## Capstone Project 2 Sentiment Analysis Report

## Twitter Sentiment Analysis on Kanye West

**Silas Lee**

**Introduction and Objective:**

The Client: In this project the client would be Kanye West's PR. To better understand the trends of the public (on twitter) sentiment of Kanye West.

**What is Sentiment Analysis?**

Sentiment analysis is essentially figuring out how a user feels about a certain topic. Commonly used to sort thru reviews, for brand management, polling for political parties, and any other reason you would want to know how the general public feels about something.

**What makes Twitter a good/bad medium for sentiment analysis?**

Tweets are limited to 280 characters (used to be 140), which means rarely more than one full thought is expressed per tweet. This gives us a very useful format where each tweet is usually one thought/ expression about something. Twitter is used for news, by users professionally and socially all over the world and across all demographics. Twitter also is very easy to work with, and tweets are easily scraped. The downside of using twitter for sentiment analysis lies in the fact that so much slang, emojis, pictures, memes and sarcasm is used throughout the platform, which is extremely hard to make up for in machine learning. Because of this, I decided to use a lexical approach to the tweets instead of a Machine Learning because it would also include slang, emojis, and acronyms that are so commonly included in social media such as tweets.

**About Twitter:**

Twitter is an online news and social networking site that limits each tweet to 280 characters and is extremely easy to scan through. Twitter is also very popular for celebrities. As of May 2018, Twitter is 4th in social media audience in the world. Fig. 1 shows users in millions.

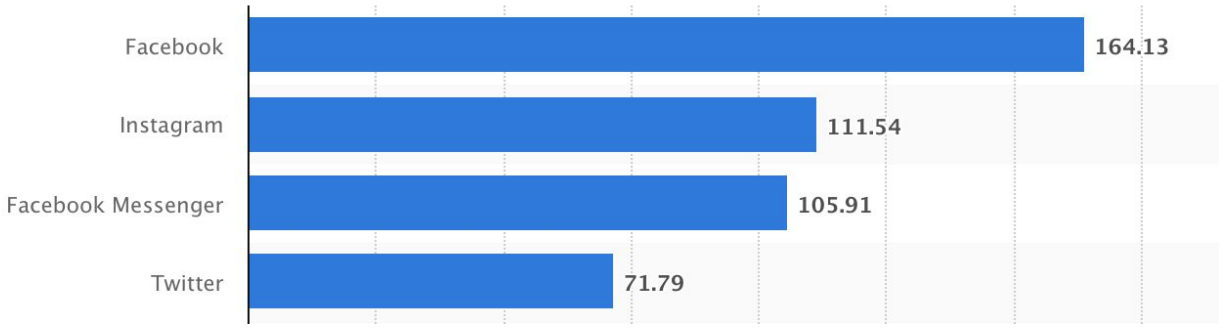



Fig. 1

Fig 2. Shows number of twitter users broken down by age:

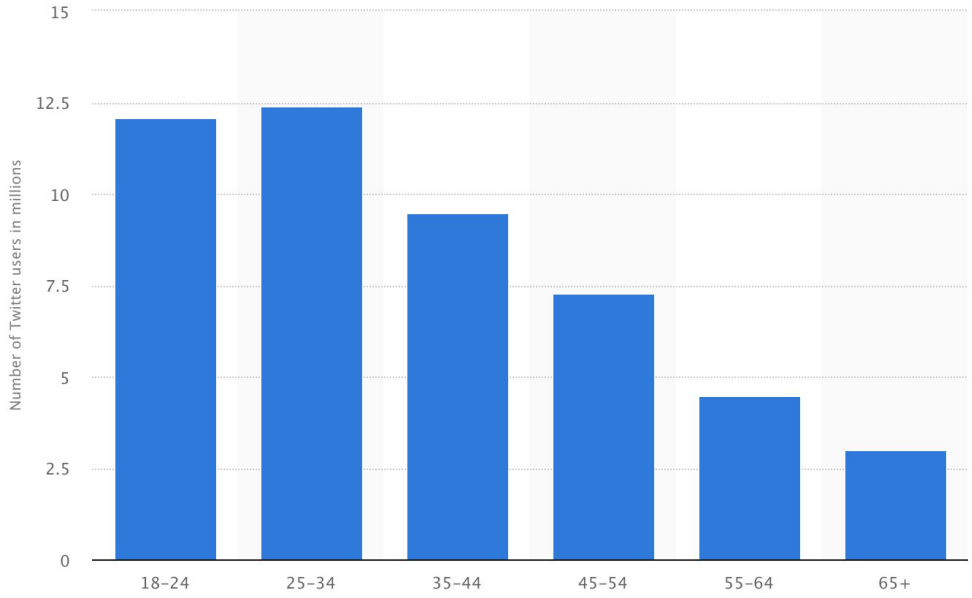



Fig. 2

Most twitter users are between the ages of 18 and 44, so all analysis of sentiment are skewed to show sentiment of the younger population. This is likely the intended audience of artists like Kanye West.

**Mining Twitter Data:**
Creating a twitter app to access the twitter api using tweepy. Using tweepy, tweets are found by searching for specific hashtags or from specific users, etc. For more information, see Appendix A.

**Goal:**
Building off the VADER sentiment analysis tools (Valence Aware Dictionary and sEntiment Reasoner) and applying it to tweets mined through twitter API. After tweets about Kanye are gathered, and sentiment analysis will be performed on gathered data to analyse positivity/negativity over time, and corresponding to different public events. Other trends such as location, time of day, Iphone users vs android in sentiment, mobile vs pc, Length of tweet, etc will be considered.

**Model:**
I decided to use a lexical approach to the tweets instead of a Machine Learning one because it would also include slang, emojis, and acronyms that are so commonly included in social media such as tweets. VADER is a useful python library that breaks down the sentiment into positive, neutral and negative sentiment. Then a compound score is given as a total sentiment value, a higher positive number correlates to a stronger positive sentiment, neutral score is calculated as

zero, and a lower negative score means stronger negative sentiment. I will be using this compound score to evaluate tweets.

**Tools:**

Tweepy - Data Mining
NLTK Toolkit, including VADER - For actual Sentiment Analysis of text
Pandas and Numpy - Algebra, data processing, CSV file I/O
String and Re - Data Cleaning
Matplotlib and Seaborn - Plotting
Scipy - Statistical Analysis

**Sentiment Scores:**

VADER's sentiment scores range from -1 to 1 and are based off of a dictionary of words evaluated by numerous human graders. It takes into account punctuation (such as exclamation points to increase sentiment intensity) and capitalization ( "AMAZING" would be ranked at a higher intensity than "amazing"). There also are a long list of degree modifiers such as "sort of". This impressive open-source library is ideal for twitter sentiment analysis. For more information on VADER, see Appendix B.

**Data:**

-All scraped Via Tweepy-
-Tweets were gathered based on creation date, included tweets are from 8/7/18 to 8/17/18.-

Kanye.csv - Collection of 3692 recent public tweets with the hashtag #Kanye
Allwest.csv - Collection of 31682 recent public tweets with the hashtag #Kanyewest (including duplicates*)
Allkanye.csv - 35323 Kanye and Kanyewest combined with duplicates* removed
Fromk.csv - 180 Public tweets from Kanye's account
Fromdt.csv - 176 Public tweets from Donald Trump's account
Fromjk.csv - 181 Public tweets from Jimmy Kimmel's account

> *Duplicates are any tweet repeated exactly from the same user more than once, only the first is kept. This also weeds out duplicates created when a tweet includes both #Kanye and #Kanyewest.

Each Dataset includes:

> Id - Unique ID number of each tweet
> Text - Text of the tweet
> Created_at - Date and time tweet was published
> Retweet_count - Number of times tweet was retweeted
> Favorite_count - Number of times tweet was favorited
> Source - Type of app used to send tweet

Country - Country of tweet sender (Only 273 tweets out of 35,000 included Country)
User_id - Unique ID number of user
User_screen_name - Screen name or hande of Twitter user
User_name - Name of Twitter user
User_created_at - Date Twitter users account was created
User_followers_count - Number of followers for user
User_friends_count - Number of friends for user
User_location - Location of user (fill-in, roughly ⅔ of tweets included location, not always accurate)

**Data Cleaning**

Each Tweet went through a process including:

- Stored into a DataFrame
- Labeled each gathered piece of data for each tweet
- Text of tweet converted to a string and hashtags, @'s, rt, and links all removed
- Organized into different groups, and all duplicates removed (eg, tweets that include both searched hashtags and were gathered more than once, or if users posted exact same tweet more than once)
- Length of tweet text added
- Converted times to datetime and set as index

**Sentiment Scoring**

Each tweet text after cleaning was then sent through VADER's Sentiment Intensity Analyzer, receiving a score between -1 and 1.

## Data Analysis

**Tweets with #Kanye and #Kanyewest**

**Over Time**

When the mean and median score every 4 hours is taken for any tweet with the hashtags #Kanye or #Kanyewest, the result is shown in the next page on figures 3 and 4. As you can see, most of the sentiment is slightly above neutral, with spikes both negative and positive. Late on August 8 there seems to be a short significant drop in sentiment, then a large spike over a couple days starting midday on August 9.

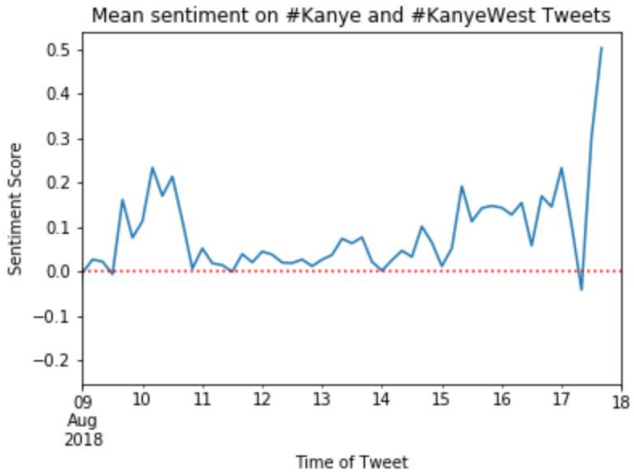



Fig. 3

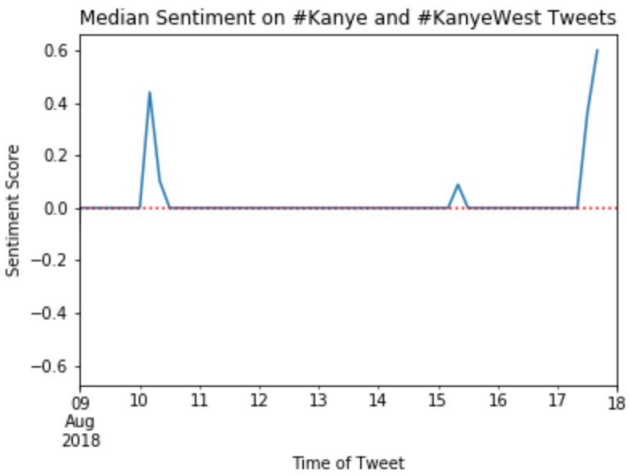



Fig. 4

One of the big controversies during this time period was the appearance on Jimmy Kimmel's show on the night of Aug 9 at 11:35 pm EDT or Aug 10 at 3:35 UTC (which all the twitter data is in). Kanye was asked if Donald cared about black people but was unable to reply before Kimmel sent it to a commercial break. Sentiment is low just prior to Kanye's appearance and then drops right back down to neutral just following the show. Kanye later tries to explain himself on twitter and it becomes a popular topic on the social media site. Some twitter users come to his defence, and some show disapproval.

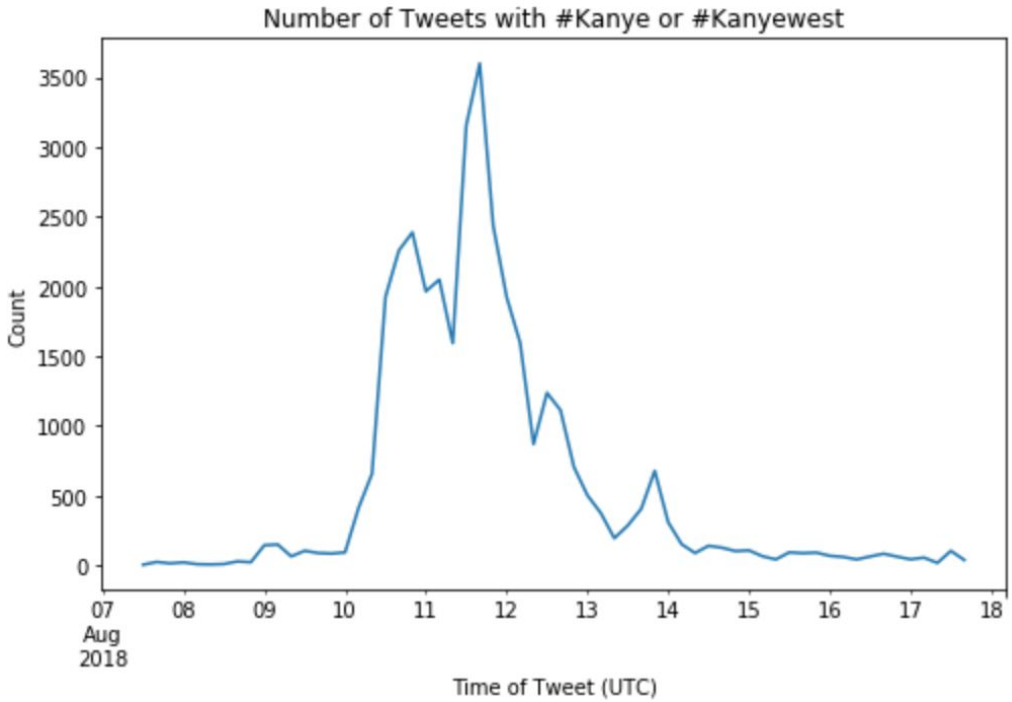



Fig. 5

As Fig. 5 shows, if all publicity is good publicity then Kanye should be quite pleased with himself after the interview. The number of people tweeting about him increased exponentially for a day and remained higher than usual for a couple of days after before returning to a normal amount.

**Top Tweets**

The most common score given was a completely neutral 0.0, most likely largely because of tweets that are used to spread news articles. Relatively objective news headlines receive a neutral score and much of twitter is used to spread news.

The most common tweet? Originally by @ChecoKicks, retweeted over 23,000 times just in this short time span was:

*“KanyeWest tries to stand up for trump and JimmyKimmel shuts em up with facts”*
Given score: 0.0

While this tweet was given a neutral score, one could argue that this tweet is negative towards Kanye, and if counted so would drastically bring down the mean and median sentiment scores. However, this tweet included a news article and is just considered news, so the neutral score is to be kept at 0. There is no indication that users retweeting this agreed or not.

Another popular tweet with a positive score and over 300 retweets:

*“The media: Black lives (and voices) matter. Also the media: Hahaha! We silenced a black man.”*
Score: 0.6114

All sentiment is subjective, but while the overall sentiment of this tweet might be seen as negative commentary on the media, it is defending Kanye and his actions during the interview and contributes to a positive score for Kanye.

**Tweet Sources:**

There are dozens of ways to tweet, from many different sources. The most popular by far is using the iPhone app, with the Android app in second. Shown in Fig. 6 is the number of tweets from the top 8 sources: iPhone, Android, Twitter Web Client, Twitter Lite, iPad, Instagram, HispterTweets, TweetDeck. Remaining sources amount to less than 0.1% of all the gathered tweets.

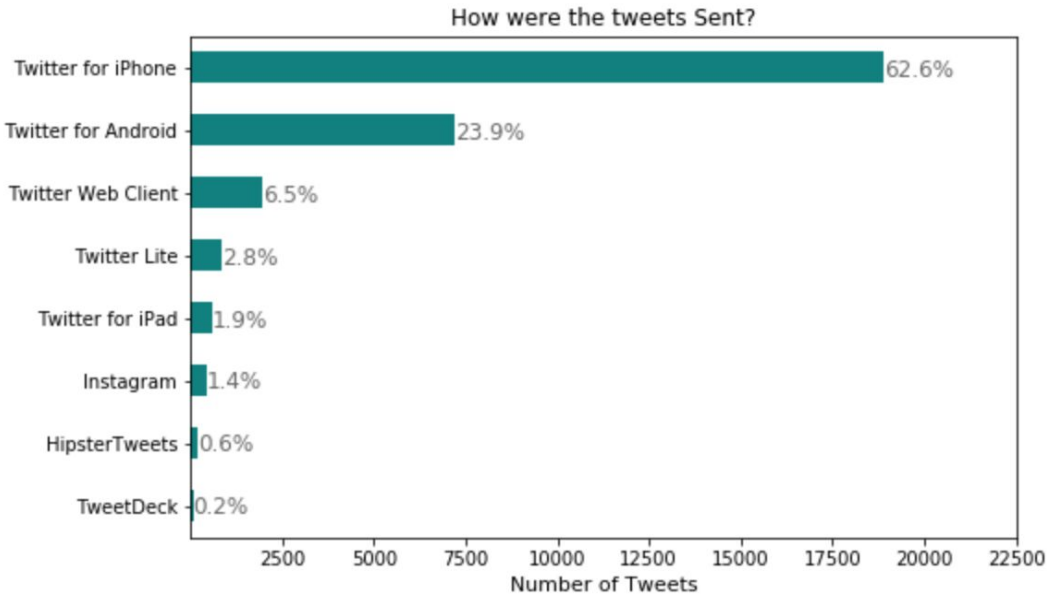
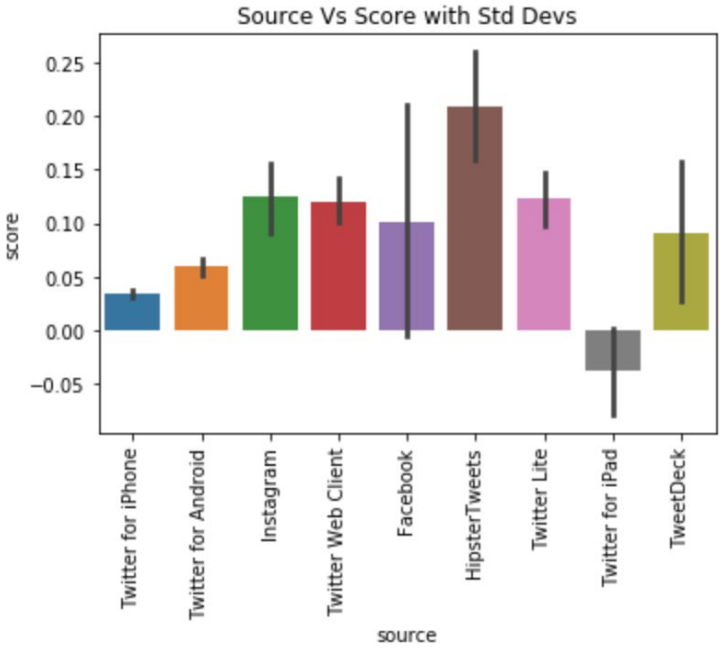



Fig. 6

How do scores differ based off of tweet source? It looks like all but one of the most popular sources were positive on average. Ipad users (only 1.9% of taken tweets) had a rounded average score of -0.01. The most common source, the iPhone, has the second lowest average at 0.04. HipsterTweets users only made 0.6% of the taken tweets but has the highest average sentiment score at 0.21. The average score across all is about 0.0578.

.




```
iPhone mean score 0.03
Android mean score 0.06
Twitter Web User mean score 0.12
Twitter Lite mean score 0.12
iPad mean score -0.04
Instagram mean score 0.13
HipsterTweets mean score 0.21
TweetDeck mean score 0.09
```

Fig. 7

Are the distributions of the top two sources actually statistically different?
In order to test a null hypothesis that there is no difference in distributions and an alternate hypothesis that there is a difference, we will run 2 sample t-test between the Twitter for iPhone and Twitter for Android distributions. The results were:

```
t-statistic = -24.434 pvalue = 0.0000
Different distributions (reject H0)
```

The t-statistic indicates that there is an extreme difference and the p-value being essentially zero means it is very likely that this is due to chance and the differences in distribution is indeed statistically significant.

**Location:**

User location on twitter is not required and can be anything, there is no set format and a lot of missing data. From what we do have, as shown in Fig 8, it looks as though most of these tweets have been made in the US.

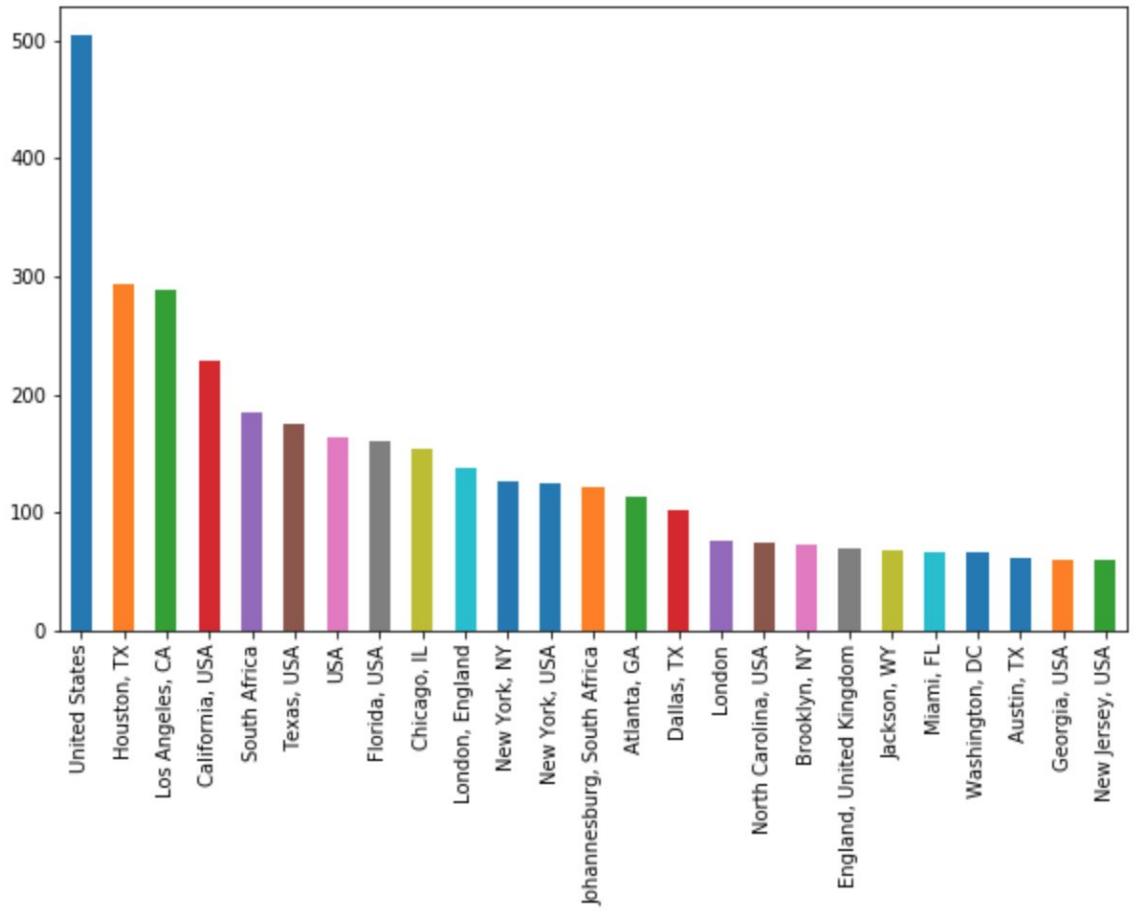



Fig. 8

One interesting thing to note is that the city with the most tweets is Houston, TX. Significantly more than even Kanye's hometown of Chicago. The most common outside of the US are from South Africa and England. This may be strongly affected by only gathering tweets in English. Also, Jackson, WY has more tweets than from Miami, FL.

Chicago is Kanye's Hometown, how does it compare to cities with similar number of tweets?

| user_location | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Chicago, IL** | 174.0 | 0.034112 | 0.285155 | -0.8779 | 0.0 | 0.0 | 0.000 | 0.9377 |
| **London, England** | 156.0 | 0.038486 | 0.207812 | -0.5661 | 0.0 | 0.0 | 0.000 | 0.8595 |
| **New York, NY** | 141.0 | 0.125126 | 0.252069 | -0.7418 | 0.0 | 0.0 | 0.296 | 0.8595 |

As a whole, London and Chicago seem to have roughly similar distributions, while New York seems to have a mean score almost three times that of Chicago, with a smaller standard deviation. However, it should be noted that Chicago has slightly more tweets and a higher maximum score. Similar to the distribution of all gathered tweets, the bulk of the tweets from each city has a neutral score.

Fig 9 shows the average score between Aug 10 and 17 for the three cities. Over time, New York had all positive averages, London had mostly with one sharp dip on Aug 16th, and Chicago dipped below zero a few times.

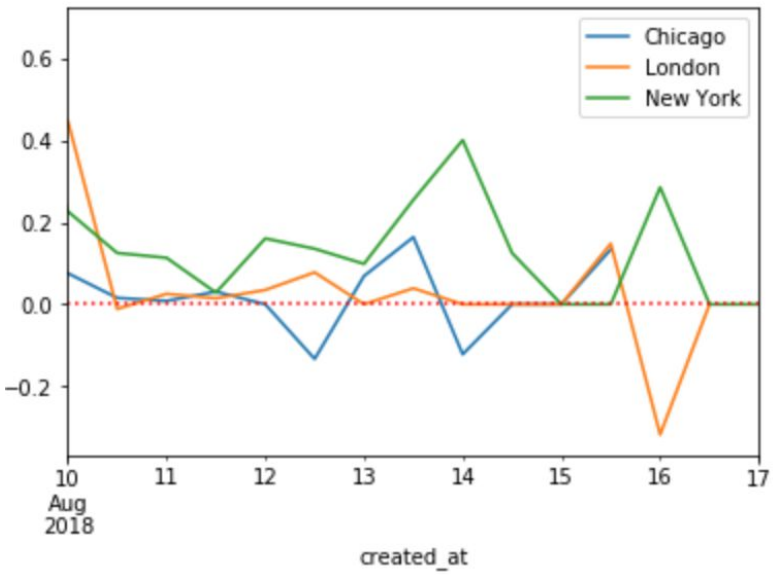



Fig. 9

Is this difference between cities statistically significant?

Null hypothesis will be that the cities distributions are not different, while the alternate is that they are different. When a 2 sample t-test is run between the Chicago and New York datasets, the result is:

```
t-statistic = -3.004 pvalue = 0.0029
Different distributions (reject H0)
```

The datasets are slightly different and because p< 0.01, you can suggest that the means are statistically different and likely not due to chance. Chicago and London were not significantly different.

**#Kanye vs #Kanyewest**

Because tweets were gathered using the query for both #Kanye and #Kanyewest, we are able to split the two and compare the use of each hashtag. There were 3692 public tweets with the hashtag #Kanye and 31682 public tweets with #Kanyewest. Some tweets are in both because they contain both hashtags.

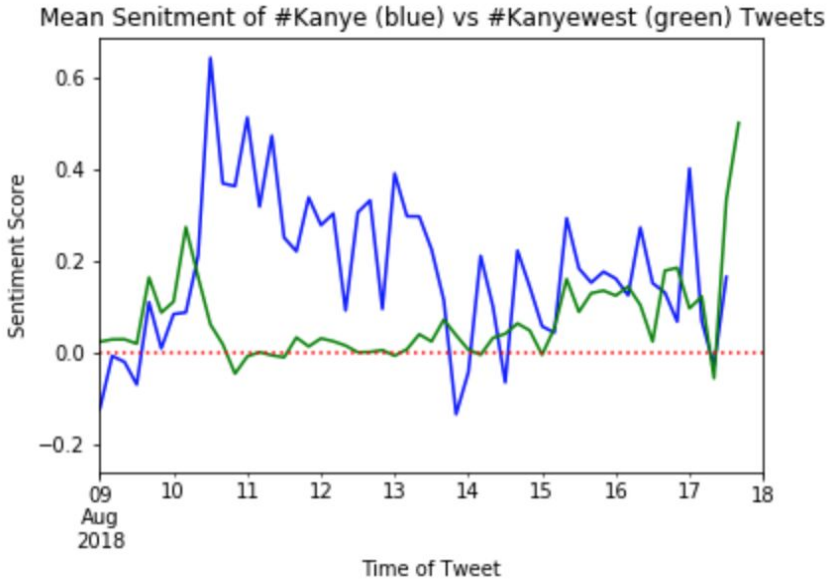



Fig. 10

Fig 10 shows each hashtag separately over time, and overall #Kanye tweets had a higher average score. This is strongly influenced by that top frequency tweet having #Kanyewest but not #Kanye, driving the score closer to zero after 8/10.

The number of tweets for each hashtag over time is shown in Fig. 11, where there is an overall increase in tweets after the Jimmy Kimmel interview, but #Kanye tweets go back to the baseline rather quickly and #Kanyewest tweets bounce up and down for a few days.

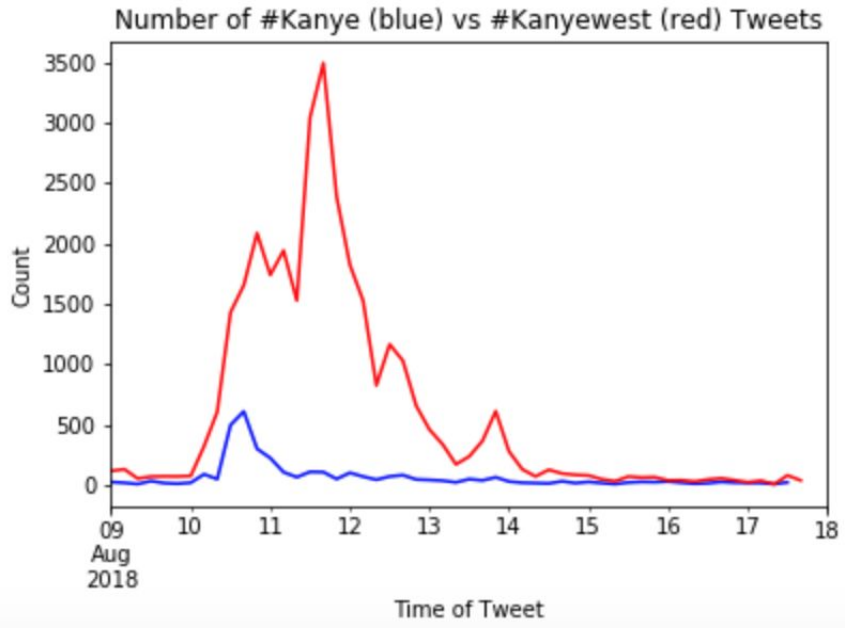



Fig. 11

## A Look at the Data without Top Tweet

As mentioned previously, 23271 of the 35323 tweets gathered are actually a retweet of one tweet and has the score of 0. Because of this, the score means and medians are much closer to zero than if the top tweet is excluded. Fig 12 and Fig 13 show this exploration.

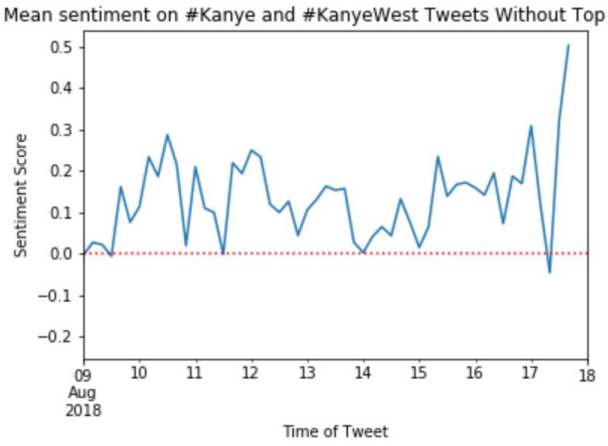



Fig. 12

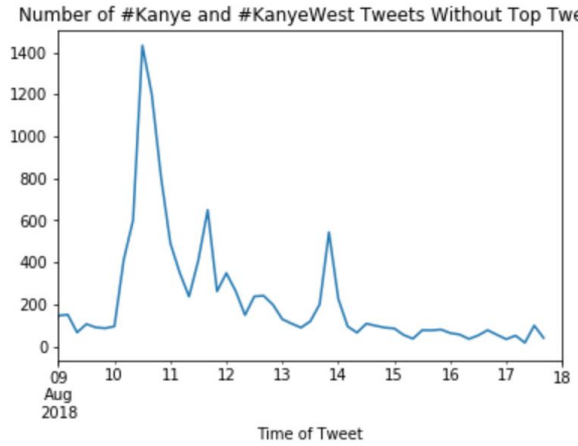



Fig. 13

The mean score is similar, but much higher between Aug 11 and 14, when that top tweet was most viral. The count is a similar shape, but much smaller number of tweets per hour as expected.

**Tweets from Kanye and other Celebrities**

Is Kanye more or less negative on twitter than Jimmy Kimmel or Donald Trump?

To answer this I gathered roughly 180 tweets from each celebrity to compare.

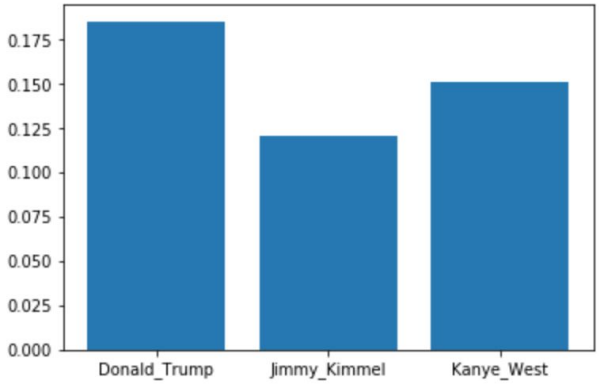



Fig. 14

The average score for each celebrity is shown to the left in Fig. 14. Donald Trump has the highest sentiment score, with Kanye West just under him and Jimmy Kimmel in third. Because of Jimmy Kimmel's profession as a comedian, his sentiment scores may not be as accurate. This is explored further in Accuracy Checks.

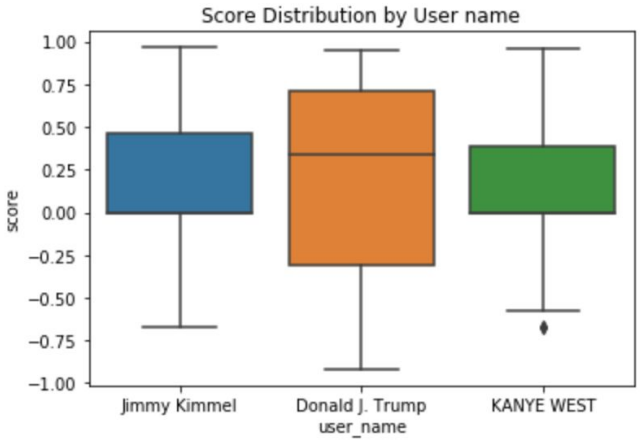



Fig. 15

In Fig. 15, It looks like Kanye's median tweet is actually completely neutral and the middle 50% are either neutral or positive. Donald Trump has the widest range in tweet scores and the highest median.

## Summary and Trends

Overall, it looks like Kanye's interview on Jimmy Kimmel Live was a very hot topic on twitter and had a wide range of reactions. The average sentiment was slightly positive. Tweets with the hashtag #Kanyewest generally had a higher sentiment score and significantly higher number of tweets over the same time period than ones with the hashtag #Kanye.

The viral tweet that was tweeted immediately after the interview and was retweeted 23,000 times in the following week had a sentiment score of 0. Without this tweet, the average sentiment score is higher.

Most of the tweets were made from the US, the most commonly reported city being Houston, TX. Between Chicago (Kanye's hometown) and two cities with similarly sized samples, London England and New York, NY, Chicago actually had the lowest average sentiment score of the three.

There was a significant difference in sentiment scores based on source of tweet. Tweets sent from an iPhone proved to have lower sentiment scores than those on an Android phone.

The average sentiment from Kanye's personal twitter account was positive, though not as positive as Donald Trump's. However, Trump did have a much larger range in scores.

## Recommendations for Further Analysis

1. Find a way to also include replies or replies included in retweets. Often people retweet a news article and then also give an opinion on the same thread.
2. Gather tweets from a wider range of dates, going back further.
3. Conduct further analysis to determine if a tweet's score is correlated to the user's information such as followers/ friend count.
4. Clustering to discover more trends about twitter users and their sentiments.

## Accuracy Checks

**All sentiment is subjective and true accuracy is nearly impossible.**

Sentiment in general can be hard to detect even by humans, so there's even more room for improvement when it comes to scoring tweets that use satire and sarcasm, and also media. A great example of this is in s tweet made by Jimmy Kimmel on August 16, 2018:

```
Tweet:  An important message from your innocent friends at Apple
Score:  0.743
```

The real tweet looks like this:


This is all the text in the tweet, but there was an accompanied video (as many of his tweets include) that includes news about Apple making a statement. This text was meant to be satire and is implying that Apple is not innocent, but this might not have been picked up on even by a human without the attached video. This issue is shown more with Jimmy Kimmel, where his career is founded on being satirical. Kanye West and Donald Trump (though debatable) are less focused on tweeting satirical and humorous tweets.

Example Tweets from Kanye, all from August 13, 2018:

```
Tweet:  deprogram
Score:  0.0
Tweet:  Yeezy slides on vacay
Score:  0.0
Tweet:  we’re no longer fighting for change we’re simply changing things
Score:  -0.5719
```

The real tweets look like:


When satire, especially with media (memes, pictures, video) are absent, the sentiment analysis is relatively accurate.

**Resources and Links**

- Twitter API - (https://apps.twitter.com) to create your own app and get auth passcodes
- tweepy cursor documentation (http://www.tweepy.org)
- tweepy and twitter tutorial can be found at (https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/)
- Vader, etc (https://github.com/cjhutto/vaderSentiment)
- For more about NLTK, (https://textminingonline.com/dive-into-nltk-part-i-getting-started-with-nltk)

**Figure List**

Fig. 12 - Mean sentiment scores over time for both #Kanye and #Kanyewest tweets excluding most popular tweet, combined from Aug 9, 2018 to Aug 18, 2018, in UTC.

Fig. 13 - Number of tweets over time for both #Kanye and #Kanyewest tweets excluding most popular tweet, combined from Aug 9, 2018 to Aug 18, 2018, in UTC.

Fig. 14 - Bar chart showing mean scores for tweets from Donald Trump, Kanye West, and Jimmy Kimmel for comparison.

Fig. 15 - Score distribution by user name, for Donald Trump, Kanye West, and Jimmy Kimmel