

All the Feels :
Sentiment Analysis
Between Emoji and Text

Alexandra Plassaras





Columbia University

QMSS G4010:
Theories and Methodologies
in the Social Sciences

Table of Contents

Introduction	3
Research Problem	4
<i>Motivation</i>	4
<i>Hypotheses</i>	5
<i>Previous Research</i>	7
Research Design	12
<i>Data</i>	12
<i>Sentiment Analysis</i>	14
Text	14
Emoji	15
<i>Data Processing</i>	16
Discussion	9
<i>Limitations</i>	9
Conclusion	16
Appendices	20
Appendix A – QMSS G4063 Information	20
Appendix B – All variables associated with a single tweet	20
Appendix C – Unicode List of Emoji	20
Appendix D – Hu and Liu’s Lexicon for Text Analysis	21
Appendix E – Text Sentiment Analysis Visualizations	21
Appendix F – Emoji Sentiment Lexicon	22
References	23

Introduction

   and  . These are just two novels that have in the past few years been translated into emoji. Can you guess which classic novels these emojis are referring to? If you guessed Alice in Wonderland (“Author Translates All”, 2016) and Moby Dick (“Emoji Dick”, n.d.) then you are correct. In recent years we have seen a new way to express ourselves in online and mobile communication. As of 2015, emojis have become the world’s fastest growing language in all forms of communications – social media, text messaging and various messaging apps and even email (Emogi Research Team, 2015). A survey conducted by TalkTalk Mobile, a British mobile retailer, found that 72% of 18 to 25 year olds stated that emojis were easier to use to express their feelings than text (Doble, 2015). Knapp and Hall made the claim that emojis which serve as nonverbal conversational cues “help to communicate ideas, manage interactions and disambiguate meaning to improve the efficiency of the conversation” (2010).

Since support for emoji became available to major mobile operating systems iOS and Android in the US in 2011 there has been an increase of emoji usage (Grady, 2016). Originally created for use in Japanese mobiles in the late 1990s, emojis have slowly made their way into mainstream communication. As of 2016, research had suggested that emojis have already taken over emoticons on social media most likely due to their flexibility in expressing not only facial expressions but food, religion, activities and even various cultures (Miller, Thebault-Spieker, Chang, Terveen and Hecht, 2016).

Prior to the widespread use of emojis, emoticons were widely used to express feelings, moods and emotions. An emoticon is shorthand for a facial expression – such as

: -) or : - (. Emojis are emoticons on steroids – instead of using alphanumeric, punctuations and logic symbols (Walther and D’Addario, 2003), emojis are graphic symbols that represent facial expressions as well as concepts and ideas (Novak, Smailović, Sluban and Mozetič, 2015). Prior to the introduction of emojis it was not possible to convey ‘wine glass’ or ‘Sweden’ using emoticons. Now however, actions, religions, cultures, animals and plants can all be expressed using emojis. Given the newfound prevalence of emoji in our every day communication, the focus of this paper is to explore the sentiments that emojis attempt to convey using sentiment analysis of Twitter data.

Research Problem

Motivation

The motivation for this topic is my interest in the use and meaning of emojis in online communication. Emoji originated in Japan in the late 1990s well before smartphone operating systems in the US allowed their usage. Having lived in Japan and having used Japanese phones and emojis to communicate with friends well before their release to the US market, I was excited when for the first time US phones allowed users to communicate with emojis in 2011. I noticed however that when I used certain emojis or combinations of emojis in text messages to American friends sometimes there was miscommunication of what I was trying to convey. Last year I learned about sentiment analysis on text which I found quite interesting. Now that emojis are more common here in the United States I wanted to see if I could use sentiment analysis (also known as opinion mining) on emoji. Given that often when people communicate with each other,

particularly on social media platforms like Twitter, emoji is not used alone I wanted to compare sentiment analysis of emojis with text to see what relationship the two types of online communication had with one another.

Having a large Twitter data set at my disposal I wanted to utilize this data, since it can either be time or money intensive to obtain data sets large enough to do sentiment analysis. Additionally I wanted to learn more about how to implement various machine learning techniques to real life data sets. As I intend on working with real-life data sets like this one in the future this is a valuable experience for me to gain hands-on analysis experience.

Hypotheses

This study will focus on three hypotheses. The first hypothesis is that there will be a significant difference between sentiment analysis of text compared to the sentiment analysis of emojis. This is because I believe that there will be a higher amount of sarcasm used on this social media platform. While the text “Make America Great Again” may be coded as having positive sentiment, I believe that the emojis will not always share the same sentiment and in the case of this example will perhaps be negative (e.g. using sad, crying or angry faces). I am also making the assumption here that there are more urban, liberal and coastal Twitter users than there are rural, conservative users who live in the center of the country and thus tweeting the sarcastic example text and emojis above would be more likely by the young urban, coastal and liberal Twitter users. Boia *et al.* looked at emoticons and their relationship with text sentiment and came to the conclusion “that the sentiment conveyed by an emoticon generally agrees with the sentiment of the entire Tweet” (Boia, Faltings, Musat and Pu, 2013). However I argue that given the vast

number of emojis that exist, many of which are not faces and instead food emojis, inanimate objects and shapes that the relationship between emojis and text will be different than the relationship between emoticons and text.

The second hypothesis is that in general tweets referencing Republican candidates (e.g. Trump, Cruz and Rubio) will contain more negative sentiment for emojis than tweets referencing Democratic candidates (e.g. Clinton and Sanders). In the example I mentioned above I am hypothesizing that people discussing Republican candidates who use emojis are more likely to be young users who I am assuming are more liberal and more likely to use emojis in a sarcastic and negative manner.

The third hypothesis is that tweets that mention Democratic candidates are more likely to have higher negative text sentiment than tweets mentioning Republican candidates. I am hypothesizing that Twitter users who reference a candidate are more likely to use harsher and more negative language if they are referring to a Democratic candidate. This might be because of 1) the strong affiliation of some Democrats who were pro-Sanders to talk about Hillary Clinton in a negative light such as those who supported the 'Never Clinton' Campaign (Foran, 2016) or 2) Republican voters who were either against Sanders or more likely against Clinton. After the recent rise in hate crimes (Yan, Sgueglia and Walker, 2016) after the popular vote that named Donald Trump President-Elect this November I am assuming that people who were more likely to support Trump were more likely to use vocabulary similar to Trump's rhetoric of "losers," "total losers," "haters," "dumb," "idiots," "morons," "stupid," "dummy" and "disgusting" (Shafer, 2015).

Previous Research

Much research has been done on text sentiment analysis ranging from subjectivity and sentiment analysis (Liu, 2010) to detecting sarcasm in sentiment analysis (Maynard and Greenwood, n.d.). As the focus of this study is on emoji sentiment analysis and how it compares to text sentiment analysis, the majority of this section will focus on previous research done concerning understanding emoji. Given that emojis were introduced to Americans in 2011 on a large scale when Apple, existing studies primarily on emoji are quite limited. Additionally as Lu *et al.* mentions, it is harder to come across large data sets of emoji usage (2016).

The research uncovered so far includes a global analysis of emoji used on smartphones via the Kika Emoji keyboard, one of the most popular third party keyboards on Android smartphones (Lu, Ai, Liu, Li, Wang, Huang and Mei, 2016). This study looked at over 400 million emoji-contained messages from users in 212 different countries and showed that there is a difference in emoji usage based on country and region. Another study looked at how people interpret different emojis as well as the same emojis on different platforms (Tigwell, Flatla, 2016). Subjects were surveyed from various countries including the US, the UK, Canada, Brazil and Germany and were recruited from social media. This study made the claim that people do in fact interpret emojis differently on an individual basis and not just from a cultural and country basis.

Students from Stanford University also used sentiment analysis for the 2016 Presidential candidates. Their study developed an operationalization of five different emotions – happy, sad, fear, laughter and anger – based on the emoji used in their tweets (Chinn, Zappone and Zhao, 2016). This study looked at over 300,000 tweets with

keywords such as “politics”, “political candidates” or the full names of the presidential candidates. They then used three different models to classify their tweets – Naïve Bayes, Support Vector Machine and Nearest Neighbors. The focus on this study was to expand upon that traditional polarity analysis of positive, neutral or negative to include five unique emotions.

Previous work done by Novak *et al.* significantly affects the scope of this project because their study created the first known emoji sentiment lexicon, referred to as the Emoji Sentiment Ranking (2015). This study labeled over 1.6 million tweets in 13 different European languages (including English) and created a polarity measure for each emoji off of the 4% of the tweets that contained emojis (roughly 64,000 tweets). What they found was that the majority of emojis were positive and that the sentiment of tweets with and without emojis varied greatly. The emoji lexicon created through this work will be used as the emoji lexicon of this study.

Previous research has not focused on comparing sentiment of text with emoji and has typically been either text or emoji. Furthermore, the research done on emoji has been conducted on either a small scale (under 30,000 instances) or using platform specific data from mobile carriers or specific applications. There have not been studies that have looked at sentiment analysis of emojis in Tweets. Thus, this paper hopes to add insights on the differences between text and emoji sentiment found in Tweets.

Discussion

Limitations

Limitations to this study include the potential for sampling error, difference between British and American English, time restraints, misunderstanding of emoji, lack of context for emoji analysis, accuracy of sentiment analysis and the use of non-standard words used during this campaign cycle. The first limitation to this study is the large possibility of sampling error. According to Pew Research Center in 2015 it was estimated that only 23% of all internet users and 20% of the entire adult population in the US use Twitter (Duggan, 2015). As such this data set is not representative of all Americans nor is it representative of Democrats and Republicans. Another study conducted by Pew Research Center estimated in 2012 that of the 16% of internet users that used Twitter, 12% of users were estimated to be Republicans and around 18% were estimated to be Democrats (Smith, 2013). Additionally, this study only looks at Tweets and not other types of social media or e-communication like Facebook, text messages or emails etc. Since this data set is not representative of the population intended to be analyzed – Americans who use electronic communication – there is the high possibility of selection bias.

The second limitation of this study is that the lexicon used to analyze emoji sentiment was created for a separate study that looked at the sentiment of emojis in the UK. Jack *et al.* suggests that people's interpretation of facial expressions and thus emojis differ between cultures (Jack, Blais, Scheepers, Schyns, and Caldara, 2009). Another study that looked at emoji usage of smartphone users ranked the top 10 emojis in the top 10 countries to see which emojis were most often used in different places (Lu, Ai, Liu,

Li, Wang, Huang and Mei, 2016). France was the only country in which all 10 emojis that were most used contained a heart somewhere in the emoji (See Appendix E for the table results). Park *et al.* discusses the fact that “easterners and westerners prefer different style of emoticons” (Park, Barash, Fink and Cha, 2013) but perhaps there are more differences between countries than just being considered eastern or western. As a result, the true sentiment of the emojis in this American data set may not be fully represented when using a lexicon built for British emoji use. However the same study that showed the top 10 French heart emojis result also made the claim that “countries sharing similar emoji usage patterns are more likely to share common language or geo-region” (Lu, Ai, Liu, Li, Wang, Huang and Mei, 2016). Perhaps, then, the difference between using a British based lexicon for an American data set might not have that much of an effect but it is something to consider throughout this study.

The third limitation is one of time constraints. Given that this data set contains only 56 days worth of data and the fact that this data was specifically capturing only tweets pertaining to the 2016 primary elections, Lu *et al.* suggests in their paper that these events may have lead to “unrepresentative user moods and behaviors” that could have affected how users chose emojis (2016).

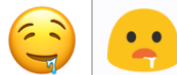
The fourth limitation of this study is that it does not account for the misunderstanding of emojis by various users. Tigwell and Flatla state that two common reasons for users to misunderstand emoji are 1) the definition and use of emoji and 2) the different emoji designs on different platforms (Tigwell, Flatla, 2016). People’s opinion on how emojis should be used and what they represent can vary greatly. For example, what exactly is the emoji in Figure 2 below and should it be used as a positive emoji or a

negative emoji? Additionally, depending on whether a person is using an iOS or an Android phone, emojis can look very different and perhaps even have different meaning entirely. Figure 3 below shows the same Unicode emoji that looks slightly different between iOS and Android platforms. It is possible that a user might think the Android emoji on the right is more negative than the iOS emoji is more positive. Both emoji have the label “drooling face” but the emotions they convey might not be the same for every user.

Figure 2 – “Smiling face with open mouth & cold sweat” Emoji



Figure 3 – “Drooling face” Emoji



iOS (left), Android (right)

The fifth limitation of this study is that context is not accounted for in the sentiment analysis for the emoji. The analysis of both the text and the emoji are done in two different silos and as a result the polarity of the overall tweet is not accounted for. It could be that a Tweet that has very negative text but very positive emojis may actually be either negative or positive as a whole but since the text and emoji are not calculated together we would not be able to discern this.

The sixth limitation is the failure of sentiment analysis to account for sarcasm. Sarcasm in general is difficult to analyze, for both humans and machines. In order to identify and understand sarcasm the context of the situation, cultural norms and topical information must be known (Maynard & Greenwood). This amount of information is

almost impossible for a machine to account for and then analyze. While algorithms have been created to detect high success rates of sarcasm as in the French company Spotter (Kleinman, 2013), the analysis used in this paper is not as robust in its analysis. Thus the polarity of a tweet's text might not be accurately depicting a user's sentiment.

The last limitation discussed in this paper that might differ between data sets is that there may exist in the text corpus words that are not recognized in the text lexicon used that may be of importance to this analysis such as balloonomania, nanity, questmonger, (Schott, 2016) words not found in the English dictionary such as braggadocious (Stack, 2016) or internet slang such as lol, gr8, jk or nsfw (Brown, 2014).

Research Design

Data

The data used for this project will be data I have previously collected via social media. More specifically this data consists of Tweets with specific hashtags collected from February 7th, 2016 to April 2nd, 2016 that referred to the presidential primaries that were taking place earlier this year. This data set was collected for QMSS's Data Processing and Data Visualization class (QMSS G4063) last semester. For more information on the data collected, see Appendix A. No experiments or surveys will be created and implemented, nor will any field research be conducted. Instead I have chose to use already available data for this project due to its low (as in non-existent) cost and because I have prior access to this large data set.

Approximately 303.62 MB of tweets were collected during this time period. Of the tweets that were collected, 1,816,475 were captured that contained geo-location data.

The geo-location attribute will be useful to filter out all tweets that were not sent from the United States. As this study is looking at American sentiment in both text and emoji making this assumption ensures a higher likelihood that the tweets we will be looking at are coming from Americans. Further analysis is needed to determine 1) how many tweets contain emojis and 2) what kinds of emojis are being used. The tweets used in this study were scraped via Twitter's Streaming API during a 56 day period and information regarding both meta data about the tweets as well as the tweet and user data were collected. A total of 50 unique variables associated for each tweet have been collected but this research will only look at a small subset of these variables. For the full list of variables collected for each tweet instance refer to Appendix B.

The variables that will be used in this study are unique identifiers for each tweet, usernames, tweet contents and locations of each tweet. The tweets were selected from Twitter's API based on their reference to one of any of the following candidates – Hillary Clinton, Bernie Sanders, Ted Cruz, Donald Trump, and Marco Rubio. Nicknames and references to particular candidates were also included such as Trumpf, Hillary and Cruz. Figure 1 below shows the full list of identifiers used to pull tweets from Twitter's API.

Figure 1 – Breakdown of Identifiers for each Candidate

Candidate's Full Name	Key words associated to locate Tweets
Hillary Clinton	Clinton, clinton, Hillary, hillary, Hillaryclinton, hillaryclinton, Hillary Clinton, hillary clinton
Bernie Sanders	Berniesanders, berniesanders, Bernie Sanders, bernie sanders, Bernie, Bernie, Sensanders, sensanders
Ted Cruz	Cruz, cruz, Ted, ted, Tedcruz, tedcruz, Ted Cruz, ted cruz
Donald Trump	Donaldtrump, donaldtrump, Donald Trump, donald trump, Trump, trump, Donald, Donald, Trumpf, trumpf
Marco Rubio	Marcorubio, marcorubio, Marco Rubio, marco

	rubio
--	-------

To identify the emojis currently in the dataset a full list of all available emoji will be used. This full list comes from the Unicode Consortium and will need to be scraped from their website which can be found in Appendix C. The reason that this list of emojis will be used is because this list is a complete list of all globally recognized emojis, which Twitter also uses. Choosing this globally recognized list guarantees that all emojis in the data set will be accounted for because users are unable to input emojis that are not recognized by the Unicode Consortium. There are a total of 2,389 recognized emojis within this list. As described below in further detail in the sentiment analysis section, the lexicon for emoji sentiment is not as complete as this full list so there may be emojis in the data set that do not have a sentiment score.

Sentiment Analysis

Text

As defined by Taboada *et al.* sentiment analysis refers to a method of extracting subjectivity and polarity from text (2011). The polarity of the text is on a scale of positivity, neutrality or negativity. Mathematically represented by:

$$c \in \{-1, 0, +1\}$$

While there are other scales on which to measure polarity for text (e.g. scales ranging from -5 to 5), this one will be used, as it is the same scale used in the polarity lexicon for the emojis. Two main methods of sentiment analysis exist – the lexicon-based approach and the text classification approach (Pang, Lee, and Vaithyanathan 2002). This study

utilizes the lexicon-based approach and calculates the orientation of text from the semantic orientation of the words in the tweet (Turney 2002). To do this, the lexicon developed by Liu and Hu that contains 4,818 negative sentiment words and 2,041 positive sentiment words (Hu and Liu, 2004) will be used as a basis on which to calculate text sentiment. See Appendix D for more information on this lexicon. Previous data analysis on this data set conducted last semester calculated both the sentiment score and sentiment ratio for each tweet. The sentiment score was the combination of all of the words in a tweet and their positive or negative score. The sentiment ratio score was calculated as:

$$(Positive\ Score + Negative\ Score) / Total\ Score.$$

This equation provides a ratio of scores between -1 and 1 that predicts the sentiment of tweets from very negative to very positive. The text data was found to be predominantly negative. This could be due to a general notion of negativity online that is well described by various authors and researchers (Wakefield, 2015). For the benefit of analyses the score was normalized around the zero mean and was scaled for the minimum and maximum of -1 and 1. For more information on the sentiment analysis already conducted on the text data see Appendix E. Note that previous analysis was done on an individual candidate level. For this study, sentiment will be calculated at a political party level. Additionally, a different method of calculating sentiment will be used and is discussed in the Data Processing section below.

Emoji

To calculate the sentiment of emojis in the data set the method that will be used is the same as the method for the text sentiment analysis and is described below in the

Data Processing section. The difference is that the lexicon will be the emoji lexicon developed by Novak *et al.*, which contains 751 emojis that occurred at least 5 times in their data set of 70,000 tweets with emojis. (2015) These emojis were cross-referenced with emojiTracker, a website that monitors in real-time the use of emojis on Twitter, in June of 2015 (“EmojiTracker”, n.d.). The study then looked at the Pearson correlation for emojis with $N \geq 5$ and determined that they were highly significant at the 1% level, confirming that the list of emojis chosen for this lexicon were representative of their general use on Twitter’s platform. More information on the emoji lexicon can be found in Appendix F.

Data Processing

In order to conduct sentiment analysis on both the text and emoji in the data set the following steps must be taken. Re-tweets and repetitive tweets must be removed from the data set. If they are not removed then the likelihood that they will skew the results is increased. In order to determine the sentiment for each tweet’s emojis a dictionary must be created from the Unicode’s full list of emoji and from the lexicon described in the sentiment analysis section. This will most likely need to be scraped from the web and then saved in a format that R can process. Once this is done then each lexicon (text and emoji) can be used to analyze the individual tweets and compute polarity scores for both the text and emoji.

Lastly, in order to finalize the data processing stage I will need to modify the data I have by normalizing the sentiment scores of the text and emoji data sets in order to compare the results I find with each data set. Note that for conducting sentiment analysis each tweet is separated into two distinct data sets – one for text and one for emoji. They

are not analyzed as a single entity. To normalize my text and emoji data – and thus be able to compare and contrast results between the two various means of communication – three methods are discussed below with their advantages and disadvantages on how to convert counts of positive and negative words and emojis into percentages. The first method is the absolute proportional difference method, which looks at how many positive and negative words or emojis exist in a tweet divided by the total number of text or emoji existing in the tweet (Lowe, Benoit, Mikhaylov and Laver, 2011). This score ranges from 0 to 1. Mathematically this method is calculated by:

$$Sentiment = \frac{Positive - Negative}{Total Number}$$

The disadvantage of using this method is that a tweet's score can be heavily affected by non-sentiment-related content. For example if there are words that are not given a polarity score in the lexicons used for the analysis they could skew the polarity of the tweet. The same issue can exist for calculating sentiment with emoji since the lexicon only contains 751 emojis out of 2,389. The second method is the relative proportional difference method that calculates a score that ranges from -1 to 1 (2011). Mathematically it is calculated by:

$$Sentiment = \frac{(Positive - Negative)}{Positive + Negative}$$

Here sentiment is calculated by only using words recognized as sentiment in the lexicons used for this analysis. The disadvantage of using this method is that a sentence's score may tend to strongly cluster and the end points of the scale given that the tweet content may be primarily or exclusively positive or negative. The third method is the logit scale

method, which can range from negative to positive infinity. Mathematically this can be calculated by:

$$Sentiment = \log(Positive + 0.5) - \log(Negative + 0.5)$$

Where the 0.5 helps smooth the results and prevents $\log(0)$ (or an undefined result) from occurring. The benefit of this method is that it is symmetric around zero and when compared to the other two methods has the smoothest properties (2011). After conducting all three analysis it is most likely that the third method of analysis will be used because the logit model focuses on the proportional changes on a symmetrical positive-negative scale.

Once each tweet has a sentiment score calculated for both text and emoji, a cross-tabulation table that looks at the text sentiment by emoji sentiment would be created like the example below in Figure 2. Additionally, a variety of descriptive statistics will be calculated to look at the average sentiment scores among the text data as well as among the emoji data and the difference between text and emoji sentiment scores for each tweet. This section of analysis would need to further filter tweets that would contain both text and emoji in a single tweet in order to test the first hypothesis that there will be a significant difference between sentiment analyses of text compared to emojis.

Figure 2: Text Sentiment by Emoji Sentiment

Emoji Sentiment	Text Sentiment		
	Negative	Neutral	Positive
Negative	%	%	%
Neutral	%	%	%
Positive	%	%	%
Total	100%	100%	100%
	(#)	(#)	(#)

To test the second and third hypotheses both the text and emoji data sets will need to be subset into two groups – those that contain tweets references Republican candidates and those that contain Democratic candidates. From there the distribution of emoji and text sentiment can be analyzed for both the Republican and Democratic groups.

Conclusion

The main purpose of this project is not the data itself but more the relationship between text sentiment and emoji sentiment. Described above were the reasons for this subject topic as well as the various methods and processing that will need to be done to conduct this analysis. For further study, it would be interesting to see the polarity of tweets between Democrats and Republicans. Since this information is not available via Twitter, this study would look at states with the highest proportion of voters who voted for Donald Trump, the Republican nominee, and compare Twitter sentiment with states with the highest proportion of voters who voted for Hillary Clinton, the Democratic nominee. Beyond the scope of this data set in the future I would like to expand this research by looking at other larger, and ideally larger data sets of social media or mobile communication data to see whether sentiment differences follow the same patterns between both modes of communication or if patterns are due to the nature of the data itself. The next step would be to obtain multiple large data sources and conduct the same analysis with them to see whether my first hypothesis is correct.

Appendices

Appendix A – QMSS G4063 Information

Below are links to Data Processing and Data Visualization's github page as well as a link to the original JSON data containing all tweets scraped.

- https://github.com/hassanpour/QMSS_G4063
- https://www.dropbox.com/sh/zyy9tsvibr14d63/AAQ6D3h0Kksxb8EeVH2RSSAa/tweets_geo_all.json?dl=0#

Appendix B – All variables associated with a single tweet

```
> colnames(tweetsUS)
[1] "X.1"           "id_str"
[3] "idx"           "text"
[5] "created_at"    "screen_name"
[7] "user_lang"     "truncated"
[9] "retweeted"     "favorite_count"
[11] "verified"      "user_id_str"
[13] "source"        "followers_count"
[15] "in_reply_to_screen_name" "location"
[17] "retweet_count" "favorited"
[19] "utc_offset"    "statuses_count"
[21] "description"   "friends_count"
[23] "user_url"      "geo_enabled"
[25] "in_reply_to_user_id_str" "lang"
[27] "user_created_at" "favourites_count"
[29] "name"          "time_zone"
[31] "in_reply_to_status_id_str" "protected"
[33] "listed_count"  "place_lon"
[35] "expanded_url"  "place_id"
[37] "full_name"     "lat"
[39] "country_code"  "place_name"
[41] "url"           "country"
[43] "lon"           "place_type"
[45] "place_lat"     "X"
[47] "Y"             "STATEFP"
[49] "NAME"          "COUNT"
```

Appendix C – Unicode List of Emoji






Below is a screenshot of the Unicode Consortium's list of all available emojis. To view more, visit:

<http://unicode.org/emoji/charts/emoji-list.html>

Emoji Data, v4.0

This chart provides a list of the Unicode emoji characters, with single image and annotations. (See also the [full list](#).) The ordering of the emoji and the annotations are based on [Unicode CLDR data](#). Emoji sequences have more than one code point in the **Code** column.

For information about the images used in these charts, see [Emoji Images and Rights](#). For details about the format and fields, see [Emoji Chart Index](#). Support of emoji is not required for conformance to the Unicode Standard — for more information about emoji, see [UTR #51 Unicode Emoji](#). To propose a new emoji, see [Submitting Emoji Character Proposals](#).

No	Code	Browser	Sample Name	Keywords
1	U+1F600		 grinning face	face grin
2	U+1F601		 grinning face with smiling eyes	eye face grin smile
3	U+1F602		 face with tears of joy	face joy laugh tear
4	U+1F923		 rolling on the floor laughing	face floor laugh rolling
5	U+1F603		 smiling face with open mouth	face mouth open smile
6	U+1F604		 smiling face with open mouth & smiling eyes	eye face mouth open smile
7	U+1F605		 smiling face with open mouth & cold sweat	cold face open smile sweat
8	U+1F606		 smiling face with open mouth & closed eyes	face laugh mouth open satisfied smile
9	U+1F609		 winking face	face wink
10	U+1F60A		 smiling face with smiling eyes	blush eye face smile

Appendix D – Hu and Liu’s Lexicon for Text Analysis

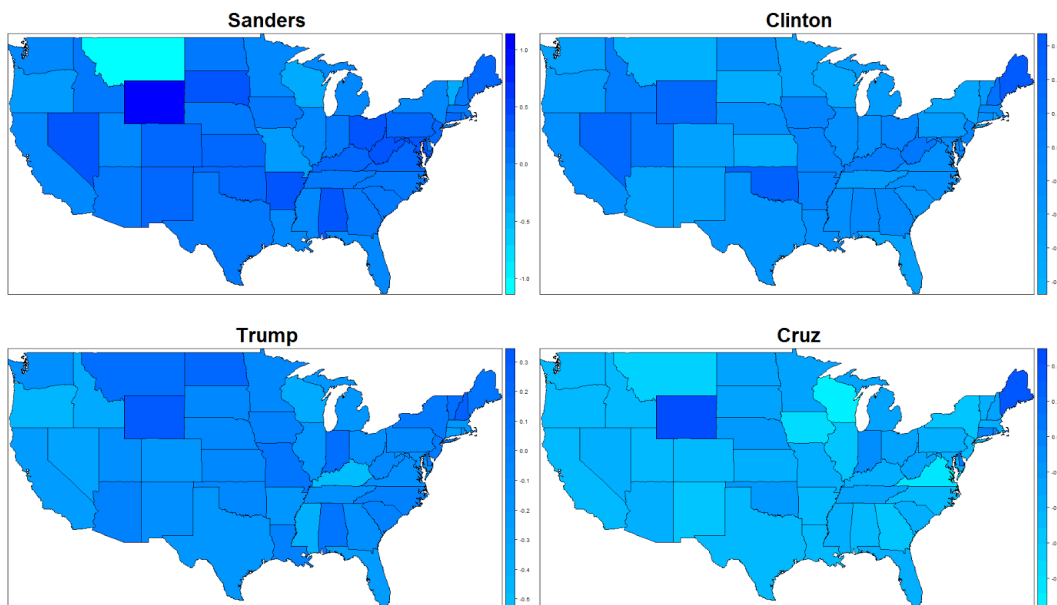
All files and papers can be downloaded via this link:

<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

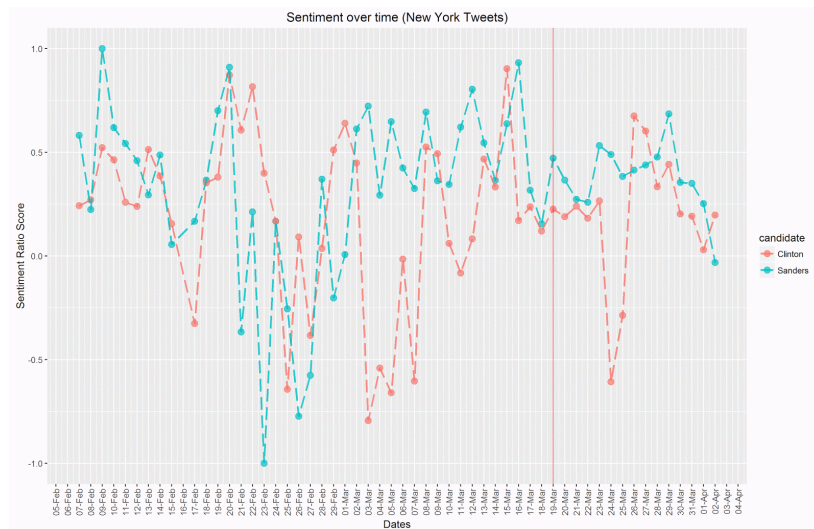
Access to the positive and negative lexicons can be found on the following github page:

<https://github.com/mjhea0/twitter-sentiment-analysis/tree/master/wordbanks>

Appendix E – Text Sentiment Analysis Visualizations



The above maps show the sentiment (degree of positivity) of tweets disaggregated by candidates.



The above visualization shows sentiment scores over time between both Democratic candidates in the state of New York.

Final report from last semester's QMSS G4063 class :

<https://docs.google.com/document/d/1le66GaGXa4XwhVyuRrHOG2oMvraBjWU7apfi-4e4tfl/edit>

Appendix F – Emoji Sentiment Lexicon

Below is a screenshot of the information found on

http://kt.ijs.si/data/Emoji_sentiment_ranking/.

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
😂		0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
❤️		0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
♥️		0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
😍		0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
😭		0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
😘		0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons

References

- Author Translates All of 'Alice in Wonderland' into Emojis | The Creators Project. (n.d.). Retrieved December 02, 2016, from <https://thecreatorsproject.vice.com/blog/author-translates-all-of-alice-in-wonderland-into-emojis>
- Boia, M., Faltings, B., Musat, C. C., and Pu, P. 2013. A :) is worth a thousand words: now people attach sentiment to emoticons and words in Tweets. In Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM2013. 345-350.
- Brown, L. (2014). Twitter - 30 Must-Know Twitter Abbreviations and Acronyms. Retrieved November 25, 2016, from <http://marketing.wtwhmedia.com/30-must-know-twitter-abbreviations-and-acronyms/>
- Chinn, D., Zappone, A., & Zhao, J. (2016). Analyzing Twitter Sentiment of the 2016 Presidential Candidates. Retrieved November 18, 2016, from <https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf>
- Doble, A. (2015, May 19). UK's fastest growing language is... emoji. Retrieved December 1, 2016, from <http://www.bbc.co.uk/newsbeat/article/32793732/uks-fastest-growing-language-is-emoji>
- Duggan, M. (2015, August 19). The Demographics of Social Media Users. Retrieved December 02, 2016, from <http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/>
- Emoji Dick;. (n.d.). Retrieved October 16, 2016, from <http://www.emojidick.com/>
- Emogi Research Team. (2015). Emoji Report. Retrieved November 21, 2016, http://emogi.com/documents/Emoji_Report_2015.pdf
- Emoji Sentiment Ranking v1.0. (n.d.). Retrieved November 02, 2016, from http://kt.ijs.si/data/Emoji_sentiment_ranking/
- Emojitracker: Realtime emoji use on twitter. (n.d.). Retrieved November 17, 2016, from <http://www.emojitracker.com/>
- Foran, C. (2016, May 5). The 'Never Clinton' Campaign. Retrieved December 01, 2016, from <https://www.theatlantic.com/politics/archive/2016/05/hillary-clinton-bernie-sanders/481389/>
- Grady, S. (2016, May 13). How can emoji analysis improve results in social analytics and social listening tools? Retrieved November 12, 2016, from <https://blog.rocketsoftware.com/2016/05/can-social-analytics-social-listening-tools-measure-quantify-emoji/>
- Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., and Caldara, R. 2009. Cultural confusions show that facial expressions are not universal. *Current Biology* 19, 18 (2009), 1543-1548
- Hu, M., and Liu, B. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA

Knapp, M., and Hall, J. A. 2010. *Nonverbal communication in human interaction* (7th ed.). Cengage Learning

Kleinman, Z. (2013). Authorities 'use analytics tool that recognises sarcasm' Retrieved December 02, 2016, from <http://www.bbc.com/news/technology-23160583>

Liu, Bing. "Sentiment Analysis and Subjectivity." An chapter in *Handbook of Natural Language Processing*, Second Edition, (editors: N. Indurkha and F. J. Damerau), 2010.

Lowe, William. Benoit, Kenneth. Mikhaylov, Slava. and Laver, Michael. (2011) "Scaling Policy Preferences From Coded Political Texts." *Legislative Studies Quarterly* 26(1, Feb): 123-155.

Lu, X., Ai, W., Liu, X., Li, Q., Wang, N., Huang, G., and Mei, Q. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 770-780. DOI: <http://dx.doi.org/10.1145/2971648.2971724>

Maynard, D., & Greenwood, M. A. (n.d.). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. Retrieved from <https://gate.ac.uk/sale/lrec2014/arcomem/sarcasm.pdf>

Miller, H., Thebault-Spieker, J., Chang, S., Terveen, L. and Hecht, B. "Blissfully Happy" or "Ready to Fight": Varying Interpretations in Emoji. the 10th Conference on Web and Social Media, (2016).

Novak, K., Smailović, P., Sluban, J., Mozetič, B. (2015) Sentiment of Emojis. *PLoS ONE* 10(12): e0144296. doi:10.1371/journal.pone.0144296

Pang, B., Lee, L., and Vaithyanathan. S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in NLP*, pages 79–86, Philadelphia, PA.

Park, J., Barash, V., Fink, C., and Cha, M.. 2013. Emoticon style: interpreting differences in emoticons across cultures. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013*.

Schott, B. (2016). Ben Schott on the Strange New Vocabulary of the 2016 Election. Retrieved December 02, 2016, from <http://www.townandcountrymag.com/society/politics/news/a7989/ben-schott-clinton-trump-vocabulary/>

Shafer, J. (2015, August 13). Donald Trump Talks Like a Third-Grader. Retrieved December 01, 2016, from <http://www.politico.com/magazine/story/2015/08/donald-trump-talks-like-a-third-grader-121340>

Smith, A. (2013, April 25). Civic Engagement in the Digital Age. Retrieved November 30, 2016, from <http://www.pewinternet.org/2013/04/25/civic-engagement-in-the-digital-age/>

Stack, L. (2016). 'Braggadocious?' In Trump and Clinton Debate, Words That Sent Viewers to the Dictionary. Retrieved December 02, 2016, from

<http://www.nytimes.com/2016/09/28/us/braggadocio-in-trump-and-clinton-debate-words-that-sent-viewers-to-the-dictionary.html>

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267-307. doi:10.1162/coli_a_00049

Tigwell, G. W., and Flatla, D. R.. 2016. 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)*. ACM, New York, NY, USA, 859-866. DOI: <http://dx.doi.org/10.1145/2957265.2961844>

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, PA.

Wakefield, J. (2015, March 26). Why are people so mean to each other online? Retrieved April 20, 2016, from <http://www.bbc.com/news/technology-31749753>

Walther, J. and D'Addario, K. P.. 2003. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 5. 2(2003), 119-134

Yan, H., Sgueglia, K., & Walker, K. (2016, November 29). 'Make America White Again': Hate speech and crimes post-election. Retrieved December 01, 2016, from <http://www.cnn.com/2016/11/10/us/post-election-hate-crimes-and-fears-trnd/>