

## Twitter Sentiment Analysis

### Analysis Overview

A sentiment analysis can be defined in various ways, but it can essentially be defined as a computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisals, affects, views, emotions, etc., expressed in text. Sentiment analyses are more widely used in the industry. As a result, for this assignment, a sentiment analysis was conducted on Twitter against the query, "Donald Trump". Prior to running a sentiment analysis on Twitter for this query, I created a new application on Twitter Apps and obtained a Customer\_key, Customer\_secret, Access\_token, and Access\_token\_secret for downloading the tweets from Twitter. As shown in Figure 1, once the script to download the tweets regarding "Donald Trump" was executed, I was left with a dictionary (dataset variable) with a total of 8,000 tweets.

Donald Trump  
8000  
Sleeping for 15 minutes

Figure 1

To view the dataset in greater detail, a breakdown of the downloaded tweets was provided in a DataFrame table by using the Pandas library. As you can see in Figure 2, the tweets were broken down into columns based on each tweet's creation date, text, hashtags, username, and user followers count. The table also served as a source of verification to make sure the text from each tweet included the query of "Donald Trump". From the information that was provided from the downloaded tweets, a sentiment analysis was conducted on these tweets using both the NLTK library and Text Blob library.

	created_at	text	hashtags	username	user_followers_count
0	2018-04-01 00:47:36	@MartyWill1173 @dubs_run_nba @Gazpickup1 @Ent...		Nicknock97	2959
1	2018-04-01 00:47:35	RT @activist360: You know the absurdity of Tru...		SlyMagellen	2212
2	2018-04-01 00:47:35	RT @activist360: You know the absurdity of Tru...		TAPOUTWATCHER	124
3	2018-04-01 00:47:35	RT @jessasella: Donald Trump is a rapist. #Tru...	Trump,TrumpOut,NotMyPresident	ChrisKosowski1	1781
4	2018-04-01 00:47:34	RT @ZlatanFacts: Donald Trump has had plans or...	ZlatanFacts	frdausse_	403
5	2018-04-01 00:47:34	RT @mymbc15: A majority of young people believ...		ATM25	545

Figure 2

### NLTK Vader Sentiment Analysis

According to Wikipedia, the Natural Language Toolkit (NLTK) can be defined as a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language (Wikipedia, n.d.). To run a sufficient sentiment analysis on the dataset of tweets using NLTK, several steps were taken to train the classifier. Each document was converted to represent a tuple and each sentence was tokenized so that it could be represented by a list of strings. Each subjective and objective instance was then split separately to keep a balanced uniformed class distribution in both training and testing sets. I then used the simple unigram word feature to handle negation and applied features to obtain a feature-value representation of the dataset. As a result, the classifier was able to be trained based on the training set and output the evaluation result. I was then able to continue the sentiment analysis by using the nltk.sentiment.vader to analyze pieces of text and classify sentences on the text of each tweet that was downloaded.

As displayed in Figure 3, a polarity score that included calculations for compound, negative, neutral, and positive sentiments was printed for each tweet. After analyzing the tweets, I found that the majority of the tweets were given a neutral sentiment. In addition, there were a few tweets that

```
In [23]: #NLTK sentiment analysis
for sentence in dataset['text']:
    print(sentence)
    ss = sid.polarity_scores(sentence)
    for k in sorted(ss):
        print('{}: {}'.format(k, ss[k]))

pos: 0.0,
RT @talbertdwan: @HopeRad1 @realDonaldTrump @FoxNews Donald Trump is a moronic, self-absorbed, incompetent, insec
ure, racist, demagogue w...
compound: -0.9274,
neg: 0.553,
neu: 0.447,
pos: 0.0,
RT @kimhunter22: @WillResistAdmin @DaleLee919 @RogueSRAdvisor Also me. Equally sickening was the Christian defens
e of it.
https://t.co/b...
compound: -0.4404,
neg: 0.18,
neu: 0.741,
pos: 0.079,
RT @my_dickind: In the era of Trump, some days it feels like our country is running out of irony and hypocrisy. h
https://t.co/RND377K40y
compound: 0.3182,
neg: 0.051,
neu: 0.844,
pos: 0.105,
```

Figure 3

had a compound of positive and negative sentiments. If I had to provide an approximate percentage, I would propose that 95% of the tweets had an overall neutral sentiment, 4% had an overall negative sentiment, and 1% had an overall positive sentiment. After further analyses of the tweets and their corresponding polarity, I believe the NLTK sentiment analysis was more so accurate. For example, the tweet displayed in Figure 3 that has a red arrow beside it had an overall negative sentiment analysis. It can be confirmed that the sentiment is accurate because the tweet includes words that have a negative connotation in the English language such as: incompetent, racist, and moronic. In addition, the tweets that have a yellow arrow beside them in Figure 3 were given an overall neutral sentiment. I believe these sentiments are accurate because an overall positive or negative sentiment cannot be assumed from the two tweets. For example, one tweet read, "In the era of Trump, some days it feels like our country is running out of irony and hypocrisy." It is difficult to gather a positive or negative sentiment from this tweet because the tweet does not heavily relay strong emotions such as joy or anger. Some could view the tweet as positive, while others could view the tweet as negative. To further analyze the sentiments of each tweet I also conducted a sentiment analysis using the TextBlob library.

### TextBlob Sentiment Analysis

TextBlob is a Python library for processing textual data that provides a simple API for diving into common natural language processing tasks. The sentiment analysis that was run using the TextBlob library included the sentiment property that returned a named tuple of the form "Sentiment (polarity, subjectivity)". The polarity score was a float within the range [-1.0, 1.0] and the subjectivity was a float within the range [0.0, 1.0] where 0.0 was very objective and 1.0 was very subjective (TextBlob, n.d.).

After analyzing the sentiments of tweets using the TextBlob library, I found quite a few discrepancies in the accuracy of the polarity of each tweet. In my opinion, I believe there were a large number of polarities that had been misclassified because the overall emotion and opinion of the tweet was misunderstood. For example, in Figure 4, the tweet outlined in red was given a negative polarity, but there was no negative vocabulary used that signified that the overall tweet should be classified with a negative polarity. I believe the classifier took the term "hard" to emphasize negative polarity, while the term "hard" was not used in a negative context for this specific tweet. Rather, the term "hard" was coupled with the phrase "working hard" to strengthen a compliment. In my opinion, the polarity should have been neutral. In contrast, there were sentiment results that were very accurate. For example, in the tweet outlined in yellow in Figure 4, the polarity was classified as negative. I agree with this polarity because it includes the signifiers, "idiot" and "no". This further justifies that the overall opinion and emotion of the tweet was one of anger and possibly frustration.

```
RT @StorminSpartin: What an Idiot. He has no clue as to the Importance
a Community College plays in a Family budget when you have 6 kids to...
Sentiment(polarity=-0.8, subjectivity=0.8)
-----
RT @SethAbramson: 3/ One thing this *does* suggest to me is that Muelle
r agrees that Roger Stone's repeated protestations that in the final...
Sentiment(polarity=0.0, subjectivity=0.0)
-----
RT @bessbell: Donald Trump declaring it National Sexual Assault Awarene
ss Month is like Donald Trump declaring it National Sexual Assault A...
Sentiment(polarity=0.5, subjectivity=0.8333333333333334)
-----
RT @CREWcrew: Donald Trump complimented Ivanka at an event for "working
hard." She must be working pretty hard--the Trump Organization pai...
Sentiment(polarity=-0.0208333333333333343, subjectivity=0.77083333333333
33)
```

Figure 4

### NLTK Vader Results vs TextBlob Results

After running a sentiment analysis on my collected tweets using both NLTK and TextBlob, I found that an analysis of the accuracy between the two libraries could be reported. Overall, both

libraries did not have completely perfect classifications. Although there were several sentiments and polarities that seemed accurate, there were a large number of sentiments and polarities that seemed extremely inaccurate. In general, it is both easy and difficult to classify tweets. On one hand tweets are short and straight to the point, but on the other hand, it is difficult to analyze the sarcasm found in tweets. I believe the difficulty in analyzing the sarcasm of the tweets caused a lot of sentiment misclassifications. With this in mind, I believe the NLTK library provided results that were more accurate when it came to positive, negative, and neutral polarity. For the most part, the calculation of polarities for both libraries were in close proximity, but there were some polarity results that showed a huge difference between the two libraries.

For example, as displayed in Figure 5, there was a difference between the polarity found from the NLTK library and TextBlob library. The NLTK library provided a polarity of approximately 0.7184 for the displayed tweet, while the TextBlob polarity provided a polarity of approximately -0.02083. As

```
RT @CREWcrew: Donald Trump complimented Ivanka at an event for "working
hard." She must be working pretty hard--the Trump Organization pai...
compound: 0.7184,
neg: 0.0,
neu: 0.76,
pos: 0.24,
RT @CREWcrew: Donald Trump complimented Ivanka at an event for "working hard." She must be
working pretty hard--the Trump Organization pai...
Sentiment(polarity=-0.0208333333333333343, subjectivity=0.7708333333333333333)
```

Figure 5

previously mentioned, I believe the difference in polarity was caused by the complexities of identifying sarcasm within tweets and how well the classifier was trained. As we discussed in the TextBlob Sentiment Analysis section, I believe the classifier within the TextBlob library took the term "hard" to emphasize negative polarity, while the term "hard" was not used in a negative context for this specific tweet. After further investigation, I believe the NLTK library provided more accurate results for the collected tweets because the classifier was trained specifically for vocabulary and sentence structures used in tweets. The classifier was not only trained on the English dictionary, but the classifier was also trained on tricky sentences specifically related to tweets that other classifiers would misinterpret as a different sentiment.

## Summary

Overall, the dataset variable that was analyzed for this assignment resulted in 8,000 tweets that were based on the input query of "Donald Trump". After running a sentiment analysis on the tweets using the NLTK and TextBlob libraries, it was found that the majority of the tweets were mostly accurate and classified as having a neutral sentiment. Although neither of the libraries had perfect accuracy for the sentiment classifications, in my opinion, the NLTK library provided more accurate results in polarity than TextBlob. As a sentiment analysis is conducted on tweets, it is always critical to consider the sarcasm behind a tweet as it may cause sentiment misclassification.

### Works Cited

TextBlob. (n.d.). Tutorial: Quick Start. Retrieved from TextBlob:  
<https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>

Wikipedia. (n.d.). Natural Language Toolkit. Retrieved from Wikipedia:  
[https://en.wikipedia.org/wiki/Natural\\_Language\\_Toolkit](https://en.wikipedia.org/wiki/Natural_Language_Toolkit)