Big Data: COSC 592

Final Project

Twitter Sentiment with Emojis

William Gillespie

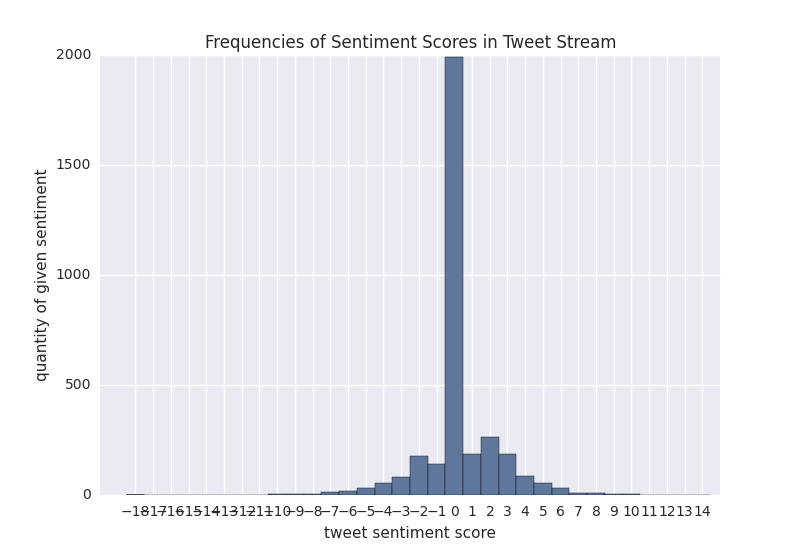
Watson Tong

Greg Moldovan

When we analyzed twitter sentiment using keywords in the AFINN-111.txt file, many of the tweet sentiments were zero. This was disappointing because we wanted a larger proportion of the tweets to have sentiment. A low proportion of tweets with sentiment could mean that most tweets truly do not have an identifiable sentiment, or that we are not capturing the sentiment.

Our goal of this project was to capture more sentiment in tweets. We decided to do this by analyzing tweet sentiment using emojis. Emojis are the faces that express some kind of emotion. We thought that this would be an effective way of capturing some sentiment that was missed when we only analyzed the words. What follows is a description of the data, the methods we used to analyze our data, and the results of our work. Because the group members each did a different analysis, the methods used and results will be separate for each group member.

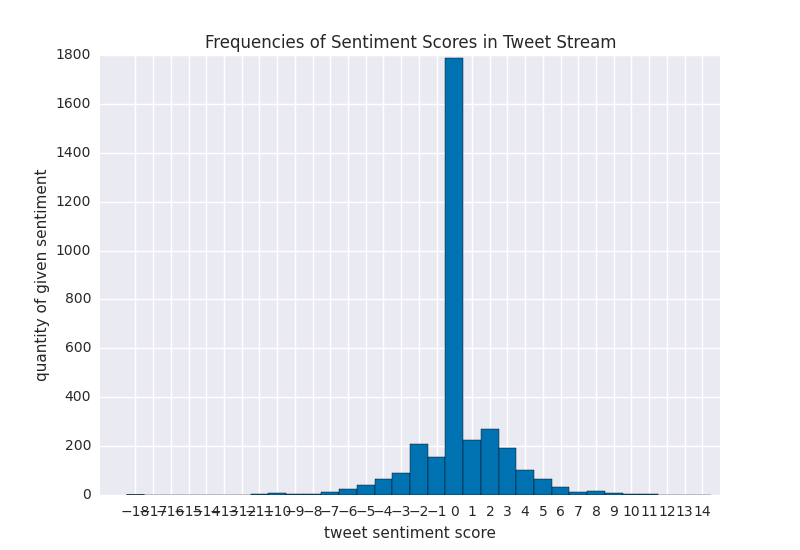
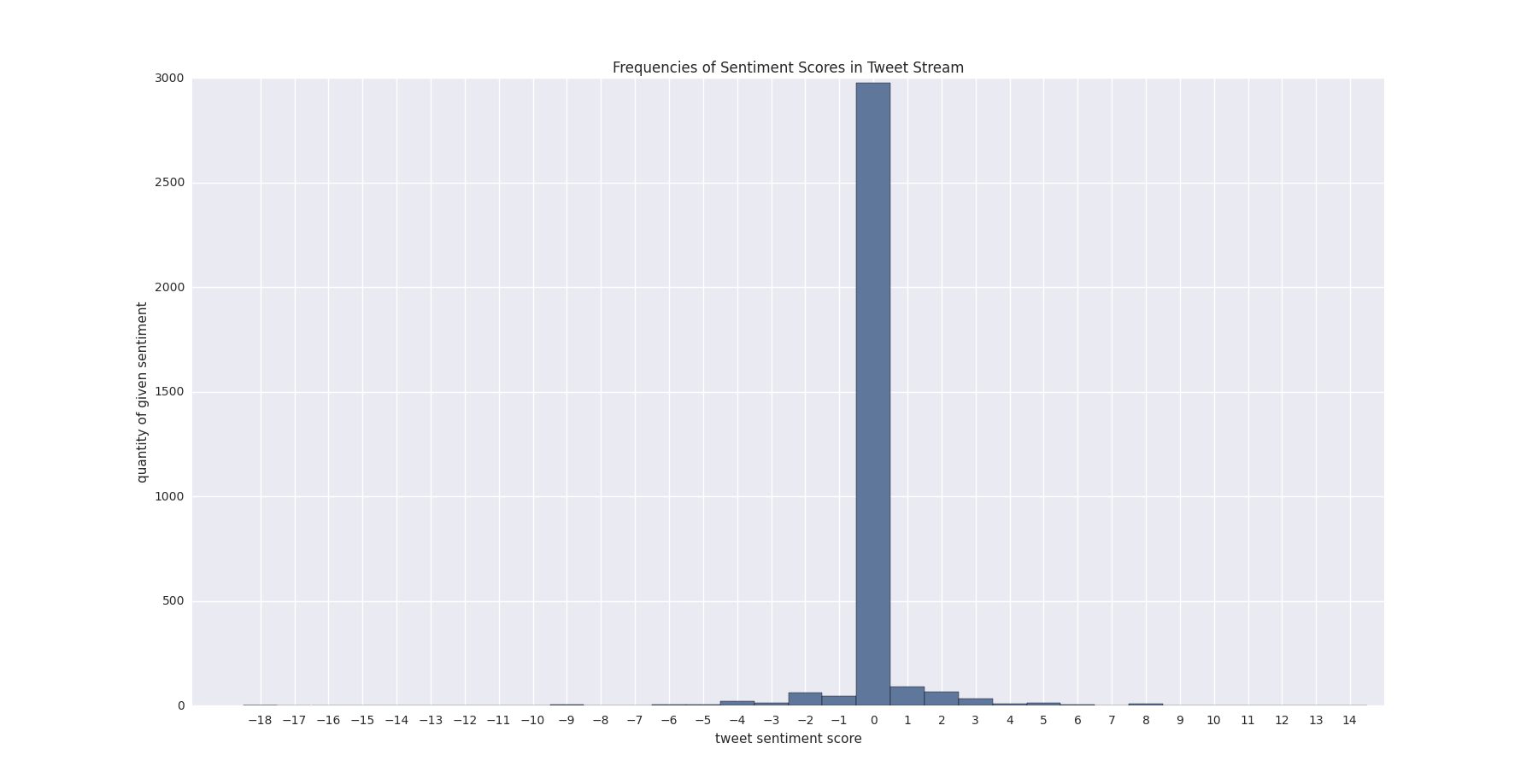
**Description of Data**

 The twitter sentiment data is very weird data. A analysis of the frequency of sentiment scores for a given number of tweets shows us that the majority of the tweets do not have any sentiment. This is true for the word sentiment analysis and the emoji sentiment analysis. If we look at the distribution of the word sentiment scores we see the giant bar that represents zero, and a normal-like distribution around it. When analyzing a twitter stream of 3359 tweets, we found 1944 (58%) tweets with no sentiment and 1415 (42%) with a positive or negative sentiment score. The histogram of word sentiment scores is shown below.

Tweet Count: 3359, Sentiment Frequency: 1415,

mean: 0.244, Standard Deviation: 2.137

When analyzing the emoji sentiment, we found the same kind of distribution, but the emoji sentiment captured far less tweets as having sentiment. In the same sample, only 396 (12%) of the 3359 had a positive or negative emoji sentiment, compared to the 42% found in the word sentiment calculation. A distribution of the emoji sentiment score frequency is shown below.

 These data suggest that the word sentiment calculation captures more tweets as having sentiment than the emoji sentiment calculation method. If we combine our methods and calculate an emoji+sentiment score, we get a distribution similar to the ones above. The new combined sentiment calculator yields a distribution that captures more tweets as having sentiment than either method independently. In our sample, it added about 200 tweets as having sentiment.

**Tweet Count: 3359, Sentiment Frequency: 1632, mean: 0.286, Standard Deviation: 2.518**

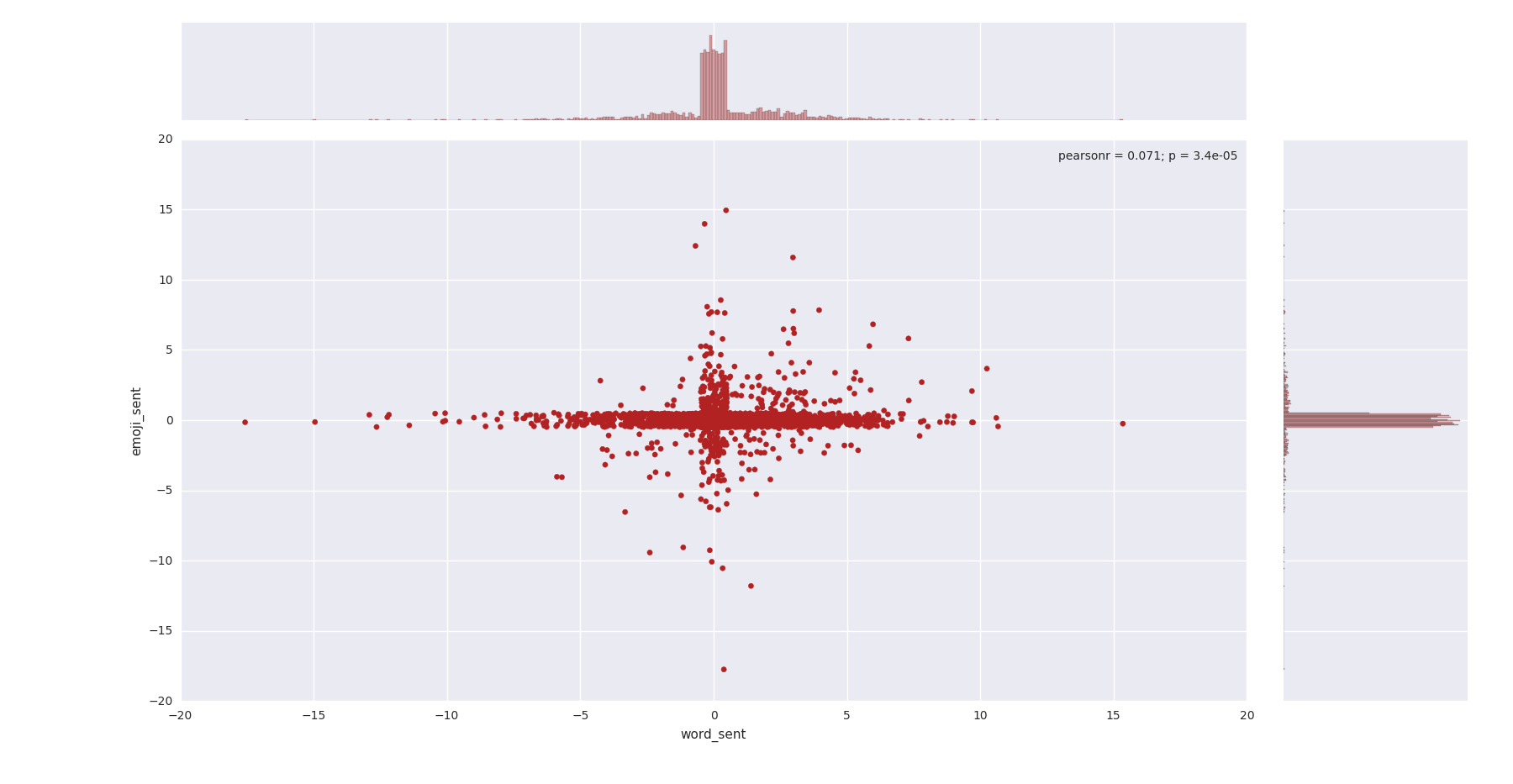
**Tweet Count: 3359, Sentiment Frequency: 396, mean: 0.043, Standard Deviation: 1.169**

Now that we have described what our data looks like, we will present each member’s research question, methods of analysis, and the results of that analysis.

**William Gillespie**

I had two research questions:

1. Do Emoji give additional information about the sentiment of tweets compared to words?
2. How similar are the sentiments expressed in words and Emoji?

To answer the first question, I wrote a script that would calculate the emoji sentiment and word sentiment of each tweet for all tweets in the sample and store each score separately. I then created a jittered scatterplot where the word sentiment was on the x-axis and the emoji sentiment was on the y-axis. I graphed the same sample shown in the histograms above and got the following scatterplot.

**Sample 1**

**Emoji\_mean: 0.04, stdev: 1.17**

**word\_mean: 0.24, stdev: 2.14,**

**t-statistic: 4.96, p-value: 7.42e-07**

In this scatterplot, note the correlation coefficient (0.07) and the p-value (0.000034) in the top right corner of the plot. For this sample data, there is a weak but statistically significant, positive correlation between word sentiment and emoji sentiment.

When interpreting this plot, it is important to note that the points in quadrants 2 and 4 indicate tweets where the word sentiment and emoji sentiment disagreed. It is in these quadrants where we may find that the emoji sentiment tells us additional information (or contradictory information) about the sentiment of a given tweet.

An excellent example of this is in another twitter stream taken at a different time than the one we have seen so far in this paper.



**Sample 2**

**Emoji\_mean: 5.0, stdev: 14.4**

**word\_mean: 0.26, stdev: 2.29,**

**t-statistic: -24.8, p-value: 1.86e-129**

In this scatterplot, we get very different information. The correlation coefficient is weakly, but statistically significantly, negative. The reason for this is the large cloud of points in the second quadrant. This quadrant has tweets with a negative word sentiment but a positive emoji sentiment. This shows us that emojis can tell us additional information about the sentiment of a tweet.

I examined what was going on by printing out the tweets with negative word sentiment but positive emoji sentiment. What I found was a trend in which people were sending tweets with many swear words followed by a blushing, smiling face. The tweets consisted of trash talking and jokes, many of which I found humorous. The swear words increased the negative word sentiment of these tweets and the spamming of happy emoji increased the positive emoji sentiment of them. The problem is that these tweets were really not negative; most of them were cheerful jokes with swear words. In this example the emoticons can tell us a better sentiment of the tweet. This is partly because of a problem with the word sentiment calculation: namely that swear words receive negative sentiment. Due to this we can say that the emoji can provide us with additional information about the sentiment of tweets beyond what the word sentiment can tell us.

Now, let’s answer my second research question: how similar are the sentiments expressed in words and emoji? To answer this question I conducted a paired t-test on both samples: the original sample used in the histograms (sample 1) and the sample with the high amount of swearing and happy emoji spamming (sample 2). I set my H0 to 0, assuming that there is no difference, and my HA assumed that there was a difference. The t-statistic for sample 1 was 4.96 with a p-value: 7.42e-07. This indicates that the word and emoticon scores were different. For sample 2, the t-statistic was -24.8 and the p value was 1.86e-129. For this sample, the sentiments scores for the tweets were significantly different, with emoji scores being higher than word scores. The excessively high t-statistic in sample 2 was caused by the disagreement of scores in the second quadrant of the plot.

In summary, I could answer both research questions. First, emoji do give additional information about the sentiment of tweets compared to words, especially in settings where negative word scores are generated from large amounts of swearing. Second, the sentiments expressed in words and emoji are often not the same. In two files collected at different times of the same day, we found words to have significantly higher sentiment scores than emoji in the first sample, but lower scores in the second sample.

Future research could examine trends in word and emoji scores over time, which from our small sample appear to have wide swings. Trends by time of day or phases of the moon would be of interest. We could also consider transformation of the heavy-tailed sentiment score distributions to facilitate statistical analyses. Other extensions of this methodology could consider modifying the positive and negative score values for different words, or expanding or contracting the sentiment word/emoji dictionaries, with the goal of bringing the scores closer to the sentiment expressed in the entire tweet.