EM-624 Final Project

Fall 2017

Bias Analysis in Three of the Top Ten Daily Newspapers in the US Using Python Text Analysis Libraries

by

Marshad M. Almarshad

December 2, 2017

Table of Contents

1. Project Goals and Conditions

This main objective of this project is to find out the whether the newspapers in the US use biased language when covering stories or incidents that are, supposedly, similar but take place in different locations. Because of the unfortunate mass killing incidents that happened in different parts of the world in 2017, I built an interest in analyzing how the newspapers in the US reacted to those incidents. The incidents selected as a sample for this study took place in Orlando, Las Vegas, Barcelona and Egypt. Table 1 shows the statistics of these incidents.

Table 1. Statistics of the four incidents

Incident Date	Location	# Killed	# Injured
11/24/2017	Sinai, Egypt	305	128
10/01/2017	Las Vegas, Nevada, USA	59	441
06/12/2016	Orlando, Florida, USA	50	53
08/17/2017	Barcelona, Spain	13	130

The analysis considers how the same newspaper reports those different stories and, also, compares how the same story is reported by three different newspapers. Accordingly, the study tries to answer the following questions:

1- Did all the different incidents get the same amount of space and coverage?

2- Did the newspaper use the same words to describe the different incidents?

3- What words were used by the different newspapers to describe the same incident?

4- Was there inconsistence in the overall tone of the articles in each newspaper?

Since the study depends on extracting text directly from the newspapers' websites, the results are objected to changes due to the fact that some newspapers modify the articles as more information is obtained. Also, some newspapers apply time based restrictions on accessing their database while others don't allow direct extracting from their websites. Thus, these limitations were considered while conducting this study and as they directly affected my choice of the newspaper. However, the selected newspapers are still among the top 10 most popular newspapers in the US according to cision.com (https://www.cision.com/us/2014/06/top-10-us-daily-newspapers/).

2. Business Understanding

In order to answer the research questions, the words and phrases from the articles had to be extracted, visualized and analyzed. Newspapers websites contain so many headers, advertisements and text which are not of our interests and might affect the results of the study. Therefore, web scraping process ought to take this issue in account by cleaning the extracted article and using the correct class.

Another issue is pertaining to getting the articles URL. Many newspapers websites don't allow extracting links from their website when using their search engine. In this study, I am looking for immediate reports during the first two days of the incident. To get these articles, the only practical way was to use search engines. Consequently, I had to manually find the articles and save them in an Excel file.

The strategy was to extract the articles and save them in lists after removing stop words. Then, different libraries, such as nltk, wordcloud and collections, were used to extract the information from the articles. The results were presented as plots, wordclouds and regular prints. The final step was to interpret the results to answer the research questions.

3. Data Understanding

The data was collected from the newspapers official websites. The newspapers selected are the New York Times, the New York Post and Chicago Tribune. For the four incidents, URLs of ten articles, except articles from the New York Post on the Egyptian case where only 6 articles were found, were collected and saved in the attached Excel file. The total number of links are 56. Using each newspaper's search engine, the articles are selected to be within two days from the incident and most relevant to it to insure the quality of the data.

4. Data Preparation

The first step was to make sure that only the article was extracted from the website. This had to be done by studying the classes used by developers of the websites. It was found that NY Times developers use 'story-body-text story-content' class to refer to the article, Chicago Tribune developers use 'trb_ar_page' while 'article-header' is used in NY Post website. Yet, the articles still contained some links and headers that were added to the stop word lists.

The Excel file was prepared to be used by Pandas library which can easily map through the file and get the links in order.

5. Data Representation

5.1. Overview

The first step in the analysis was to have an overview of how much space dedicated by each newspaper to each incident and how deep the newspaper went to explore the stories. This can be obtained by finding the total number of words written by each newspaper. Figure 1 shows that the incidents did not get the same level of coverage. The NY times gave more focus on the case of Las Vegas while much fewer words were used to write about the incident that took place in Barcelona and Egypt. Similarly, the New York Post's coverage of international incidents is relatively low compared to the incidents of Las Vegas and Orlando. The Chicago Tribune, on the other hand, showed more interest in the incidents of Egypt and Barcelona as more words were used to report those two incidents.

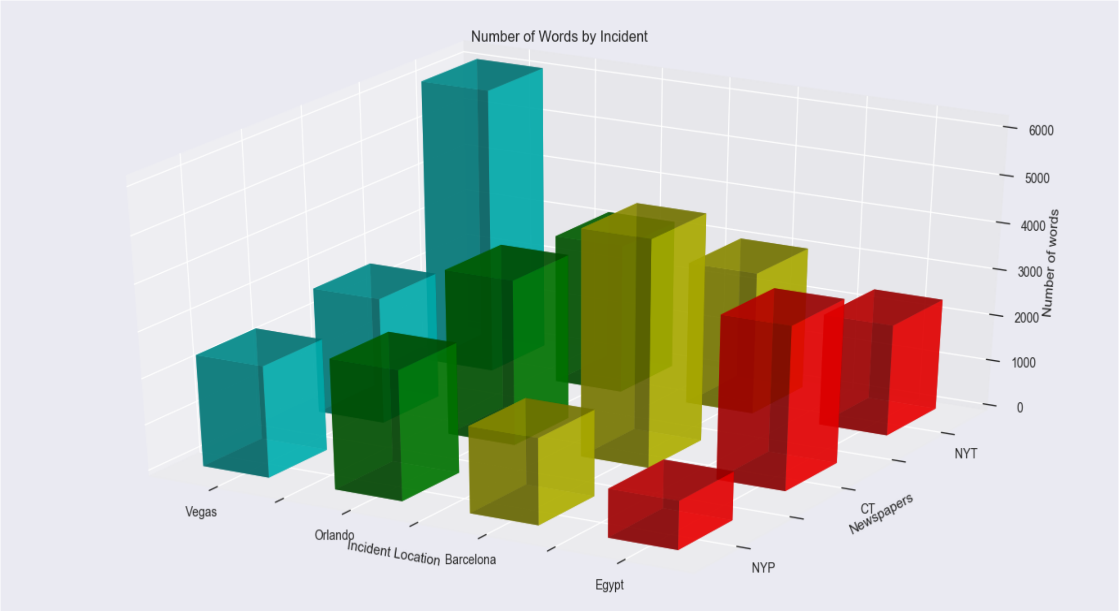

Figure 1. The total number of words used by the newspapers to cover the four incidents by

5.2. Comparison Between the Different Newspapers on the Same Incident

The second approach was to compare how the three newspapers reacted to the same incident. This was analyzed by finding the most frequent words in the three newspapers and creating wordclouds of the common words. As figure 2 shows for the incident in Vegas, the three newspapers put more emphasis on the scene where the shooting took place, such as “outdoor”, “crowd”, “concert” and “festival”, the shooter’s name “paddock” and the action “shot”.

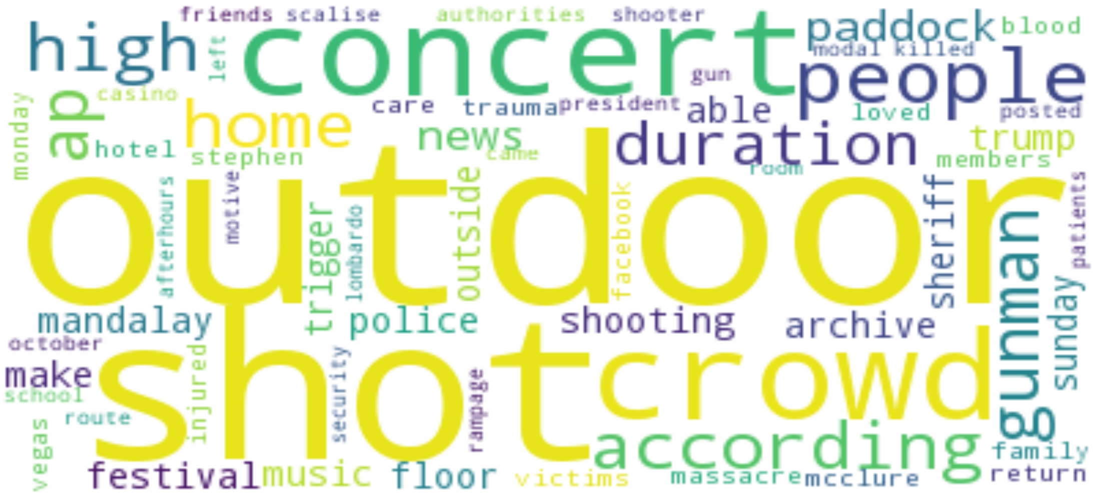

Figure 2. Wordcloud of the common frequent words (Vegas)

Unlike the case in Vegas, the incident in Orlando was influenced by the presidential election in 2016 as “trump” and “clinton” were more common in the articles than the name of the shooter” mateen” and the location “nightclub”, as shown in figure 3. This is a clear bias toward the political situation at that time.

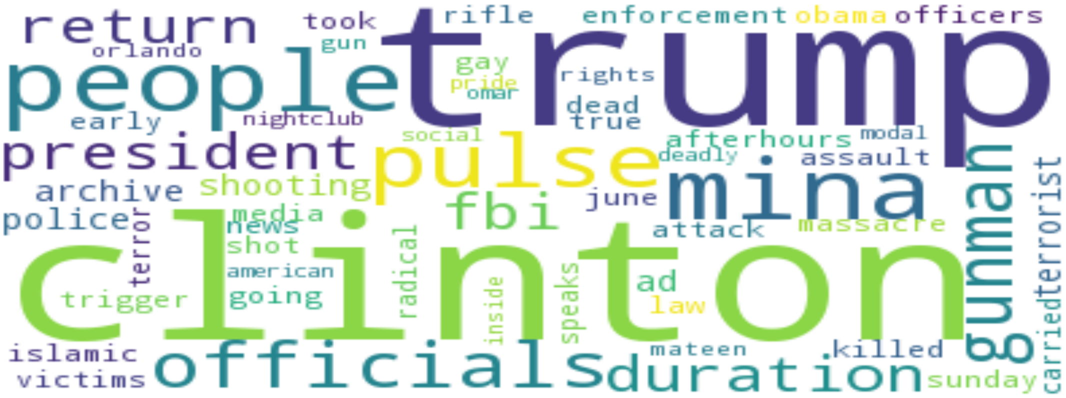

Figure 3. Wordcloud of the common frequent words (Orlando)

Figure 4 shows that the three newspapers described the incident in Barcelona as “attack” done by “terrorist”. Focusing more on the attacker, the newspapers frequently mentioned words like “suspect” and “extremist”. The name of the American president “trump” was very common in the articles as well. Less emphasize was given to religious background “islamic” but still among the most frequent words.

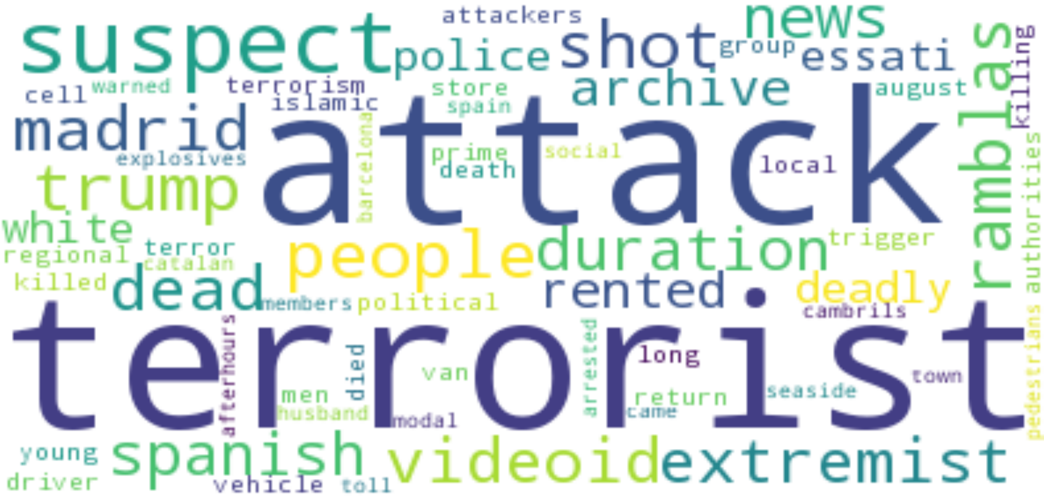

Figure 4. Wordcloud of the common frequent words (Barcelona)

In the fourth incident, the common words include describing the incident as “attack” done by “militant”. The victims “sufi” where described as “worshipers” as the incident took place in a “mosque” in “egypt”. Again, the president’s name is very frequently mentioned as shown in figure 5.

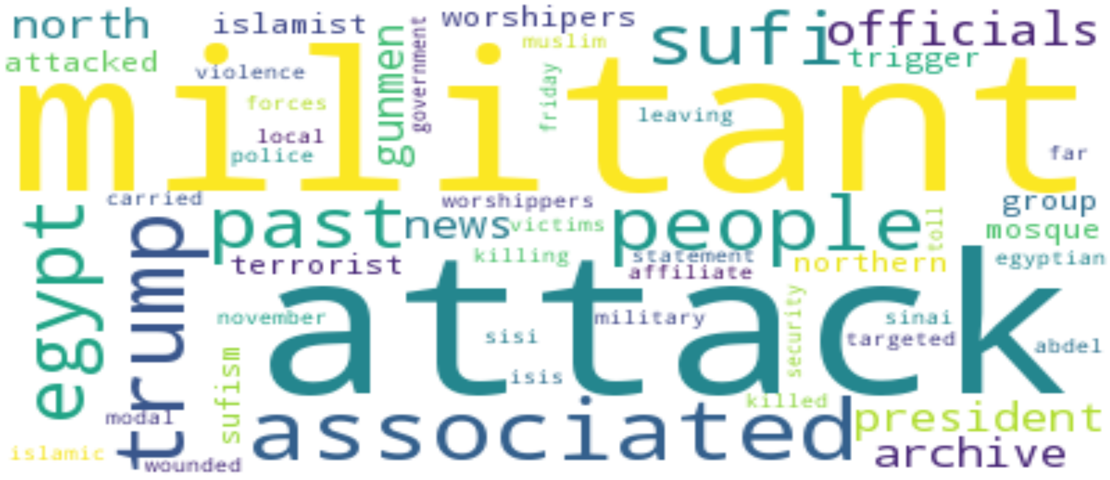

Figure 5. Wordcloud of the common frequent words (Egypt)

In general, the newspapers distinguished between incidents took place in the U.S. and incidents around the world in terms of the used language and vocabulary. The newspapers tend to get very aggressive when describing events not in the U.S. while more neutral words are used to describe the incidents took place in the U.S. This could be related to internal issues that the newspapers try to avoid such as gun control.

5.3. How Each Newspaper Described the Different Incidents

Even the same newspaper had different looks at the four incidents. The NY Times described the incident in Vegas as a “shooting” with a” gun” done by “paddock”. The rest of the top 30 words in figure 6 does not include any other word related to the attacker, his background or the victims despite the fact that this incident is the worst in the history of the United States. On the other hand, the most frequent words in the incidents of Orlando, Barcelona and Egypt are “killed”, “attacks”, “islamic” and “assault”.

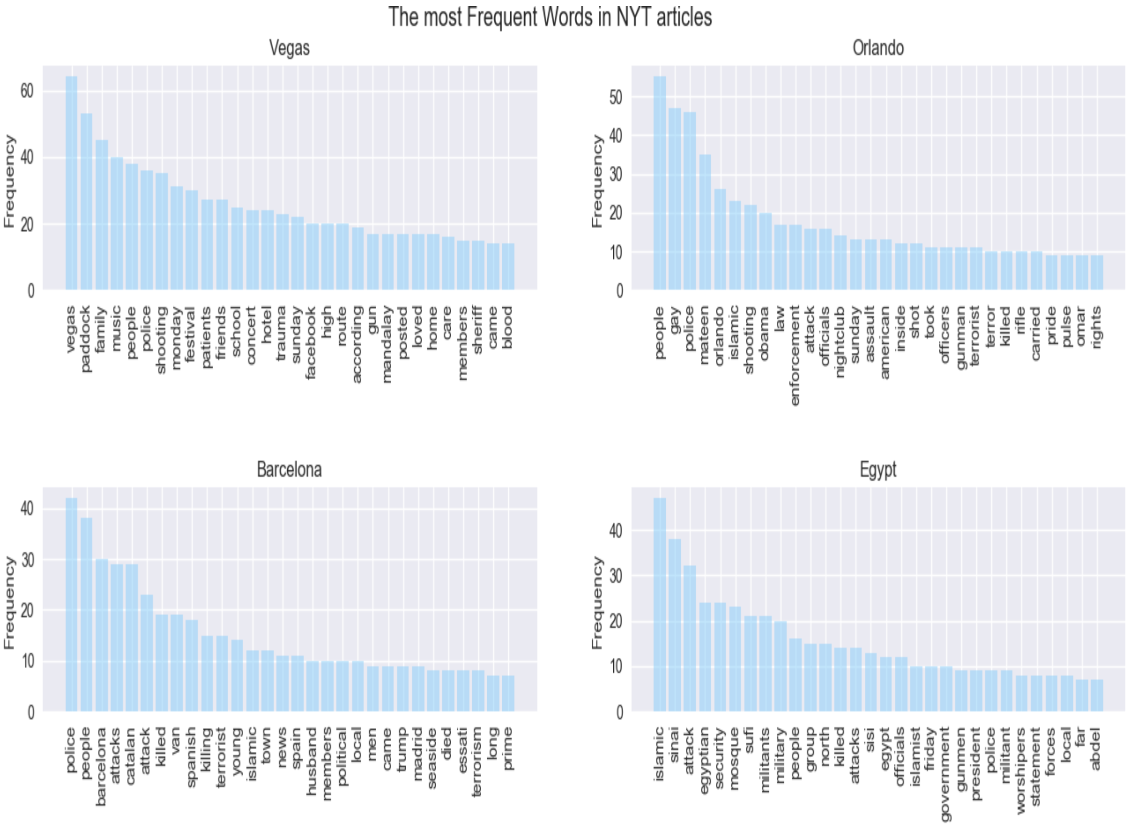

Figure 6. Most Frequent Words in NY Times

The NY Post was relatively less biased than the NY Times. NY Post described the incident in Vegas as “shooting”,” massacre” that “killed” many “victims”. The name of the shooter and the location were very frequently mentioned as well. Also, almost no mention of the shooter background or the gun control laws. On the other hand, the words “terrorist”, “terrorism”,” death” and “islamic” are the most frequent (figure 7).

Lastly, the Chicago Tribune used the same words to describe the four incidents as did the NY Times and the NY Post. (figure 8)

Overall, the three newspapers had inconsistent approaches when describing the very similar incidents but happened in different locations by different criminals.

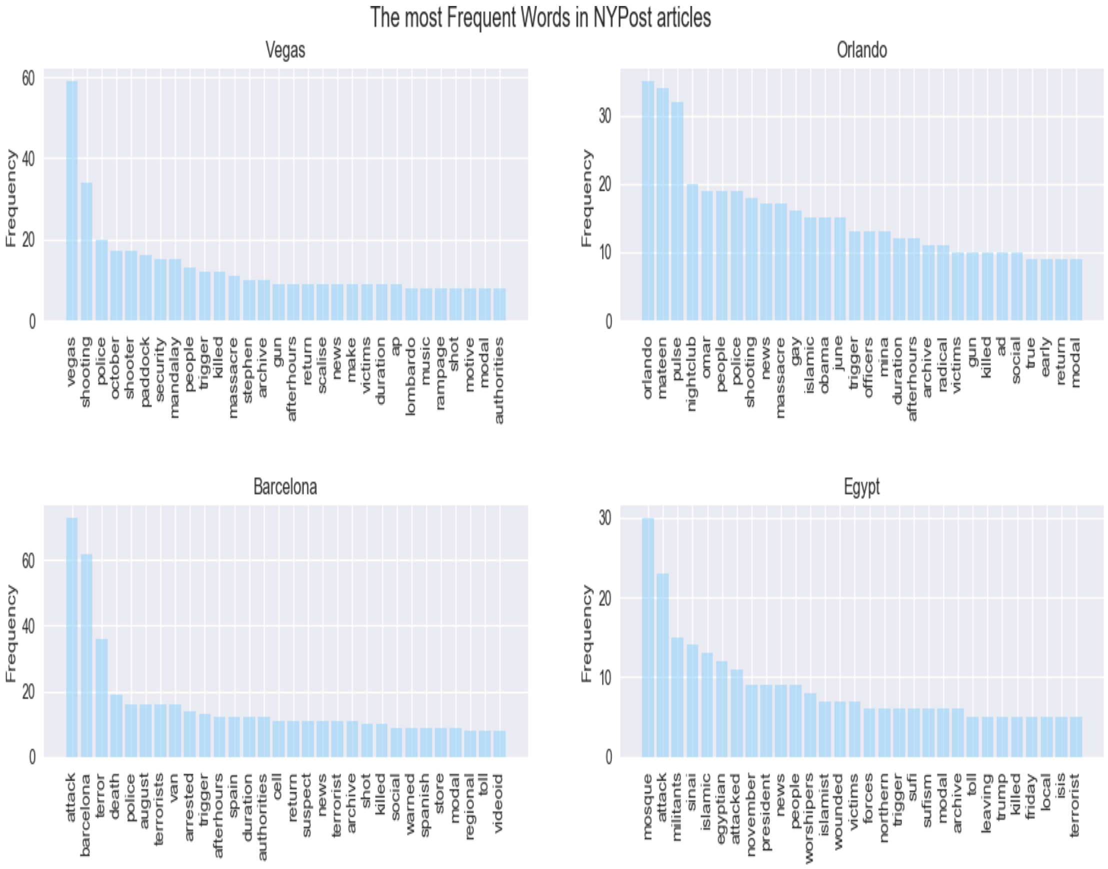

Figure 7. Most Frequent Words in NY Post

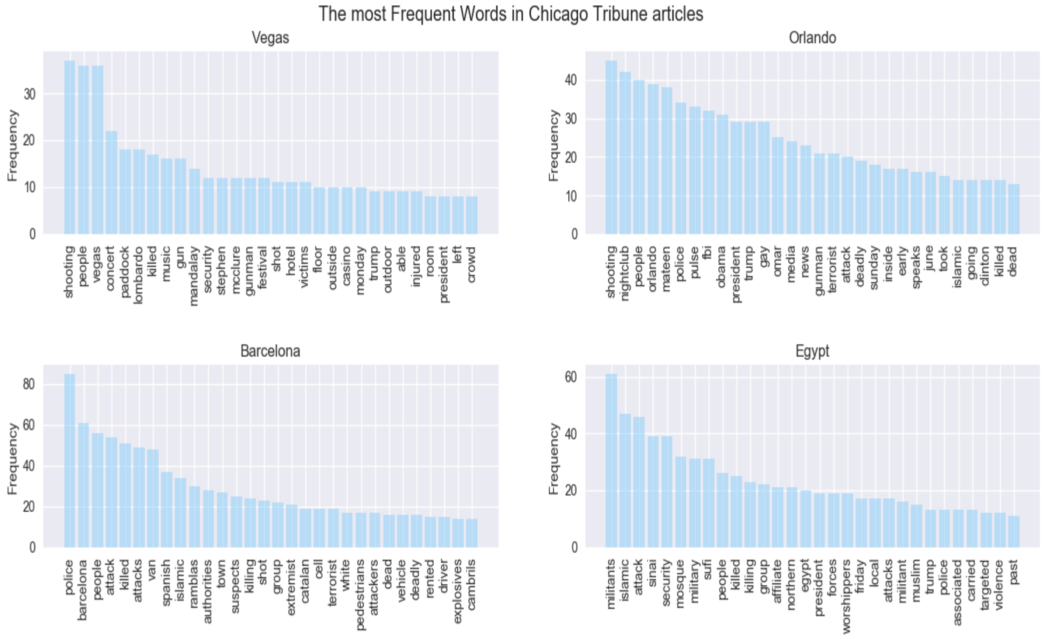

Figure 8. Most Frequent Words in Chicago Tribune

5.4.Sentiment Analysis

The calculated sentiment scores of the articles support the findings in the above sections that the three newspapers tended to use more positive tone when describing shooting incidents happened in the United States. Figure 8 shows that the Chicago Tribune sentiment scores were negative in the case of Barcelona and Egypt but positive in the case of Las Vegas and Orlando. The same behavior was adopted by the NY Post while the NY Times was leaning toward the neutral area in the case of Barcelona.

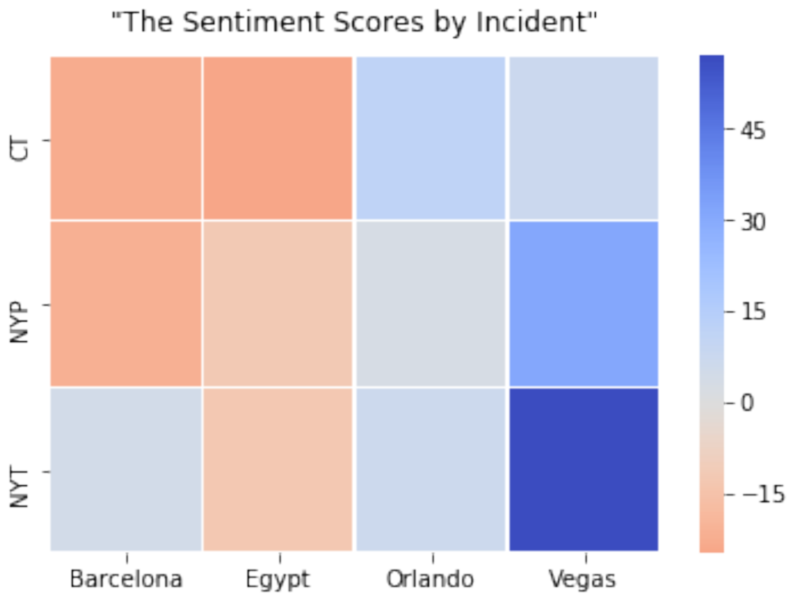

Figure 8. Sentiment Analysis of the 56 articles

6. Conclusion

In this study, 56 articles were analyzed using different libraries in python. The process involved collecting the data from the newspapers, preparing the data to be clean and ready to explore and extracting meaningful information. The major findings from this study are:

1- The newspapers were selective when covering the shooting incidents. For example, the NY Times wrote more than 6000 words on the incident in Vegas as opposed to fewer than 3000 words on the incidents in Orlando.

2- The newspapers distinguished between incidents took place in the U.S. and incidents around the world in terms of the used language and vocabulary. The newspapers tend to get very aggressive when describing events not in the U.S. while more neutral words are used to describe the incidents took place in the U.S.
3- The three newspapers had inconsistent approaches when describing the very similar incidents that happened in different locations by different attackers. For example, the word "terrorist" was almost never used in the case of Las Vegas while it was the most common word in the other cases.

With that said, I think this study can be further improved by increasing the sample size to include more of the top 10 newspapers like the Washington Post. More advanced web scraping methods can be utilized to solve the problem of having access to more articles. Also, the sentiment analysis can be further improved by using more advanced techniques than SentiWord.