All the Feels :

Sentiment Analysis

Between Emoji and Text

Alexandra Plassaras

Columbia University

QMSS G5999:

Master’s Thesis

Table of Contents

Introduction 3

Research Problem 4

Motivation 4

Hypotheses 5

Previous Research 7

Research Design 8

Data 8

Sentiment Analysis 11

Text 11

Emoji 12

Data Processing 12

Results 15

Overall 15

Model 1 17

Model 2 22

Model 3? 26

Comparison of Models 27

Discussion 30

Limitations 30

Further Study 33

Conclusion 33

Appendices 35

Appendix A – QMSS G4063 Information 35

Appendix B – All variables associated with a single tweet 35

Appendix C – Unicode List of Emoji 35

Appendix D – Hu and Liu’s Lexicon for Text Analysis 36

Appendix E – Text Sentiment Analysis Visualizations 36

Appendix F – Emoji Sentiment Lexicon 37

Appendix G – Additional Visualizations 38

References 40

# Introduction

Macintosh HD:Users:alexandraplassaras:Desktop:Screen Shot 2016-12-02 at 9.42.49 PM.pngandMacintosh HD:Users:alexandraplassaras:Desktop:Screen Shot 2016-12-02 at 9.42.57 PM.png. 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, punctuation 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 and how this compares to text sentiments.

# Research Problem

## 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). This paper is 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 used 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 this paper argues 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 mentioned above this paper hypothesizes that people discussing Republican candidates who use emojis are more likely to be young users who might be 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. This paper hypothesizes 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 this paper is 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](http://lastnighton.com/2015/08/02/jimmy-kimmel-spotlights-donald-trumps-love-of-the-word-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 with 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.

Previous work done by Novak *et al.* significantly affects the scope of this project because their study created the first known open-source 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.

# Research Design

## Data

The data used for this project 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. 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.

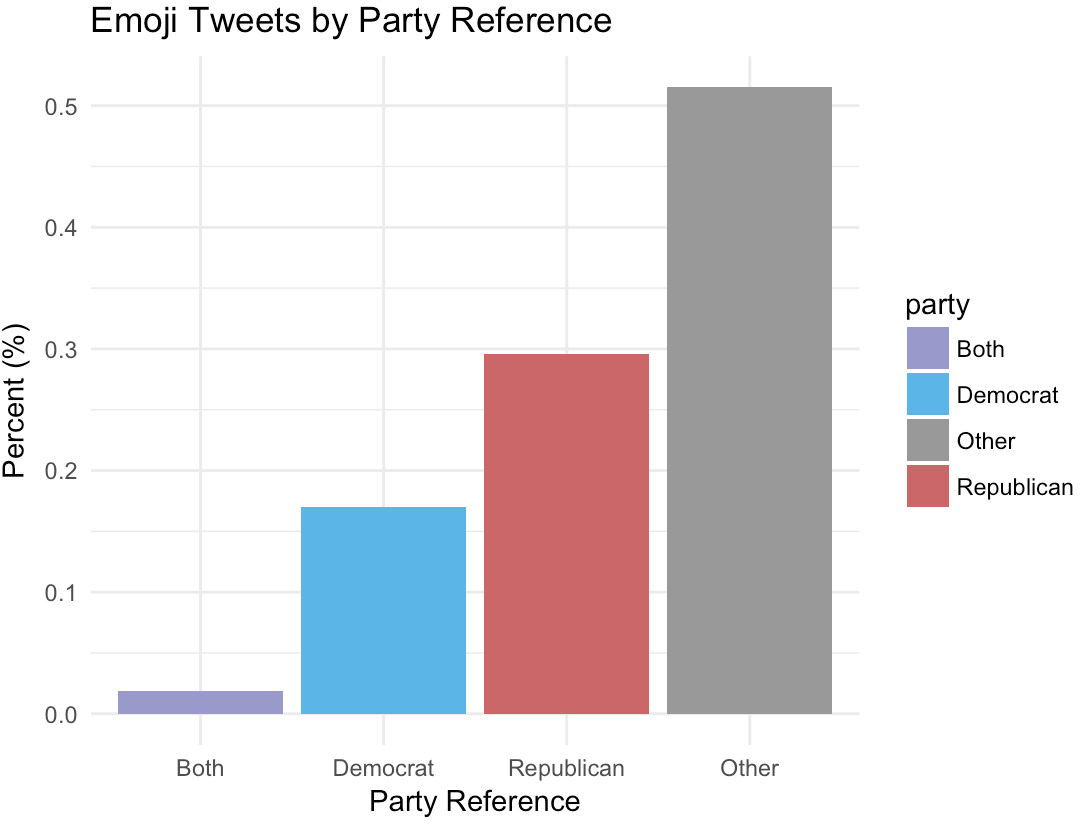
The tweets used in this study were scraped via Twitter’s Streaming API during a 56 day period and information regarding both meta data about the tweets as well as the tweet and user data were collected. Of the initial 1,816,475 Tweets in the data set, only 86,258 (~4.75%) contained emojis. A total of 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 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. For more information on the data collected, see Appendix A.

|  |  |
| --- | --- |
| **Candidate’s Full Name** | **Key words associated to locate Tweets** |
| Hillary Clinton | Clinton, clinton, Hillary, hillary, Hillaryclinton, hillaryclinton, Hillary Clinton, hillary Clinton, HillaryClinton |
| 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, realDonaldTrump |
| Marco Rubio | Marcorubio, marcorubio, Marco Rubio, marco rubio |

**Figure 1 –** Breakdown of Identifiers for each Candidate

Furthermore only 41,834 (~2.30% of the original data set) contained references to the Democratic or Republican party. As two out of the three hypotheses for this analysis deal with sentiment across political parties, the following results are based on the 41,834 tweets that contain emojis and references to a political party. Figure 2 below shows the distribution of tweets as categorized into political parties. The largest group of tweets are those classified as “other” at 51.5% meaning that they were not classified into any political party. The second largest group was the Republican party at 29.6% whereas the Democratic party consisted of 17% of all tweets. A final fourth category contained tweets referencing both Democrat and Republican parties at 1.9%. The other category being the largest category is not particularly surprising given the limitations of subsetting tweets using key words like the primary candidates and the parties because there may be other references to a particular party that are more nuanced and difficult for an algorithm to decipher. What is surprising is that there are more tweets with emojis referencing the Republican party than there are the Democratic party. This may be because of the use of sarcasm towards Republicans or it might mean that people who Tweet about the Republican might be more emoji-literate. The further study section will discuss ways to improve the classification of tweets into either democratic or republican parties.



**Figure 2** – Emoji Tweets by Party Reference

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 which was scraped from their website. More info on the Unicode Consortium 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:

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 Finn Årup Nielsen called Afinn which contains 2477 words will be used as a basis on which to calculate text sentiment.[[1]](#footnote-1) With this lexicon words were given a scale of polarity of . By reducing the range to by dividing every sentiment score by 5, the range becomes the same range that the emoji lexicon follows – more on the emoji lexicon below.

Another popular lexicon that was considered in this study was Liu and Hu’s lexicon which contained 6,859 words. This lexicon focused on positive/negative polarity and identified only 24,457 unique words from the tweet corpus used in this study while the Afinn lexicon identified 26,026 unique words. Thus Liu’s lexicon was not used in this study. Instead the Afinn lexicon was chosen as it focused on sentiment strength instead of positive/negative polarity. Additionally the Afinn lexicon was designed for sentiment analysis of microblogging media platforms such as Twitter and is therefore an appropriate lexicon for use in this study. As a result of focusing in microblogging sites the lexicon has added polarity scores to common phrases found on the internet such as “lol” (laughing out loud) as well as strong curse words into it’s lexicon.[[2]](#footnote-2)

The Afinn lexicon was originally created in 2009 to look at sentiment analysis of tweets focused on the United Nation Climate Conference (COP15).[[3]](#footnote-3) The original lexicon was referred to as AFINN-96 contained 1,486 unique words with a small amount of phrases. A second version was released in 2011 that contained 2,477 unique words as well as additional phrases not included in the original release.[[4]](#footnote-4) Some notable additions to the lexicon that were not included in other lexicons include “lol”, “wtf” and “rofl”. The sentiment strength for each word in the Afinn lexicon was calculated by taking the sum of the word polarity divided by the number of words represented.

### Emoji

The emoji lexicon used in this study was developed by Novak *et al.* andcontains 751 unique 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 >= 5 and determined that they were highly significant at the 1% level, confirming that the list of emojis chosen for this lexicon was representative of the emoji’s 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 were taken. Re-tweets and repetitive tweets were removed from the data set. If not removed the likelihood that they would skew the results would have increased. In order to determine the sentiment for each tweet’s emojis a dictionary was created which identified each emoji character to it’s unique identifier and it’s name. Multiple open source emoji dictionaries such as twitterEmojiProject[[5]](#footnote-5), Prismoji’s emoji tutorial[[6]](#footnote-6) and Twimoji[[7]](#footnote-7) were used to create a more extensive emoji list.

**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:**

**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:**

**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:**

**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.**

# Results

## Overall

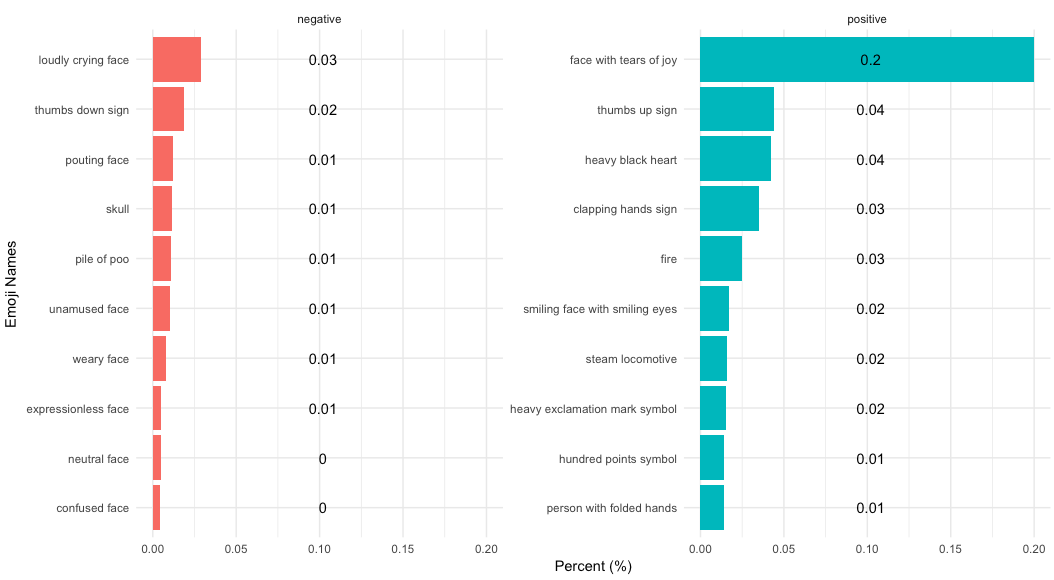
The models that follow contain smaller subsets of the 41,834 tweets that were categories as either referencing the Republican, Democratic or both parties because the text lexicons used in each model identified sentiment for different words and thus the number of tweets with sentiment differed among models.

For this model 41,834 tweets were identified as having polarized emojis consisting of 658 unique emoji characters.

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

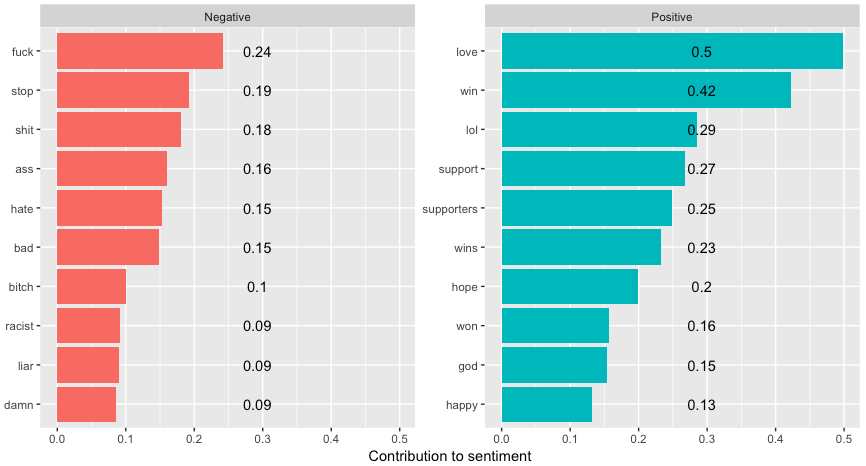
**Figure 2:** Text Sentiment by Emoji Sentiment

Using this somewhat limited range for sentiment Figure 4 shows that the emoji named “face with tears of joy” was the most commonly used emoji in tweets with some sort of party affiliation (20%). The remaining top five emojis used in tweets are the “thumbs up sign” (4.4%), the “heavy black heart” (4.2%), the “clapping hands sign” (3.5%) and the “loudly crying face” (2.9%). Within the top five most used emojis four out of five emojis were polarized as having positive sentiment while the “loudly crying face” emoji was polarized as having negative sentiment. Note that there is also a large gap of 15.6 percentage points between the most used emoji and the second most used emoji.



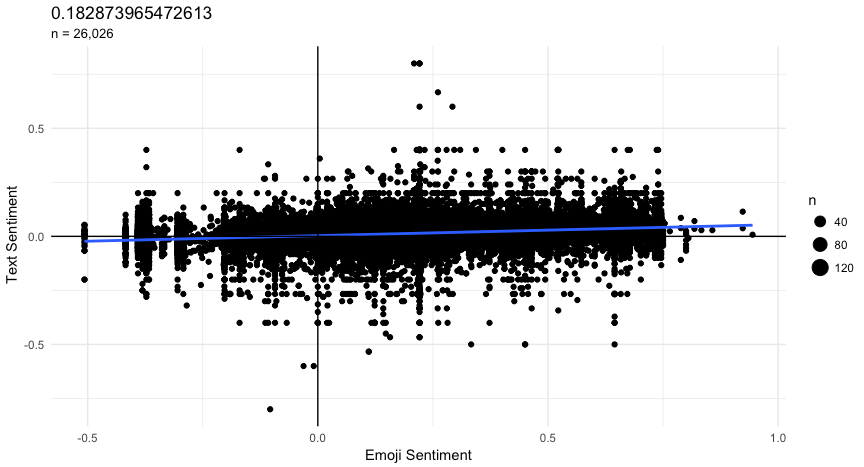
**Figure 4** – Top 20 Emojis used within Tweets with Party References by Polarity

When looking at the text results we see that the top five most used words were “love” at 50%, followed by “win” at 42%, “lol” at 29%, “support” at 27% and “supporters” at 25% as shown in Figure 10. Since the last two words which have the same stem were not combined, it is one of potentially many examples were stemming in inappropriate or ineffective for natural language processing. Additionally, note the presence of “lol” in the Afinn dictionary with correctly polarized sentiment is potentially a sign of how valuable this lexicon is over other lexicons that do not include such words in their analysis. Had another lexicon been used the third most commonly used word would not have been identified. Looking at the top most used words all five contain positive sentiment. It is important to note that the sixth most common word, “fuck” with 24%, has a negative polarity.



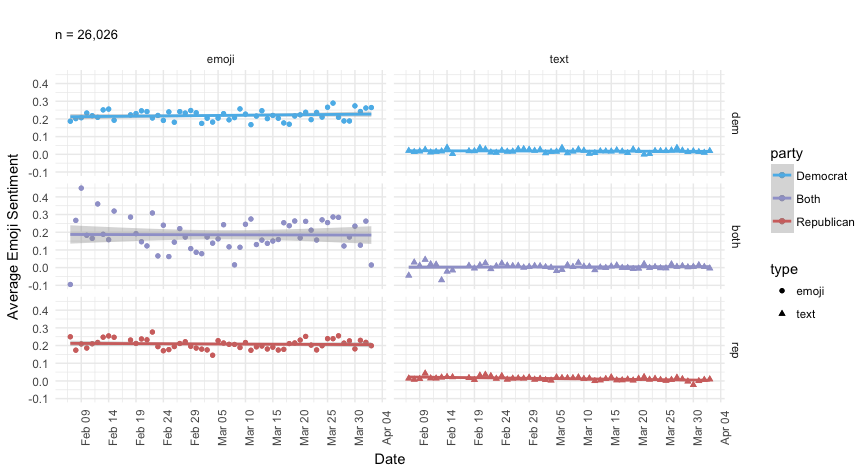
**Figure 10** – Top 20 Words used within Tweets with Party References by Polarity

Figure 11 looks at the correlation between emoji and text sentiment. There is to be a correlation of 0.183 between text and emoji sentiment. While this correlation is quite weak it is important to note that this does not prove or disprove any causal relationships between text and emoji sentiment. In order to make claims of statistically significant correlation a regression model would need to be implemented.



**Figure 11** – Sentiment of Text vs. Emoji

Looking at the average sentiment of text and emoji over time by political party, Figure 12 below shows that emoji sentiment remains, on average, positive over time by political party reference. There appears to be a slight increase in positive sentiment over time for tweets referencing democrats compared to the slight increase in negative sentiment over time for tweets referencing republicans. When looking at text sentiment over time the average sentiment remains around zero for all three political party reference groups.



**Figure 12** – Sentiment of Text and Emoji over Time by Political Party

Digging deeper in the distribution of tweets by type of character we see that the mean score for text sentiment is 0.014 whereas emoji sentiment is at 0.206. Figure 13 below shows that for emoji sentiment, 81.2% of tweets were categorized as having positive polarity, 17.5% for negative polarity and only 1.3% for neutral polarity. For text sentiment 60% of tweets were categorized as having positive polarity, 35.9% for negative polarity and only 4.1% for neutral polarity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Positive %** | **Neutral %** | **Negative %** | **Total %** |
| **Emoji** | 81.2 | 1.3 | 17.5 | **100** |
| **Text** | 60 | 4.1 | 35.9 | **100** |

**Figure 13** – Polarity of All Tweets by Type of Character

Figure 14 shows the polarity of tweets by political party reference and type of characters used. There are more positive emojis in tweets referring to Republicans (49.9%) compared to positive tweets referring to Democrats (27.9%). Within the text sentiment we see that there are more positive tweets referring to Republicans (35.7%) followed by the second highest category, which was negative sentiment for Republicans. For additional information about the distribution of tweets refer to Appendix G Figures 3.a and 3.b which show a right skew to the distribution of emoji sentiment and what appears to be a normal distribution for text sentiment.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Emoji** | **Text** |
| **Party** | **Polarity** | **Percent** | **Percent** |
| **Democrat** | **+** | 27.9 | 21.9 |
|  | **0** | 0.3 | 1.3 |
|  | **-** | 6 | 11 |
| **Both** | **+** | 3.4 | 2.5 |
|  | **0** | 0.1 | 0.2 |
|  | **-** | 0.9 | 1.7 |
| **Republican** | **+** | 49.9 | 35.7 |
|  | **0** | 0.9 | 2.6 |
|  | **-** | 10.6 | 23.2 |
| **Total** |  | **100** | **100** |

**Figure 14** – Polarity of Tweets by Political Party and Type of Character

## (ATTEMPT AT) Model 3

In Model 3 the text lexicon was modified to include n-grams to account for double negatives such as “not angry” or “wasn’t horrible”. As a first attempt only bi-grams were attempted to see how many bi-grams could be successfully polarized given the Afinn lexicon. Classifying the text corpus into bi-grams resulted in 215,087 unique bi-grams. Figure 15 shows that the most common bi-gram is “donald trump” at 24.84% followed by “bernie sanders” at 0.84%. Note that the top ten bi-grams are either references to the individual people running in the presidential primaries or they contain either one or two common stop words typically removed when dealing with individual words. In the further study section further discussion on how or whether to combine both processing techniques will be discussed.

|  |  |
| --- | --- |
| Word | Percent |
| donald trump | 24.84 |
| bernie sanders | 0.84 |
| ted cruz | 0.56 |
| vote for | 0.5 |
| is a | 0.44 |
| trump is | 0.44 |
| in the | 0.44 |
| hillary clinton | 0.42 |
| of the | 0.33 |
| is the | 0.33 |

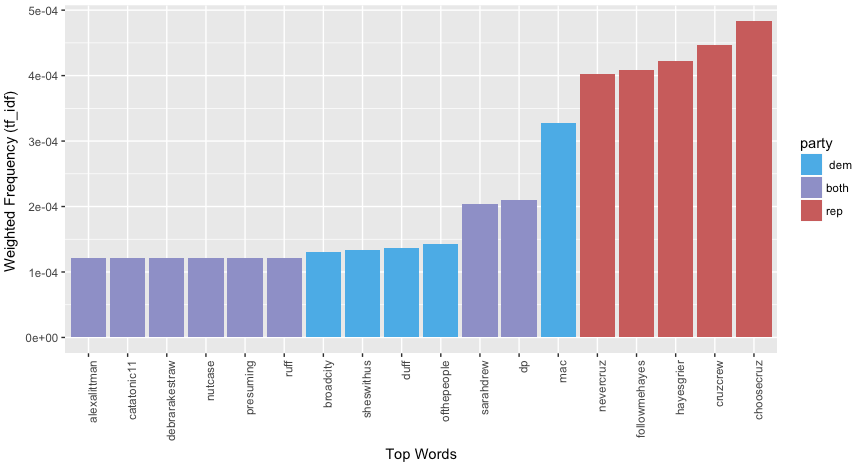
**Figure 15** – Model 3: Top Ten Most Common Bi-Grams

When an attempt to polarize the 215,087 bi-grams was made only nine bi-grams were successfully polarized as seen in Figure 16. Given that only 0.0041844% of bi-grams were polarized it does is not feasible to include polarized bi-grams into this model. Next steps will be discussed in the further study section.

|  |  |  |
| --- | --- | --- |
| Word | Score | Count |
| can’t stand | -0.6 | 32 |
| fed up | -0.6 | 26 |
| not good | -0.4 | 16 |
| not working | -0.6 | 6 |
| screwed up | -0.6 | 2 |
| some kind | 0.0 | 2 |
| dont like | -0.4 | 2 |
| no fun | -0.6 | 1 |
| right direction | 0.6 | 1 |

**Figure 16** – Model 3: Polarized Bi-Grams

As a next attempt, Model 3 attempted to calculate text with a weighted frequency by looking at an individual word’s inverse document frequency (idf), which “decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.” (Silge, Ch. 3) By combining this with term frequency to create *tf-idf*, Model 3 attempts to adjust for how rarely a word is used within the realm of all tweets used in this stage of analysis, 26,026. As *tf-idf* measures how important a word is to a document in a corpus of documents, Model 3 attempted first to treat each tweet as a single document and all tweets as the document corpus. The result was that the diversity within a single document, or tweet, was so low that every word was given a high *tf-idf* score. In the second attempt, the entire tweet corpus was used and the result was that all of the terms were given *tf-idf* scores that were so small the results were displayed as zero for every word. As a third attempt the tweets were categorized into political party reference groups, which provided more reasonable *tf-idf* results as shown in Figure 17. The top five most frequently weighted words are all from Republican references – “choosecruz”, “cruzcrew”, “hayesgrier”, “followmheyes” and “nevercruz”.



**Figure 17** – Model 3: Top Weighted Frequency of Words by Political Party

After calculated adjusted weight frequencies for the text corpus attempts were made to integrate this information with Model 2’s Afinn lexicon. And it is here that I have run into a roadblock.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| # of tweets (polarized emoji) | 41,834 | 41,834 |  |
| # of tweets (polarized text) | 24,457 | 26,026 |  |
| Unique emojis | 658 | 658 | 658 |
| Unique words | 2,348 | 1,453 |  |
| # of comparison tweets | 24,457 | 26,026 |  |
| Sentiment Correlation | 0.111 | 0.183 |  |
| Average Emoji Sentiment | 0.477 | 0.206 |  |
| Average Text Sentiment | 0.016 | 0.014 |  |

**Figure 18** – Comparison of Models

# 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, the use of non-standard words used during this campaign cycle and the accuracy of the text lexicon. 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.

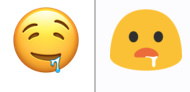
The second limitation of this study is that the lexicon used to analyze emoji sentiment was created for a separate study that looked at the sentiment of emojis in the UK. Jack *et al.* suggests that people’s interpretation of facial expressions and thus emojis differ between cultures (Jack, Blais, Scheepers, 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 for 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 led 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



iOS (left), Android (right)

**Figure 3** – “Drooling face” Emoji

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. **Perhaps using this study we can begin using the difference between text and emoji as a way to capture sarcasm.**

**The seventh limitation 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 gr8, jk or nsfw (Brown, 2014). The Afinn lexicon does include some slang words as well as curse words but this lexicon was created in 2011 so more recent internet slang may not be accounted for.**

The last limitation discussed in this paper is the accuracy of the current text lexicon used in this study.

Afinn lexicon limitation

The word list have a bias towards negative words (1598, corresponding to 65%) compared to positive words (878).[[8]](#footnote-8)

## Further Study

1. classification algorithm for political parties.
2. Figure out how to remove stop words from bi-grams or not remove stop words at all…
3. Given that only 0.0041844% of bi-grams were polarized it does is not feasible to include polarized bi-grams into this model.

# 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 were used to conduct this analysis. For further study, it would be interesting to see the polarity of tweets between identified Democrats and Republicans instead of political party reference.

**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/zyy9tsvibrl4d63/AAAQ6D3h0Kksxb8EeVH2RSSAa/tweets\_geo\_all.json?dl=0#](https://www.dropbox.com/sh/zyy9tsvibrl4d63/AAAQ6D3h0Kksxb8EeVH2RSSAa/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"

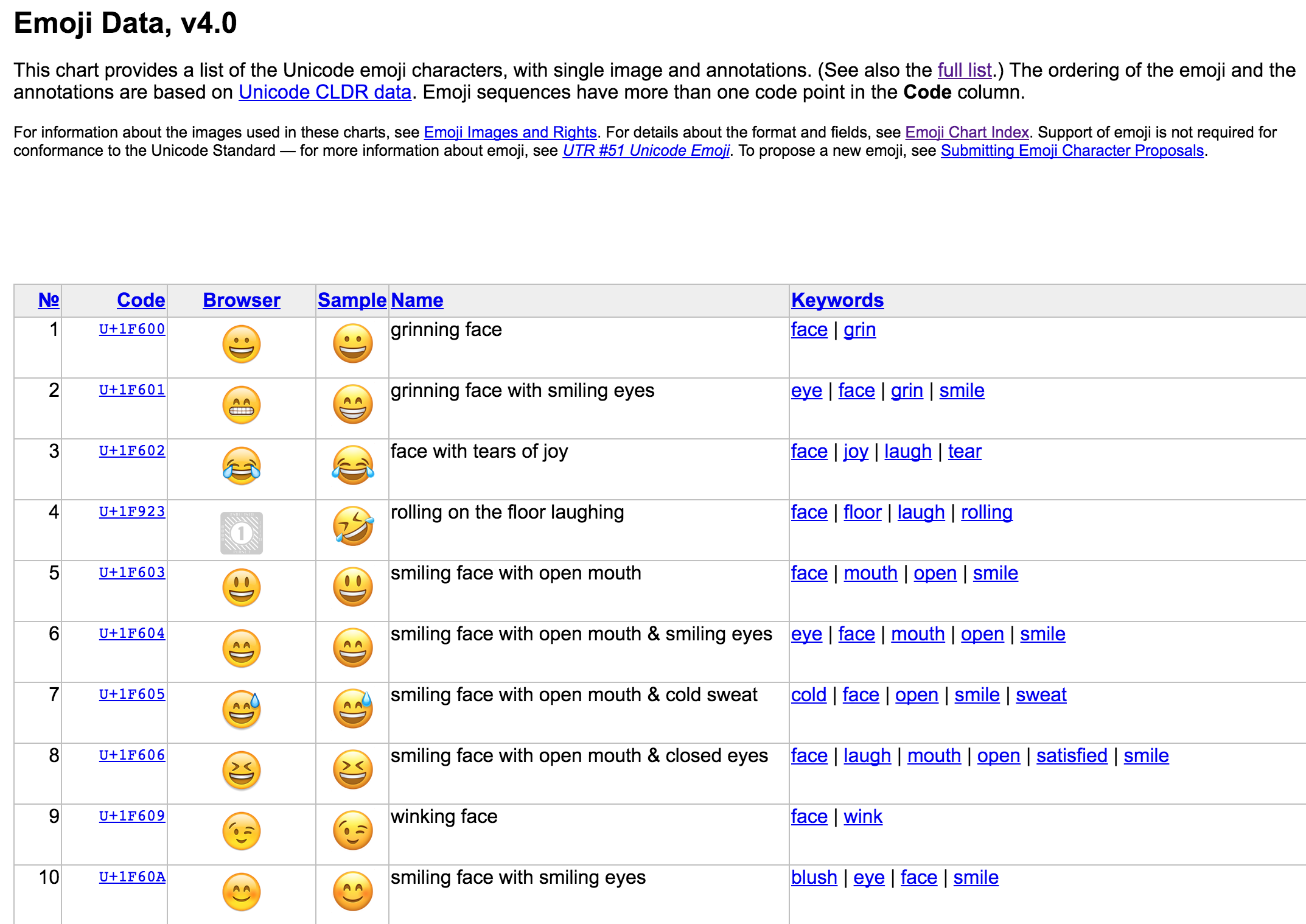
[47] "Y" "STATEFP"

[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

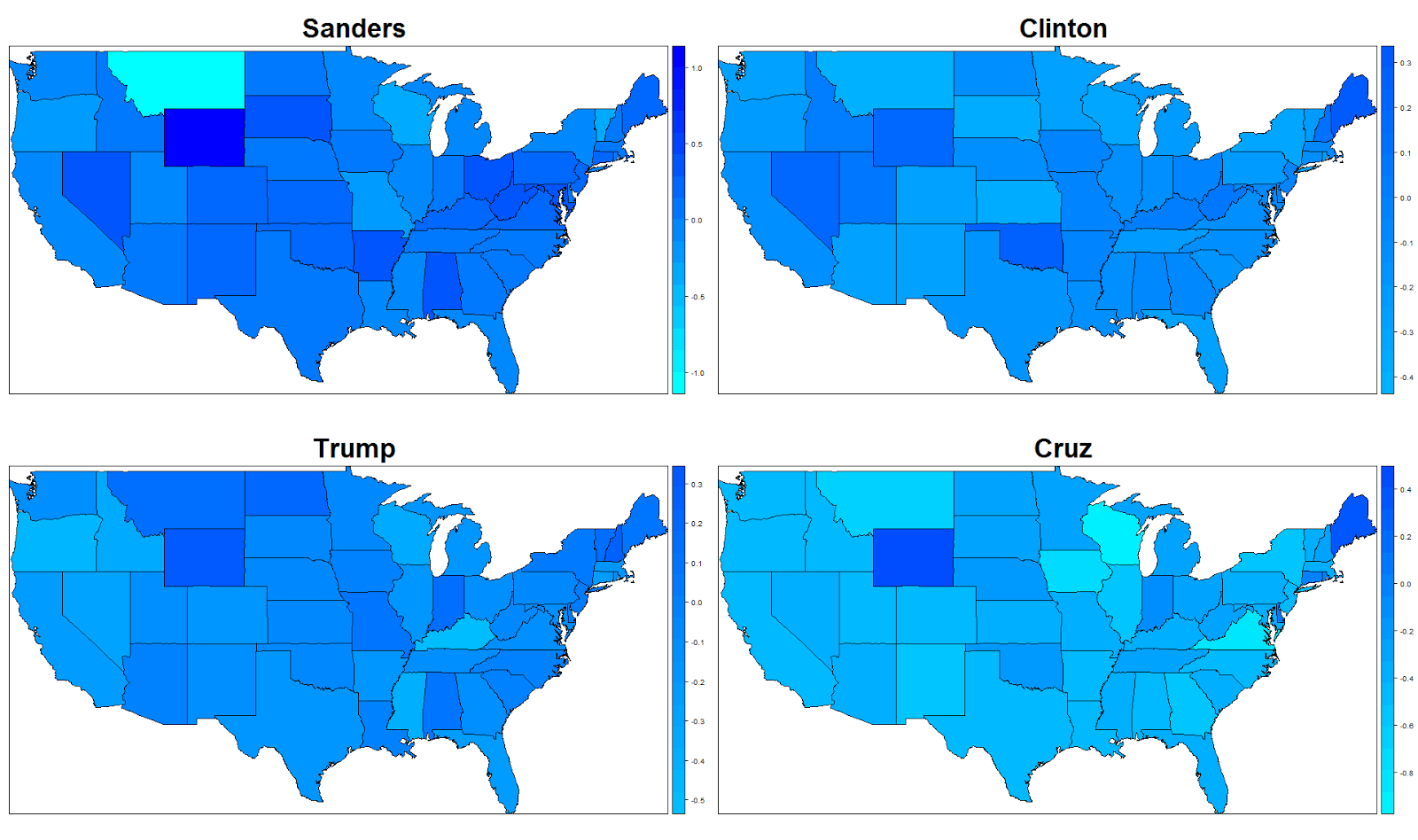


### Appendix D – Hu and Liu’s Lexicon for Text Analysis

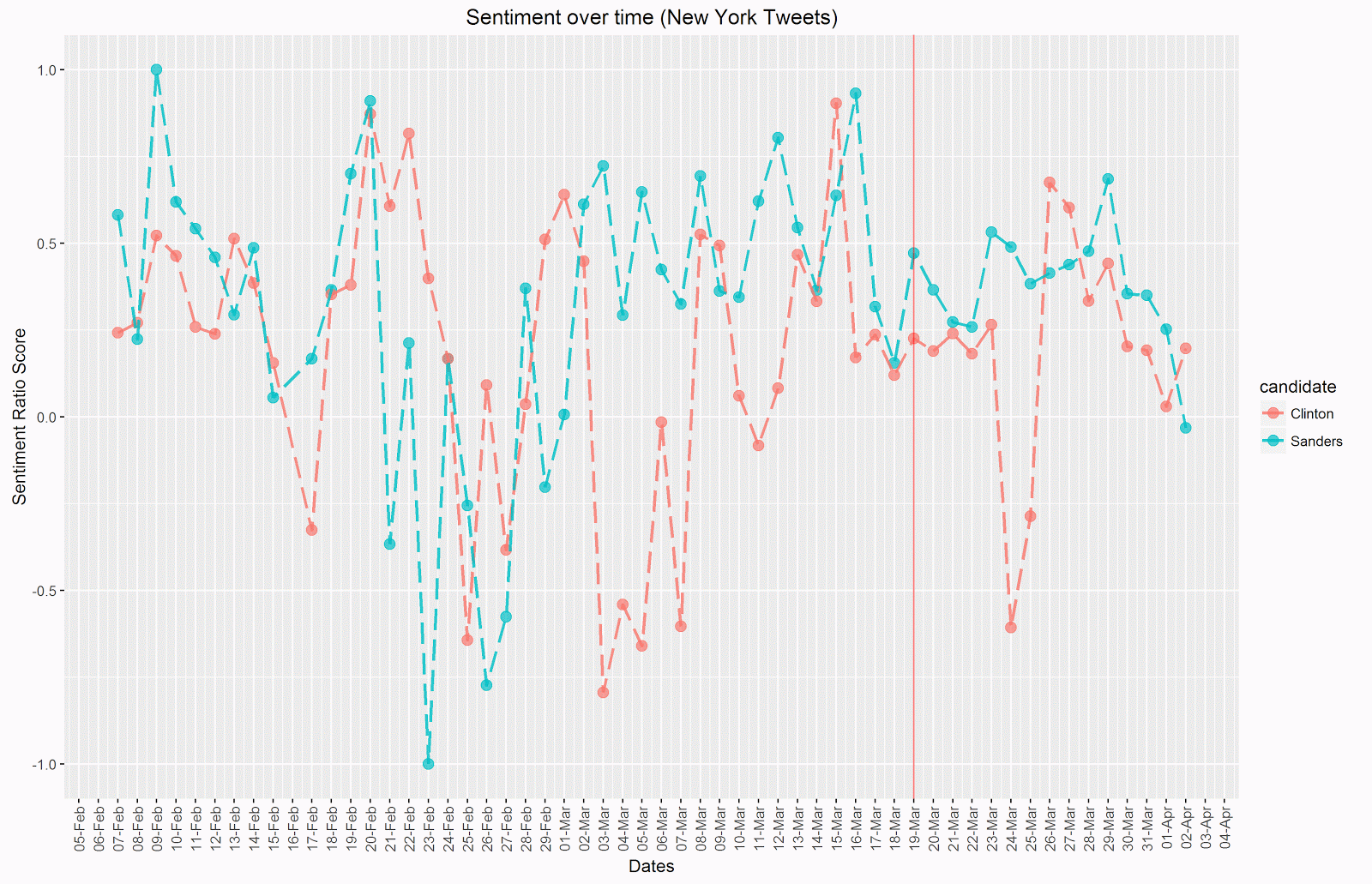
All files and papers can be downloaded via this link: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

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.



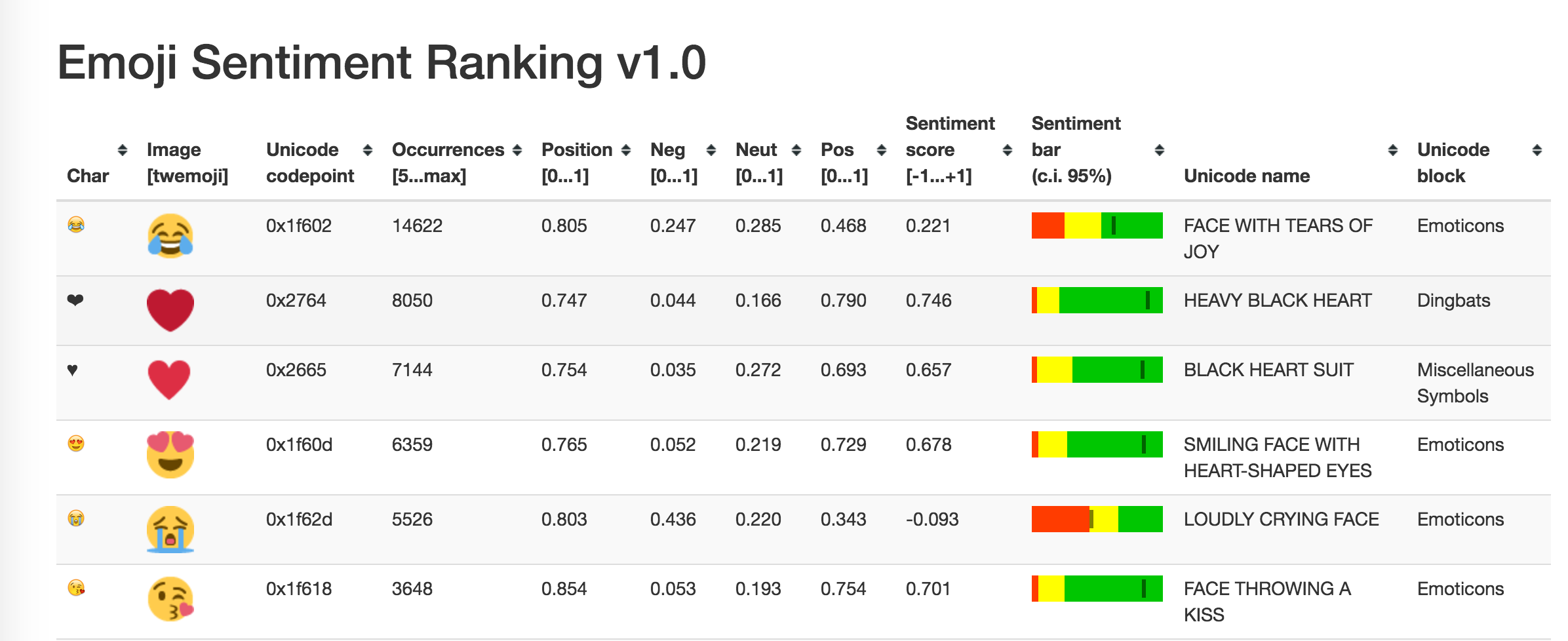
The above visualization shows sentiment scores over time between both Democratic candidates in the state of New York.

Final report from last semester’s QMSS G4063 class :

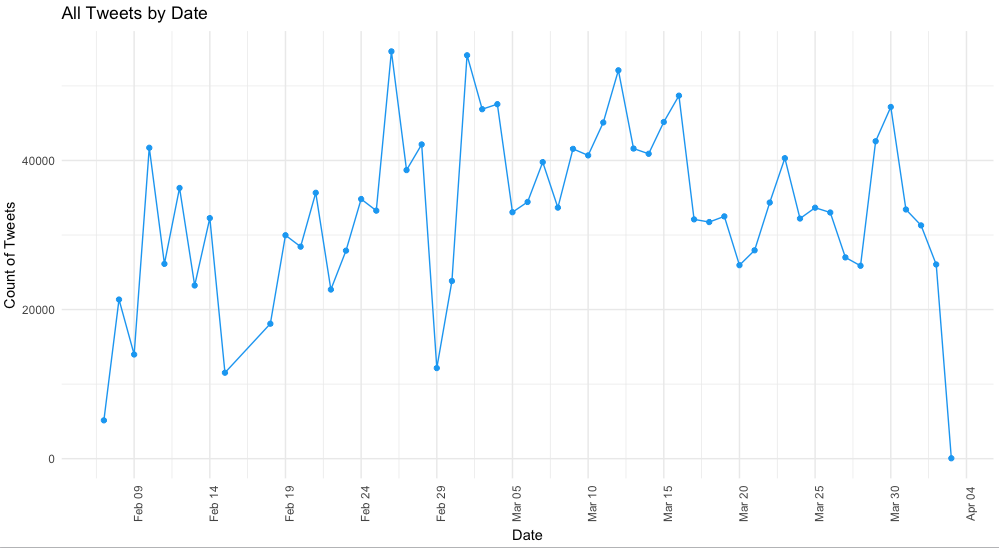
<https://docs.google.com/document/d/1le66GaGXa4XwhVyuRrHOG2oMvraBjWU7apfi-4e4tfI/edit>

### Appendix F – Emoji Sentiment Lexicon

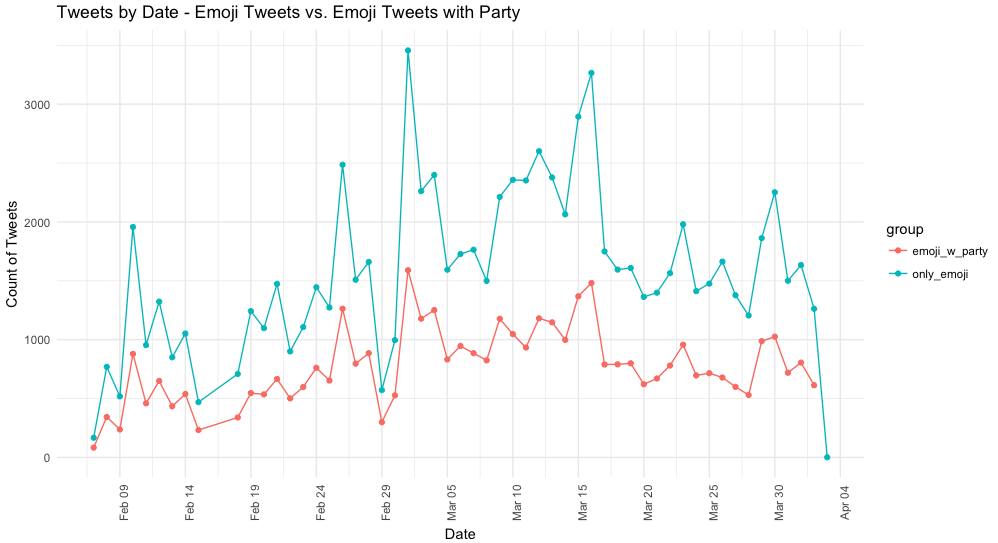
Below is a screenshot of the information found on <http://kt.ijs.si/data/Emoji_sentiment_ranking/>.



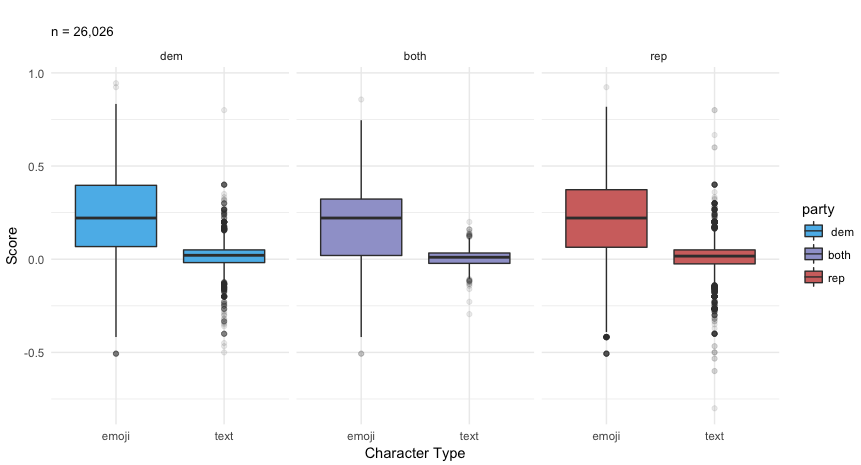
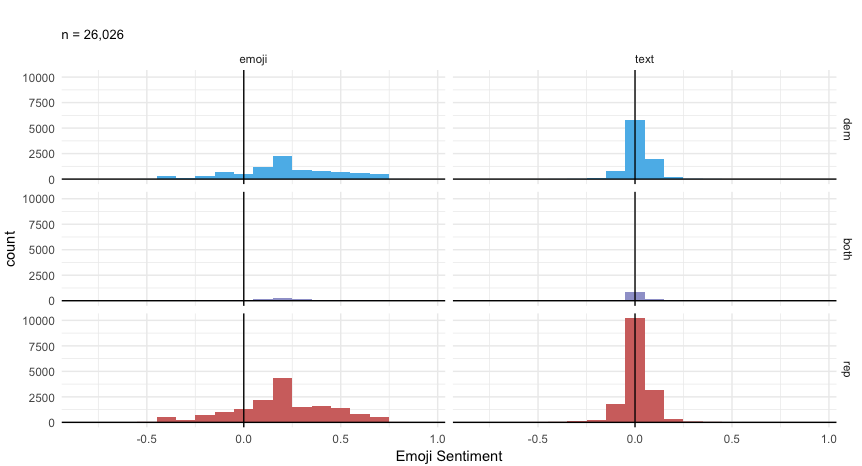
### Appendix G – Additional Visualizations



**Figure 1.a** – Count of Tweets by Day



**Figure 1.b** – Count of Tweets by Day – Emoji Tweets vs. Emoji Tweets with Political Party References



**Figures 3.a & 3.b** – Sentiment of Text and Emoji over Time by Political Party (Left – Histogram, Right - Boxplot)



# 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. Indurkhya 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/>

TO ADD:

<http://tidytextmining.com/tfidf.html>

1. Afinn lexicon (<http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010>) [↑](#footnote-ref-1)
2. https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/ [↑](#footnote-ref-2)
3. https://arxiv.org/pdf/1103.2903.pdf [↑](#footnote-ref-3)
4. ibid [↑](#footnote-ref-4)
5. https://github.com/laurenancona/twimoji/tree/master/twitterEmojiProject [↑](#footnote-ref-5)
6. https://github.com/PRISMOJI/emojis/tree/master/2017.0206%20emoji%20data%20science%20tutorial [↑](#footnote-ref-6)
7. http://miningthedetails.com/blog/r/IdentifyEmojiInTweets/ [↑](#footnote-ref-7)
8. https://arxiv.org/pdf/1103.2903.pdf [↑](#footnote-ref-8)