US News and World Report Ranking using Social Media Sentiment of UCF and other universities

Alshahrani, Lamia; Beema, Umapathy; Sreedhar, Shriram; Waddington, Nicholas
Supervisors: Dr.Nassif, Edwin; Dr. Yousefi, Niloofar

ABSTRACT

Sentiment analysis is one of the useful natural language processing technologies to know and analyze people's opinions on a particular topic. We observed sentiment analysis on data scraped from Twitter which is one of the powerful social media platforms where most people express their opinion, share their daily activities or thoughts. This helped us to achieve our goal which is to analyze the sentiment of the users on different universities around the United States and find if there are links between that and the US News ranking.

KEYWORDS

Sentiment Analysis, Tweepy, and Word-Cloud.

1 Introduction:

Twitter is a social media platform that allows users to communicate and share their opinions about any topic or trend which makes it useful for so many businesses to see the user's feedback about their products. In this project, we have studied the sentiment analysis of four different universities compared to the University of Central Florida in ranking. The data in this project are scraped from Twitter on daily basis for one semester-long based on certain hashtags and keywords for each university separately. Since this is big data, we used *AWS* to store the data in the cloud. The sentiment analysis was applied to the text of the tweet after cleaning the text and making sure it does not have any unwanted characters. Sentiment analysis was applied using *the TextBlob* library. It is a library that calculates the polarity between[-1,1] if it is 1 means it is positive whereas if it is -1 it is a negative sentiment as well as it calculates the subjectivity of a text. Also, we apply some visualization such as word-cloud to see the most frequent words in the text and find the Bigrams and Trigrams of the text.

2 Background and Related work:

This section show previous researches that have been done to analyze similar topics using similar approaches or methods. Looking at some previous papers that helped us understand the methodology of the project and the sentiment analysis idea. Reviewing this paper[1] which was describing a similar concept of our project. The paper[1] helped us to understand how opinion mining works which is sentiment analysis for specific businesses and how it can matter. The paper studied the sentiment analysis of Twitter data for two restaurants KFC and McDonald's to find out which one is better and popular using their sentiment analysis from the customers. This has the same idea as our project since we do have five different universities to study their sentiment and to find the correlation between that and their rating. This paper[2] used different machine learning techniques on Twitter data that was already manipulated and labeled as positive, negative, and neutral it uses *Naïve Bays*, *SVM*, and *Semantic Analysis* (WordNet). It is very insightful to review this paper to look at how they used multiple Machine learning approaches to find the highest accuracy which was Semantic Analysis (WordNet). The last paper[3] was very interested since it uses the same library that we decided to use which is *TextBlob* to test the sentiment. Along with that, it uses the scraped Twitter data and clusters the results to group them into three different categories. To approach this clustering *K-mean* was appliead. Studying these three papers helped us to know how to approach our project and help students to think about how to solve the problem logically and to know which algorithms can be more suitable for the problem.

3 Data Processing :

The section will show steps and processes that have been done to use a clean and accurate dataset for the analysis. First, collecting the datasets depends on the keywords and hashtags, and the idea of deciding which keywords to use is by creating different buckets like research, sports, social life, and so on and collecting the hashtags from Twitter based on that. Then we start cleansing the text by removing any unwanted characters from the scraped data such as punctions, and emojis. Also, applying lemmatization, stemming, and removing the stop words. Now the data is ready to be used and analyzed.

4 Methodology:

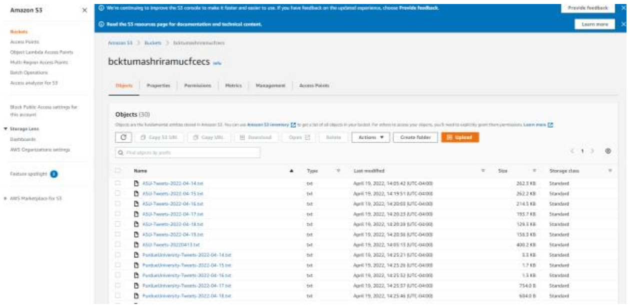
To answer the research questions, we implemented different algorithms and methods on the Twitter sentiment analysis data such as Sentiment analysis, Word-cloud, Bigrams, and Trigrams. Along with tweets scraping and AWS configuration. In this analysis, we scraped five different universities' tweets which are the University of Central Florida, University of Florida, University of South Florida, University of Arizona, and Prude University based on six different categories which was Research, Academic, Alumni, Sports, Faculty, and Crime based on that the data were collected and gathered.

4.1 Tweepy:

This analysis used the data that were scraped from Twitter. We used *Tweepy* in this project to scrape the data. We scraped it day by day based on the keywords and hashtags. The scraped tweets were a JSON file that contain a metadata that has all the important information about any tweet. At this stage of the project, we only do use the text of the tweet.

4.2 AWS configuration:

AWS is a powerful cloud computing tool. We used AWS due to the high volume of data that has been collected daily for one semester long. In this analysis, we created an *EC2* instance and used *S3* buckets to store the text file for the data. Automated this task by using a python script to scrape data using Tweepy to gather data day by day and used these text files in the further analysis. Figure[1] shows the S3 storage and the text files for these data.



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Read the S3 resources page for documentation and technical content.

bcktumashriramucfcecs

Objects (30)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
ASU-Tweets-2022-04-14.txt	txt	April 19, 2022, 14:05:42 (UTC-04:00)	262.3 KB	Standard
ASU-Tweets-2022-04-15.txt	txt	April 19, 2022, 14:19:51 (UTC-04:00)	262.2 KB	Standard
ASU-Tweets-2022-04-16.txt	txt	April 19, 2022, 14:20:03 (UTC-04:00)	214.5 KB	Standard
ASU-Tweets-2022-04-17.txt	txt	April 19, 2022, 14:20:23 (UTC-04:00)	195.7 KB	Standard
ASU-Tweets-2022-04-18.txt	txt	April 19, 2022, 14:20:28 (UTC-04:00)	129.5 KB	Standard
ASU-Tweets-2022-04-19.txt	txt	April 19, 2022, 14:20:56 (UTC-04:00)	158.3 KB	Standard
ASU-Tweets-20220413.txt	txt	April 19, 2022, 14:05:13 (UTC-04:00)	400.2 KB	Standard
PurdueUniversity-Tweets-2022-04-14.txt	txt	April 19, 2022, 14:25:21 (UTC-04:00)	3.3 KB	Standard
PurdueUniversity-Tweets-2022-04-15.txt	txt	April 19, 2022, 14:25:26 (UTC-04:00)	1.7 KB	Standard
PurdueUniversity-Tweets-2022-04-16.txt	txt	April 19, 2022, 14:25:32 (UTC-04:00)	1.3 KB	Standard
PurdueUniversity-Tweets-2022-04-17.txt	txt	April 19, 2022, 14:25:37 (UTC-04:00)	754.0 B	Standard
PurdueUniversity-Tweets-2022-04-18.txt	txt	April 19, 2022, 14:25:46 (UTC-04:00)	604.0 B	Standard

Figure 1

4.3 Sentiment Analysis:

Sentiment analysis is an algorithm that can be used to describe a topic that is subjective and objective to find if it is positive, negative, or neutral results from a specific topic. The ultimate goal of this project is to know the sentiment or in other words the feedback of the people on these universities and if this can affect the US news ranking. Based on six different buckets/categories. Sentiment Analysis was the primary algorithm that we used in this analysis. After scraping the data from Twitter. We focused on the text column and the timeframe as the scraped data would bring the metadata for any hashtag or keyword inserted. Some cleansed techniques were applied to the text column such as removing the spaces, RT before any retweets, and removing punctuations. Also, stemming, lemmatization, and converting the text to lowercase. Then saved this result into a new column called text. After we were done with text cleansing, we start calculating polarity, subjectivity, Sentiment negative-positive, and neutral for each one of these we created a new column in the data frame. Figure[2] shows a plot of a simple pie chart with the keyword Go Knight which belongs to UCF to see the sentiment of this word on Twitter. However, the data in this pie chart was only for 10k tweets for one week so it did not have a high percentage. After that, we started to see and compare the results for other universities in different areas and compare them to each other. In figure[4] a histogram was used to see the difference between those groups. As a result, we can see in the research category we have ASU with the highest positive tweets as well as the alumni category. In the Sports category, UCF has the highest number of positive tweets. Finally, in the Crime category, we can see several negative tweets on these five different universities. We can conclude these results are varied depending on each university and it is popularity and the number of tweets for each one of them.

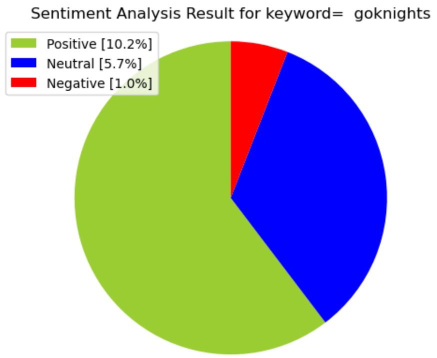


Figure 2

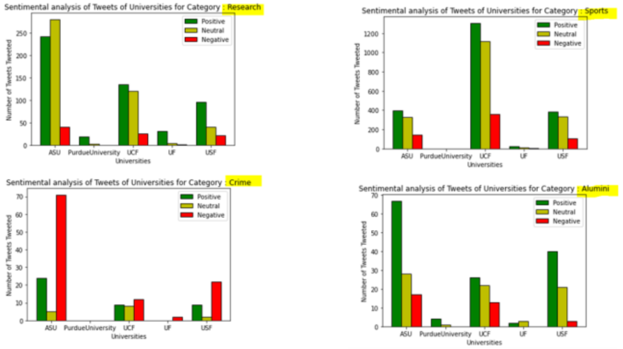


Figure 3

4.4 Word Cloud:

To form the word cloud, we import the Word Cloud library, make a data frame of the cleansed text and make a series of that text then apply a word cloud of the frequent words that we have. For this paper, I show some samples of three different keywords of different categories. Figure[1] shows a word cloud of UCFCECS which is a college of Engineering and computer science at the University of Central Florida. In Figure[2] shows a word cloud of different categories which is UCF sports and Finally in Figure [3] it was for UCF alumni. The goal of the Word cloud function is to visualize the most prominent words that were repeated in the tweets for a given keyword or hashtag. Word cloud is an effective way to visualize the most important words used in tweets based on their frequency.

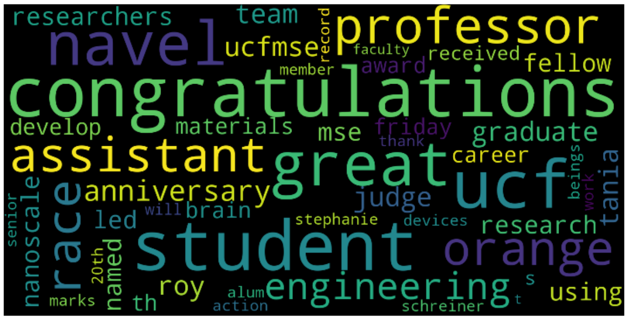


Figure 4

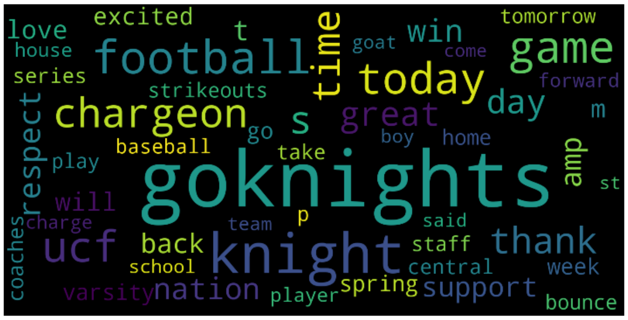


Figure 5

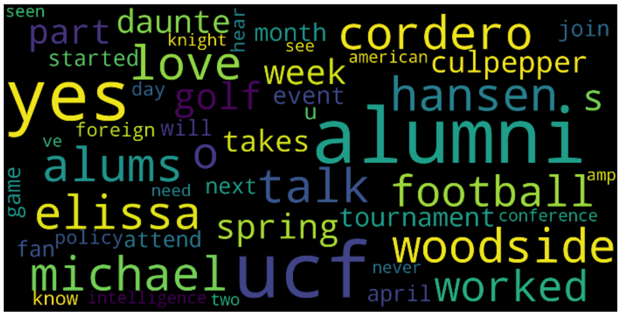
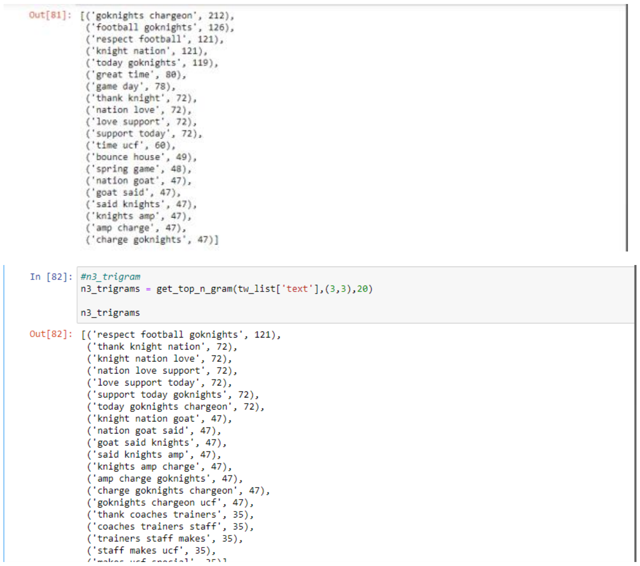


Figure 6

4.5 Bigrams and Trigrams:

Bigrams and Trigrams are effective ways to analyze a text data and find what are the words that commonly appears together as word associations. Bigrams would look at the combination of two words while trigrams are the combination of three words. In this analysis, we did filter out the stop word from the text which are the words that are commonly used in English, but it does not have significant importance such as are, they, is, etc. Since we used the library *TextBlob* there was an interesting way to see Bigrams and Trigrams. First, we used the function *CountVectorizer* which takes two parameters *stopwords* list and n-gram which can be 2 and 3 in our case 2 for Bigrams and 3 Trigrams this function returns a matrix of n-grams. Then we calculate it is own frequency by applying the sum on this matrix after converting it to an array. In figure[7] we can see a list of Bigrams on the first output and a list of Trigrams on the second output of the keyword “Go knights” which belongs to the category Sports.



```
Out[81]: [('goknights chargeon', 212),
          ('football goknights', 126),
          ('respect football', 121),
          ('knight nation', 121),
          ('today goknights', 119),
          ('great time', 80),
          ('game day', 78),
          ('thank knight', 72),
          ('nation love', 72),
          ('love support', 72),
          ('support today', 72),
          ('time ucf', 60),
          ('bounce house', 49),
          ('spring game', 48),
          ('nation goat', 47),
          ('goat said', 47),
          ('said knights', 47),
          ('knights amp', 47),
          ('amp charge', 47),
          ('charge goknights', 47)]

In [82]: #n3_trigram
         n3_trigrams = get_top_n_gram(tw_list['text'],(3,3),20)

         n3_trigrams

Out[82]: [('respect football goknights', 121),
          ('thank knight nation', 72),
          ('knight nation love', 72),
          ('nation love support', 72),
          ('love support today', 72),
          ('support today goknights', 72),
          ('today goknights chargeon', 72),
          ('knight nation goat', 47),
          ('nation goat said', 47),
          ('goat said knights', 47),
          ('said knights amp', 47),
          ('knights amp charge', 47),
          ('amp charge goknights', 47),
          ('charge goknights chargeon', 47),
          ('goknights chargeon ucf', 47),
          ('thank coaches trainers', 35),
          ('coaches trainers staff', 35),
          ('trainers staff makes', 35),
          ('staff makes ucf', 35),
          ('makes ucf special', 35)]
```

Figure 7

5 Conclusion and Future Work:

This paper highlighted one of the powerful text mining algorithms that analyze the opinions or feedback from the customers/users which is sentiment analysis of Twitter scraped data by using keywords and hashtags based on a logical way of thinking and some categories that belong to the topic. Sentiment analysis was applied to five different universities to see the correlation between the sentiment and the US ranking for these universities. This paper also shows some visualizations such as word-cloud to see the most frequent words for any particular hashtag or keyword. It also shows a Bigram and Trigram of the words that usually appear to gather which is one of the ways to analyze text. The future work of this project will be to use as much as we can of data which can improve the accuracy results of the analysis and use different algorithms and pipelines for further analysis. In addition, we can use clustering algorithms to find what are the universities that can be on the same rating and save it in the pipeline then apply the sentiment based on that.

REFERENCES

[1] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1-4, DOI: 10.1109/ICCISci.2019.8716464. (Link)

[2] G. Gautam and D. Yadav, "Sentiment analysis of Twitter data using machine learning approaches and semantic analysis," 2014 Seventh International Conference on Contemporary Computing (IC3), 2014, pp. 437-442, DOI: 10.1109/IC3.2014.6897213. (Link)

[3] S. Ahuja and G. Dubey, "Clustering and sentiment analysis on Twitter data," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), 2017, pp. 1-5, DOI: 10.1109/TEL-NET.2017.8343568. (Link).