Twitter Sentiment Analysis

Using Kanye West as a topic

Introduction

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?

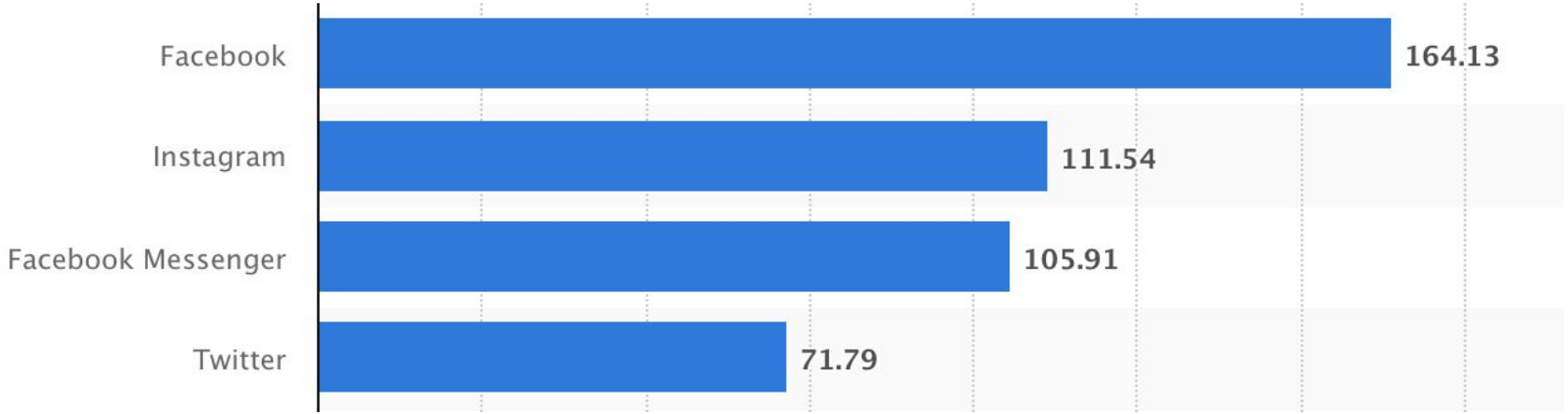
Sentiment analysis is essentially figuring out how a user feels about a certain topic. Commonly used to sort through reviews, for brand management, polling for political parties, and any other reason you would want to know how the general public feels about something.

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

Pros	Cons
Short format, usually only one thought per tweet	Tweets often include Video or Pictures that cannot be easily evaluated
Twitter is used for news, by users professionally and socially all over the world and across all demographics.	Slang and emoijs are common
Twitter also is very easy to work with, and tweets are easily scraped	Sarcasm and humor can be hard to accurately score

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. Below is the number of users in millions for each platform.



Approach

- 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 analyze 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.

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

About the 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.

More about the datasets

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
- Sent through VADER's Sentiment Intensity Analyzer

Relevant Current Events

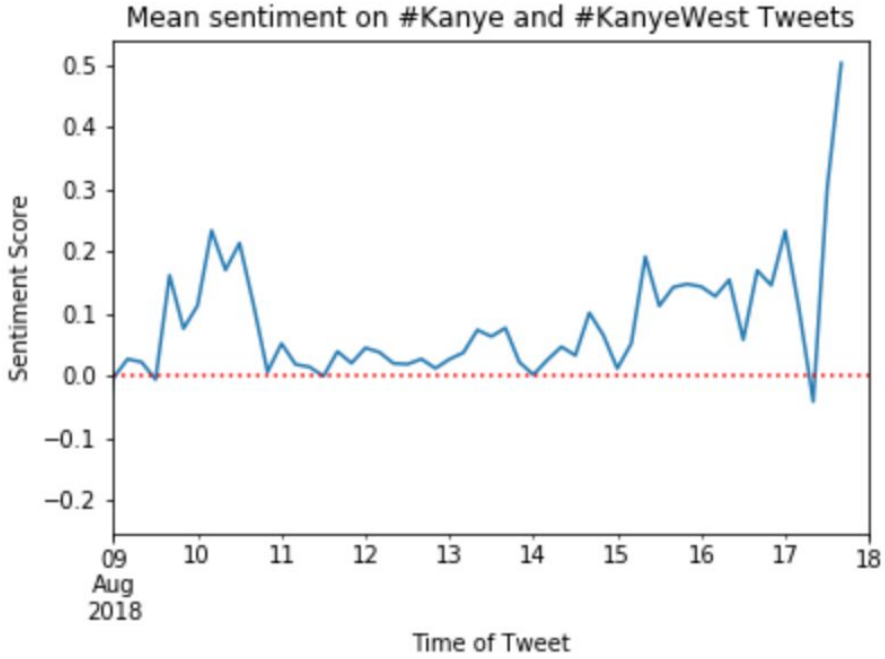
One of the big controversies during this time period was the appearance on Jimmy Kimmel's show on the night of Aug 9 at 11:35 pm EDT or Aug 10 at 3:35 UTC.

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 jumps up the days 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.

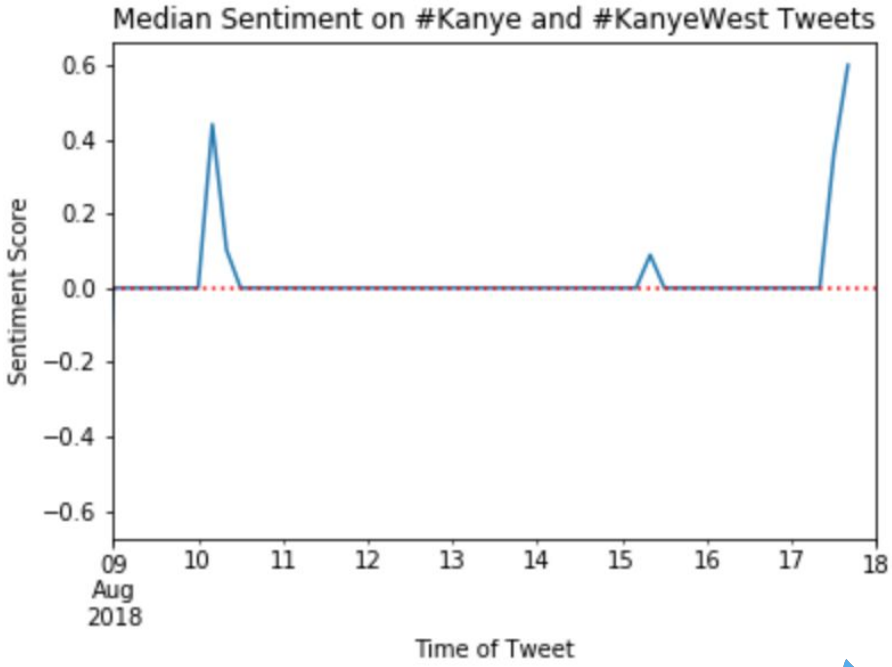
Average Sentiment Scores

When the mean score every 4 hours is taken for any tweet with the hashtags #Kanye or #Kanyewest, most of the sentiment is slightly above neutral, with spikes both negative and positive. There is a larger score over a couple days starting midday on August 9, then also on Aug 15 and late Aug 17.



Median Sentiment Scores

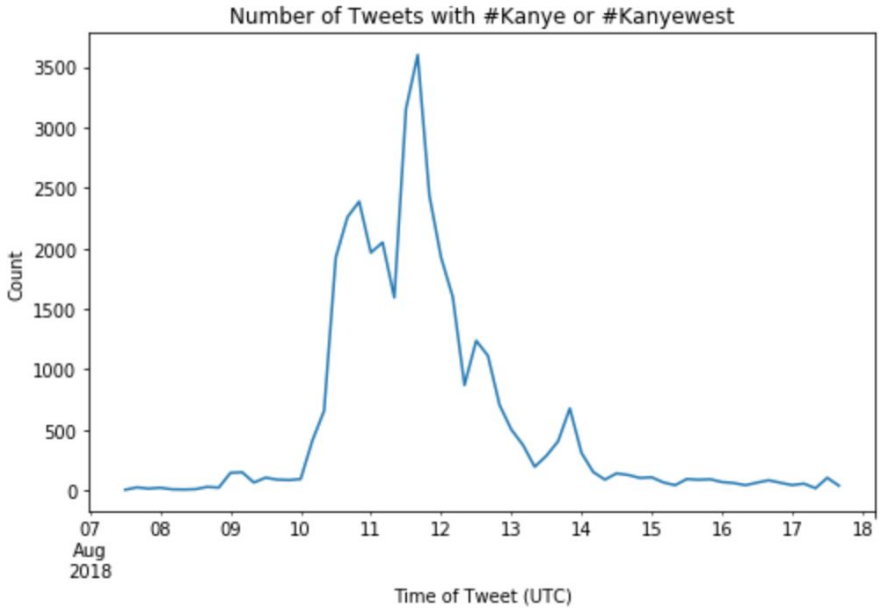
The median scores take on a similar, yet much more neutral pattern, as almost all median scores were exactly 0.0, except for the Aug 10, Aug 15 and Aug 17 spike.



Actual number of Tweets

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 Tweet

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

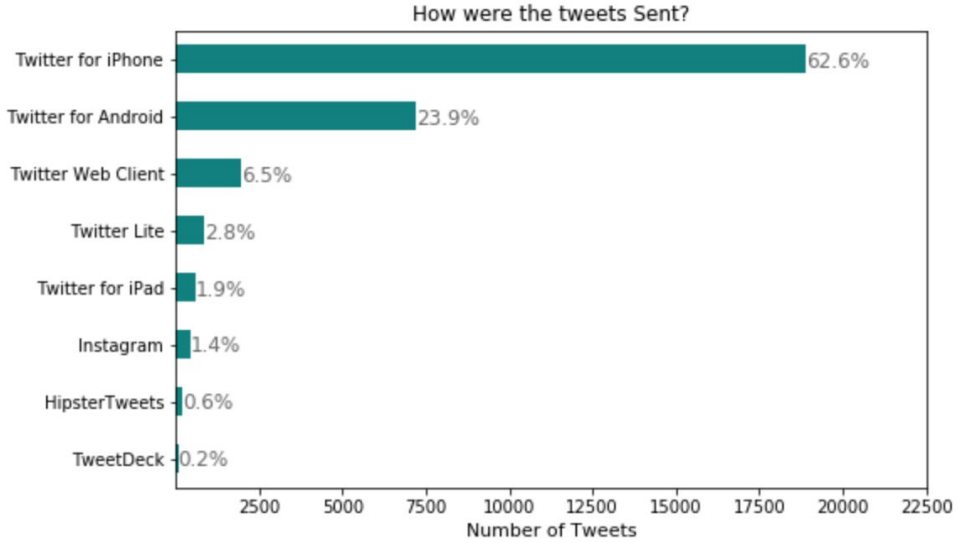
“KanyeWest tries to stand up for trump and JimmyKimmel shuts em up with facts”
Given score: 0.0

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.

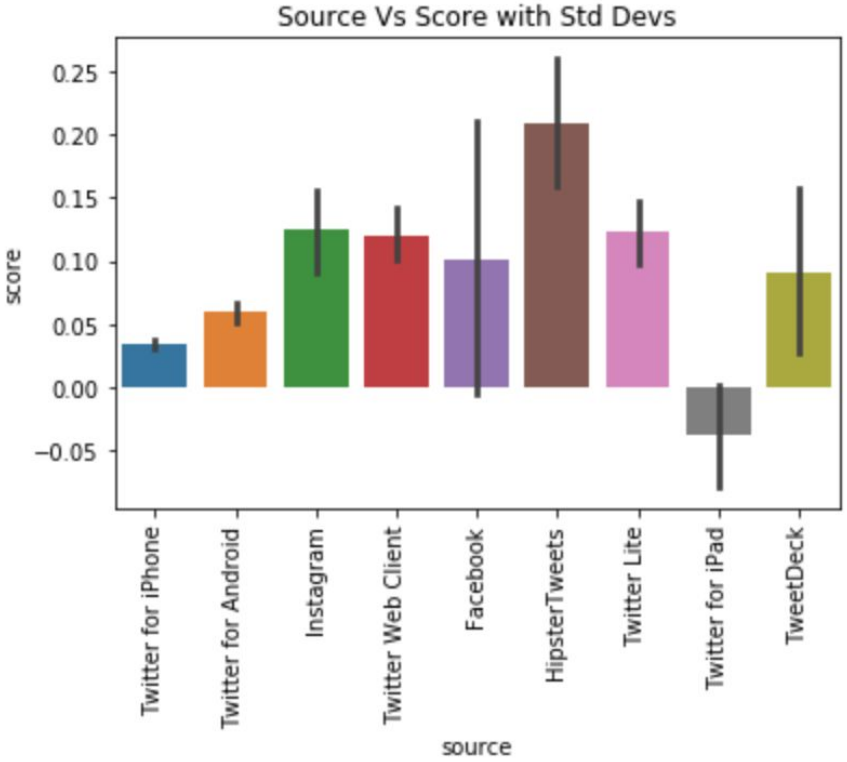
We will explore how this one tweet affects the whole dataset later on.

Tweet Sources:

The most popular way to tweet by far is using the iPhone app, with the Android app in second. Shown in below 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.



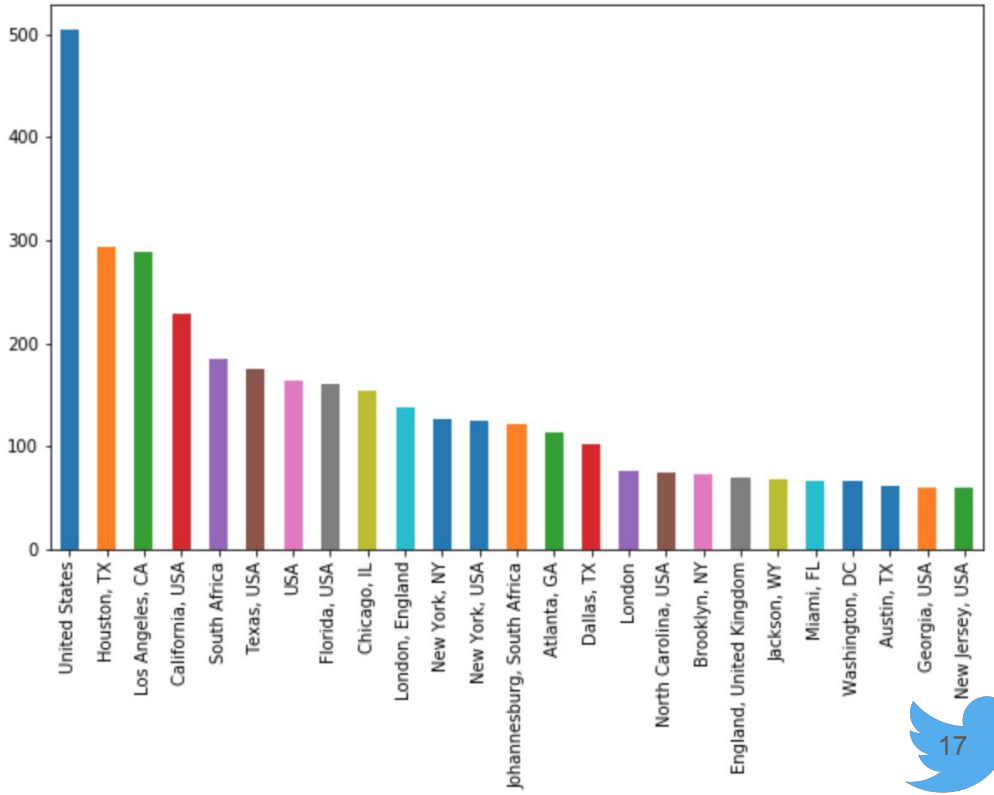
How do the different sources of tweets compare?



- It looks like all but one of the most popular sources were positive on average.
- Ipad users (only 1.9% of taken tweets) had such a small sample size, we aren't too concerned with it's low mean.
- The most common source, the iPhone, has the second lowest average at 0.04.
- The average score across all is about 0.0578.

Do the Tweets differ by Location?

- For user location, there is no set format and a lot of missing data. Most reported locations are in the US.
- 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.

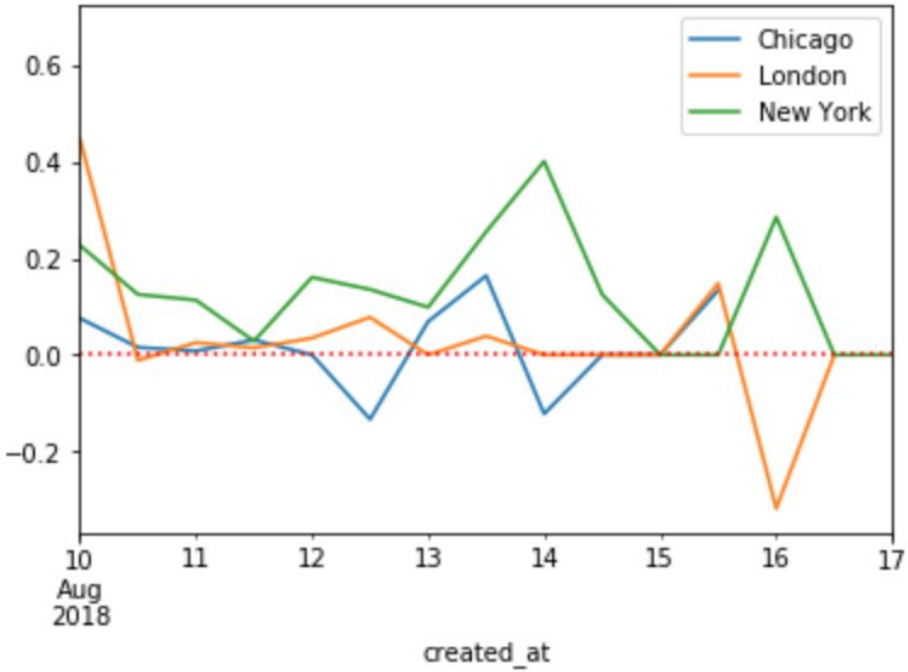


How does Chicago, Kanye's Hometown, Compare to Other Cities?

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.

Comparing Cities over Time



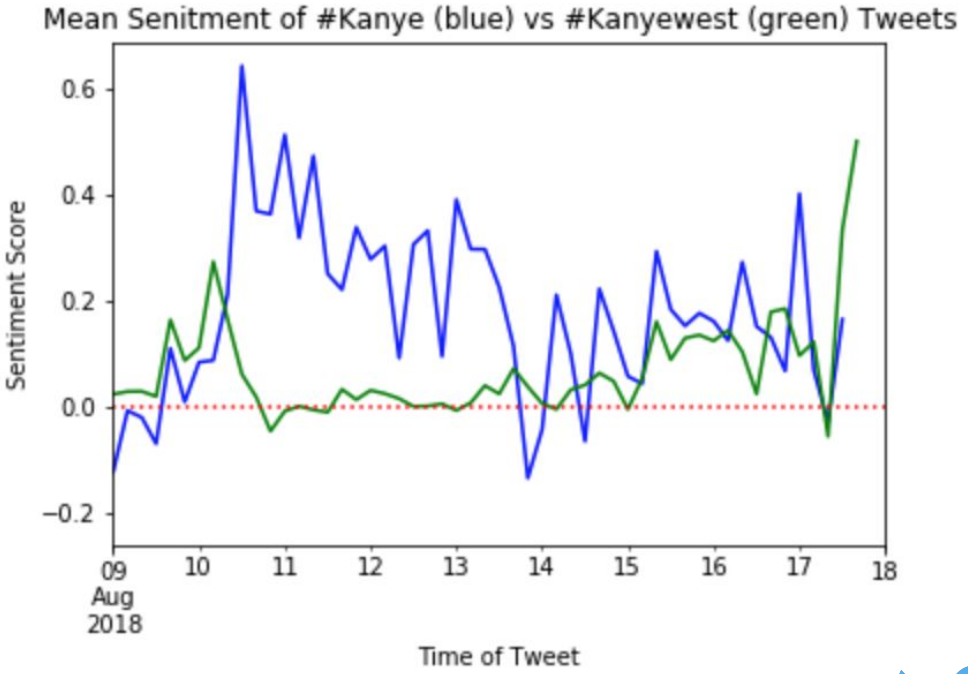
To the left shows the average score between Aug 10 and 17 for the three cities. Over time, New York had all positive averages, London had mostly with one sharp dip on Aug 16th, and Chicago dipped below zero twice.

#Kanye vs #Kanyewest

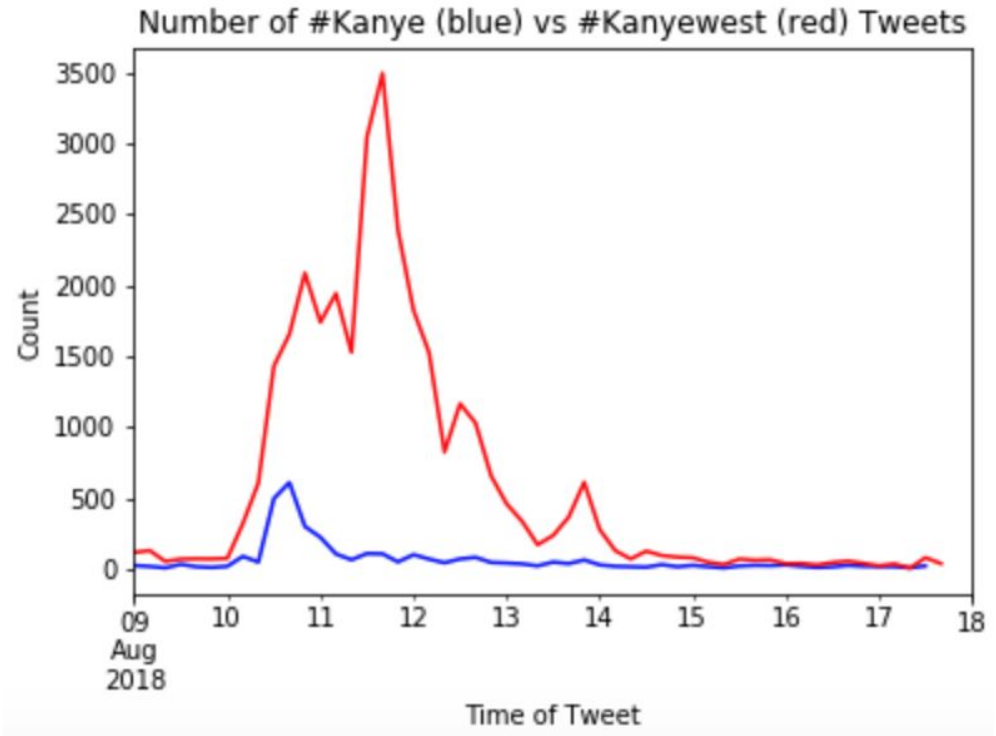
Because tweets were gathered using the query for both #Kanye and #Kanyewest, we are able to split the two and compare the use of each hashtag. There were 10x as many tweets with the hashtag #Kanye than with #Kanyewest.

To the right 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 Aug 10.



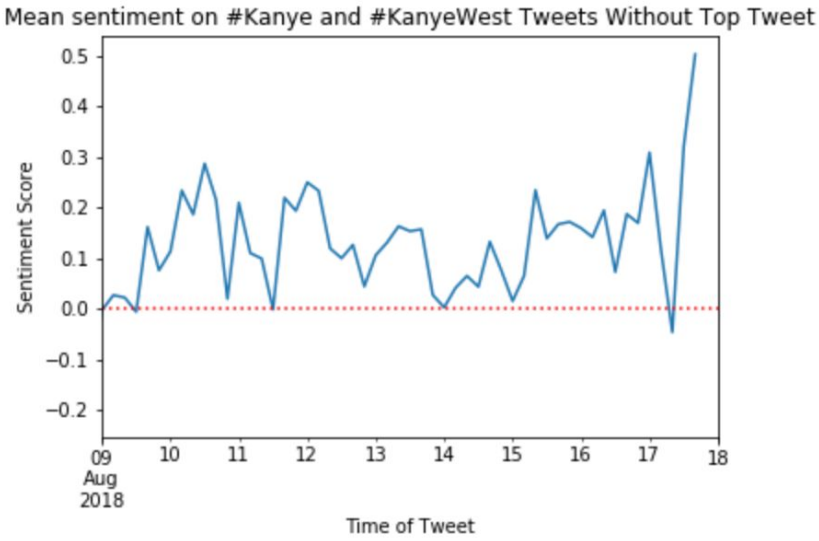
#Kanye vs #Kanyewest

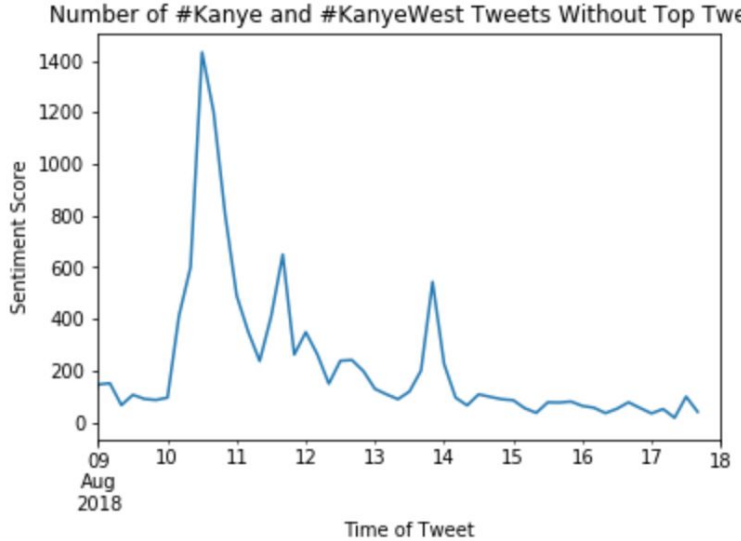


The number of tweets for each hashtag over time is shown to the left, where there is an drastic increase in tweets after the Jimmy Kimmel interview, but #Kanye tweets go back to the baseline rather quickly and #Kanyewest tweets bounce up and down for a few days.

A Look at the Data without Top Tweet

As mentioned previously, almost ⅔ of the tweets gathered are actually a retweet of one tweet and has the score of 0. Because of this, the score means and medians are further to zero than if the top tweet is excluded. The mean score is higher between Aug 11 and 14, when that top tweet was most viral. The count shows a similar shape, but much smaller number of tweets per hour as expected.



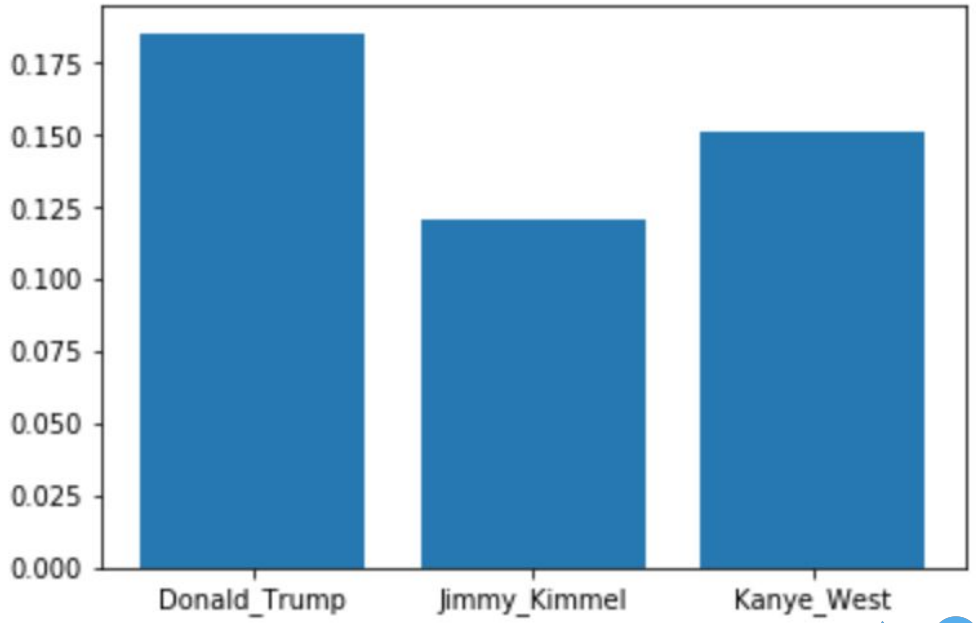


Tweets from Kanye and other Celebrities

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

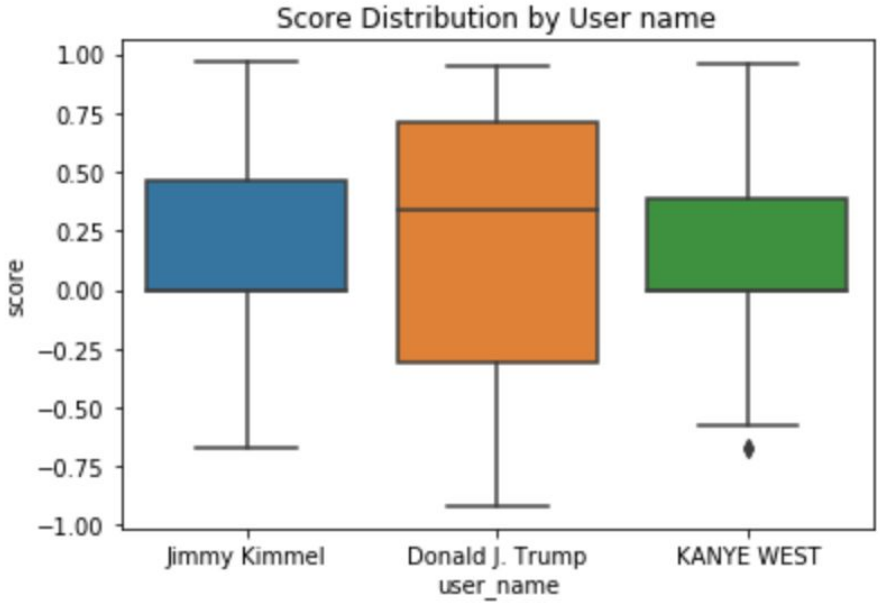
To answer this roughly 180 tweets from each celebrity were gathered to compare.

The average score for each celebrity is shown. 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.



Tweets from Kanye and other Celebrities

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.

Summary Pt 2

- 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.

- Accuracy of sentiment scores are extremely hard to test, and even people can disagree on sentiment. Tweets often include satire, sarcasm, media, and slang that can all throw off the score.

The full report and code can be found at:

https://github.com/SilasNeptune/Sentiment-Analysis