_tweet)
print(emoji_frequencies)
```

```
Counter({'🤣': 3, '🔥': 2, '🤔': 1, '💗': 1})
```

Figure 3.3: Emoji-extractor example.

From there, four new datasets were created, one for each location, using emoji as features, the 10000 users from each location as instances, the gender and ethnicity of each user as classification target and the emoji counts from each user as feature variables. Figure 3.4 show an example of the structure of the prepared datasets.

Emoji Count

User

	💩	😀	😎	👍	...	😊	😍	😂	gender	ethnicity
0	0	0	0	0	...	0	5	0	female	white
1	0	0	0	0	...	0	15	36	female	white
2	0	0	0	0	...	0	0	0	female	white
3	0	0	0	0	...	0	0	8	male	asian
4	0	0	1	0	...	0	0	0	male	white
5	0	0	0	0	...	0	10	18	female	hispanic

Figure 3.4: Prepared dataset example structure.

Since we had do deal with large amounts of tweets, we were limited in the amount of available random access memory and execution time for the preparation of the dataset. As a result, smaller random sample of data was used to work through the problem before applying it to the whole data and certain control flows such as nested iterations were avoided when possible, saving us both execution time and memory.

3.3 Analysis Process

The data analysis process consists of methods used to analyse tweet emoji distribution, how most frequent emoji by each group were extracted and different techniques on identifying distinctive emoji for each group of persona.

First of all, we analysed the frequency of tweets consisting emoji in their content. This was done using the initial unprepared dataset provided. For each of 40,000 users

the frequency of all tweets as well as the frequency tweets including emoji in their content were counted. From there, we were able to get the minimum, maximum, mean, and standard deviation of tweet occurrences in users' timeline, for both all tweets and tweets accommodating emoji. Figure 3.5 shows in more detail what information could be derived and stored into a dataframe.

	Tweets	Tweets_with_Emoji
mean	344.421425	72.725075
std	627.965717	197.683637
min	0.000000	0.000000
25%	8.000000	0.000000
50%	83.000000	7.000000
75%	343.000000	50.000000
max	3267.000000	3254.000000

Figure 3.5: Tweets in users' timeline summary extracted into dataframe.

Plotly[9] (Python graphing library) was used to display the acquired information into box plots. Box plot is a typical way of displaying the distribution of data based on a five number summary: {minimum, first quartile, median, third quartile, maximum}. With box plots we could see how symmetrical our data is and identify possible outliers lying outside upper and lower fences. In order for us to be able to draw the box plots, the following calculations needed to be made:

$$UpperFence = Q3 + (1.5 * InterquartileRange) \tag{3.1}$$

$$LowerFence = Q1(1.5 * InterquartileRange). \tag{3.2}$$

where,

$$InterquartileRange = Q3 - Q1 \tag{3.3}$$

and,

Q1 (first quartile) is the middle number between the smallest number and the median of the dataset, Q3 (third quartile) is the middle value between the median and the highest value of the dataset; and median is the middle value of the dataset.

After we finished summarising the data, our next step was to identify most frequent emoji by each group of persona. Using the four datasets prepared in Section 3.2, one for each location, we were able to get the frequencies of each emoji and store them into dataframes, sorted by emoji with highest counts for each group respectively. Figure 3.6 shows an example of how emoji frequencies were stored for female users from all four locations (sorted by emoji with highest frequency).

[9] https://plot.ly/python/

Emoji	Count
🔥	51490
😍	113112
🤣	115111
❤	136339
😭	194472
😂	621533

Figure 3.6: Dataframe showing the 6 most frequent emoji by female users and their counts.

With the dataframes created we were able to acquire the top used emoji for each group, including gender, race and location of tweets. Matplotlib[10] (Python 2D plotting library) was used to plot bar charts, showing the 15 most frequent emoji on the x-axis and their frequency values on the y-axis for each group respectively. Placing emoji on the bar chart is not yet supported by plotting libraries because of the outdated Unicode fonts with limited emoji. Thus, we have used Python-emojipedia[11] (Python library including emoji data from Emojipedia.com) to load emoji images as rendered in Twitter platform and customly place them in our bar charts.

Our next step, was to identify the percentage usage of each emoji by each group. Since, we already knew the occurrences number of each emoji by each group, the percentage usage of each emoji was calculated as follows:

$$e_i|x = \frac{count(e_i|x)}{count(e|x)} \tag{3.4}$$

where $(e_i|x)$ is the frequency emoji i in group x and $(e|x)$ is the frequency of all emoji in group x. For example, $e_i|female$ is the frequency of this emoji within the timelines of female users and $e|female$ are the total frequencies of all emoji in female timelines.

Then, customised functions using Matplotlib and Python-emojipedia were used to plot the percentage usage of emoji for each of the four locations (Random Worldwide, Johannesburg, London, New York City) into pie chart representation, as we have done for bar chart representation.

Considering that most frequent emoji turned out to be similar for almost all groups we needed a way to spot emoji that were distinctively used by different groups.

One approach was to investigate the differences in percentage emoji usage. We calculated the most distinctive emoji set for each gender and race by measuring the difference in percentage of usage between a given group and the others. For an emoji e_i, the distinctiveness of e in group x, $D(e_i|x)$ is calculated as follows:

$$D(e_i|x) = \frac{count(e_i|x)}{count(e|x)} - \frac{count(e_i|\bar{x})}{count(e|\bar{x})} \tag{3.5}$$

[10] https://matplotlib.org/index.html
[11] https://github.com/bcongdon/python-emojipedia

where $count(e|x)$ is the count of all emoji in x, and $\bar{x}$ represents all the other groups excluding group x. For example, $D(e_i|female)$ is the percentage of usage of this emoji within the timelines of female users minus its usage in males timelines. Similarly, $D(e_i|black)$ is the percentage of usage of e within the timelines of black users minus its usage percentage in timelines of While, Asian, and Hispanic users combined.

The most distinctive set of emoji (with the highest difference in usage percentage) was extracted for each group. Similarly to most frequent emoji plots, we plotted bar charts, showing the 15 most distinctive emoji on the x-axis and their difference in usage percentage on the y-axis for each group respectively.

Capturing distinctive emoji by calculating the difference of frequency percentages among two sets was not reliable in some cases. For example, if $e_1|female$ was 11% and $e_1|male$ was 3%, $e_2|female$ was 8% and $e_2|male$ was 0%, our measure will treated $D(e_1|female)$ (11%-3%) and $D(e_2|female)$ (8%-0%) to be equal, meaning that emoji_1 and emoji_2 will have the same distinctive power, however, emoji_2 should have been more distinctive as it only appears in female timelines.

Another approach to determine which emoji are more distinctively used by each group was to use Chi-squared (χ^2) feature selection. The Chi-squared test, calculates chi-square statistics between every feature variable and the target variable and observes the existence of a relationship between the variables and the target as follows:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{3.6}$$

where,

$$Expected = \frac{RowTotal * ColumnTotal}{OverallTotal} \tag{3.7}$$

and,

$$DegreeOfFreedom = (Rows - 1) * (Columns - 1) \tag{3.8}$$

In our case, we have used the Chi-Square test to find which "emoji" variables have an association with the "gender" and "race" variable respectively. If the null hypothesis is rejected, then the "emoji" variable it's an important variable for target group variable.

For Chi-Square test, we have used emoji as attributes, users as instances, the gender and race of each user as classification target, and the percentage usage of emoji from each user as feature variables.

By setting the significance level to 5%, our function calculates the Degree of Freedom and χ^2 and returns the P-values for each "emoji" variable. To reject the null hypothesis, the calculated P-Value for each emoji needed to be below a predetermined threshold of 0.05, meaning that our experimental results would be highly unlikely to occur if there was no real connection between a specific emoji and our target group (gender and race

variables). The smaller the P-value of each emoji, the more important the emoji is for the target group.

As we have done in the percentage difference approach, each group was compared to the rest of the groups when feed to our χ^2 function. For example, when we needed to see how important an emoji was for Black race, χ^2 of an emoji was calculated using "Black" and "Not_Black" (White, Hispanic, Asian) target variables.

For all two gender and four race groups, respectively, a graph was plotted, showing the 20 most distinctive emoji (with the lowest P-values) on the x-axis and their Log P-values on the y-axis. A red dotted line was also plotted to show the significance level and which emoji P-values fall beneath it, rejecting the null hypothesis.

Our final approach for selecting distinctively emoji for each group was to calculate feature (emoji) importance by training a Random Forest [Ho95] with our data. Random Forests are excellent models for many data science projects. Most machine learning methods require prepossessing before trained. Random Forest is not that case, and can take both categorical and numerical variables as input.

Most of the times, these methods are used for prediction, however, Random Forests and decision trees in general have a helpful feature call "feature importance". By looking at the feature importance, we can know which features are important for making the model predictions. In many cases, this feature is used to create new features and drop out noise features. In our case we have used it to see which features (emoji) are more important for predicting the target group.

Decision trees in general work by splitting data into subsets, and continue to split until a noticeable relationship between the variable and the target is identified. They determine which splits will be the most important for distinguishing the target class. The more often a feature is chosen as a splitter, the higher its "importance" for the classification. More information on how Random Forest works can be seen in Algorithm 1 and Algorithm 2 of Section 3.4.

For this experiment we have used Scikit-Learn[12] (Python library with machine learning tools for data mining and data analysis) and fitted a Random Forest model on data containing 2579 emoji as features, users as instances, the gender or race of each user as classification target and the percentage usage of emoji from each user as feature variables. Part of the model fitting and feature importance extraction code can be seen in Appendix D.

A 5-Fold Cross-validation (re-sampling procedure for evaluating models) was used to determine a suitable value d (maximum depth of the tree), achieving the best "F1 Score" for gender classification and best "Accuracy Score" for race classification. More information on "Cross-validation", how "F1 Score" and "Accuracy Score" are calculated can be found in Section 3.4.

The "feature_importances" attribute in our trained Random Forest model could extract an array of each feature's importance. To find the corresponding emoji importance for each target group (eg. Male or Female for gender classification) we have used a custom

[12]https://scikit-learn.org/stable/

function that takes as inputs the features importance array and the data used to train the model; and returns a dictionary of the emoji importance for each target group.

Finally, we plotted bar charts, showing the 15 most distinctive emoji (with highest importance values) on the x-axis and their model prediction importance on the y-axis for each group respectively.

3.4 Classification Process

For all classification tasks we have used 2579 emoji as features, users as instances, the gender and/or race of each user as classification target and the percentage usage of emoji from each user as feature variables.

For gender classification we have used 16,005 users, which were confidently gender labelled and used emoji in their tweets. User's class counts used can be seen in more detail on Table 3.3. 80% of these users were randomly chosen to be used as part of train set and 20% for testing our classifiers.

Table 3.3: Counts of users used for gender classification.

Class	Count
Male	7521
Female	8484

For race classification we have used 14,316 users, which were confidently race labelled and used emoji in their tweets. User's class counts used can be seen in more detail on Table 3.4. Again, 80% of these users were randomly chosen to be used as part of train set and 20% for testing our classifiers.

Table 3.4: Counts of users used for race classification.

Class	Count
White	7047
Black	6161
Asian	780
Hispanic	328

Since our classes were not balanced, meaning that we had more instances of a class compared to other classes, we tried different methods to deal with it. Firstly we have used the original data, to see how low instances of "Asian" and "Hispanic" users can affect our classification results. Then we tried to both oversample and undersample the train set, respectively.

The first approach was SMOTE oversampling with 4 neighbours on our train set before training our classifiers. SMOTE (Synthetic Minority Over-sampling Technique)

[CBHK02] creates new synthetic data points by taking the vector between the current data point and one of its k-neighbours; and by multiplying this vector by a random number **x** which lies between 0, and 1. We have chosen 4 neighbours since we had 4 different races to be predicted (White, Asian, Black, Hispanic). A visual representation of resulting train set (after oversampling) can be seen in Figure 3.7, having balanced counts of users for all 4 races, equal to the counts of the majority race (White) of the original data.

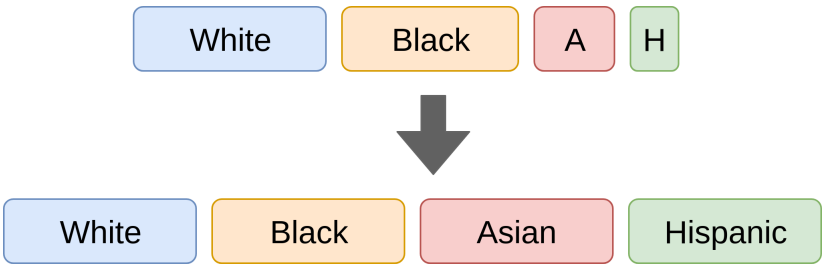


Figure 3.7: Oversampling Train Set.

The second approach was to undersample the whole data set using Tomek's links and a random undersampler. Tomek's links are pairs of very close points, but of opposite classes. Tomek's links were used to remove points of the majority class of each pair and increase the space between our classes, smoothing the classification process [Tom76]. Then, we undersampled the majority classes (White, Black, Asian) by randomly choosing users from each class. A visual representation of resulting data set (after undersampling) can be seen in Figure 3.8, having balanced counts of users for all 4 races, equal to the counts of the minority race (Hispanic) of the original data.

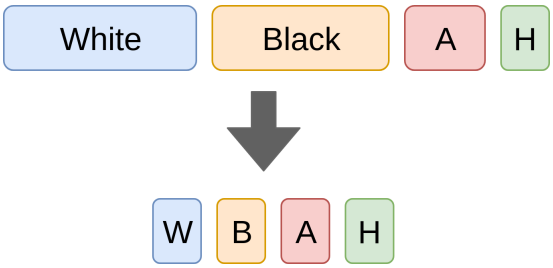


Figure 3.8: Undersampling Data Set.

For the combined race and gender classification we have used 14,210 users, which were confidently gender and race labelled and used emoji in their tweets. User's class counts used can be seen in more detail on Table 3.5. As previously done, 80% of these users were randomly chosen to be used as part of train set and 20% for testing our classifiers.

Table 3.5: Counts of users used for combined gender and race classification.

Class	Count
White Female	3843
White Male	3153
Black Female	3197
Black Male	2920
Asian Female	298
Asian Male	446
Hispanic Female	188
Hispanic Male	140

For all classification methods we have used Scikit-Learn (Python library with machine learning tools for data mining and data analysis). In addition, we have used Imbalanced-learn [LNA17] (Python library with tools that help deal with imbalanced data) for both SMOTE oversampling and Tomek's links with random undersampling techniques.

For evaluating our gender classification models we have used "F1 Score" as well as "Accuracy Score" on test set. "F1 Score" is the harmonic mean of precision and recall, calculated using:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.9}$$

where,

$$Precision = \frac{TP}{TP + FP} \tag{3.10}$$

and

$$Recall = \frac{TP}{TP + FN} \tag{3.11}$$

The terms positive (P) and negative (N) refer to the classifier's prediction, and the terms true (T) and false (F) refer the observation.

On the other hand,"Accuracy Score" is ratio of correctly predicted observations to the total observations, calculated using:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.12}$$

For evaluating our race classification models as well as the combined gender and race classification models, we plotted the confusion matrix on test set. This time our classification was not binary and visualising performance with a confusion matrix was made easier. We could conveniently look at all correct and incorrect predictions since each row of the matrix represents the instances in the predicted class while each column represents the instances in the actual class.

Multiple machine learning classification techniques were used to classify both the gender and race of each user:

- Prior Dummy Classifier.
- Gaussian Naive Bayes.
- K Nearest Neighbour.
- Random Forest.
- Linear SVM.
- SVM with Chi-squared kernel.

At first we have used a Prior Dummy classifier as our baseline model. Prior always predicts the class that maximises the class prior, which is similar to choosing the most frequent class. This classification method helped us to compare other algorithms used and see how well they perform.

Then we have used Gaussian Naive Bayes algorithm for classification, which is a simple and popular classification technique both for binary (Male, Female) and multi-class (Black, White, Asian, Hispanic) problems. Naive Bayes methods are one of the most powerful probabilistic classifiers used in machine learning and perform very well, especially in text documents classification, also used for sentiment analysis [GBH09] and spam filtering [MAP06].

Given class variable **y** (for gender classification: male or female; and for race classification: White or Hispanic or Asian or Black) and dependent feature vector $\mathbf{x_i}$ through $\mathbf{x_n}$ (emoji), the Bayes algorithm works as follows:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{3.13}$$

where, the likelihood of the features is assumed to be Gaussian since our feature take up a continuous values (percentage usage) and are not discrete:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3.14}$$

and, the parameters $\sigma_\mathbf{y}$ and $\mu_\mathbf{y}$ are estimated using maximum likelihood.

After that, we tried K Nearest Neighbour classification. K-NN predicts by computing from a majority vote of the nearest neighbours of each point. It computes the distance $\mathbf{D}(\mathbf{x}, \mathbf{x_i})$ to every training example $\mathbf{x_i}$, selects **k** closest instances $\mathbf{x_{i_1}} \ldots \mathbf{x_{i_k}}$ and their classes $\mathbf{y_{i_1}} \ldots \mathbf{y_{i_k}}$ (for gender classification: male or female; and for race classification: White or Hispanic or Asian or Black), and outputs the class **y** which is most frequent in $\mathbf{y_{i_1}} \ldots \mathbf{y_{i_k}}$.

A 5-Fold Cross-validation on the training set was used to determine a suitable value for **k** parameter, achieving the best "F1 Score" for gender classification and best "Accuracy Score" for race classification.

Cross-validation is a re-sampling technique that is used to evaluate machine learning models. The fold parameter called **K** refers to the number of groups that a given data sample is to be split into. For example, a 5-Fold Cross-validation, randomly splits the data into 5 sets, validates on each in turn (train on 4 others) and averages the results over 5 folds. It is primarily used in applied machine learning to estimate the skill of a machine learning model in order to optimise the values of its parameters. Figure 3.9 shows how data is split using a 5-Fold Cross Validation on training set.

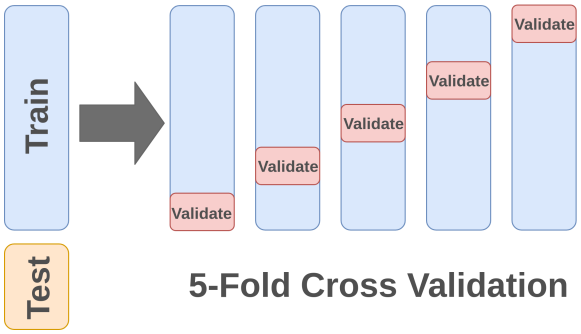


Figure 3.9: How data is split on a 5-Fold Cross Validation on Train Set.

Similarly to one of our approaches in Section 3.3 to identify distinctive emoji for each group, we tried Random Forest for gender and race classification, a state-of-the-art method that performs well in many domains. Previous studies on spam detection on Twitter have shown that Random Forest is more successful than Naive Bayes and K-NN on tweet classification [MC11]. The Random Forest algorithm works as follows:

Algorithm 1 Building Random Forest

1: Randomly select **K** features (emoji) from total of **m** features, where **K** $<<$ **m**
2: Among the **K** features, calculate the node **d** using the best split point
3: Split the node into child nodes using the best split
4: Repeat steps **1** to **3** until **l** nodes are reached
5: Repeat steps **1** to **4** for **n** times, to build forest with **n** trees

Algorithm 2 Predicting with Random Forest

1: Take features in test set and use rules of each tree created to predict the outcome and store predicted outcome
2: Calculate votes for each predicted outcome
3: Consider the highest voted predictions as the final prediction from random forest algorithm

As done with K Nearest Neighbour approach, a 5-Fold Cross-validation on the training set was used to determine a suitable value of **l** (maximum depth of the tree).

Finally we tried another form of classifier: the SVM (Support Vector Machine) [CV95], which commonly performs well for text based classification and achieves good error

rates for unusual types of data, especially when the number of features is relatively high (in our case 2579 emoji) [HHEQ08], showing increased performance over other classification methods such as Naive Bayes or K-NN. An advantage of Support Vector Machines is that they can be kernelised, meaning that they can use a set of functions to transform input data into the required form before making predictions.

SVMs compute the decision boundary between the two classes (eg. for gender classification: male or female) as the line furthest from any training point. This is done by building a wide margin between classes. The decision boundary (hyper plane or line) is defined only by the support vectors (training points closest to each other from opposite classes), the rest points are not taken into consideration. A visual representation showing how the decision boundary is defined can be seen in Figure 3.10.

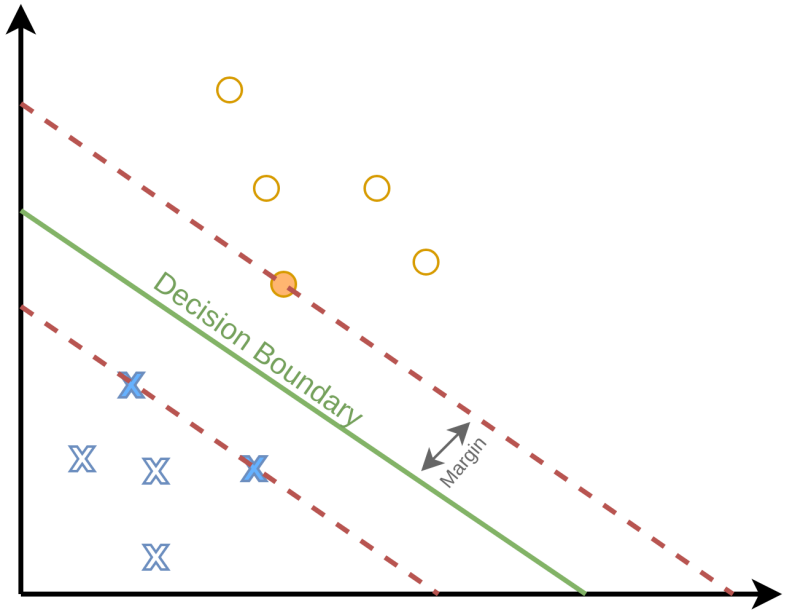


Figure 3.10: Defining decision boundary by SVM example.

Using a "Linear" kernel, this boundary is a line, but it can take more complex shapes when other kernels are used. In our study we tried two different kernel approaches.

Our first approach was to use "Linear" kernel, where classification decision is made based on the value of a linear combination of attributes. The reason we have chosen the "Linear" kernel first is that solving the optimisation problem is much faster than using any other kernels, meaning that we could see how well a SVM can perform without spending extra time using other kernels. If **x** and **y** are column vectors, their linear kernel calculated using:

$$k(x,y) = x^\top y \tag{3.15}$$

As with K-NN and Random Forest techniques, a 5-Fold Cross-validation on the training set was used to determine a suitable value of penalty parameter **C** of the error term, a parameter required by the "Linear" kernel that is used for controlling the outliers.

Our second approach was to use "Chi-square" kernel, which outperforms other kernels for histogram (bags) of visual words features, is very popular for computer vision applications and achieves descent results for various classification tasks, such as disease diagnosis [Dal13]. The chi squared kernel is given by:

$$k(x,y) = \exp\left(-\gamma \sum_i \frac{(x[i]-y[i])^2}{x[i]+y[i]}\right) \tag{3.16}$$

All six classification models were fitted data in the form "X_train" and "y_train", where "X_train" represents the set of independent features (emoji percentage usage) and "y_train" the set of dependent variables (target group) in train set. The dependent variables used for gender classification were: {male, female}, for race classification: {white, black, asian, hispanic} and for the combined gender and race classification: {white_male, white_female, black_male, black_female, asian_male, asian_female, hispanic_male,hispanic_female}. Part of the code used to train all six classification models using Scikit-Learn built in functions can be seen in Appendix E.

Race classification as well as the combined gender and race classification are tasks that include more than two classes in comparison to gender classification which is a binary classification task (male or female). Many machine learning methods used in this research are naturally adapt to multi class problems:

- Gaussian NB: probabilistic classifier that does not need any modification to accept multi-classes.
- K-NN: generally incorporates multi classes by voting between the closest **K** neighbours in a **D** dimensional space.
- Random Forest: inherits multi classes naturally, since each leaf refers to one of different classes.

However, a transformation to a binary problem was needed for our SVM models since it cannot naturally inherit multi class problems. Therefore, "One vs Rest" decision function was used, where one class contains only samples from a class **c** and the other class contains samples from the rest $\bar{\mathbf{c}}$ classes (representing all the other classes excluding class **c**).

In this chapter an explanation of different approaches and techniques used for data collection, preparation, analysis and classification has been provided. More information regarding the scripts used and the implementation of the code can be found in our version control repository, listed at the beginning of Section 3.

Chapter 4

Analysis Results

In this chapter, we summarise the statistics of the data used in our research. We delve into the top used emoji by each group of users according to their gender, race and location. Apart from that, we provide results of the three different approaches for deriving the difference in emoji usage in user timelines according to their demographics, including gender and race.

4.1 Tweet Emoji Distribution

Initially we examined the distribution of tweets in our collection. We wanted to know the number of tweets in user profiles as well as the number of tweets containing emoji in user profiles. From 40,000 timelines, we derived that there are on average 344 tweets in a user's profile, while the mean number of tweets having emoji in their content was 73. Besides that, the maximum number of tweets presented on a user's timeline was 3267, where the maximum number of tweets consisting emoji on a user's timeline was 3254. We also had some users who didn't posted any tweets at all, meaning that our minimum number of tweets in user's profile was 0. Moreover, we saw how spread tweets counts were by determining the standard deviation for all tweets and tweets having emoji in their content to be 628 and 198 respectively. More details for tweet frequency in users' profile are shown using box plots in Figure 4.1.

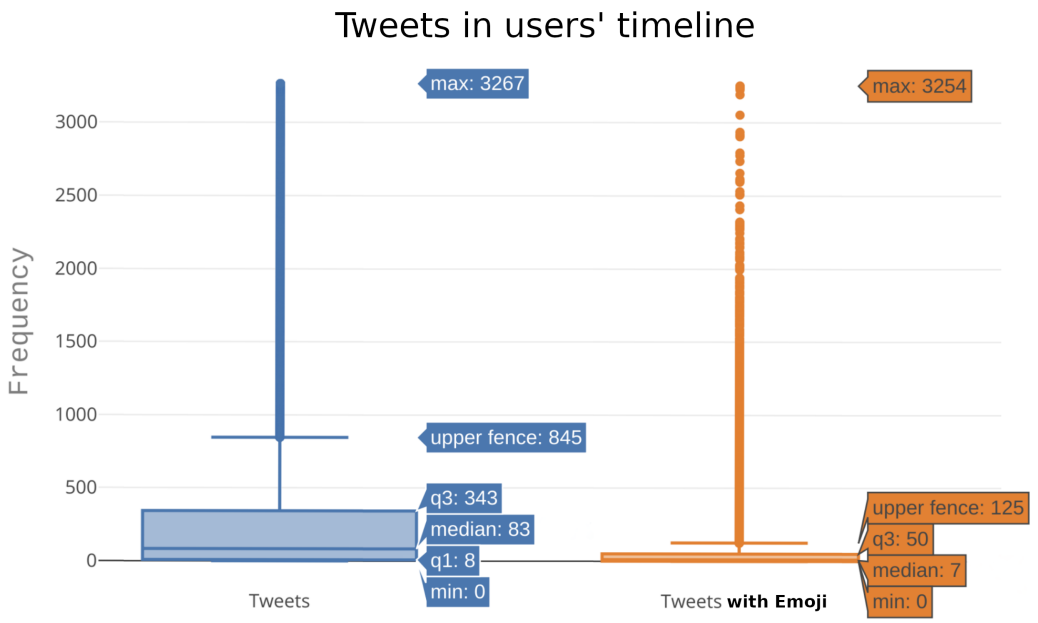


Figure 4.1: Box plots showing frequency of all tweets and tweets having emoji in their content in users' timeline.

Our datasets distribution can be seen in the box plots above including minimum, first quartile, median, third quartile and maximum number of tweets in user's profile. As seen from the graph, the upper fence for tweets in user's timeline was 845 and for tweets consisting emoji was 125, meaning that timelines including more tweets than those numbers could be considered as outliers (plotted as little dots on the box plots).

4.2 Most Frequent Emoji by Demographics

In this section we explore the top used emoji by each group of users according to their gender and race, by simply counting the most frequent emoji in timelines, as well as getting the percentage usage of most frequent emoji from each of the four locations (Worldwide, Johannesburg, New York City, London).

Figure 4.2 shows the most frequent emoji used by users in general from our collection.

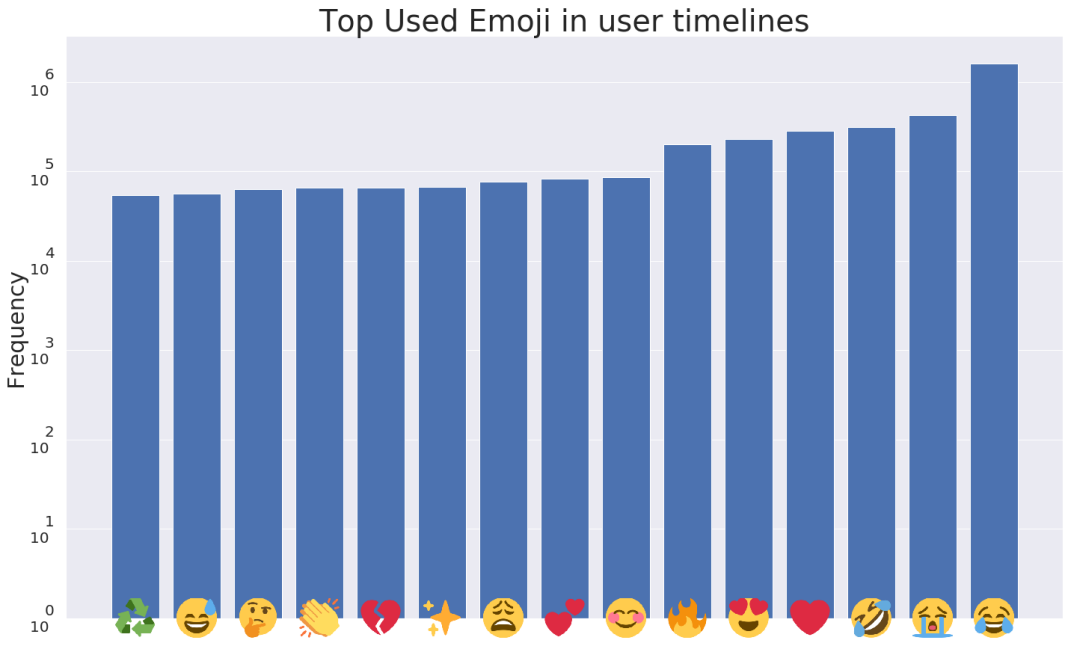


Figure 4.2: Most frequent emoji in general.

As shown, the "Face with Tears of Joy" 😂 is the most frequent emoji with 1.6 million counts for all users, which aligns with previous studies and reports [LF16]. Followed by the "Loudly Crying Face" 😭 with 0.4 million uses and the "Rolling on the Floor Laughing" 🤣 with 0.3 million uses. Unpredictably the "Recycling Symbol" ♻️ is found to be the 15th most popular emoji, where in related studies is found to be the 3rd most frequent emoji [Hur18], representing most of the times "Share".

Figure 4.3 shows the most frequent emoji used by 16,005 gender labelled users who included emoji in their tweets, where 7521 are males 8484 are females.

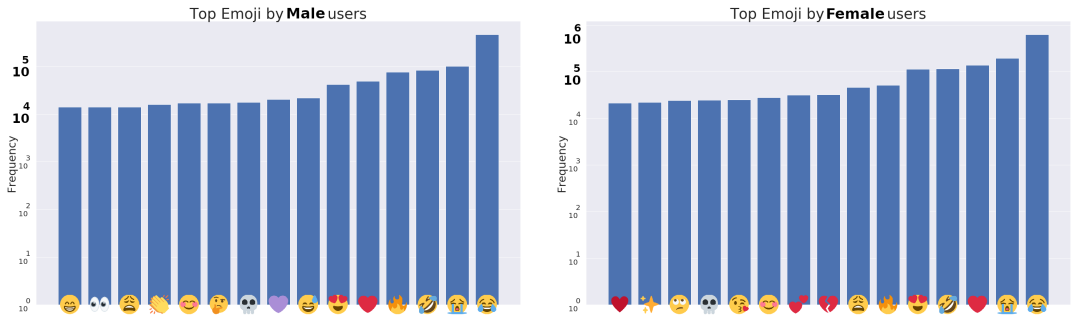


Figure 4.3: Top used emoji by user timelines in our collection according to gender.

As seen to our most frequent emoji in general results, the "Face with Tears of Joy" 😂 and the "Loudly Crying Face" 😭 are at most used emoji by both genders. We can see a priority on the "Red Heart" ❤️ by female users and a priority on the "Fire" 🔥 by the male users. We can clearly see some emoji appearing to be mostly used by one gender and not the other, such as the "Broken Heart" 💔 and the "Two Hearts" 💕 for female

users; the “Purple Heart” 💜 and the “Grinning Face With Sweat” 😅 for male users. Because of these variations, it can be derived there might be different kinds of usage according to gender.

Figure 4.3 shows the most frequent emoji used by 14,341 race labelled users who included emoji in their tweets, where 7047 are White, 6161 are Black, 780 are Asian and 328 are Hispanic users.

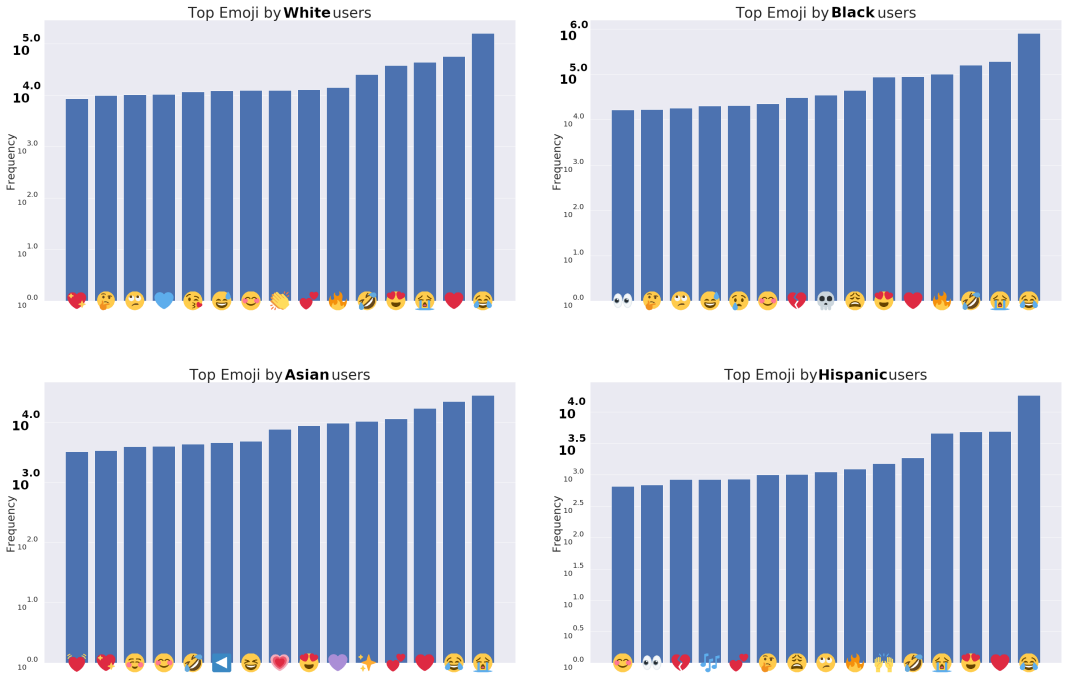


Figure 4.4: Top used emoji by user timelines in our collection according to race.

From the figure above we can see that the “Face with Tears of Joy” 😂 is the most used emoji by all races except Asian users where the “Loudly Crying Face” 😭 takes the first place. Some variations could be noticed for the later most frequent ones. Some examples are the “Clapping Hands” 👏 for White users, the “Scull” 💀 for Black users, the “Raising Hands” 🙌 for Hispanic users and the “Sparkles” ✨ for Asian users. These variations indicate that there might be different patterns of usage according to race.

We further explore our analysis by finding the percentage usage of most frequent emoji by users for each of the 4 locations in our collection. Figure 4.5 shows percentage usage of most popular emoji posted from London, New York City, Johannesburg and randomly worldwide.

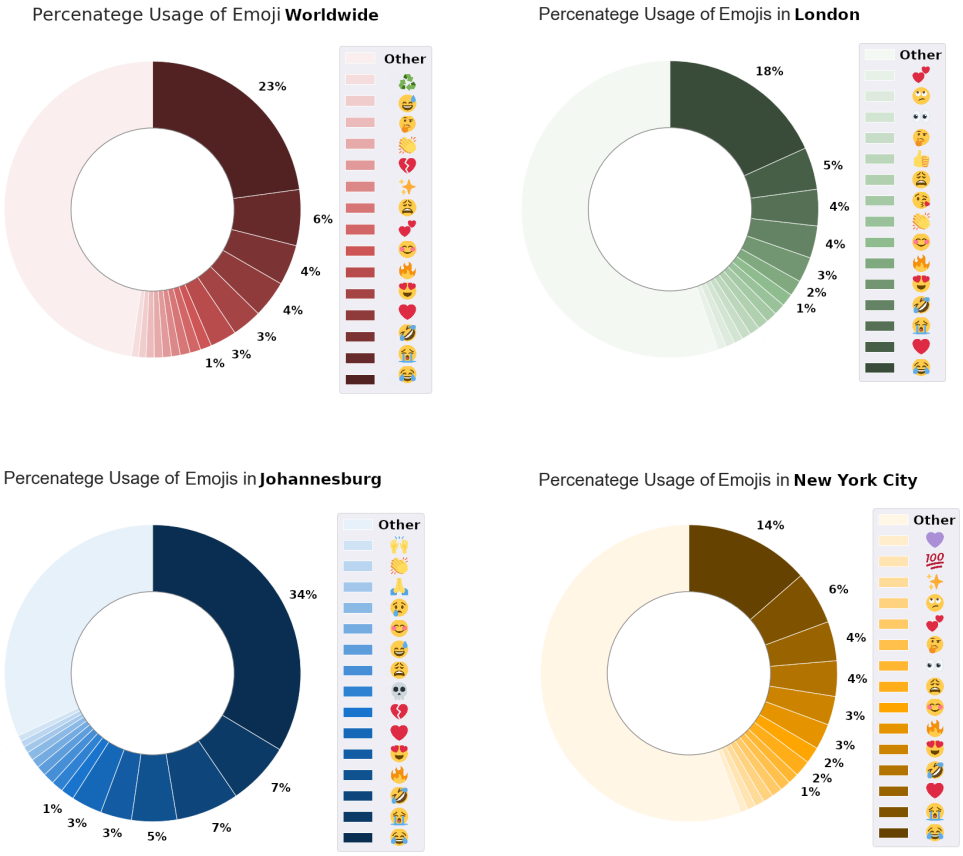


Figure 4.5: Most frequent emoji percentage usage by location.

As seen from the pie charts above, the "Face with Tears of Joy" 😂 continuous to be the leading emoji for all four locations. It can be observed, that some emoji appear to be mostly used by one location and not the rest, such as the "Skull" 💀 in Johannesburg, the "Sparkles" ✨ in New York City and the "Face Blowing a Kiss" 😘 in London , however the percentage usage of those emoji is close to 1% for each location respectively.

4.3 Most Distinctive Emoji for each Gender and Race

To further investigate the differences in emoji usage by each group, we calculated the most distinctive emoji set for each gender and race by using three different approaches: 1) Difference in percentage usage, 2) Chi square (χ^2) feature selection and 3) Random Forest feature importance.

4.3.1 Difference in percentage usage

Our first approach measures the the difference in percentage of usage between a given group and the others. The most distinctive set of emoji (with the highest difference in usage percentage) are extracted for each group.

Figure 4.6 reports the top distinctive emoji set for male and female users.

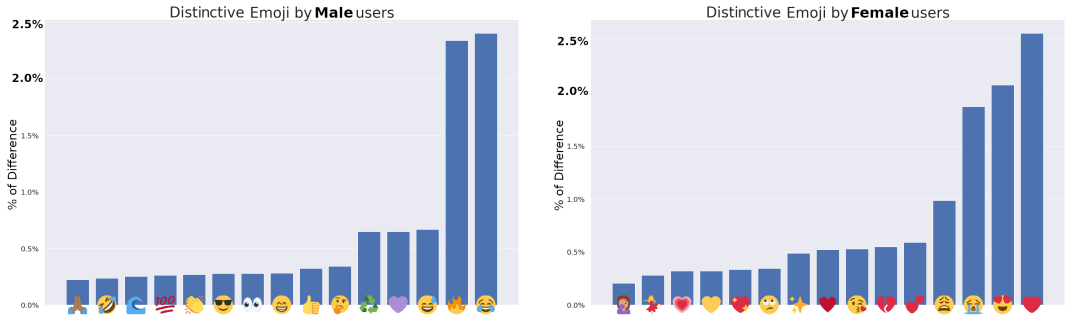


Figure 4.6: The most distinctive emoji on x-axis with highest percentage difference on y-axis between females and males.

As shown in the figure above, the most distinctive emoji for males were the "Face With Tears of Joy" 😂 with 2.4% difference in usage than females, the "Fire" 🔥 with 2.37% difference and the "Grinning Face With Sweat" 😅 with 0.67%, as we already noticed in Section 4.2.

It is clear that emoji representing "Hearts" are significantly more popular within female users than males, as we have seen in previous findings [BCC18], where the most distinct emoji for females were the "Red Heart" ❤️ with 2.56% difference in usage than males, the "Smiling Face with Heart-shaped Eyes" 😍 with 2.07% and the"Loudly Crying Face" 😭 with 1.87%. These observations show that there is clear difference in the percentage of emoji usage between males and females.

Figure 4.7 shows distinctive emoji with highest difference in usage percentage by race.

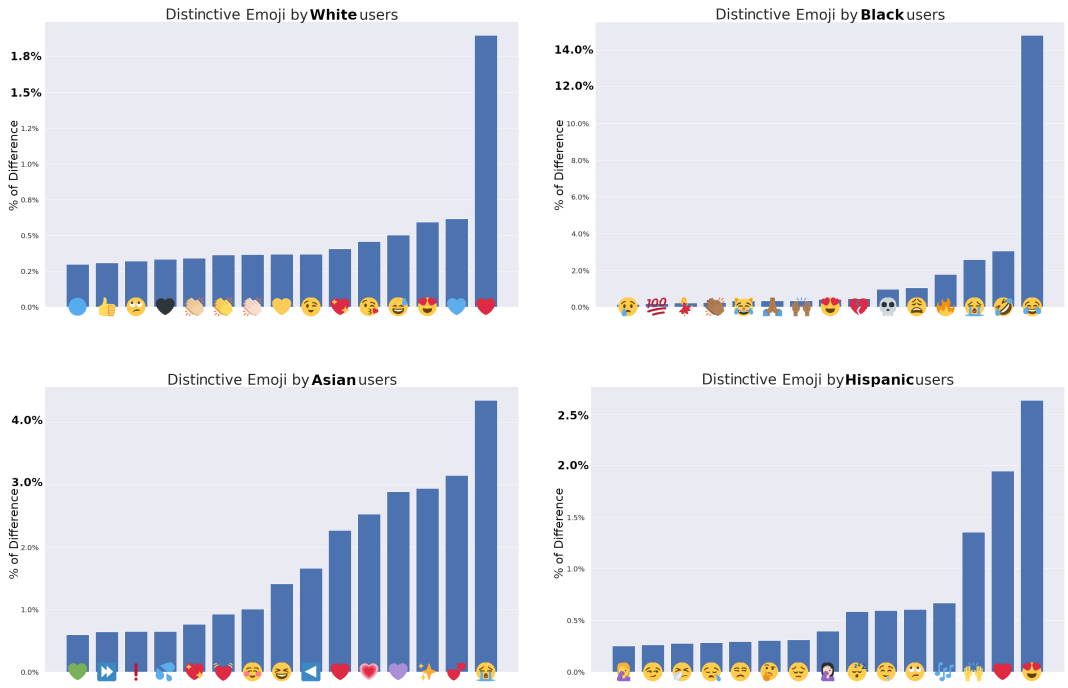


Figure 4.7: The most distinctive emoji on x-axis with highest percentage difference on y-axis between the four race groups.

As shown from the figure above, the "Red Heart" ❤️ is more used by White users than any other race with 1.9% difference in usage than other groups. The "Face With Tears of Joy' emoji 😂 is the most distinctive emoji for Black Users with a superior 14.79% usage more than the other race groups. It could be also noticed that for Black users, emoji with dark skin tones are more commonly used than other races, similarly for White users and blonde emoji, which aligns with findings in [RMG18].

The most distinctive emoji for Hispanic users was the "Smiling Face with Heart-shaped Eyes" 😍 with 2.63% difference than all other race groups, where the most distinctive emoji for Asian users was the "Loudly Crying Face" 😭 with 4.35% differnece, already noticed in Section 4.2.

As mentioned in Section 3.3, capturing distinctive emoji by calculating the difference of frequency percentages among two sets was not reliable in some cases. In the following section we explore the results of our second approach.

4.3.2 Chi square (χ^2) feature selection

In our second approach we used chi-square test to find which emoji features reject the null hypothesis using 5% significance level for each different group. The most distinctive emoji (with the lowest P-values) are extracted for each group and a red dotted line was plotted to show the significance level, clarifying which emoji P-values fall beneath it, rejecting the null hypothesis.

Figure 4.8 reports distinctive emoji with lowest P-values extracted for male and female users using chi-square test.

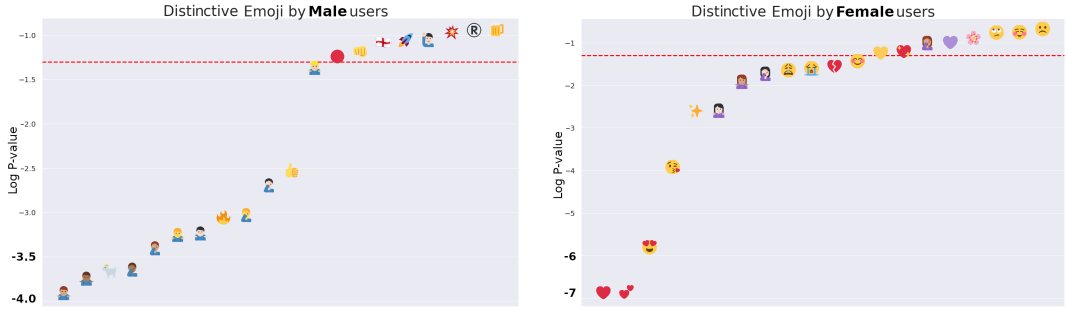


Figure 4.8: The most distinctive emoji on x-axis with lowest Log P-values on y-axis between females and males.

As shown in the figure above, the most distinctive emoji for males were the "Man Shrugging: Medium Skin Tone" 🤷🏽‍♂️ with 1.2×10^{-4} P-value, the "Man Shrugging: Medium-Dark Skin Tone" 🤷🏾‍♂️ with 1.8×10^{-4} P-value and the "Goat" 🐐 with 2.2×10^{-4} P-value, often used in relation to the phrase "Greatest Of All Time" (G.O.A.T.) [Blo16].

It is clear that emoji representing "Man" are fairly used by male users than female; "Hearts" and "Woman" emoji are more popular within female users than males, where the most distinct emoji for females were the "Red Heart" ❤️ with 1.3×10^{-7} P-value, the "Two Hearts" 💕 with 1.4×10^{-7} P-value and the "Smiling Face with Heart-shaped Eyes" 😍 with 1.6×10^{-6} P-value. Observations above show that there is clear difference in emoji usage between males and females.

Figure 4.9 shows distinctive emoji with lowest P-values extracted by race using chi-square test.

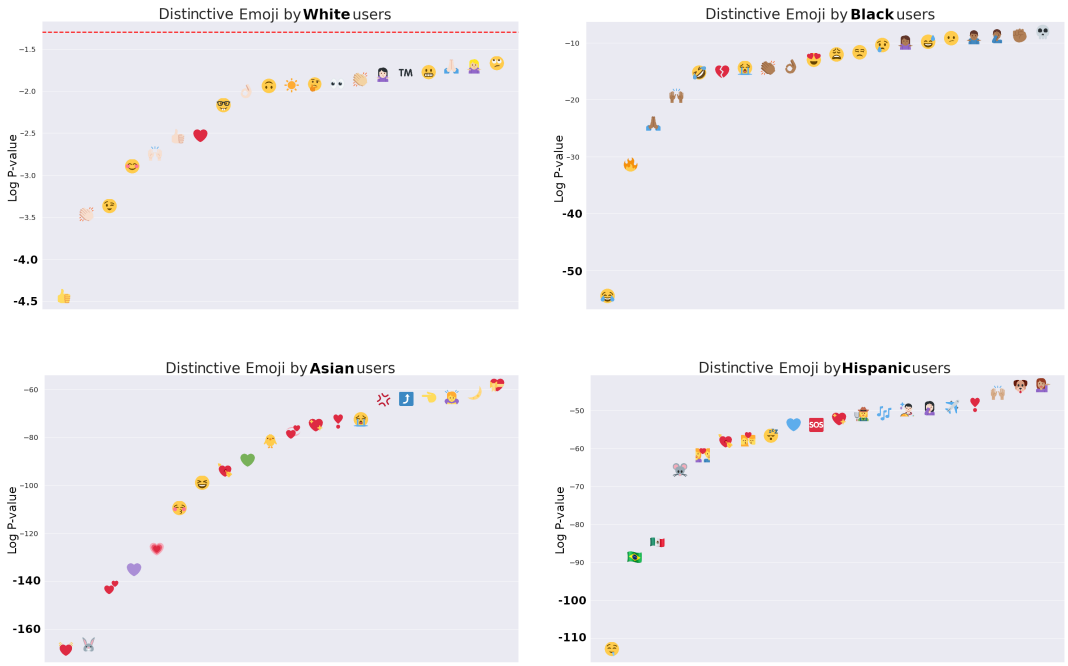


Figure 4.9: The most distinctive emoji on x-axis with lowest Log P-values on y-axis between the four race groups.

From the figure above it can be noticed that the "Thumbs Up" 👍 and "Clapping Hands: Light Skin Tone" 👏 are used by White users more than any other race. The "Face With Tears of Joy' emoji 😂 is the most distinctive emoji for Black Users, as similarly seen to our previous approach results. Repeatedly, we can see that for Black users, emoji with dark skin tones are more common, where for White users, emoji with white skin are more frequent.

The most distinctive emoji for Hispanic users was the "Drooling Face" 🤤. It was interesting to see flags of Brazil 🇧🇷 and Mexico 🇲🇽 as one of the most distinctive emoji for Hispanic race, which indicates that this race represents citizens of Latin America. For the Asian users it could be observed that the "Beating Heart" 💓 and the "Rabbit Face" 🐰 were the most distinctively used emoji. Additionally, it can be noticed that emoji representing "Heart" appear to be distinctive for Asian users. From the results above we can see that there exist a difference in emoji usage between difference races.

4.3.3 Random Forest feature importance

Our third approach to capture distinctive emoji by each demographic was to train a Random Forest model based on the collected data. Since emoji were used as features and demographics of each user as classification target, we were able to extract features (emoji) with the highest importance for each target group respectively.

Figure 4.10 reports the top distinctive emoji set for male and female users, having the highest weights on the Random Forest model.

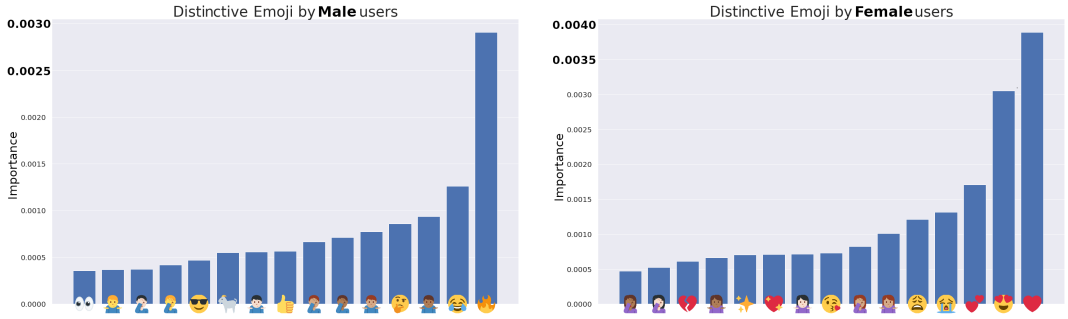


Figure 4.10: The most distinctive emoji on x-axis with highest importance on the y-axis between females and males.

As shown in the bar charts above, the most distinctive emoji for males were the "Fire" 🔥 with 0.0029 importance and the "Face with Tears of Joy" 😂 with 0.0012. As seen in our previous approach, emoji representing "Man Facepalming" 🤦🏾‍♂️🤦🏻‍♂️ and "Man Shrugging" 🤷🏾‍♂️🤷🏻‍♂️ with multiple colour skin tones appear to be very important for male user classification using the Random Forest model.

Moreover, the most distinctive emoji for females were the "Red Heart" ❤️ with 0.0039 importance and the "Smiling Face with Heart Shaped Eyes" 😍 with 0.0031. Repeatedly, we can seen that emoji representing "Female Facepalming" 🤦🏾‍♀️🤦🏻‍♀️, "Female Shrugging" 🤷🏾‍♀️🤷🏻‍♀️ of different skin tones and "Heart" are crucial in predicting female users by our Random Forest model. Our observations using the Random Forest feature importance approach show that there is obvious difference in emoji usage between male and female users.

Figure 4.11 reports the top distinctive emoji extracted by each race, having the highest weights on the Random Forest model.

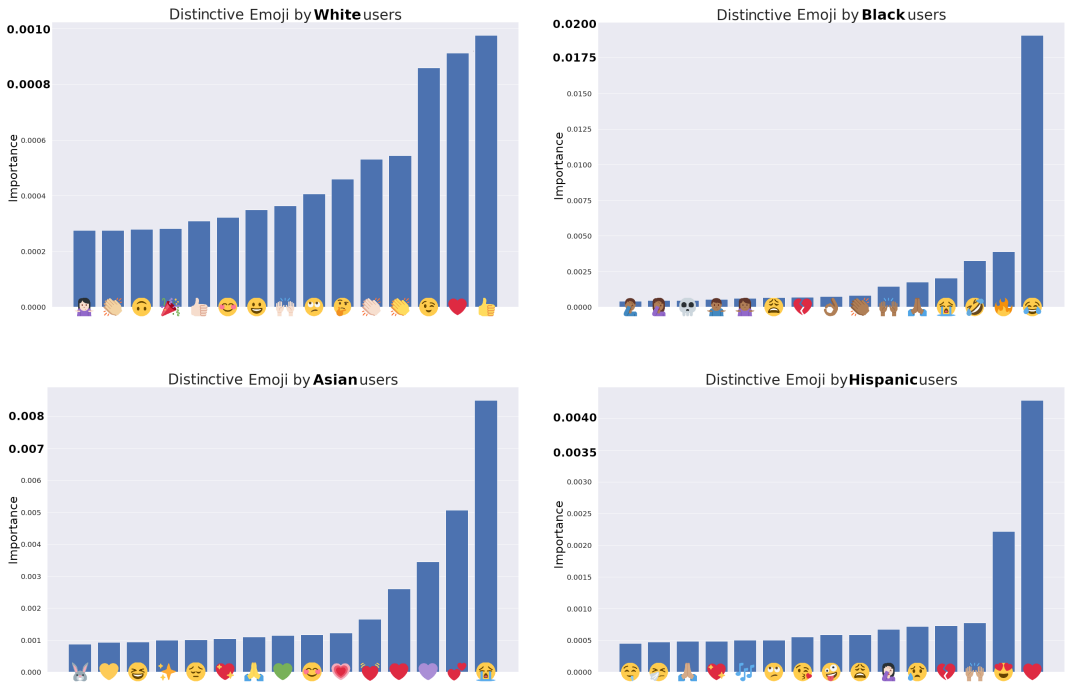


Figure 4.11: The most distinctive emoji on x-axis with highest importance on the y-axis between the four race groups.

As we already seen in the chi-square test approach, Figure 4.11 shows the "Thumbs Up" 👍 to be the most important feature for White users classification, with 0.00098 weight and the "Face With Tears of Joy' 😂 the most distinctive emoji for Black users classification, with 0.0191 weight. Moreover, from the bar charts above we can conclude that white skin tone emoji appear to be more important for White user classification, simultaneously emoji with dark skin tones are more critical for Black users classification.

The most distinctive emoji for Hispanic users was the "Red Heart" ❤️ with 0.0041 importance, where for the most important emoji for Asian user classification was the "Loudly Crying Face" 😭, as we found out to be the most distinctive emoji for the same group in our first approach. Again, we can observe that emoji representing "Heart" appear to be relatively important for Asian user classification using our Random Forest model.

Our previous analysis, using all of the three approaches, highlights the main differences in emoji usage between different gender and race groups, which was shown to be significant in some cases. In the following chapter, we explore the effectiveness of using these difference to automatically detect the user gender and/or race.

Chapter 5

Classification Results

In this part, we explore the results of our classifiers build for gender and race detection, as well as results combining both. As mentioned in Section 3.4, a set of 2579 unique emoji were used as our feature set (each emoji represents a feature), users as instances, the gender and/or race of each user as classification target and the percentage usage of emoji from each user as feature variables. For our experimentation, we split our dataset into 80/20% for training and testing respectively. Multiple machine learning classification techniques were examined. The following results are based on predictions on testing set.

5.1 Gender Classification Results

For gender classification we have used 16,005 users, which were confidently gender labelled and used emoji in the tweets. From those, 7,521 are male and 8,484 are female users. Table 5.1 reports the accuracy and macro F1-score for different models used for gender classification. A more visual representation of our classifiers accuracy can be seen in Figure 5.1.

Method	**Accuracy**	**F1 Score**
Prior Dummy	0.521	0.343
Gaussian NB	0.629	0.602
K-NN	0.667	0.574
Random Forest	0.778	0.762
SVM (Linear)	0.759	0.744
SVM (χ^2 kernel)	**0.782**	**0.766**

Table 5.1: Gender classification models performance using accuracy and macro F1-Score

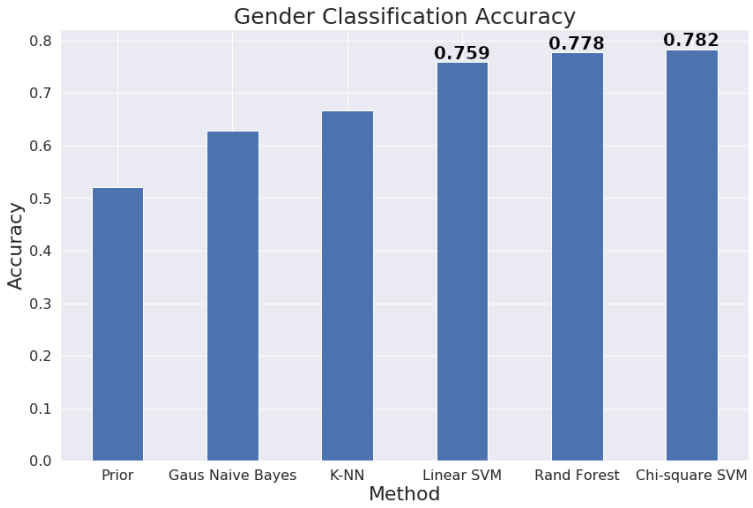


Figure 5.1: Gender classification accuracy by each model.

As shown from the results above, the majority-class baseline using the prior dummy classifier achieved the lowest performance since it was predicting all test set as males. In addition Gaussian Naive Bayes and K-Nearest Neighbour classifiers achieved low performance compared to other classifiers. The best performing classifier was the SVM with χ^2 kernel, which achieved a 78.2% accuracy. The second most performed classifier was the Random Forest (with tree depth = 55) with 77.8% accuracy followed by the Linear SVM (with penalty parameter C = 10) with 75.9% accuracy. These results illustrate the significant difference in usage of emoji between males and females that allowed detecting of the users' gender from the used emoji in their timeline solely.

5.2 Race Classification Results

For race classification we have used 14,316 users who were trusty race labelled and posted emoji in their tweets. From these users 7047 were White, 6161 were Black, 780 were Asian and 328 were Hispanic. Our results for race classification are divided into three different approaches, using the original data set of users specified above, an oversampled train data and an undersampled data set. The reason we tried the additional two approaches was that our original data was imbalanced, meaning that we had limited number of Asian and Hispanic users compared to White and Black users in order to sufficiently train our classifiers.

5.2.1 Original Data

The achieved accuracy per race category for race classification using the original data stated above can be seen on Table 5.2. Additionally, Figure 5.2 illustrates the average prediction accuracy by each classifier.

Method	White	Black	Asian	Hispanic	Accuracy
Prior Dummy	1.00	0.00	0.00	0.00	0.51
Gaussian NB	0.36	0.47	**0.38**	**0.33**	0.41
K-NN	0.7	0.79	0.04	0.00	0.68
Random Forest	0.88	0.79	0.05	0.00	0.78
SVM (Linear)	0.82	0.82	0.03	0.00	0.76
SVM (χ^2 kernel)	**0.89**	**0.80**	0.19	0.00	**0.80**

Table 5.2: Race classification models performance on each of the four race categories, and the overall model accuracy.

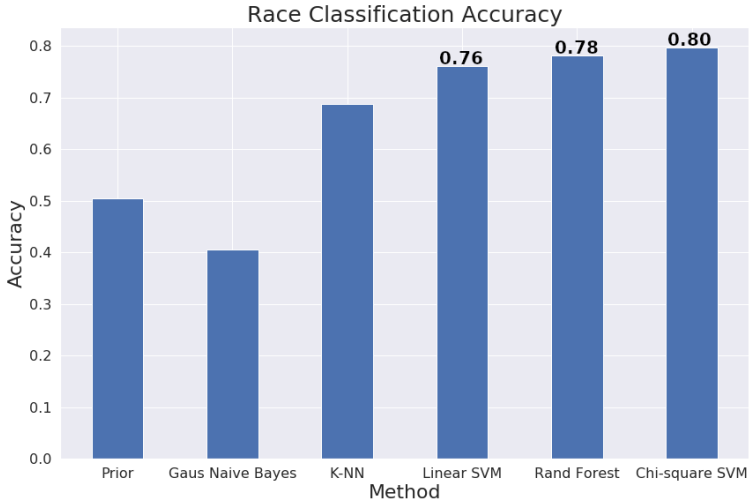


Figure 5.2: Race classification accuracy on original data by each model.

Most of the classifiers failed dramatically with the Asian and Hispanic race categories, where actually none of the classifiers (except Gaussian Naive Bayes) could detect any of the Hispanic race. While Gaussian Naive Bayes achieved the best performance for detecting the two rare classes, it achieved the worst overall performance, because of it's poor performance with the White and Black categories which constitute over 90% of the data population. Actually, Gaussian Naive Bayes achieved an overall accuracy that is worse than the dummy majority-class baseline that assigns all predictions to the White category. The best performing classifier, similar to gender detection, was the SVM with the χ^2 kernel, where it achieved an 80% accuracy for race detection. However, its performance with the Asian and Hispanic classes was poor. Repeatedly, the second and third performing classifiers were the Random Forest (with tree depth = 144) and Linear SVM with 78% and 76% average accuracy respectively.

To better understand the performance on race classification using emoji, we constructed the confusion matrix for the best performing model (SVM with χ^2 Kernel), illustrated in Figure 5.3.

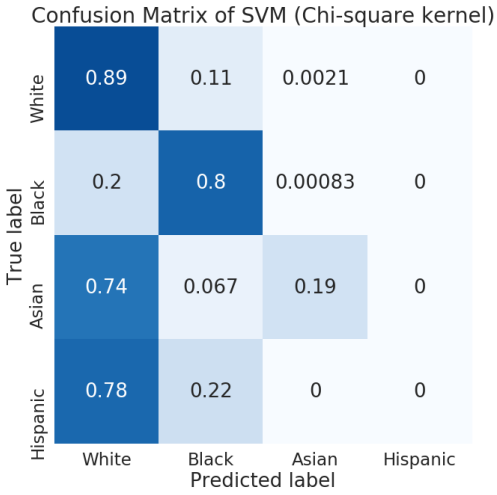


Figure 5.3: Confusion matrix of SVM (χ^2 kernel) for race classification on original data.

From the confusion matrix above we can see that 74% of Asian users were misclassified as White, while 78% and 22% of Hispanic users were misclassified as White and Black respectively. To enhance our classification for the two rare races we tried to oversample our training data.

5.2.2 Oversampled Train Data

As our classification results for the minority races (Asian, Hispanic) were unsatisfactory , we oversampled our train data using SMOTE oversampling with 4 neighbours before training our classifiers. Using this oversampling technique we were able to develop a balanced train set of equal user counts from each class to the majority class (White).

Table 5.3 shows the reached accuracy per race category for race classification on testing set using the oversampled train set. Also, Figure 5.4 illustrates the average prediction accuracy by each classifier.

Table 5.3: Race classification models performance on each of the four race categories, and the overall model accuracy using over-sampled data

Method	White	Black	Asian	Hispanic	Accuracy
Prior Dummy	1.00	0.00	0.00	0.00	0.51
Gaussian NB	0.38	0.47	0.38	**0.32**	0.42
K-NN	0.54	0.60	0.36	0.30	0.55
Random Forest	0.82	0.79	0.29	0.00	0.76
SVM (Linear)	0.62	0.78	**0.68**	0.20	0.68
SVM (χ^2 kernel)	**0.85**	**0.80**	0.44	0.03	**0.79**

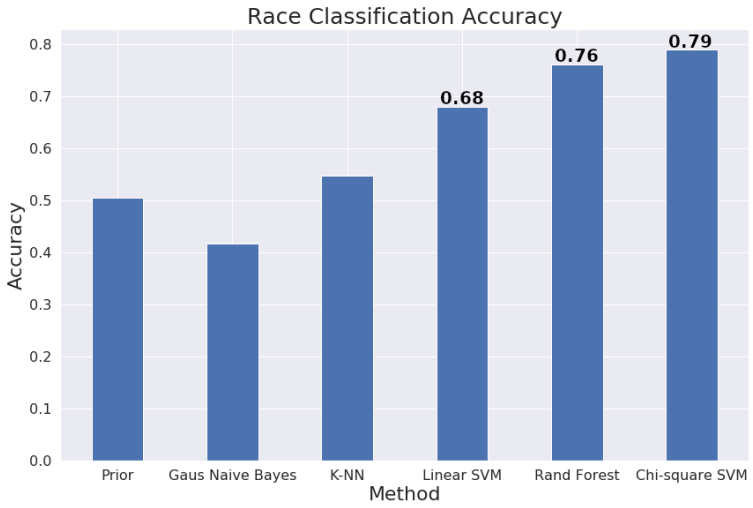


Figure 5.4: Race classification accuracy on oversampled data by each model.

This time almost all classifiers performance, especially of Linear SVM classification has increased for Asian race prediction, reaching a 68% accuracy. However, as we have seen on the previous approach, the classifiers failed thoroughly with Hispanic race category, where only Gaussian Naive Bayes and K Nearest Neighbour could detect a limited amount of the Hispanic race. Again, Gaussian Naive Bayes achieved the best performance for detecting the Hispanic race, and worst overall performance out of all classifiers. The best performing classifier was repeatedly the SVM with the χ^2 kernel, where it achieved an 79% accuracy for race detection, with better Asian race detection than using the original data. However, its performance with Hispanic class remained poor. Followed by the Random Forest (with tree depth = 144) and Linear SVM (with penalty parameter C = 100) with 76% and 68% average accuracy respectively.

The confusion matrix of our best performing model (SVM with χ^2 Kernel) on oversampled train set can be seen in Figure 5.5.

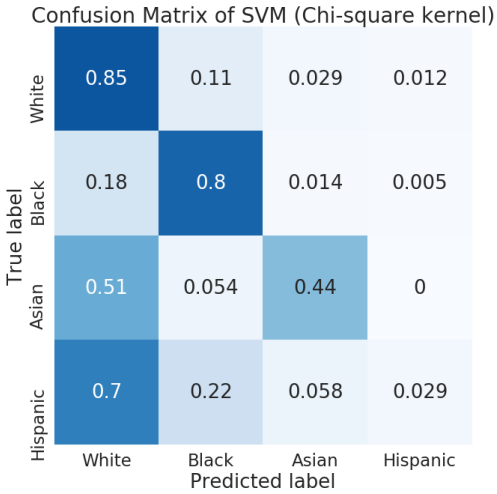


Figure 5.5: Confusion matrix of SVM (χ^2 kernel) for race classification on oversampled train data.

From the confusion matrix above we can see that this time only 51% of the Asian users were misclassified as White compared to the 74% using original train set, while 70% and 22% of Hispanic users were misclassified us White and Black respectively. Using oversampled train data, we were not able to achieve any better overall results, however a gain in Asian race prediction could be noticed. In the following section a third approach is undertaken to improve our classifiers predictions by undersampling our data set.

5.2.3 Undersampled Data

From the previous two approaches we saw that our classifiers struggle to correctly predict the minority races from our dataset, especially the Hispanic race. In this section we try to to undersample our data using a combination of two techniques, the Tomeks links & a random undersampler. Using these two techniques we were able to develop a balanced dataset of equal user counts from each class to the minority class (Hispanic).

Table 5.4 shows the achieved accuracy per race category for race classification on testing set using undersampled dataset. In addition, Figure 5.6 illustrates the average prediction accuracy by each classifier.

Table 5.4: Race classification models performance on each of the four race categories, and the overall model accuracy using under-sampled data

Method	White	Black	Asian	Hispanic	Accuracy
Prior Dummy	1.00	0.00	0.00	0.00	0.25
Gaussian NB	0.36	0.64	0.52	0.29	0.45
K-NN	0.19	0.67	0.58	**0.52**	0.49
Random Forest	**0.62**	0.72	0.70	0.39	**0.61**
SVM (Linear)	0.54	**0.75**	**0.71**	0.36	0.59
SVM (χ^2 kernel)	0.54	0.70	0.62	0.36	0.55

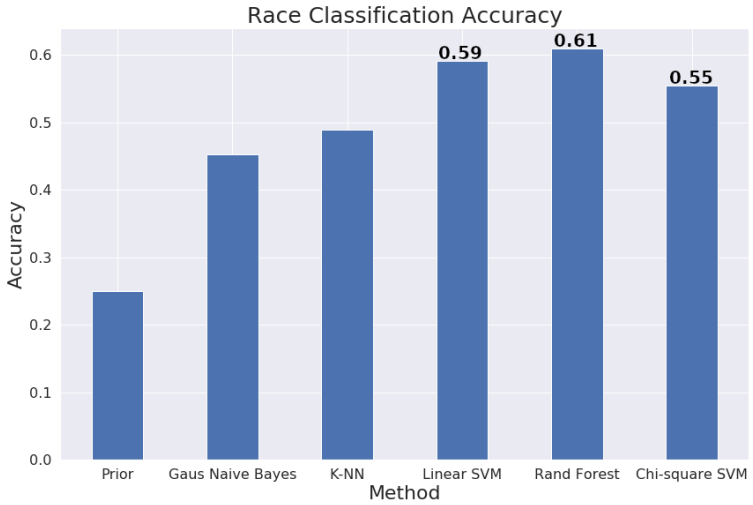


Figure 5.6: Race classification accuracy on undersampled data by each model.

Using the undersampled dataset we were able to achieve better classification accuracy for both Hispanic and Asian races, with K Nearest Neighbour (with neighbours k = 50) reaching 52% accuracy for Hispanic race; and Linear SVM (with penalty parameter C =10) achieving 71% accuracy on Asian race, outperforming our two previous approaches on predicting the minority classes. However, a noticeable decrease in overall performance by almost all classifiers can be seen, especially for SVM with the χ^2 kernel and Linear SVM classifications, having fairly bad results for White and Black race categories predictions compared to previous approaches. The best performing classifier was the Random Forest (with tree depth = 21) with 61% overall accuracy, followed by the Linear SVM with 59%.

The confusion matrix of our best performing model (Random Forest) on undersampled data can is illustrated in Figure 5.7.

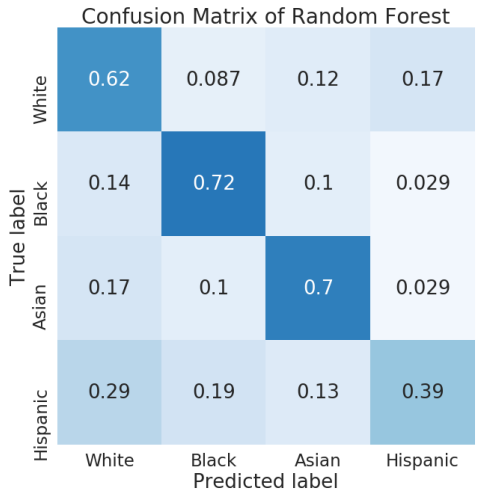


Figure 5.7: Confusion matrix of Random Forest for race classification on undersampled data.

From the confusion matrix above we can see that Hispanic and Asian race categories predictions were much better compared two the two previous approaches, having 39% and 70% correct classification respectively. Only 17% of the Asian users were misclassified as White and 10% as Black, while 29% and 19% of Hispanic users were misclassified us White and Black respectively. However, 12% and 17% of White users were missclassified as Asian and Hispanic respectively. Repeatedly, using undersampled data, we were not able to achieve any better overall results. The insignificant number of users from the two rare races in our sample is considered one of the limitations of the race classifiers performance. In the following section we explore the combination of both gender and race as classification targets.

5.3 Gender & Race Classification Results

For the combination of both gender and race classification we have used 14,185 users who were both confidently gender and race labelled and posted emoji in their tweets. From these users 3843 were White females, 3153 White males, 3197 Black females, 2920 Black males, 298 Asian females, 446 Asian males, 188 Hispanic females and 140 Hispanic males.

The achieved accuracy per race and gender categories for our combined gender and race classification can be seen on Table 5.5. Additionally, Figure 5.8 shows the average prediction accuracy by each of the six different classifiers.

Table 5.5: Gender and Race classification models performance on each of the four race and two gender categories, and the overall model accuracy

Method	White		Black		Asian		Hispanic		Accuracy
	M	F	M	F	M	F	M	F	
Prior Dummy	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.271
Gaussian NB	0.26	0.27	0.33	0.39	**0.22**	**0.20**	**0.33**	**0.16**	0.304
K-NN	0.29	0.65	0.41	0.64	0.02	0.01	0.00	0.00	0.470
Random Forest	**0.64**	0.72	0.59	0.71	0.00	0.01	0.00	0.00	0.616
SVM (Linear)	0.57	0.68	**0.70**	**0.73**	0.07	0.03	0.00	0.00	0.619
SVM (χ^2 kernel)	0.63	**0.75**	0.67	0.71	0.07	0.01	0.00	0.00	**0.642**

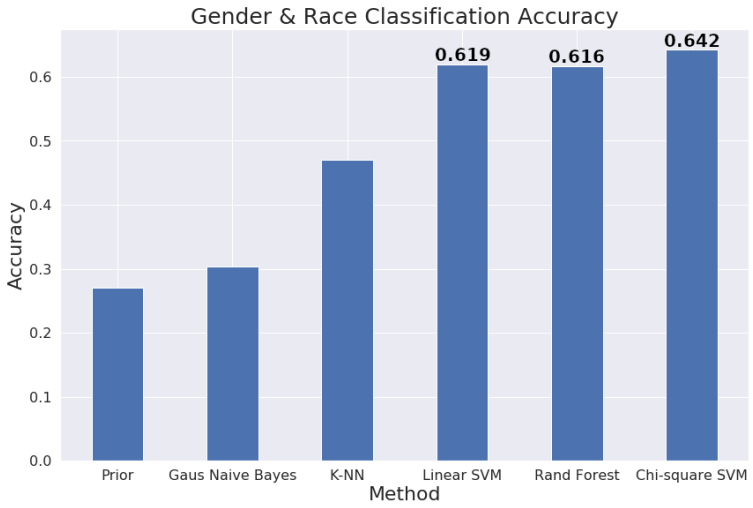


Figure 5.8: Gender and Race classification accuracy by each model.

As we have previously seen on our race only classification, most of the classifiers failed dramatically with the Asian and Hispanic race categories for both female and male users, where again only the Gaussian Naive Bayes could detect any of the Hispanic and Asian race for both genders. The best performing classifier, was the SVM with the χ^2 kernel, where it achieved an 64.2% accuracy for both gender and race detection.

However, its performance with the Asian and Hispanic classes was poor for both genders. Followed by the Linear SVM (with penalty parameter C = 10) and the Random Forest (with tree depth = 34) with 61.9% and 61.6% overall accuracy respectively.

As we have done for the race only classifiers, we created a confusion matrix of the best performing model (SVM with χ^2 Kernel) to better visualise the performance on gender and race classification using emoji, illustrated in Figure 5.9.

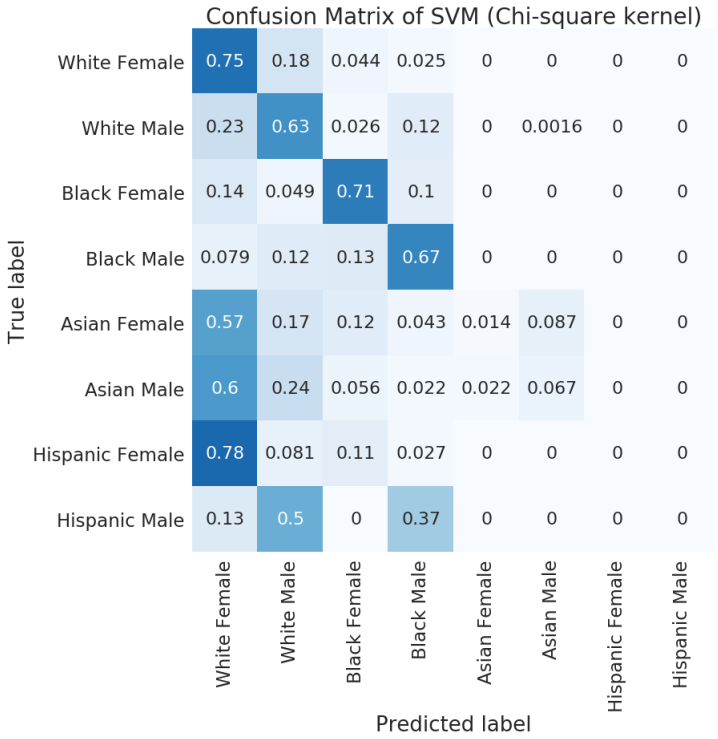


Figure 5.9: Confusion matrix o of SVM (χ^2 kernel) for both gender race classification.

From the confusion matrix above we can see that 78% of Hispanic female users were missclassified as White female users, where 50% and 37% of Hispanic male users were wrongly predicted as White male and Black male respectively. Moreover, we can see that 57% of Asian female users were missclassified as White female and 60% of Asian male users were wrongly predicted to be White male. However, it is noticed that White and Black races of both genders achieved reasonable accuracy ranging from 63% to 75%.

As we have previously seen, the minority races in our sample appear to be affecting our classification results for both gender and race classification, since gender was most of the times correctly predicted. Nevertheless as mentioned earlier, the main objective of building this classifiers is not creating the state-of-the-art race and/or gender classifiers, which could be achieved through integrating more features, such as lexical and network features. Rather, the main objective is to demonstrate the significance in emoji usage variation between different race and gender of users, which we have successfully demonstrated for males and females, and for black and white racial groups.

Chapter 6

Conclusion

Currently, social media platforms are expanding, handling thousands of public posts every second, so there is large amount of data to be analysed. Since most posts are limited to a certain amount of specified characters (Twitter allows 280 characters per tweet), emoji are more commonly used to express ideas and feelings. This is helpful identifying user demographics including race and gender by analysing emoji usage, and then match those emoji to commonly used emoji by similar demographic groups, rather than analysing a bucket of words and predicting demographics from keywords, as done in related studies.

In this study we have demonstrated, through analysis of several million user tweets, that there are differences in emoji usage between males and females, as well as between four racial groups. These results hint at the possibility that emoji are not a "universal language" and, what is more, this is significant enough to allow us to predict certain demographics, in particular, gender and the most common racial groups, with a high degree of accuracy. Crucially, the only data required to perform this classification is knowledge of the distribution of a user's emoji usage on Twitter, with no other text features or meta-data required.

6.1 Future work

In this paper we have done an in-depth analysis of the connection between emoji and two demographic features, gender and race. Additionally, classification models were trained using emoji distribution of users as the only features for predicting gender and race. However, different use cases can also be applied to further explore the emoji variation by different demographics in social media.

The most important factor in our study was the dataset used, since all analysis, experimentation and results were based on it. About 40% of the tweets analysed in this paper are from Johannesburg, New York City and London, were the rest of them were from random locations. It is highly possible that our analysis might be affected by different behaviours of users based on the areas they live. A wider dataset can be used, contain-

ing more defined geographical locations in order to better explore the usage of emoji across different areas.

In addition, user profile photos were used to identify a users gender or race for the dataset used. It is likely that a great amount of these profile pictures could be fake. As part of future work will be to detect fake user accounts and remove them from the dataset, or use other ways to validate user demographics. An approach could be to detect a user's online presence in the same or across different social network [KMW12].

Moreover the dataset we have used in this study includes tweets posted from March to October 2018. We do not know for sure how different times of the year can affect emoji usage in general. A larger dataset can be used, including a whole's year data to better understand the emoji usage in tweets for a larger period of time. Also, emoji data recorded from tweets posted on specific international days or events, might altered both our analysis and classification results. An example can be the "Jack-O-Lantern" 🎃, which is excessively used on Halloween's day. Getting rid of these noise in our dataset can lead into better results on the analysis of emoji usage by demographics, as well as for the classification models implemented based on emoji.

As seen in Section 5.2, classification of Asian and Hispanic groups based only on emoji was a difficult task for our classifiers due to the limited amount of labelled users from these two race groups in our dataset. If we further expand our dataset achieving a balanced number of users from all four races, consequently we can accomplish better classification accuracies for Asian and Hispanic race groups. On the other hand, instead of using a new dataset, different imbalanced data techniques from the ones we already tried can be applied, such as ADASYN (Adaptive Synthetic Sampling) [HBGL08].

In this paper we have shown that classification of gender and race, using only a user's emoji usage on Twitter, can be achieved at a high level of accuracy. By applying our idea of incorporating emoji usage as features to traditional gender and race classifiers, might result into superior predictions. Furthermore, other classification techniques can be used for gender and race detection using only emoji, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) that are surprisingly effective in related classification tasks [ZZ17].

As part of future work to further extend the current research could also be to explore the usage of emoji by additional demographic features such as age or posting language; or extend current demographic features. Apart from that, data can also be gathered from further social media platforms other than Twitter, such as Facebook, LinkedIn and Instagram, since the platform used might affect the way people use emoji in their posts.

Bibliography

[ALL+17] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. *Untangling Emoji Popularity Through Semantic Embeddings.* AAAI press, 2017.

[Asl18] Salman Aslam. Linkedin by the numbers: Stats, demographics fun facts. https://www.omnicoreagency.com/linkedin-statistics/, Jun 2018.

[BBRS18] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. Multimodal emoji prediction. *arXiv preprint arXiv:1803.02392*, 2018.

[BBS17] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. Are emojis predictable? *CoRR*, abs/1702.07285, 2017.

[BCC18] Francesco Barbieri and Jose Camacho-Collados. 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, 2018.

[BES14] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[Blo16] The Grammarphobia Blog. G.o.a.t. (greatest of all time). https://www.grammarphobia.com/blog/2016/07/goat.html, Jul 2016.

[CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[CGN17] Nina Cesare, Christan Grant, and Elaine O Nsoesie. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

[CLS+17] Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. Through a gender lens: An empirical study of emoji usage over large-scale android users. *arXiv preprint arXiv:1705.05546*, 2017.

[CMS15] Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. Image2Emoji: Zero-shot emoji prediction for visual media. In *Proceed-*

ings of the 23rd ACM International Conference on Multimedia, MM '15, pages 1311–1314, New York, NY, USA, 2015. ACM.

[Coa18] Steven Coats. Skin tone emoji and sentiment on Twitter. *arXiv preprint arXiv:1805.00444*, 2018.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[Dal13] Mohammad Reza Daliri. Chi-square distance kernel of the gaits for the diagnosis of parkinson's disease. *Biomedical Signal Processing and Control*, 8(1):66–70, 2013.

[DE14] Mark Davis and Peter Edberg. Proposed draft unicode technical report: Unicode emoji. Technical Report 51, Unicode Consortium, December 2014.

[Dic15] Oxford Dictionaries. Word of the year 2015. https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2015, Oct 2015.

[Dor06] Jack Dorsey. just setting up my twttr. https://twitter.com/jack/status/20, Mar 2006.

[ERA+16] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, 2016.

[FMS+17] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625. Association for Computational Linguistics, 2017.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[Gro18] Michael Grothaus. Twitter's q3 earnings by the numbers. Oct 2018.

[HBGL08] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

[HGaTNL17] Tianran Hu, Han Guo, Hao Sun andThuy-vy Thi Nguyen, and Jiebo Luo. Spice up your chat: The intentions and sentiment effects of using emoji. *CoRR*, abs/1703.02860, 2017.

[HHEQ08] Ismail Hmeidi, Bilal Hawashin, and Eyas El-Qawasmeh. Performance of knn and svm classifiers on full word arabic articles. *Advanced Engineering Informatics*, 22(1):106–111, 2008.

[Ho95] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[Hur18] George Hurlburt. Emoji: Lingua franca or passing fancy? *IT Professional*, 20(5):14–19, 2018.

[KMW12] Katharina Krombholz, Dieter Merkl, and Edgar Weippl. Fake identities in social media: A case study on the sustainability of the facebook business model. *Journal of Service Science Research*, 4(2):175–212, 2012.

[KMW17] Linda K Kaye, Stephanie A Malone, and Helen J Wall. Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences*, 21(2):66–68, 2017.

[KWM16] Linda K Kaye, Helen J Wall, and Stephanie A Malone. "turn that frown upside-down": A contextual account of emoticon usage on different virtual platforms. *Computers in Human Behavior*, 60:463–467, 2016.

[Lee18] SooJin Lee. Emoji at moma: Considering the'original emoji'as art. *First Monday*, Sep 2018.

[LF16] Nikola Ljubešić and Darja Fišer. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop*, pages 82–89, 2016.

[LNA17] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[MAP06] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes? In *CEAS*, volume 17, pages 28–69. Mountain View, CA, 2006.

[MC11] Michael Mccord and M Chuah. Spam detection on twitter using traditional classifiers. In *international conference on Autonomic and trusted computing*, pages 175–186. Springer, 2011.

[MKTS+17] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. *Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication*, pages 152–161. AAAI Press, 2017.

[MLK+18] Hannah Miller Hillberg, Zachary Levonian, Daniel Kluver, Loren Terveen, and Brent Hecht. What i see is what you don't get: The effects of (not) seeing emoji rendering differences across platforms. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):124:1–124:24, November 2018.

[MLKCR17] Antonio A Morgan-Lopez, Annice E Kim, Robert F Chew, and Paul Ruddle. Predicting age groups of twitter users based on language and metadata features. *PloS one*, 12(8):e0183537, 2017.

[Mon15] Kristina Monllos. Why dominos went nuts and wrote hundreds of tweets almost entirely in pizza emojis. https://www.adweek.com/creativity/why-dominos-went-nuts-and-wrote-hundreds-tweets-almost-entirely-pizza-emojis-164732/, May 2015.

[MTSC+16] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. *"blissfully happy" or "ready to fight": Varying interpretations of emoji*, pages 259–268. AAAI press, 2016.

[NSSM15] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, Jan 2015.

[Off11] Office for National Statistics. *2011 Census aggregate data*. 2011.

[OKP+17] Anna Oleszkiewicz, Maciej Karwowski, Katarzyna Pisanski, Piotr Sorokowski, Boaz Sobrado, and Agnieszka Sorokowska. Who uses emoticons? data from 86 702 facebook users. *Personality and Individual Differences*, 119:289–295, 2017.

[PDR17] Henning Pohl, Christian Domin, and Michael Rohs. Beyond just text: semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):6, 2017.

[PP11] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288, 2011.

[PS19] Carmen Pérez-Sabater. Emoticons in relational writing practices on whatsapp: Some reflections on gender. In *Analyzing Digital Discourse*, pages 163–189. Springer, 2019.

[RMG18] Alexander Robertson, Walid Magdy, and Sharon Goldwater. Self-representation on Twitter using emoji skin color modifiers. In *Proceedings of the 12th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

[RTG15] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[RTG16] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[SMBW15] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation

and social class from twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

[soc19] International conference on social informatics (socinfo 2019). http://socinfo2019.qcri.org/, 2019.

[Sta11] Statistics South Africa. Census 2011, 2011.

[Sug15] Satomi Sugiyama. Kawaii meiru and maroyaka neko: Mobile emoji for relationship maintenance and aesthetic expressions among japanese teens. *First Monday*, 20(10), 2015.

[TF16] Garreth W. Tigwell and David R. Flatla. Oh that's what you meant!: Reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, pages 859–866, New York, NY, USA, 2016. ACM.

[Tom76] Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 11 1976.

[twe14] Twitter emoji (twemoji). https://github.com/twitter/twemoji, 2014.

[Uni10] United States Census Bureau. Census 2010, 2010.

[WG18] Sarah Wiseman and Sandy J. J. Gould. Repurposing emoji for personalised communication: Why "pizza" means "I love you". In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 152:1–152:10, New York, NY, USA, 2018. ACM.

[Wol00] Alecia Wolf. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3(5):827–833, 2000.

[WXMX16] Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao. Improving users' demographic prediction via the videos they talk about. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1359–1368, 2016.

[ZZ17] Luda Zhao and Connie Zeng. Using neural networks to predict emoji usage from twitter data, 2017.

Appendix A

Paper for submission to SocInfo 2019

The following pages include a paper prepared to be submitted to the 11th International Conference on Social Informatics.

Emoji Usage Variation by Gender and Race on Twitter

Author 1
University
City, Country

Author 2
University
City, Country

Author 3
University
City, Country

ABSTRACT

In this paper we present a quantitative analysis study on the differences of emoji usage on Twitter by gender and race. Our experiments use the timelines of a set of up to 40,000 Twitter users that are manually labeled for gender and race. Our observations show that there are clear differences in the emoji used by males and females, and also among different racial groups. To demonstrate the significance of these differences, we built gender and race classifiers that are trained solely on the emoji used by each user. Our classifiers achieved 78% and 80% accuracy on the gender and race respectively, which confirms the significant difference in the emoji used by each demographic.

ACM Reference Format:
Author 1, Author 2, and Author 3. 2019. Emoji Usage Variation by Gender and Race on Twitter. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Emoji — icons used to represent emotions, ideas, or objects — became a formally recognized component of the Unicode Standard in 2010 [Davis and Edberg 2014] and are widely used in computer-mediated communication. Their increasing use has received much popular attention in the media and there is a growing body of research into their linguistic and sociological properties. Several studies have recently analysed the usage of these emoji by users [Kaye et al. 2017], including semantic meaning [Novak et al. 2015], sentiment effect [Miller et al. 2017; Tigwell and Flatla 2016], representation of user identity [Robertson et al. 2018], and usage by location [Ljubešić and Fišer 2016]. The importance of these studies lies better understanding of online human communication, which has wide range of applications in sociolinguistics and social science in general.

This work contributes the first in-depth analysis of the connection between emoji and two demographic features, gender and race. We apply a quantitative analysis for the difference in emoji usage between different gender and race demographics. Our results demonstrate that there are distributional differences in emoji usage between male and female Twitter users, and between Asian, Black, Hispanic and White users.

To further explore the significance of these differences in usage, we explored using emoji distribution of users as the only features for training gender and race classifiers. Our hypothesis is that the distribution of emoji used in the timeline of a given user might be a sufficient signal to detect the gender and the race of this user. Our classification results show that these differences can be successfully leveraged by classifiers to automatically detect the age or race of Twitter users. Using only the emoji found in a user's post history, gender and race of a user can be predicted with accuracy of 78.2% and 80% respectively. This highlights the significant variation in emoji usage between different demographics, which allows prediction of gender and age using emoji only. All analysis and experimentation in our study are based on a set of almost 20,000 manually annotated Twitter user accounts.

The rest of the paper is organised as follows: Section 2 gives a background on the related work. Section 3 describes the data collection used for our analysis and experimentation. Section 4 examines the distribution of tweets in users' timelines. Section 5 presents the analysis of emoji usage by gender and race. Section 6 explores utilising emoji as features for gender and race detection and reports the results. Finally, section 7 concludes the work and provides possible research directions.

2 RELATED WORK

Prior work has extensively examined the range of semantic and pragmatic functions fulfilled by emoji. This has focused on rating affectivity of single emoji as positive or negative [Miller et al. 2016; Novak et al. 2015], how people interpret emoji sentiment with/without context and across different platforms [Miller et al. 2017; Miller Hillberg et al. 2018; Tigwell and Flatla 2016], how emoji establish and/or strengthen the emotional tone of a message [Hu et al. 2017; Kaye et al. 2016], examining personalised private usage of emoji based on specific appearances [Wiseman and Gould 2018], training models for generating emoji based on input images, as well as for predicting which emojis are recalled by text-based tweet messages [Barbieri et al. 2017, 2018; Cappallo et al. 2015], learning embeddings for emoji from corpora [Ai et al. 2017; Eisner et al. 2016] and extending distant supervision through emoji prediction on tweets dataset[Felbo et al. 2017] Research has also explored broad patterns of emoji usage based on skin colour [Coats 2018] and gender [Barbieri and Camacho-Collados 2018; Chen et al. 2017; Sugiyama 2015], or in more detail the connection between emoji usage and ethnic identity [Robertson et al. 2018, 2019]. Significantly more attention has been on the general task of demographic prediction. Online social media generates large volumes of data which are of interest to researchers in fields such as social science, psychology and linguistics. However, to take full of advantage of this it is generally necessary to determine social variables based on demographic information. For example, a sociolinguist exploring the use of a particular construction online will want to know at least one of age, sex, race, ethnicity or social class of the user responsible for any observed data. Without this information, it becomes difficult to

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

do anything besides simply counting observations of phenomena. This information is rarely made explicitly available by a user as part of their social media presence: while platforms like Facebook encourage users to provide a wide range of personal information, those such as Twitter and Instagram enable users to provide a more limited range. To overcome this, researchers have developed a range of methods for inferring the required variables from the data that is actually available.

The most comprehensive overview of these methods could be found in Cesare et al. [2017], a review of 60 studies that focused on detecting age, sex and race/ethnicity from text extracted from a user's online presence. This is by far the most commonly leveraged resource and there is a large body of work on precisely which source of text provides the strongest signal for which demographics. These include a user's tweets [Bamman et al. 2014], profile display name [Vicente et al. 2015], topic keywords [Wang et al. 2016], user profile text [Sloan et al. 2015] and a user's social network features [Pennacchiotti and Popescu 2011]. While text is arguably the most useful resource available for determining demographics, metadata can be extracted to improve classifier accuracy. [Morgan-Lopez et al. 2017] improved accuracy of user age prediction by including the presence of selected emoji. However, there have to date been no attempts to predict demographics using *only* emoji.

3 DATA COLLECTION

We used the Twitter Streaming API (1% sample) to sample 3.4 million tweets made by 2.6 million unique users on the 14th of March, 2018. We inspected the location field for the users in our collection, and collected a random sample of 20,000 users from three different location. The locations we focus on are Johannesburg, London and New York City. Users were considered to be based in a location if the self-provided location on their profile matched one of the three targets or, in the case of London and New York City, a borough such as Camden or The Bronx. The motivation for the selected three locations is their demographic composition, each having different proportions of ethnic groups which should provide variation in the groups for our experimentation. Johannesburg has a predominately Black population [Statistics South Africa 2011], London predominantly White [Office for National Statistics 2011]. New York City is somewhat more balanced and a large proportion (2.3 million, approximately 28% of the city) of residents identify as Hispanic or Latinx, besides their racial identity [United States Census Bureau 2010]. In addition to those three locations, we selected a fourth group of 20,000 users that were sampled randomly from the full set of users regardless to their location. However, it was ensured that none of the users in the fourth group exist in the other groups.

For all 80,000 users we attempted to retrieve their public profile photo to be used for annotating the gender and race of the user. In some cases this was not possible, because users had removed their accounts, set their profiles to private or been banned in the time since initially selecting users. Once we had done this, each group was randomly reduced to 10,000 users.

We then used the Figure Eight[1] crowd sourcing platform to hire annotators to determine the validity of profile photos, and then the gender and race of those valid photos. We provided two choices for gender: {male, female}, and four choices for race: {White, Black, Asian, Hispanic}, where we made Hispanic to represent none-of-the-above choices. Annotators were instructed to classify a photo as "invalid" if it contained multiple or no people, if not a full-colour photo, or if the photo subject's face was obscured. Instructions for valid/invalid annotation, with examples, were available for annotators to view throughout the task. Each profile photo was annotated by three annotators. Only Twitter users where at least two annotators agreed on both photo validity and gender/race were used in this study. Out of the 40,000 user accounts, only 19,382 accounts had profile pictures that were agreed on by at least two annotators on the gender and race of the user. Statistics on the annotated accounts are shown in Table 1. As shown, males and females distribution are almost even, while the distribution of race have the majority as Black and White, while Asian and Hispanic are much less.

	White	Black	Asian	Hispanic	Total
Female	5,244	4,203	509	249	10,205
Male	4,704	3,936	782	205	9,627
Total	9,948	8,139	1,291	454	19,382

Table 1: Number of users with valid photos, per gender and race.

Finally, in April 2018 we retrieved up to the most recent 3,200 tweets for each user and excluded the retweets. We repeated this process again in November 2018, in order to gather more tweets per user. Users who had since set their profile to private, or had been deleted, were removed from the final dataset. The total number of tweets collected for the 19,382 accounts to be used in our analysis are over 18 million tweets.

4 TWEET EMOJI DISTRIBUTION

We examined the distribution of tweets in our collection by analysing the number of tweets in user profiles as well as the number of tweets containing emoji in user profiles. From 40000 timelines, we derived that there are on average 344 tweets in a user's profile, while the mean number of tweets having emoji in their content was 73. Besides that, the maximum number of tweets presented on a user's timeline was 3267, where the maximum number of tweets consisting emoji on a user's timeline was 3254. We also had, some users who didn't posted any tweets at all, meaning that our minimum number of tweets in user's profile was 0. Moreover, we saw how spread our tweets count were by determining the standard deviation for all tweets and tweets having emoji in their content to be 628 and 198 respectively.

To further explain the distribution of tweets we plotted a box plot for all tweets in users' timeline and a box plot for tweets including emoji in users' timeline. Figure 1 shows the minimum, first quartile, median, third quartile and maximum number of tweets in users' profile. As seen from the graph, the upper fence for tweets in users' timeline is 845 and upper fence for tweets consisting emoji is 125.

[1] http://www.figure-eight.com

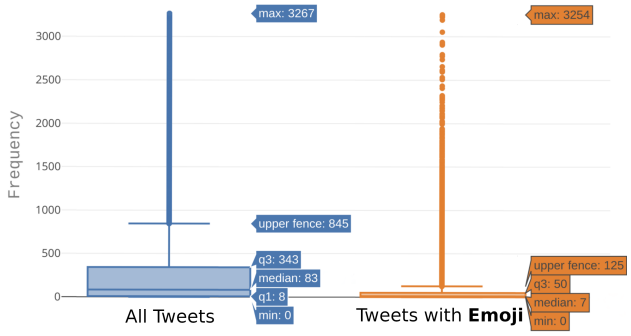


Figure 1: Box plots of all Tweets & Tweets including emoji in users' timelines

5 ANALYSING EMOJI USAGE BY DEMOGRAPHICS

In this section, we analyse the difference in emoji usage in the user timelines in our collection according to their demographics, including gender and race.

5.1 Most Frequent Emoji by Demographics

Initially, we explore the top used emoji by each group of users according to their gender and race, by simply counting the most frequent emoji in timelines. This was done using Emoji-Extractor [2], which counts the frequencies of emoji in a string, as well as frequencies of repeated emoji, since a single tweet can consist an emoji multiple times. Figure 2 shows the most frequent emoji in used by users in general and by each gender and race. As shown, the "Face with Tears of Joy" 😂 is the most frequent emoji for all genders and races, which aligns with previous studies and reports [Ljubešić and Fišer 2016]. However, some variations could be noticed for the later most frequent ones, especially for different user races. These variations indicate that there might be different patterns of usage according to race and gender.

5.2 Most Distinctive Emoji for each Gender and Race

To further investigate the differences in emoji usage, we calculated the most distinctive emoji set for each gender and race by measuring the difference in percentage of usage between a given group and the others. For an emoji e_i, the distinctiveness of e in group x, $D(e_i|x)$ is calculated as follows:

$$D(e_i|x) = \frac{count(e_i|x)}{count(e|x)} - \frac{count(e_i|\bar{x})}{count(e|\bar{x})} \quad (1)$$

where $count(e|x)$ is the count of all emojis in x, and $\bar{x}$ represents all the other groups excluding group x. For example, $D(e_i|female)$ is the percentage of usage of this emoji within the timelines of female users minus its usage in males timelines. Similarly, $D(e_i|black)$ is he percentage of usage of e within the timelines of black users minus its usage percentage in timelines of while, Asian, and Hispanic users combined.

[2]https://github.com/alexanderrobertson/emoji-extractor

The most distinctive set of emoji —with the highest difference in usage percentage— is extracted for each group. Figure 3 and Figure 4 report the top distinctive emoji set for each gender and race respectively.

As shown in Figure 3, the most distinct emoji for males were the "Face With Tears of Joy" 😂 with 2.40% difference in usage than females, the 'Fire' 🔥 with 2.37% and the "Grinning Face With Sweat" 😅 with 0.67%. For female users, it is clear the emoji representing 'hearts' are significantly more popular within female users than males, where the most distinct emoji for females were the "Red Heart" ❤️ with 2.56% difference in usage than males, the "Smiling Face with Heart-shaped Eyes" 😍 with 2.07% and the "Loudly Crying Face" 😭 with 1.87%. These observations show that there is clear difference in the percentage of emoji usage between males and females.

Figure 4 shows the distinctive emoji according to race. As shown, the "Red Heart" ❤️ is more used by white users more than any other race with 1.90% difference in usage than others. The "Face With Tears of Joy' 😂 is the most distinctive emoji for Black users with a superior 14.79% difference in usage that other race groups. It could be also noticed that for Black users, emoji with dark skin tones are more commonly used other races, correspondingly white skin tone emoji for White, which aligns with findings in [Robertson et al. 2018]. The most distinct emoji for Hispanic users was the "Smiling Face with Heart-shaped Eyes" 😍 with 2.63% difference than all other race groups, where the most distinct emoji for Asian users was the "Loudly Crying Face" 😭 with 4.35% difference.

Capturing distinctive emoji by calculating the difference of frequency percentages among two sets was not reliable in some cases. For example, if $e_1|female$ was 11% and $e_1|male$ was 3%, $e_2|female$ was 8% and $e_2|male$ was 0%. Our measure will treated $D(e_1|female)$ (11%-3%) and $D(e_2|female)$ (8%-0%) to be equal, meaning that emoji_1 and emoji_2 will have the same distinctive power, however, emoji_2 should have been more more distinctive as it only appears in female profiles. Therefore we tried an additional approach for capturing distinctive emoji for each group by calculating feature (emoji) importance by training a Random Forest [Ho 1995] model.

We fitted a Random Forest model on data containing 2579 emoji as features, users as instances, the gender or race of each user as classification target and the percentage usage of emoji from each user as feature variables. A 5-Fold Cross-validation was used to determine a suitable value d (maximum depth of the tree), achieving the best "F1 Score" for gender classification and best "Accuracy Score" for race classification. Then we extracted emoji (feature) importance for emoji having the highest weights in order for our model to predict a specified target group.

The most distinctive set of emoji (features) —with the highest importance in Random Forest model— is extracted for each group. Figure 5 and Figure 6 report the top distinctive emoji set for each gender and race respectively.

As shown in the bar charts in Figure 5, the most distinctive emoji for males were the "Fire" 🔥 with 0.0029 importance and the "Face with Tears of Joy" 😂 with 0.0012. It was interesting to see emoji representing "Man Facepalming" 🤦🏾‍♂️🤦🏽‍♂️🤦🏻‍♂️🤦🏼‍♂️ and "Man Shrugging" 🤷🏾‍♂️🤷🏽‍♂️🤷🏼‍♂️🤷🏻‍♂️ with multiple colour skin tones. They appear to be very

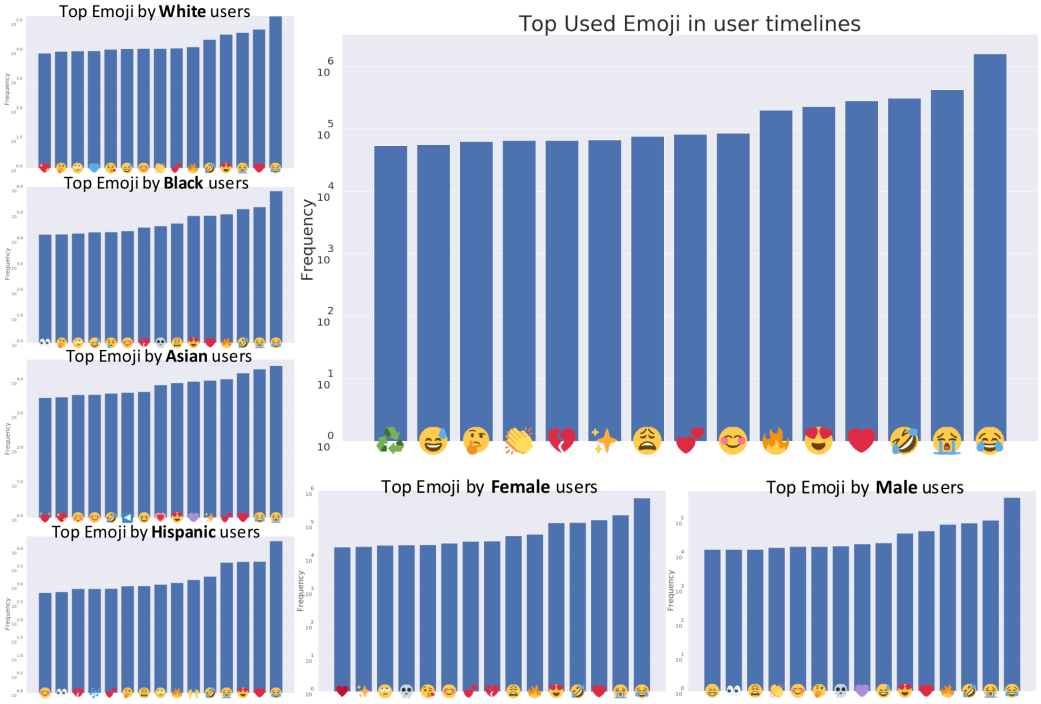


Figure 2: Top used emoji by user timelines in our collection in general and according to gender and race.

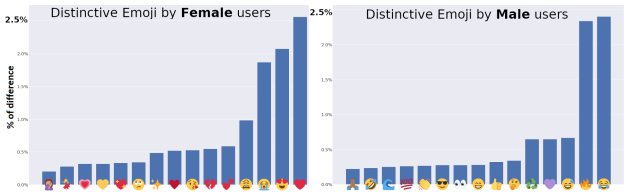


Figure 3: The most distinctive emoji with highest percentage difference between females and males. Y-axis values are different for both graphs.

important for male user classification using the Random Forest model.

Moreover, the most distinctive emoji for females were the "Red Heart" ❤️ with 0.0039 importance and the "Smiling Face with Heart Shaped Eyes" 😍 with 0.0031. Repeatedly, we can seen that emoji representing "Female Facepalming" 🤦‍♀️, "Female Shrugging" 🤷‍♀️ of different skin tones and "Heart" are crucial in predicting female users by our Random Forest model. Our observations using the Random Forest feature importance approach show that there is obvious difference in emoji usage between male and female users.

From Figure 6 we can see the "Thumbs Up" 👍 to be the most important feature for White users classification with 0.00098 weight, and the "Face With Tears of Joy" 😂 the most distinctive emoji for Black users classification, with 0.0191 weight, as we have seen in our previous approach. Again, we see white skin tone emoji appear to be more important for White user classification, and emoji with dark skin tones to be more critical for Black users classification.

The most distinctive emoji for Hispanic users was the "Red Heart" ❤️ with 0.0041 importance, where the most important emoji for Asian user classification was the "Loudly Crying Face" 😭, as we found out to be the most distinctive emoji for the same group in our first approach.

Our previous analysis highlights the main differences in emoji usage between difference gender and races, which was shown to be significant in some cases. In the following section, we examine the effectiveness of using these difference to automatically detect the user gender or race.

6 GENDER AND RACE DETECTION USING EMOJI

6.1 Gender and Race Classifiers

In this part, we build two classifiers for gender and race detection based on emoji only. Each user in our collection is represented as the set of emoji found in his/her timeline. A set of 2579 unique emoji were used as our feature set, where each emoji represent a feature, users as instances, the gender or race of each user as classification target and the percentage usage of emoji from each user as feature variables.

For gender classification we have used 16,005 users, which were confidently gender labelled and used emoji in their tweets. From those, 7,521 are male and 8,484 are female users. Then, for race

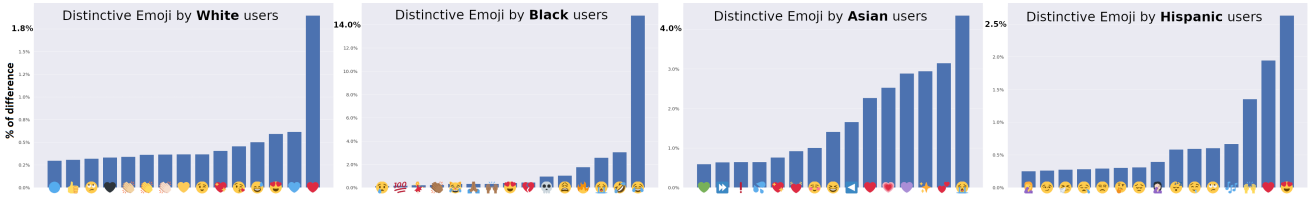


Figure 4: The most distinctive emoji with highest percentage difference among White, Black, Asian, and Hispanic users. Y-axis values are different for each graph.

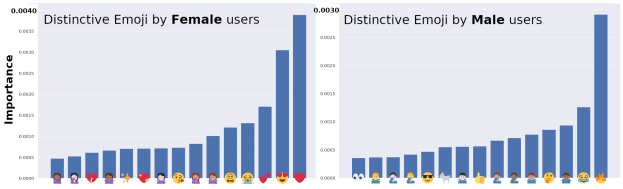


Figure 5: The most distinctive emoji with highest importance between females and males. Y-axis values are different for both graphs.

classification we have used 14,316 users who were trusty race labelled and posted emojis in their tweets. From these users 7047 were White, 6161 were Black, 780 were Asian and 328 were Hispanic. 80% of these users were randomly chosen to be used as part of train set and 20% for testing our classifiers.

Scikit-Learn, a Python machine learning (ML) library [Cournapeau et al. 2007] was used for training our classifiers. Multiple ML classification techniques were examined to classify both the gender and race of each user as follows:

- Gaussian Naive Bayes (NB).
- K Nearest Neighbour (KNN).
- Random Forest.
- Support vector machines (SVM) with linear kernel.
- SVM with Chi-squared kernel [Cournapeau et al. 2007].

In addition, we used the Prior Dummy classifier as our baseline model, which represents the majority-class baseline by choosing the most frequent class as the predicted label.

Then we have used Gaussian Naive Bayes algorithm for classification, which is a a simple and popular classification technique both for binary (female or male) and multi-class (Asian, Black, Hispanic or White) tasks. For the KNN classifier, the optimal K was selected based on a 5-fold cross-validation on the training set by selecting the value that achieves the highest F1 Score for gender classification and best accuracy for race classification. Similarly, the value of tree depth d and C parameters for the random forest and SVM classifiers respectively were determined through 5-fold cross-validation on the training set.

Both F1-score and accuracy where used for evaluating the classifiers. In addition, we report the performance of the classifiers for each of the race classes separately to have a deeper understanding of the performance in each race category because of their high imbalance in size.

Table 2: Gender classification models performance using accuracy and macro F1-Score

Method	Accuracy	F1 Score
Prior Dummy	0.521	0.343
NB	0.629	0.602
K-NN	0.667	0.574
Random Forest	0.778	0.762
SVM (Linear kernel)	0.759	0.744
SVM (χ^2 kernel)	**0.782**	**0.766**

6.2 Classification Results

Table 2 reports the accuracy and macro F1-score for different models used for gender classification. As shown, the majority-class baseline using the prior dummy classifier achieved the lowest performance since it was predicting all test set as males. NB and KNN classifiers achieved low performance compared to other classifiers. The best performing classifier was the SVM with χ^2 kernel, which achieved a 78.2% accuracy. This result illustrates the significant difference in usage of emoji between males and females that allowed detecting of the users' gender from the used emoji in their timeline solely.

Table 3 reports the classifiers performance for race classification using overall accuracy and the achieved f-score per race category. Most of the classifiers failed dramatically with the rare race categories Asian and Hispanic, where actually none of the classifiers (except NB) could detect any of the Hispanic race. While NB achieved the best performance for detecting the two rare classes, it achieved the worst overall performance, because of it poor performance with the White and Black categories which constitute over 90% of the data population. Actually, NB achieved an overall accuracy worse than the dummy majority-class baseline that assigns all predictions to the White category. The best performing classifier was the SVM with the χ^2 kernel, where it achieved an 80% accuracy for race detection. Nevertheless, its performance with the Asian and Hispanic classes was poor.

To better understand the performance on race classification using emoji, we constructed the confusion matrix for the best performing model (SVM with χ^2 Kernel). Figure 7 shows that 74% of Asian users were misclassified as White, while 78% and 22% of Hispanic users were misclassified us White and Black respectively.

Since our classes were not balanced, meaning that we had more instances of a White and Black race groups than Asian and Hispanic groups, we tried two additional approaches expecting to improve our classification accuracy.

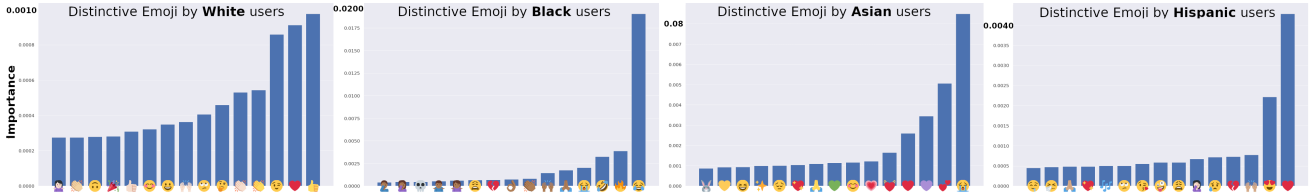


Figure 6: The most distinctive emoji with highest importance among White, Black, Asian, and Hispanic users. Y-axis values are different for each graph.

Table 3: Race classification models performance on each of the four race categories measured using F1-score, and the overall model accuracy

Method	White	Black	Asian	Hispanic	Accuracy
Prior Dummy	1.00	0.00	0.00	0.00	0.51
NB	0.36	0.47	**0.38**	**0.33**	0.41
K-NN	0.7	0.79	0.04	0.00	0.68
Random Forest	0.88	0.79	0.05	0.00	0.78
SVM (Linear)	0.82	0.82	0.03	0.00	0.76
SVM (χ^2 kernel)	**0.89**	**0.80**	0.19	0.00	**0.80**

The first approach was SMOTE (Synthetic Minority Over-sampling Technique) [Chawla et al. 2002] oversampling with 4 neighbours on our train set before training our classifiers. The second approach was to undersample the whole data set using Tomek's links[Tomek 1976] and a random undersampler. However, the best accuracy using the oversampled train data was 79%, achieved by the SVM with the χ^2 kernel; and the best accuracy using the undersampled dataset was 61%, achieved using the Random Forest classifier.

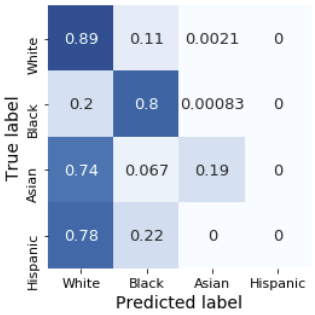
The limited amount of Asian and Hispanic race group users is considered one of the limitations of the race classifier. Nevertheless as mentioned earlier, the main objective of building this classifiers is not creating the state-of-the-art race or gender classifiers, which could be achieved through integrating more features, such as lexical and network features. Rather, the main objective is to demonstrate the significance in emoji usage variation between different race and gender of users, which we have successfully demonstrated for males and females, and for black and white racial groups.

7 CONCLUSION

We have demonstrated, through analysis of several million user tweets, that there are differences in emoji usage between males and females, and also between four racial groups. These results hint at the possibility that emoji are not a "universal language" and, what is more, this is significant enough to allow us to predict certain demographics (in particular, gender and the most common racial groups) with a high degree of accuracy. Crucially, the only data required to perform this classification is knowledge of the distribution of a user's emoji usage on Twitter - no other text features or metadata are necessary.

As part of future work, we can explore the usage of emoji by additional demographic features such as age or posting language. Apart from that, additional data can be gathered from further social media platforms other than Twitter, such as Facebook, LinkedIn and Instagram.

Figure 7: Confusion matrix of the SVM (χ^2 kernel) race classifier



In this paper we have shown that classification of gender and certain race groups, using only a user's emoji usage on Twitter, can be achieved at a high level of accuracy. By applying our idea of incorporating emoji usage as features to traditional gender and race classifiers, might result into superior predictions. Furthermore, other classification techniques can be used for gender and race detection using only emoji, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) that are surprisingly effective in related classification tasks [Zhao and Zeng 2017].

REFERENCES

Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. *Untangling Emoji Popularity Through Semantic Embeddings.* AAAI press, 2017.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

Francesco Barbieri and Jose Camacho-Collados. 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, 2018.

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. Are emojis predictable? *CoRR*, abs/1702.07285, 2017. URL http://arxiv.org/abs/1702.07285.

Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. Multimodal emoji prediction. *arXiv preprint arXiv:1803.02392*, 2018.

Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. Image2Emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1311–1314, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806335. URL http://doi.acm.org/10.1145/2733373.2806335.

Nina Cesare, Christan Grant, and Elaine O Nsoesie. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. Through a gender lens: An empirical study of emoji usage over large-scale android users. *arXiv preprint arXiv:1705.05546*, 2017.

Steven Coats. Skin tone emoji and sentiment on Twitter. *arXiv preprint arXiv:1805.00444*, 2018.

David Cournapeau et al. scikit-learn. https://github.com/scikit-learn/scikit-learn, 2007.

Mark Davis and Peter Edberg. Proposed draft unicode technical report: Unicode emoji. Technical Report 51, Unicode Consortium, December 2014. URL https://www.unicode.org/reports/tr51/tr51-1-archive.html.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, 2016.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1169. URL http://aclweb.org/anthology/D17-1169.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

Tianran Hu, Han Guo, Hao Sun andThuy-vy Thi Nguyen, and Jiebo Luo. Spice up your chat: The intentions and sentiment effects of using emoji. *CoRR*, abs/1703.02860, 2017. URL http://arxiv.org/abs/1703.02860.

Linda K Kaye, Helen J Wall, and Stephanie A Malone. "turn that frown upside-down": A contextual account of emoticon usage on different virtual platforms. *Computers in Human Behavior*, 60:463–467, 2016.

Linda K Kaye, Stephanie A Malone, and Helen J Wall. Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences*, 21(2):66–68, 2017.

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

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. *"blissfully happy" or "ready to fight": Varying interpretations of emoji*, pages 259–268. AAAI press, 2016.

Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. *Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication*, pages 152–161. AAAI Press, 2017.

Hannah Miller Hillberg, Zachary Levonian, Daniel Kluver, Loren Terveen, and Brent Hecht. What i see is what you don't get: The effects of (not) seeing emoji rendering differences across platforms. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):124:1–124:24, November 2018. ISSN 2573-0142. doi: 10.1145/3274393. URL http://doi.acm.org/10.1145/3274393.

Antonio A Morgan-Lopez, Annice E Kim, Robert F Chew, and Paul Ruddle. Predicting age groups of twitter users based on language and metadata features. *PloS one*, 12 (8):e0183537, 2017.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

Office for National Statistics. *2011 Census aggregate data*. 2011. doi: http://dx.doi.org/10.5257/census/aggregate-2011-1.

Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288, 2011.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. Self-representation on Twitter using emoji skin color modifiers. In *Proceedings of the 12th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. Emoji skin tone modifiers: Analyzing variation in usage on social media. In *Submitted*, 2019.

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

Statistics South Africa. Census 2011, 2011. URL http://www.statssa.gov.za/?page_id=3836.

Satomi Sugiyama. Kawaii meiru and maroyaka neko: Mobile emoji for relationship maintenance and aesthetic expressions among japanese teens. *First Monday*, 20 (10), 2015.

Garreth W. Tigwell and David R. Flatla. Oh that's what you meant!: Reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, pages 859–866, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4413-5. doi: 10.1145/2957265.2961844. URL http://doi.acm.org/10.1145/2957265.2961844.

Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 11 1976. doi: 10.1109/TSMC.1976.4309452.

United States Census Bureau. Census 2010, 2010. URL https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045217.

Marco Vicente, Fernando Batista, and Joao Paulo Carvalho. Twitter gender classification using user unstructured information. In *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*, pages 1–7. IEEE, 2015.

Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao. Improving users' demographic prediction via the videos they talk about. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1359–1368, 2016.

Sarah Wiseman and Sandy J. J. Gould. Repurposing emoji for personalised communication: Why "pizza" means "I love you". In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 152:1–152:10, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173726. URL http://doi.acm.org/10.1145/3173574.3173726.

Luda Zhao and Connie Zeng. Using neural networks to predict emoji usage from twitter data, 2017.

Appendix B

Twitter developer account application

Gmail

Twitter developer account application [ref:_00DA0K0A8._5004A1V0qle:ref]

S Nicoletti [redacted] 28 September 2018 at 13:21
To: developer-accounts@twitter.com

I am a fourth year student at the University of Edinburgh. The core use case of the Twitter APIs will be to stream Tweets and analyse the usage and interpretation of emojis across Tweeter for my University Undergraduate Honours Project. In this project, we want to know if emojis are used differently by different personas. The main two categories of analysis are 1) gender: males/females, and 2) age: young/old. However, this can include other factors, such as countries, languages, and popularity on social media.

For each group of persona, the following analysis to be applied:

Percentage of usage of emojis in each group
Top used emojis
Used in communication (replies, mentions) or in general posts
Topics of usage
Once the analysis is performed, the next step will be building models for recommending emojis to social media users in their posts

The work steps contain the following:

Collecting millions of tweets with emojis in English at least, (Twitter API to be used)
Applying one of the state-of-the-art methods for gender and age estimation of Twitter users
Performing statistical analysis for data
Building emojis usage models for recommendation

My project will not involve Tweeting, Retweeting, or liking content.

Twitter content will be displayed on aggregate level. Nothing personal will be displayed. As I said previously, data will be displayed in graphical form (pie chart, bar chart) eg. male vs female, young vs old, percentage usage of an emoji.
[Quoted text hidden]

Twitter developer account application [ref:_00DA0K0A8._5004A1V0qle:ref]

developer-accounts@twitter.com <developer-accounts@twitter.com> 28 September 2018 at 15:10
To: [redacted]

Your Twitter developer account application has been approved!

Thanks for applying for access. We've completed our review of your application, and are excited to share that your request has been approved.

Sign in to your developer account to get started.

Thanks for building on Twitter!

Appendix C

Streaming public tweets code

Part of the code used to stream public tweets using “Tweepy” library, including authentication of user credential to access Twitter API.

```
# Import from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

# User credentials to access Twitter API
access_token = '1044780251073310666l-ZfLYOF4jquMLDRvOGJLOj4uNqOPMaS'
access_token_secret = 'f4Wgi5OEopkfwonPuQGhgtqOsxvJTn10P5402gyxcf1jt'
consumer_key = '7gpuwvPv5YAkdPMkVUplTLDmz'
consumer_secret = '07s8856620tisy3cvRKxoMiDJ6j55kbuOpNOAvci55JqbO3qn'

# This is a basic listener that just prints received tweets to stdout.
class StdOutListener(StreamListener):

    def on_data(self, data):
        print data
        return True

    def on_error(self, status):
        print status

if __name__ == '__main__':

    # Twitter authetification and the connection to Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    # Capture tweet samples
    stream.sample()
```

Appendix D

Random Forest feature importance extraction code

Part of the code used to fit a Random Forest Model and extract emoji importance for each target group.

```
# Fit Random Forest model to the data
tree_model = RandomForestClassifier(n_estimators=500,criterion='entropy',
                                    random_state = 42,max_depth=55)
tree_model.fit(X,y)
feature_importances = tree_model.feature_importances_
```

```
# Function that returns the importance according to each target group
def class_feature_importance(X, Y, feature_importances):
    N, M = X.shape
    X = scale(X)
    out = {}
    for c in set(Y):
        out[c] = dict(
            zip(range(N), np.mean(X[Y==c, :], axis=0)*feature_importances))
    return out
```

```
# Call function to return emoji importances according to each target group
importance_by_target_group = class_feature_importance(X, y, feature_importances)
```

Appendix E

Fitting classification models code

Part of the code used to train all six classification models using Scikit-Learn built in functions is provided below. Code highlighted as green show parameters where a 5-Fold Cross-validation on the training set was used to determine suitable values for certain models. Code highlighted as yellow shows the data fitting process, where data was given to the models as "X_train" (independent features) and "y_train" (dependent variables).

```
# Prior Dummy
pd = DummyClassifier(strategy='prior').fit(X_train, y_train)

# Gaussian Naive Bayes
gnb = GaussianNB().fit(X_train, y_train)

# K Nearest Neighbour
knn = KNeighborsClassifier(n_neighbors=best_k_achieved_with_cross_validation).fit(X_train, y_train)

# Random Forest
rf = RandomForestClassifier(n_estimators=500, criterion='entropy',random_state = 42,
                        max_depth=best_depth_achieved_with_cross_validation).fit(X_train, y_train)

# Linear SVM
l_svm = linearSVC(random_state=0, tol=1e-5,
          C=best_C_achieved_with_cross_validation,max_iter=30000).fit(X_train, y_train)

# SVM with Chi-square Kernel
cs_svm = SVC(kernel=chi2_kernel, probability=True).fit(X_train, y_train)
```