Christina Hill
4.18.18

Twitter Topics: Sentiment Analysis and Topic Modeling Analysis

Analysis Overview

The provided analysis applies information extraction concepts through a thorough analysis on tweets collected from Twitter's API. The collected tweets are specifically related to the following topics: "Avengers", "Wakanda", and "Marvel". Throughout the overall analysis, various sentiment analyses and topic modeling analyses are conducted on multiple datasets to assess how dataset size, stop-word removal, word-stemming, and other factors affect topic modeling and sentiment analysis results.

Sentiment Analysis Overview

	created_at	text	hashtags	username	user_followers_count	topic
9153	2018-04-07 22:03:43	marvel knowing their place and putting wakanda...		barnesquinzels	67	Marvel
9154	2018-04-07 21:46:49	RT @LakelPruitt: Marvel: Man, a lot of people ...	BlackPanther	Larissa44919266	262	Marvel
9155	2018-04-07 21:46:21	RT @PipeRiosP: En el último spot de #InfinityW...	InfinityWar	TheTnight394	33	Marvel
9156	2018-04-07 21:41:48	'Avengers: Infinity War' Heads To Wakanda In T...		go4viral	350	Marvel
9157	2018-04-07 21:41:08	RT @LakelPruitt: Marvel: Man, a lot of people ...	BlackPanther	ThePantherGod_	1739	Marvel

Figure 1

A sentiment analysis can be defined as a computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisals, affects, views, emotions, etc., expressed in text. As it relates to this analysis, a sentiment analysis was conducted on Twitter against the following queries: "Avengers", "Wakanda", and "Marvel". A dataset size of 4,000 tweets was used to conduct the sentiment analysis for the entire dataset. A small sample of the breakdown of the collected tweets is provided in a DataFrame table in Figure 1. As you can see in Figure 1, the tweets were broken down into columns based on each tweet's creation date, text, hashtags, username, user followers count, and topic. From the information that was provided from the downloaded tweets, a sentiment analysis was conducted on these tweets using the Text Blob library.

The sentiment analysis from 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 several sentiment results within the entire dataset, I saw that the sentiment analysis results were very accurate, but the results did not contain perfect classifications.

```
RT @CcreativechicC: @HouseOfDynasty @theblackpanther @Avengers It is some what annoying whe
n you have to explain the life of @Marvel comics…
Sentiment(polarity=-0.8, subjectivity=0.9)
----------
RT @IdentSecreta: ¡El ejercito de Wakanda se prepara para la guerra contra Thanos! NUEVO SP
OT DE TV DE INFINITY WAR AQUÍ: https://t.co/o83b…
Sentiment(polarity=0.0, subjectivity=0.0)
----------
Using #Wakanda &amp;amp; #BlackPanther success to catapult #InfinityWar is brilliant! Any that
 miss BP in theaters will lik… https://t.co/dUnpKEZOJN
Sentiment(polarity=0.65, subjectivity=0.5)
```

Figure 2

As displayed in Figure 2, the tweet outlined in green was given an overall negative sentiment. I believe this sentiment result was accurate because the word "annoying" was used in a negative connotation. In addition, the tweet outlined in blue was given an overall positive sentiment. I believe the positive sentiment was accurate because the term "brilliant" was used in a positive connotation. However, I found a downside of using the TextBlob library that was based on the English language. As shown in Figure 2, the tweet outlined in red was tweeted in Spanish instead of English and was given an overall neutral sentiment. If we were to translate the tweet to English, it would translate to, "The army of Wakanda is preparing the war against Thanos! NEW INFINITY WAR TV SPOT AQUF". Although the tweet could be classified as having a neutral or positive sentiment (if we were to consider the exclamation point and all-caps wording), the TextBlob library should consider

training the analyzer on multiple languages (i.e. English, Spanish, Chinese, etc.) for more accurate results.

```
Senitments Breakdown - Entire Dataset:
Percentage of Positive Tweets: 60%
Percentage of Neutral Tweets: 27%
Percentage of Negative Tweets: 11%
```

Figure 3

Through analyzing several tweets, I was able to gather that the majority of the tweets were of positive sentiment, followed by neutral sentiment. This can further be proven by the sentiment breakdown provided in Figure 3. As displayed in Figure 3, over 50% of the tweets in the dataset were of positive sentiment; 27% of the tweets were of neutral sentiment; and 11% of the tweets were of negative sentiment. A deeper analysis can be made for each topic.

Sentiment Analysis: "Avengers" vs "Wakanda" vs "Marvel" Topics

```
I can't wait!!! ❤😍

#Marvel #Avengers #InfinityWar
#AvengersInfinityWar #Wakanda
#Starbucks https://t.co/ceFjYPEmuY
Sentiment(polarity=0.0, subjectivity=0.0)
```

Figure 4

The TextBlob library was also used to conduct a sentiment analysis on the the following three topics within the dataset: "Avengers", "Wakanda", and "Marvel". The sentiment results provided for each topic were very accurate, but I believe the results did not provide perfect classifications. For example, as displayed in Figure 4, the tweet from the "Wakanda" topic received a neutral sentiment. However, I would argue that the tweet was of positive sentiment. I believe the library gave the tweet a neutral sentiment because the tweet included the word, "can't", while also including heart emojis (which signifies positive sentiments). This could be viewed as a tweet that contains positive and negative sentiments. However, I believe the tweet was of positive sentiment because the phrase "I can't wait" also included exclamation points. A human who is accustomed to the human language would understand that the tweet is displaying excitement and positivity about the topic.

```
RT @marvelmex: ¡Vengadores, Guardianes y Wakanda unidos! Nuevo spot de Avengers: #InfinityWar.
https://t.co/6bKj5m5MSq
Sentiment(polarity=0.0, subjectivity=0.0)

¡El ejercito de Wakanda se prepara para la guerra contra Thanos! NUEVO SPOT DE TV DE INFINITY WAR AQUÍ:… https://t.co/RyiOgLduYK
Sentiment(polarity=0.0, subjectivity=0.0)

"Wakanda per sempre"? Thanos non sembra del loro stesso avviso. Quanto siete in fibrillazione per il nuovo cinecomi… https://t.co/pCRk61UjVT
Sentiment(polarity=0.0, subjectivity=0.0)
```

Figure 5

In addition, a polarity score of 0.0 was given to tweets that were tweeted in Spanish within all three of the topics. As you can see in Figure 5, each tweet was pulled from one of the three topics and they received the same neutral polarity scored. If the tweets were translated to English, they would read: "Avengers, Guardians and Wakanda united! New Avengers spot: #InfinityWar.", "The army of Wakanda is preparing the war against Thanos! NEW INFINITY WAR TV SPOT AQUF", and "Wakanda forever"? Thanos does not seem to be of their own opinion. How much you are in fibrillation for the new…". Although it could be argued that the tweets are still of neutral sentiment, I would argue that the words "not" and "fibrillation" found in the bottom tweet triggers a negative sentiment and negative connotation. Furthermore, this supports the notion that the TextBlob library should consider additional languages because the polarity scores would possibly change if the tweets were translated to English.

```
Senitments Breakdown - Avengers Dataset:
Percentage of Positive Tweets: 71%
Percentage of Neutral Tweets: 16%
Percentage of Negative Tweets: 12%
```

Figure 6

We could also take a closer look at the breakdown of sentiment percentages for a specific topic. For example, in Figure 6 we are able to see a breakdown of overall sentiments to tweets related to the "Avengers"

topic. From the breakdown, we are able to gather that over 50% of the tweets related to the "Avengers" topic was of positive sentiment. With the combination of this analysis and the percentage breakdown of the entire dataset sentiments, it could be assumed that the remaining two topics will have positive sentiments for the majority of their tweets.

Sentiment Analysis: Entire Dataset vs 3 Topics

As we can see from the sentiment results from the entire dataset and the three topics, there are several similarities and differences. The entire dataset and three topics did not contain perfect classifications, but the majority of the tweets were accurate. All of the sentiment results included tweets that were tweeted in Spanish and those tweets were given a polarity score of 0.0. In addition, the breakdown of sentiment results across all topics and the entire dataset showed the majority of tweets having positive sentiments, followed by neutral sentiments.

Top Hashtag Analysis

Table 1 provides a list of the top 5 hashtags for each topic. As you can see from the table, each topic shared over 50% of the same top hashtags. For example, all three of the topics contained 'BlackPanther', 'Marvel', and 'Avengers' listed as their top hashtags. We are also able to see that a large number of tweets did not contain a hashtag. In addition, "Avengers" and "Marvel" both had 'BlackPanther' listed as their top hashtag used among the collected tweets. It could be assumed that the topics "Avengers" and "Marvel" are closely related because they have the same top hashtag.

Avengers	Wakanda	Marvel
BlackPanther	No Hashtag	BlackPanther
No Hashtag	BlackPanther	InfinityWar
Avengers	Marvel	No Hashtag
Marvel	InfinityWar	Avengers
HeroesManga	Avengers	AvengersInfinityWar

Table 1

In addition, it could be stated that there is a relation between the topics and the top hashtags. For example, in the comic book world, Black Panther is a superhero who is a member of the Marvel team. The actor who played the character of 'Black Panther' in the movie, "Black Panther", will also be staring in the upcoming movie, "Avengers: Infinity War". With these findings in mind, it can be assumed that the inter-related topics of interests and events are the cause of such a strong relation between the topics and the top hashtags. For example, a tweet could state how the actor is going to be a great character in "Avengers: Infinity War" just like he was in "Black Panther", followed by hashtags for both of the topics. Aside from hashtag analyses, a stop-word removal analysis can be applied to the dataset and topics.

Stop-Word Removal Analysis

Stop-words can be defined as words that are frequently used in a specific language. For example, in the English language, the words "I", "is", and "she" are frequently used. From the Twitter language perspective, the word "RT" would be classified as a stop-word because it is included in every tweet that is retweeted and could be classified as "meaningless". From the analysis, it was proven that the stop-word removal process improved the results of topic modeling by removing unnecessary words and space so that the results of the topic model analysis were more accurate. For example, if "RT" were to constantly appear, the topic modeling analysis would constantly be considering a word that has no significance, which tampers with the accuracy of the results.

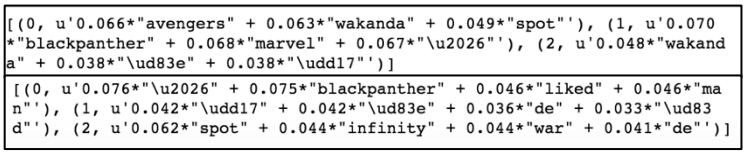

Figure 7

In addition to removing the word, "RT", I also added "wakanda", "avengers", and "marvel" as stop-words to assess the LDA results. These three words were frequently used within the tweets that related to the three designated topics. In other words, they were context-dependent stop-words. Figure 7 displays the topic modeling results where stop-words were removed (bottom results) vs stop-words not being removed (top results) during the topic modeling process. By removing the additional frequently used words, we are left with results that are cleaner and easier to assess when analyzing the results. By removing the context-dependent stop-words, we are also left with topics that are sub to the topics we are analyzing, instead of possibly providing topics that are the exact topics of the collected tweets. For example, if we were to analyze the first topic result provided in the bottom results of Figure 7, we could pose that the topic is about Black Panther being a well-liked man. The results do not include the frequently used words and it could be proven that Black Panther is sub to "Wakanda" (i.e. Black Panther is from Wakanda). However, if we considered the first topic result provided in the top result, it is more difficult to analyze the specific LDA results because the frequently used words ("wakanda" and "avengers") tamper the authenticity of the results.

Punctuation-Removal Analysis

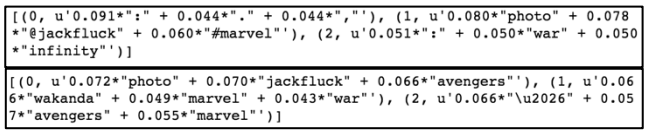

Figure 8

It can also be proven that punctuation-removal is necessary for topic modeling. As you can see in the top section's topic modeling results in Figure 8, I found that when punctuation was not removed, punctuation appeared in the topic modeling results and it was somewhat difficult to assess the overall topic because the punctuation marks did not add much meaning. For example, the first topic in the LDA results where punctuation was not removed only includes ':', '.', and ',', There was no text included in the results and a topic analysis can hardly be made on a set of punctuation marks.

However, when the punctuation was removed (as displayed in the bottom section's topic modeling results in Figure 8), the topic modeling results seemed to be more accurate and easier to assess because meaningless characters were removed and the LDA results were more correlated. For example, the first topic in the topic modeling results where punctuation was removed could be assessed as a topic relating to a Jack Fluck who is a photographer for Avenger characters. Furthermore, as displayed in Figure 8, it can be argued and proven that punctuation marks will replace critical words that are needed during topic modeling. With these findings in mind, it is necessary to remove punctuation during topic modeling.

Word-Stemming Analysis

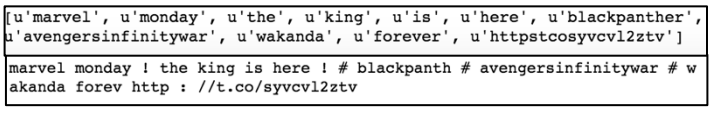

Figure 9

Word-stemming can be defined as reducing a word to its stem, which is also known as the root form of the word. As you can see in Figure 9, a word-stemming analysis was conducted on the words found in each of the collected tweets. The

bottom section in Figure 9 displays the stemmed version of the word “Blackpanther” and “forever”, where “-er” was removed from the root words. It can be compared to the top section of Figure 9, which displays the non-stemmed version of the words.

Although word-stemming frees up space within the analysis, it can be proven that the stemmed words cause difficulty in topic modeling and interpretation. The stemmed version of the words is displayed in the LDA results, which has the effect of possibly changing the context or meaning of an analyzed topic. For example, if the three words provided in a topic modeling output were “aveng”, “wakanda”, and “war”, it would be difficult to analyze how to incorporate “aveng” in analyzing the specific topic. From a quick Google search, “Aveng” is related to multiple companies (not the Avengers), which would not apply to the context of the topic. Essentially, it can be proven that word-stemming tampers the accuracy and contextual meaning provided in topic modeling.

Effect of Increased Dataset Size

```
Senitments Breakdown - Entire Dataset:
Percentage of Positive Tweets: 50%
Percentage of Neutral Tweets: 40%
Percentage of Negative Tweets: 9%
```

Figure 10

Lastly, an additional analysis can be conducted on how the dataset size affects sentiment and topic modeling results. Within the analysis, an additional sentiment analysis and topic modeling analysis was conducted on a larger dataset of 8,000 collected tweets (instead of 4,000). After analyzing the larger dataset, I saw that the percentage of negative tweets increased by 2%, while the percentage of positive tweets decreased by 10%. However, as displayed in Figure 10, 50% of the tweets were still of positive sentiment. It could be argued that an increased dataset size would train the classifier better for sentiment analysis purposes. The classifier would have a great number of tweets related to analyze the sarcasm and trick sentences that apply specifically to the tweet. The increased dataset size could also increase or decrease the overall breakdown of sentiment percentages. For example, in the smaller dataset size, the collected tweets were of 60% positive sentiments and 27% neutral sentiment, while the collected tweets from the larger dataset size were of 50% positive sentiments and 40% neutral sentiments. The increased dataset size caused the percentages to fluctuate.

```
[(0, u'0.072*"photo" + 0.070*"jackfluck" + 0.066*"avengers"'), (1, u'0.066*"wakanda" + 0.049
*"marvel" + 0.043*"war"'), (2, u'0.066*"\u2026" + 0.057*"avengers" + 0.055*"marvel"')]
[(0, u'0.078*"marvel" + 0.062*"wakanda" + 0.055*"avengers"'), (1, u'0.057*"avengers" + 0.053
*"wakanda" + 0.046*"de"'), (2, u'0.047*"\ud83e" + 0.047*"\udd17" + 0.043*"avengers"')]
```

Figure 11

In addition, the larger dataset size caused the three topics (“Wakanda”, “Marvel”, and “Avengers”) to appear in all of the LDA results, instead of only a few LDA results. As displayed in Figure 11, the top LDA results are from the smaller dataset and the bottom LDA results are from the larger dataset. From Figure 11, we are able to see how our three topics dominate the LDA results. This also proves the importance of removing frequent words that appear in tweets for stop-word removal purposes. It can be posed that the larger the dataset for a specific number of topics, the higher the possibility that the LDA results will include the actual topics (if the topics are not implemented as a stop-word).

Closing Discussion

In closing, information extraction concepts such as sentiment analyses and topic modeling provide insightful results for collected tweets and topics. From the overall analysis, we were able too see how punctuation-removal and stop-word removal is necessary in topic modeling for more accurate and clean results. If stop-word removal and punctuation removal are not

conducted, the LDA results will not provide results that can be fairly assessed. We were also able to see that over 50% of the tweets collected for the three topics and the entire dataset were of positive sentiments. It can be assumed that the topics have an overall positive connotation and can be classified as positive topics. In addition, there was a strong correlation between the top five hashtags and the three specified topics. It was posed that this could be contributed to the inter-related events and interests between the topics and hashtags. Lastly, we found that by increasing the dataset size, a classifier is able to be trained better, while topic modeling results may include the designated topics if they have not been implemented as stop-words.

Works Cited

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