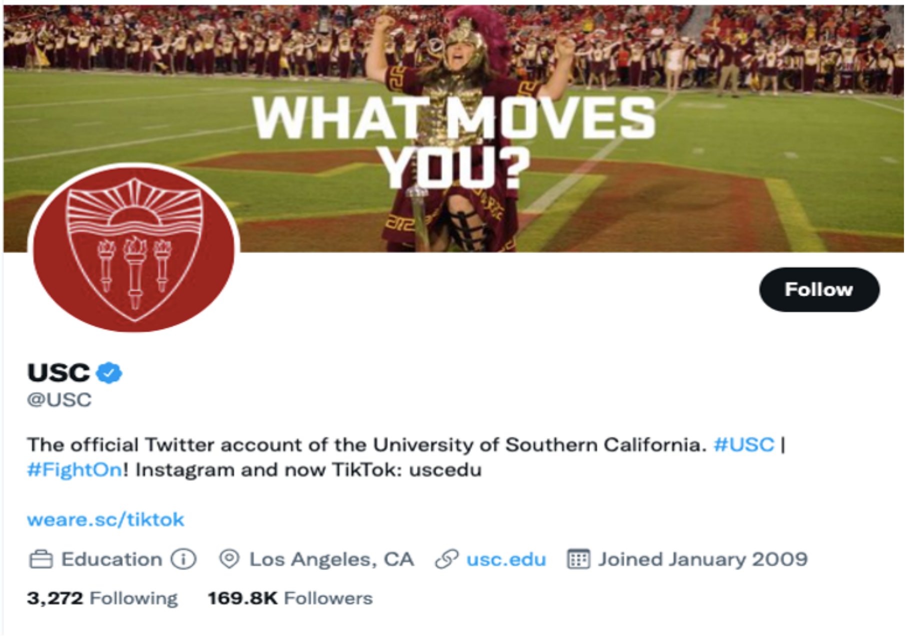
Professor Gene Lee
BAIT 508 Business Analytics Programming
Section BA1

Project Report: Social Media Analysis on USC with Twitter



USC

@USC

The official Twitter account of the University of Southern California. #USC | #FightOn! Instagram and now TikTok: uscedu

weare.sc/tiktok

Education Los Angeles, CA usc.edu Joined January 2009

3,272 Following 169.8K Followers

Group 20

Patrick Lu (luhaihao@student.ubc.ca, 29161759)

- Completed Question A1-A4 and WordCloud

Victor Chen (vchen110@student.ubc.ca, 33871997)

- Completed Question B1-B4 and Insight Analysis

Bingtao Zou(louiszou@student.ubc.ca, 42087262)

- Completed Question B5-B8 and Sentiment Analysis

Introduction

As social media platforms turn into the major outlet for people to express sentiments, it has become increasingly vital for companies and organizations to gather and analyze social media data for growth and audience engagement. For this project, we would like to explore the current key interests and attitudes towards the University of Southern California, from users on one of the most popular social media platforms, Twitter.

With Python, we were able to collect ten thousand recent tweets that contain the keyword USC. The tweets are preliminarily analyzed to locate the key indicators, such as the top 10 most popular hashtags. We then transformed the most common words that appeared in these tweets into WordCloud to provide a more powerful visual representation. Based on the results from the preliminary analysis, we calculated the polarity and subjectivity scores for the collected tweets, reflecting the online sentiment towards our keyword.

Keyword Selection and Data Collection

The keyword "USC" was selected since two members of our group graduated from the University of Southern California and we would like to take the opportunity to discover the online comments representative of the majority of social media users towards the prestigious university on the West Coast. First things first, we created a Twitter developer account to get access to huge amounts of tweets in an organized structure. With the TwitterCollector function, we managed to collect and compile 10,000 most recent tweets about USC.

To locate the most frequent user IDs on Twitter that discuss the topic 'USC', we initialized an empty list and created a for-loop to append any author id that is not already in the uniqueid list. To learn about how many unique users have tweeted about the topic, we also applied 'len' to confirm the length of the unique id list.

Collecting further information on all the authors is another important component of the preliminary analysis. On top of the unique author ID, we would also like to know other key information about the author, such as the number of followers they have, the number of tweets, if they are verified, and their user bio. To achieve this, we first created an empty list to include all the author information, after which we attempted to fetch additional author info for each author in the list using the append method. We first tried to suspend execution for 1 second after each attempt in the for loop. However, Twitter will still show a 'TooManyRequests' if the code is kept running for too long. Therefore, we adopted a 'try' and 'except' structure that allows the execution to stop for 15 minutes if TooManyRequests shows up and continue to run after the period. At the same time, we noticed there is an attribute error that interrupts the execution, which could result from one none type error. Similarly, we used a ‘try’ and ‘except’ structure to circumvent that error.

Preliminary Analysis

10 Most Popular Words with Stopwords

Interested in the key information associated with USC, we dug into the content of the tweets. Firstly, we would like to find out the ten most popular words. To do so, we started by creating a new list containing all the words in the tweet texts, which gave us a total count of 229,756. Then the ‘Counter’ function helped us determine the 10 most popular words with stop words, which are shown below.

1. the
2. RT
3. USC
4. and
5. to
6. a
7. of
8. is
9. for
10. this

As we can clearly see, a lot of these common words are meaningless. That’s why we need to remove these stopwords and dig real useful information from the data set.

10 Most Popular Words without Stopwords

We applied the stopwords library and added some of the frequent meaningless words that appear in the tweets to the list. Using the If.not function, we created a new list of words that are not on the stopwords list, called ‘words 2’. After that, the same ‘Counter’ function enabled us to find the 10 most popular words that are meaningful to our research, as demonstrated below.

1. USC
2. team

3. literally
4. people
5. women
6. getting
7. State
8. #USC
9. show
10. Black

From these ten words, we can get a bit more insight into what people are talking about when they talk about USC. Obviously, most of these tweets are about USC Football Team, especially since the team has been in great form since the start of this season. However, there's a limited amount of information that we can get from single words, hence we need to further explore the data set.

10 Most Popular Hashtags

In addition to the most popular words, another representative component of the tweets is the hashtags mentioned. Since all hashtags contain the symbol "#", we used the 'if.and' function to find the words that include '#' in them and stored them in a list. With the 'most.common' method, we ranked the frequency of the hashtagged words. The 10 most popular hashtags are the following:

1. #USC
2. #FightOn
3. #Fighton
4. #BREAKING
5. #Gophers
6. #durabarrierscam
7. #WhiteCollarCrime
8. #bernarddelvaux
9. #PerfectCrime
10. #BullsNation

We can see that the top 5 hashtags are all related to USC Football. Fight on is USC's slogan and is used in a lot of tweets related to USC athletic teams. Gophers is the name of the University of Minnesota Football team, and #breaking is commonly used in news.

10 Most Frequently Mentioned Usernames

For a better understanding of online engagement, USC should also be aware of the users that are active in USC-related discussions. To find out the most frequently mentioned usernames, we started by creating a list for all users that appeared in the tweets by finding words beginning with an '@'. With the 'most.common' method, we ranked the frequency of the most frequently mentioned usernames. The 10 most popular usernames are as following.

1. @StevieDwayne
2. @LaShea2019
3. @USC
4. @LincolnRiley
5. @TendentiousG
6. @CoachDee_USC
7. @Scott_Schrader
8. @Scott_Schrader'
9. @CBSSports
10. @CFBONFOX

The usernames include popular college football commentaries and the new head coach of USC. The fact that Twitter users frequently mention these usernames indicate that the football program is the most vulgar among other USC-related topics.

3 Most Common Sources of Tweets

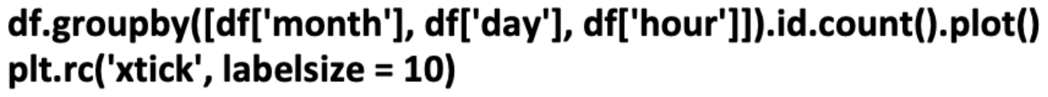
Additionally, obtaining the source of the tweets allows us to understand user behavior. Similarly, we used the 'for_in loop' function to collect and create a list of the sources of all the tweets in the data set by using the 'source' in recent_tweets['tweets']. According to the results from the 'Counter' function and 'most.common' method, the 3 most common sources are shown below.

1. Twitter for iPhone
2. Twitter for Android
3. Twitter Web App

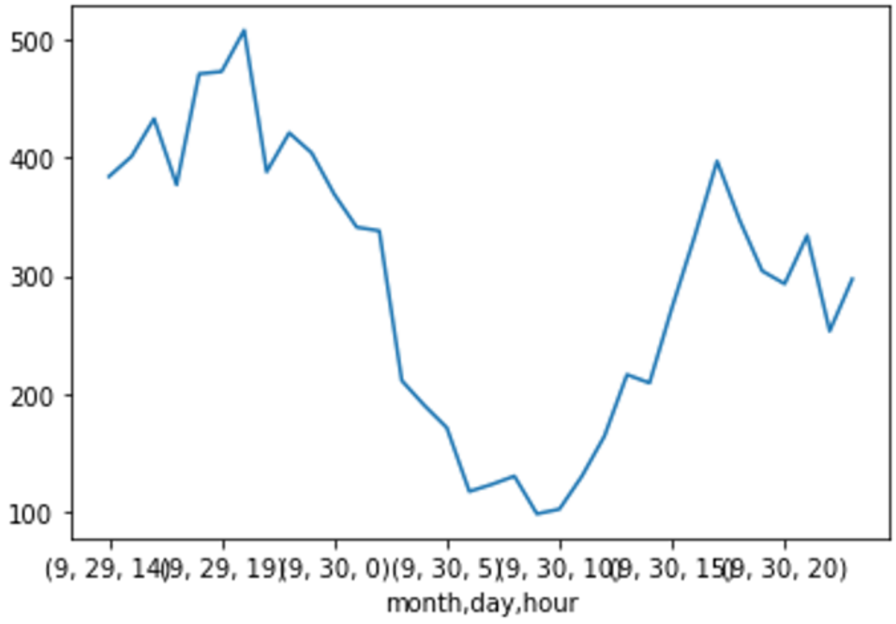
From the results, we can see that phone-based Twitter apps are most used in discussions about USC. The implications can lead USC to post more mobile-friendly content on its Twitter channel.

Time Trend of Tweet Counts

A line chart can better illustrate the relationship between time and the number of tweets during a period. Firstly, we converted recent_tweets['tweets'] into a data frame for analysis. Using the 'groupby' method and the code below, we plotted a line chart based on month, day, and hour.



```
df.groupby([df['month'], df['day'], df['hour']]).id.count().plot()
plt.rc('xtick', labelsize = 10)
```



As shown in the graph, there is a significant drop in the number of tweets about USC after 12pm pacific time, this reflects that most people discussing USC on Twitter live in North America, specifically in the pacific time zone.

3 Most Influential Tweets

The influence score of a tweet is another great tool for USC to understand user sentiment. With the input from "public_metrics" in the tweet data, our code used the "for_in" loop to create an influence score data frame. With the "sort_values" method, we were able to locate the top 3 highest influence scores. The first two tweets contain the same text.

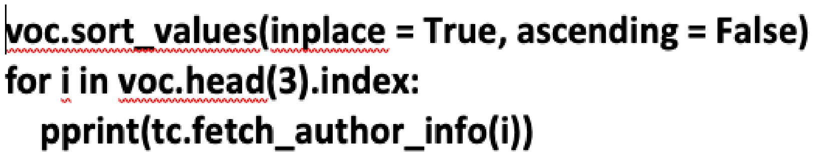
1. RT @FootbaIIism: Kids today will never understand how insane USC Reggie bush was
2. RT @FootbaIIism: Kids today will never understand how insane USC Reggie bush was

3. RT @SportsCenter: Remember that time Vince Young took it to the house against USC for a title? Only 10 days til CFB is back.

The 3 most popular tweets tell us that people are reminded of the glorious history of the USC Football Team in the past.

3 Most Vocal Authors

To find out the most frequently tweeting authors in the tweet data, we first grouped the data frame by author ID and counted the number of tweets by each of these authors. Then, we sorted the result by the count and printed the information of the first three.



```
voc.sort_values(inplace = True, ascending = False)
for i in voc.head(3).index:
   pprint(tc.fetch_author_info(i))
```

Author Username	User ID
'uscfootball'	16172976
'SharingNewDeals'	1381352768997367826
'greg_greg76'	1214052011190022144

From our analysis, the outlier ‘SharingNewDeals’ is among the 3 most vocal authors because ‘USC’ is a common part of promotion codes that are used by companies.

3 Most Influential Authors

To calculate and find the 3 most influential tweet authors, we began with converting author information into a data frame. After adding up the sum of “followers_count”, “following_count”, “listed_count”, and “tweet_count” with the “append” method, we added the list to the data frame. After sorting, the result is listed below.

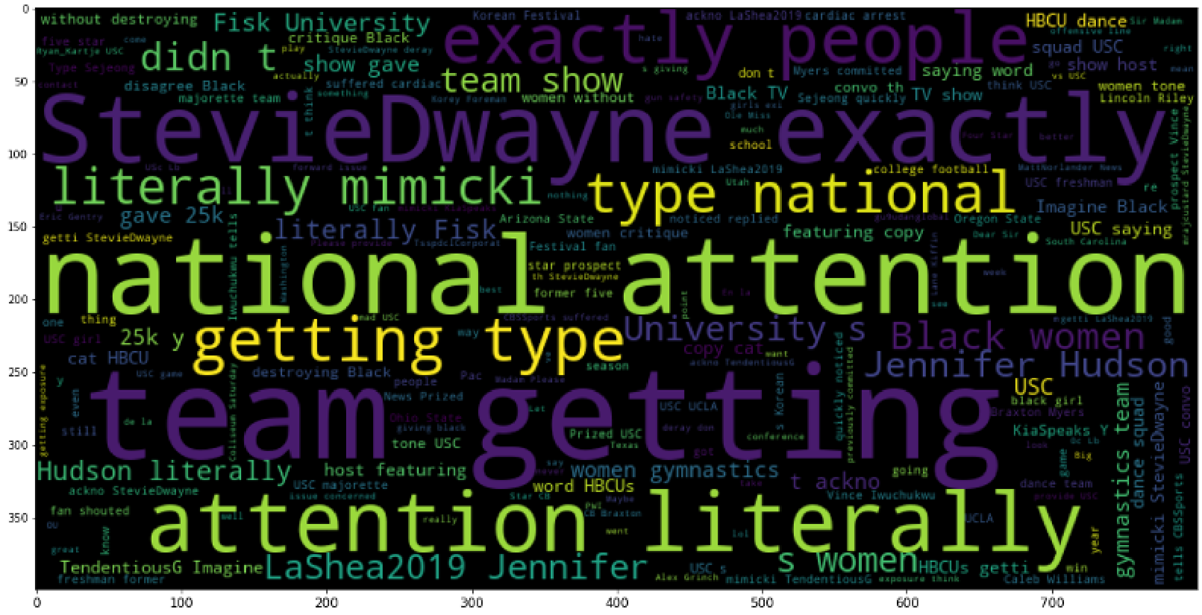
Username	Influence Score
Japan_lawson	110,128,763
suntory	21,543,016

cas_2050001283	18,603,038

Both of the two most influential authors that have mentioned USC are reply bots that automatically reply to thousands of tweets a day and have replied to users with the words USC in their ids. Because authors will appear in the list as long as they mention USC once in their tweets, it's very easy for this data to be contaminated.

Word Cloud

A Word Cloud graph can demonstrate the popularity of words or phrases associated with USC by making the most frequently used words appear larger and bolder. To create a Word Cloud graph, we first put all the words in the tweets into a single string with the previously created "words2" list. Then, using the 'Wordcloud' function, we generated a graph that illustrates the most popular words that are mentioned in USC-related tweets. For better demonstration, we configurated the width to 800 and the height to 400.



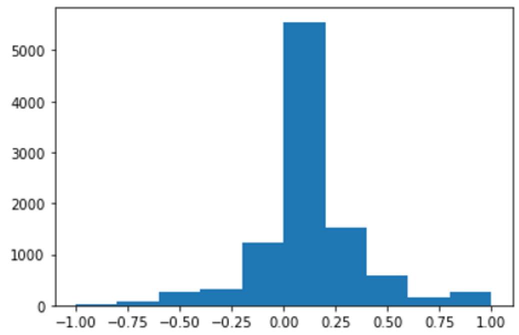
From this word cloud, we can see that one of the most common phrases that appeared in these tweets is 'national attention.' This is understandable because the USC Football team has been on an incredible run and is currently 5-0 in the season. This presents a sharp contrast with the last season when they performed poorly.

Sentiment Analysis

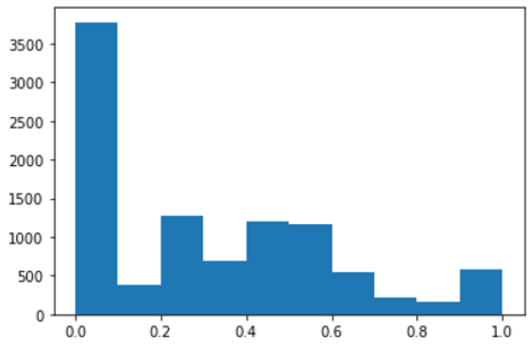
To calculate the average polarity and subjectivity scores for the tweets we collected, we initialized two empty lists to fit the polarity and subjectivity of the tweets respectively, named ‘pol’ and ‘sub’. Using ‘for_in’ loop and ‘TextBlob’, we attached the polarity and subjectivity of the tweets to the respective lists. Next, we add the lists to the data frame using the ‘df’ function. Finally, we can just print out the average polarity and subjectivity of the tweets with the ‘mean’ method in the data frame. The average polarity score is 0.083 and the subjectivity score is 0.298.

Histograms are a great way to visualize the polarity and subjectivity score distribution. Since the data frames use the scores as the X-axis and the number of tweets as the Y-axis, using the ‘plt.hist’ function, we can plot two histograms showing the distribution of polarity and subjectivity.

Polarity Histogram



Subjectivity Histogram



To find out the most positive and negative tweets about USC, polarity can be used as an indicator. First, we sorted the data frame by polarity in a descending order using the following code:

```
df.sort_values(by = ['polarity'], inplace = True, ascending = False)
```

Next, with the ‘iloc’ function, we can locate the tweets (in the 7th column) that are on the top 3 and bottom 3 of the data frame.

Positive

1. USC Hosting Another Impressive Group of Visitors at the L.A. Coliseum Saturday (10/1) ❕ 🔖 👀\n\nTwo Official Visitors ☑️\n\nT…
2. USC vs. Arizona State prediction, odds, line: 2022 college football picks, Week 5 best bets from proven model https://t.co/Xsv2v2lEj9 ▶️ https://t.co/qwq5GEGtOp #OnlineBetting #SportsBetting #NCAAF

3. USC Hosting Another Impressive Group of Visitors at the L.A. Coliseum Saturday (10/1) ❕ 🔖 👀\n\nTwo Official Visitors ☑️\n\nT…

Negative

1. @TsspdclCorporat There is no electricity in dabirpura and when its coming, the fluctuation is terrible.\nUSC-110008729
2. USC jr looks pathetic
3. Kids today will never understand how insane USC Reggie bush was\n https://t.co/m4gSO1hL6B

Most of the most positive and negative tweets are related to the USC football performance. People are happy when USC football is performing well or have a high chance of winning, and people are angry when USC loses in football games or does not perform as good as before. Interestingly, the third most negative tweet as rated by textblob was actually praising the player.

Insights

Based on the analysis we conducted, it is evident that the main topic discussed around USC is its football program, which is not surprising given the school's recent upgrade from Pac-12 to the Big 10. This trend is reflected by the most popular hashtags, most frequently mentioned usernames, most influential tweets, most vocal authors, Word Cloud, and the sentiment analysis, which either includes names of famous players on the USC football team or the previous and current coaches for the football program. The amount of sports-related information is so great that it makes up the majority of the tweets.

On the other hand, the most-influential-author metric is not appropriate to connect to our keyword. Since the top 2 authors happen to be two Japanese convenience stores that sent out automatic replies and happened to mention users that had the letters USC in their id. In this case, it is better to have a larger range for this category, such as the top 100 most influential authors, which will more accurately reflect the user profile concerning USC.

Regarding the sentiment analysis, it can be interpreted that though improvement can be made, most people generally hold neutral or slightly positive opinions towards USC, considering the 0.08 polarity score. The subjectivity score is, similarly, relatively low, at almost 0.3, meaning that the tweets contain more factual information than personal opinions. The low subjectivity is reasonable, as USC is a prestigious educational institution and enjoys decent media coverage, which usually provides true pieces of information.

Future Project Idea

Provided that we have other unstructured data, we would be able to conduct a competitor analysis with social media. Hypothetically, if we have a competitor selling a product that targets the same audience as us, we would be able to gather social media feedback on consumer sentiment on their product, which can then be used to better our product. For example, content creators on Tik Tok or YouTube often post unboxing videos and share their firsthand experiences using a product. Once we grab the keywords from their content, we can conduct sentiment analysis to find out the strengths and weaknesses of the competitor's product. The results can be used for not only marketing purposes but R&D improvements as well.