Sentiment and Content Analysis Surrounding the Pulse Nightclub Shooting

Jiacheng He*, Mohammad Isyroqi Fathan†, Sri Gayatri Sundar‡, Muhammad Saad Adnan§ and Sierra Seacat¶
*Department of Economics, University of Kansas, Kansas, USA
†‡§¶Department of Electrical Engineering and Computer Science, University of Kansas, Kansas, USA
Email: *jiacheng.he@ku.edu, †m576f167@ku.edu, ‡sri@ku.edu, §saad.adnan@ku.edu, ¶scseacat@ku.edu

***Abstract*—Data science is proving to be very useful in many fields, including Natural language processing, that has tremendously helped in understanding people's opinions and sentiments. By utilizing Natural language processing techniques, we are investigating on the media articles and twitter tweets related to the Orlando Pulse Nightclub shooting incident. In this paper, we mine for influencing, popular topics talked about during the incident and discuss how topic popularity changes with time. This paper also aims to capture the changing sentiment over time and tries to attribute the changes to a particular event that happened around that time. By utilizing Topic Modelling techniques, clustering algorithms and Sentiment Analysis, we aim to answer our research questions and provide significance metrics for our results. In future, we also plan on taking this forward by analysing other major mass shooting events and comparing them to find similarities and differences.**

***Index Terms*—Sentiment Analysis, Orlando shooting, Topic Modeling**

I. Introduction

On June 12, 2016, Omar Mateen, a security officer at G4S Secure Solutions, open fired at people at The Orlando Pulse Nightclub, resulting in 53 injured and 49 killed. [1]. The National Public Radio has reported, "What happened in Orlando was the deadliest mass public shooting in the modern U.S history" [2]. Adding on to that, the Orlando shooting incident has also been called out by The New York Times as *the most extreme act of violance against the L.G.B.T.Q community in the US history* [3]. The Washington Post has also tagged this tragic incident as "the deadliest terrorist attack on US soil since September 11, 2001" [4].

Given a terrifying incident of such mass scale, a lot of people, celebrities and news media groups, expressed their grievances, varied opinions and raised voices against many different important matters associated with this incident. Interested in the public opinions and hot topics of discussion, we decided to focus our analysis on news articles published by various media groups which were collected in the LexisNexis database and in addition, we considered tweets posted on Twitter regarding the incident. Using Data Science and Natural Language Processing techniques, we aim to answer the following research questions:

1) *What are the popular, sentiment-influencing key topics discussed by people and media?*
2) *Did the popularity of these key topics change over time? If so, How?*
3) *How does the sentiment expressed in tweets and news article change over time?*
4) *Can we attribute any sharp, positive or negative change in sentiment with any key event or happenings related to the event, that occurred around that time?*

II. Related Work

The world has been resorting to Twitter to express their sentiments, opinions, concerns and love. Twitter is the most popular microblogging community and has also been very popular in a lot of sentiment analysis studies, natural language processing and various data analysis studies. With various sorts of events that happen all around the world, there are some events that trend the most while some completely ignored because of distractions. However, terrorist attacks have usually made it to the trending news on twitter and therefore, many researchers have been interesting in analysing public pattern and sentiments of such events. Minyoung Chong from the University of North Texas has done a similar research by analysing the Twitter tweets from the Paris terrorist attack [5]. The study also extracts discussion topics from the twitter tweets, while this paper summarizes results to do the same. There have also been several works done on sentiment analysis using Twitter data [6] [7]. Similarly, a lot of work has involved performing Topic Modelling to twitter tweets to predict trending hastags and also extracting topics of discussion [8] [9] [10].

III. Data

There were two datasets that were used for the incident; Twitter and LexisNexis database.

A. Twitter

The initial Twitter data set covers the period of 6/18/2016 - 7/6/2016; one week after the shooting. They are extracted based on the criterion of containing one of the two hashtags: #orlandostrong, and #orlandounited. There are 36,783 tweets in total. After removing hashtags and urls, The number of words in these tweets ranges from 1 to 31, with a median of 17 words. The total number of Twitter users in the dataset is 30,147. The average number of tweets per user is 1.22 tweets.

B. LexisNexis

Initial exploratory analysis of the Twitter data revealed that the data wasn't sufficient enough to perform extensive analysis. Based on the number of words in the tweets, which ranges from 1 to 31, it is not sufficient to perform topic modeling such as Latent Dirichlet Allocation. Furthermore, a major proportion of the tweets (about 33%) talked about Justin Bieber, who was having a concert in Orlando on June 30th, 2016. This can result in biases when we perform topic modeling, since Justin Bieber is the most frequent topic discussed.

Limitations in the Twitter API made it more difficult to collect more tweets. So, it was not possible to extend the Twitter dataset to include tweets that might capture different aspect discussed about the incident. These limitations caused the need for a wider range of data. Therefore, news articles from the LexisNexis database were used as an alternative.

LexisNexis database is a database containing news and articles sources from various media. Documents were queried with criteria of the co-occurrence of "Orlando" and "Shooting" within 5 words or "Pulse" and "Shooting" within 5 words in the document's title. The start date and end date for the queried documents were set to be from the day of the incident (June 12th, 2016) until June 23rd, 2017. This allowed us to analyze the data from the day of the incident as well as the one year anniversary of the incident. The documents were further hashed to minimize the chance of having duplicate documents on the dataset. This is because some news were published in various media such as news letter, websites, and many more. The features used as input for the hashing functions are the title, date published, full text, and highlight.

C. Comparison

The resulting LexisNexis dataset contains 1,258 documents with a median length of 437 words. This enables us to perform topic modeling on the dataset. The comparison between the Twitter dataset and the LexisNexis dataset is shown in Table I.

TABLE I
SUMMARY OF DATA SETS

	Twitter	LexisNexis
# of Documents	36783	1258
# of Authors	30147	536
Median Length	17	437
Start Date	6/18/2016	6/12/2016
End Date	7/6/2016	6/23/2017

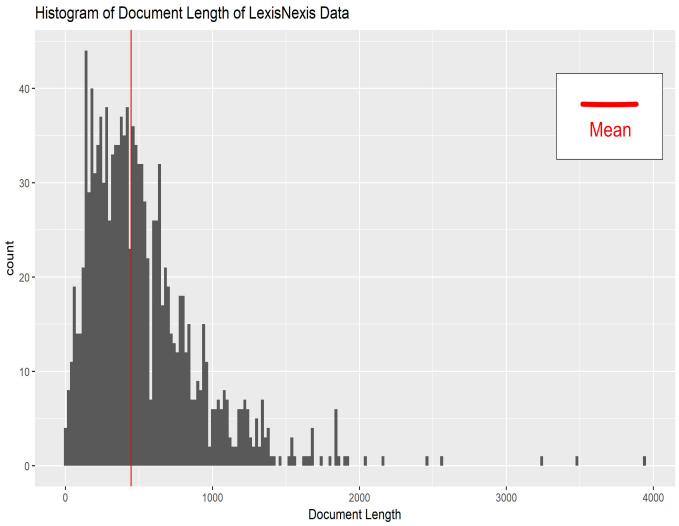

Fig. 1.

TABLE II
NUMBER OF OBSERVATIONS BY MONTH

	Twitter	LexisNexis
2016		
June	28471	1094
July	8312	49
August	0	13
September	0	4
October	0	3
November	0	4
December	0	10
2017		
January	0	9
February	0	1
March	0	6
April	0	2
May	0	0
June	0	62
July	0	0

IV. METHOD

Based on the research questions, we divided our approach to analyze the data into two parts, topic modeling and sentiment analysis. The topic modeling approach is used to identify popular sentiment-influencing key topics that arise from the incident. The second part of our analysis is sentiment analysis. Here, we aim to analyze how sentiment changes over time and key events that might have influenced these changes by measuring the sentiment values of the text.

A. Topic Modeling

We used Latent Dirichlet Allocation (LDA) [11] with "topicmodels" package in R [12] to model the topics of the news articles from LexisNexis[1]. We extracted uni-gram,

[1]The Twitter data have small sample size problem. And we cannot aggregate the tweets into author level or hashtage level either. So we did not perform topic modeling analysis on the Twitter data.

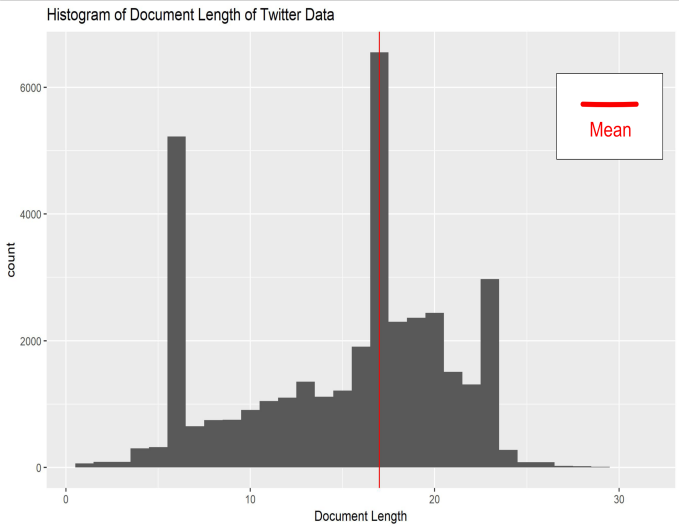

Fig. 2.

removed stop words, performed stemming[2], and removed words with tf-idf below the 10% quantile. A central tuning parameter in LDA is the number of topics.

Researchers proposed many statistical technique to determine the optimal number of topics, such as likelihood and density method [13] [14] [15] [16]. We try these statistical criteria. But different criteria gave us different answers of the optimal number, and all of them are hard to interpret. Thus we decided to embrace human judgment. We run LDA with topic number ranged from 2 to 15 then eyeball examined the result. Then we finally use 9 topics because the result of 9 topics is the most interpretable.

Figure 3 shows the topic-term classification results of the LDA. Based on the most frequent terms within each topic, we named the topics as: Family and Support, Gun and Politics, Medical and Financial Support, Victims, Incident General Information, LGBTQ, Terrorism, Shooter, and Celebrities.

B. Sentiment Analysis

We chose a simple bag of words with sentiment dictionary model to measure the sentiment values of the text. This model makes an assumption that each word in the text are independent of each other. Using the sentiment dictionary, our model measures the sentiment values by counting the number of words that fall into the specified criteria or values.

There are two dictionaries used for the sentiment analysis; ”Afinn” [17] and ”Bing” [18]. Afinn dictionary contains 2477 English words and phrases rated for valence with an integer between -5 and 5. Bing dictionary contains around 6,800 English words annotated with binary classification of positive and negative. The comparison between the two dictionaries are shown on the table II.

[2]We used ”SnowballC” package in R for stemming.

TABLE III
SENTIMENT LIBRARY

Library	# of Terms	Range of Value
AFFIN	2,477	[-5,5]
Bing	6,800	Positive & Negative

We used both the Twitter LexisNexis dataset in our sentiment analysis. We start by performing data preprocessing on the dataset. For Twitter dataset, we first removed hashtags, user mentions, URL, and pre-defined stop words from the tweets. The predefined stop words come from tidytext [19] package in R, which contains most common stop words in English such as ”a”, ”the”, etc. In addition, we removed ”Justin” and ”Bieber”, since these words are very frequent in most of the tweets. For LexisNexis dataset, we also removed URL and email address that might be in the text.

The cleaned dataset was then further processed by removing words with the highest 10% and lowest 10% TF-IDF value. This is because words that are too frequent or too less frequent in all documents are usually not interesting. Unlike in the topic modeling analysis, we did not perform stemming on the data because we are using dictionary based approach. Performing stemming on the data reduce the words to their base forms, which reduces the number of overlaps between the dictionary and the words. This might affect how expressive the sentiments of the documents and the tweets that we can capture. So, we decided not to perform stemming on the data.

From here, we measured the sentiment values of the documents and tweets by counting the number of negative and positive words from bing dictionary. To compare the sentiment between the twitter dataset and the lexis nexis dataset, we plot the sentiment change over time for the period where the Twitter dataset and LexisNexis dataset overlap.

We also plotted the key events surrounding the incident and try to see if these key events affected the sentiments of news media in LexisNexis dataset.

V. RESULTS AND DISCUSSIONS

A. Results

Using LDA, we were able to classify the documents into 9 topics.

Family & Support. Documents classified in this topic talked about the families of the victims. Top words: victim, community, lgbt, gai, people.

Guns & Politics This is a combined topic that mentioned guns and politics. Top words: gun, attack,control, trump, republican, obama, senate.

Medical & Financial Support Now moving on, we have documents talking about supporting the victims. Top words: donate, blood, victim, community, gai.

*Victims*Information about he victims themselves. Top Words: gun, victim, mateen, lui, pride, family.

General Information information from the investigation into the event that got released over time. Top words: police, mateen. people, club, friend.

LGBTQ the LGBTQ community, who were the primary victims of this tragedy. Top words: victim, community, lgbt, gai, rainbow.

Terrorism Articles talking about the incident as an act of terrorism. Top words: community, attack, american, terrorist, attack, violence.

Shooter Articles that give details about the shooter. police, mateen, attack, kill, gunman.

Celebrities various celebrities, who helped support the victims. Top words: music, award, club, host.

Figures 3 & 4 on the next page shows the topic popularity and the sentiment of these topics over time.

B. Discussion

Using the results, we were able to answer the following research questions:

1) What are the popular, sentiment-influencing topics discussed by people and media?: We were able to classify our documents into 9 topics, as mentioned in the previous section. With our model, 9 gave us the most distinct number of topics.

Shooting events tend to re-spark the national debate on gun control. Therefore, We were interested in look at the relationship between gun control and politics, however these two areas could not be separated into separate topics, regardless of the number of topics used. This may imply a correlation, but we'll need to do more extensive analysis to determine the relationships, as correlation does not imply causation.

2) How does a topic popularity change over time: The popularity of topics vary over time. Figure 3 shows the popularities of the different topics. Overall, the popularity of topics decreases over time, with various events bringing up the popularity. During the days following event, the popularity is high with all topics being discussed. Over time, the popularity decreases. However, some events may trigger a response in other topics, such as release of new information about the event. While the exact topics and their popularity may differ from experiment to experiment, this pattern would still be observable.

Within our set of results, Guns & Politics and Terrorism were the first topics to lose their popularity, as various writers seemed to turn their attention towards the victims. In the weeks following the event, victims and Family & Support are talked about as new news about the victims get released over time. Towards the end of September, the last injured victim was released from the hospital, resulting in the spike in the victims topic around that time.

Celebrities got mentioned almost periodically throughout the year, as they held events dedicated to the victims. Some of these topics seem to coincide with general information of the event that got released over time as the investigation continued. Other events, such as announcements about building a memorial for the victims sparked news articles about the victims and their families.

3) how does topic sentiment change over time: Figure 4 shows the sentiment of topics over time. Overall there is a net negative sentiment towards the topics. but release of good news resulted in a more positive sentiment towards the topics.

4) What are the key events that bring about a sharp negative change in topic: We added key events related to the shooting to the sentiment time series plot. This events ranged from information, such as police footage, being released, to news about creating a memorial for the victims. From the graph, it appears that many of them are not correlated with changes in the sentiment. One particular event that appeared to bring about a negative change in sentiment was the release of the 911 call logs on June 30, 2016. These transcripts made evident the terror that the patrons inside the nightclub were experiencing during the shooting [20]. There was also a noticeable increase in both positive and negative sentiment around the one-year anniversary of the shooting. Also around this time, many of the victims who survived the shooting were sharing their stories, which could have affected the sentiment of the news articles. The following are the main events that brought a negative change in sentiment:

- 6/30/16: Emergency call logs released – these showed the panic and terror inside the nightclub when the shooting was occurring
- 7/12/16: One month anniversary of the shooting
- 7/14/16: Police release the first video from the shooting
- 8/1/16: Owners of the nightclub announce it will reopen as a memorial
- 6/1/17: Police release body cam footage from the shooting

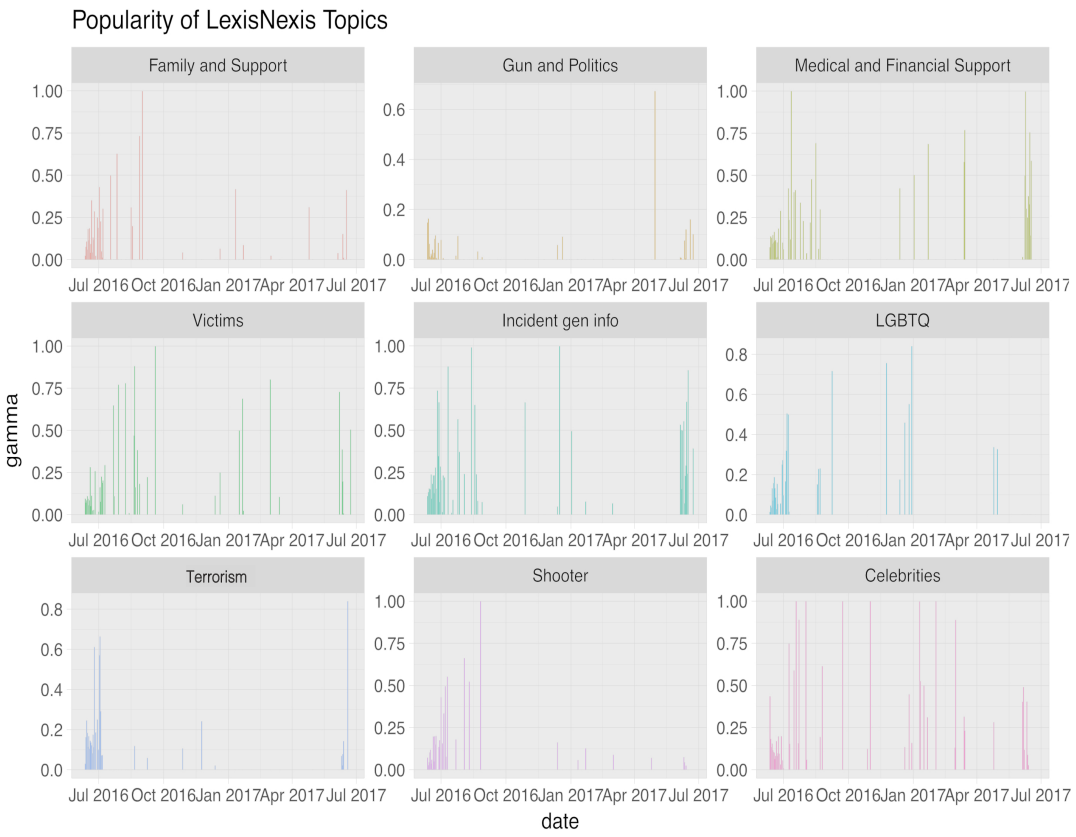

Fig. 3. LDA Result

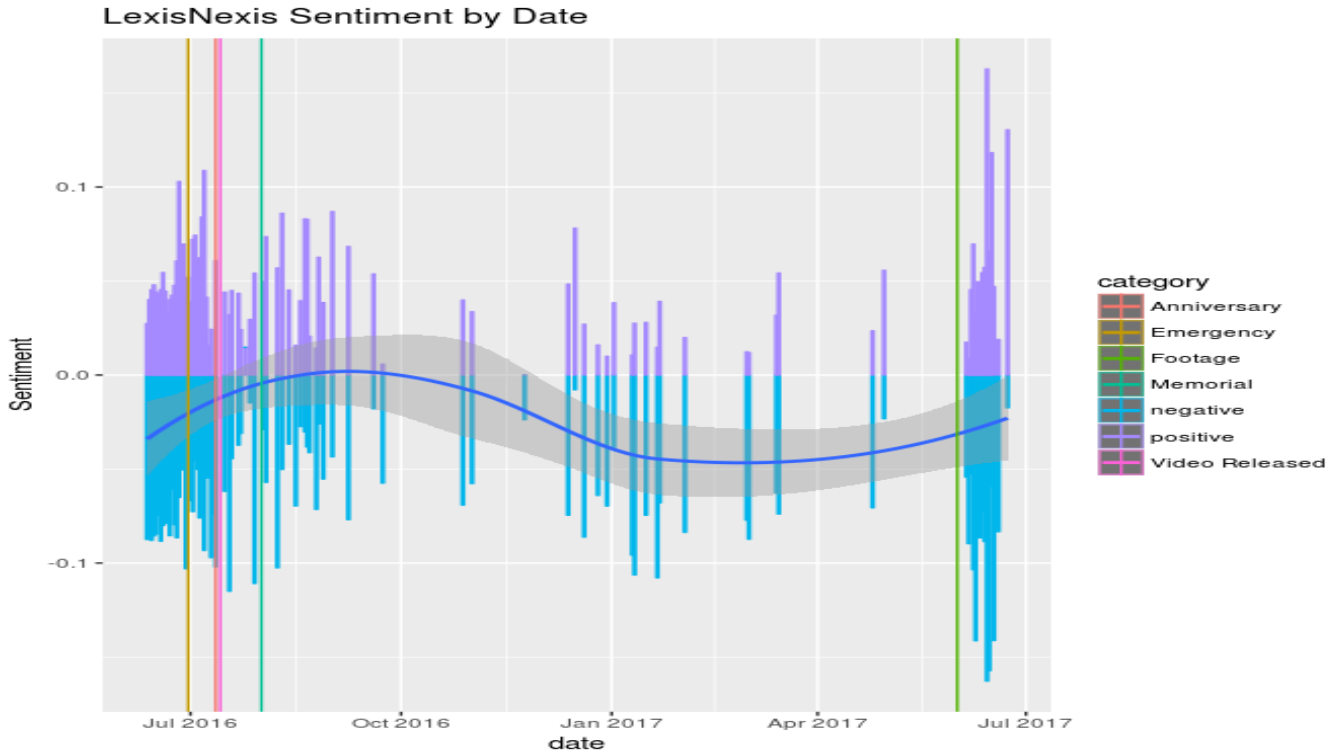

Fig. 4. LexisNexis Sentiment Time Series

- 6/12/17: one year anniversary of the shooting

VI. Limitations

A. Analysis

Language is not always used in its literal sense, and that is a major limitation in the analysis methods. The simple bag of words with dictionary based model that we chose for sentiment analysis cannot capture the context in which the words are used, due to the assumption that the words are independent. In addition, the dictionaries used were based on uni-gram model, which also failed to capture the sentiment of compound words and negation. The dictionary based approach also limits the number of words that were actually measured, since the words must overlap with the entries in the dictionary. Hence, small dictionary size such as Afinn actually failed to capture quite a lot of words, which may or may not affect the overall sentiment of the text. Complex use of words and phrases such as sarcasm and metaphor were also not captured by the model.

Conclusion and Future work

In conclusion, by performing Topic Modeling, we were able to identify sentiment-influencing, key topics that were the main focus for discussion in Twitter and news articles. We had pretty interesting result that showed how the topic of discussion kept changing with time, with 'Family and Support' peaking closer to the event, discussions about victims peaking around two months after the incident, talks on 'Terrorism' were more prevalent during the one year anniversary of the shooting, and of course, 'Celebrities' being the constant topic of discussion throughout the year. We also performed Sentiment Analysis to analyze the changing sentiment of Twitter tweets and news articles, over time. The analysis was performed by normalizing by document length and taking the average for each day. The resulting graph showed us that the Twitter tweets were more positive closer to the event that news articles which reversed roles in about a month after the incident. Due to Twitter data limitations, we only analyzed the Lexis Nexis data to map events that brought a change in sentiment over time. We could identify how emergency call log release had a significant negative sentiment change and how the release of the last victim had a significant positive change in sentiment, while the one year anniversary stirred up both positive and negative sentiments as people were remembering the terrifying attack and at the same time, trying to support the victims and pay tribute to those who lost their lives in the incident.

We plan to take this research work forward by performing a similar analysis on the Las Vegas Shooting incident. Though both the events had a different shooter motives, they both stand at the top of the list for the most deadliest mass shootings in the United States. Therefore it would be very interesting to compare the incidents to find any similarity or differences in sentiment trends as well as topics of discussions. Another reason that this analysis would be interesting is because, the two incidents happened during different presidential reign. Comparing them could give raise to topics like, did a change in presidency affect public sentiment related to mass shootings? Did it affect people's views on topics like terrorism and gun control?

In future, we would also like to perform community detection techniques since the Orlando incident involved three different communities namely, the L.G.B.T.Q, the Latinos and the Muslim community.

Acknowledgment

We would like to thank our Dr. Hyunjin Seo for providing us with the Twitter tweets dataset that helped us a lot with our research. We also thank Dr. Nicole Beckage for giving us the opportunity to perform this research work and guiding us throughout the process.

References

[1] F. K. A. Fantz and E. C. McLaughlin, "49 killed in florida nightclub terror attack," 13-Jun-2016.
[2] E. Peralta, "Putting 'deadliest mass shooting in u.s. history' into some historical context," 13-Jun-2016.
[3] L. Stack, "Before orlando shooting, an anti-gay massacre in new orleans was largely forgotten," 14-Jun-2016.
[4] A. Swanson, "The orlando attack could transform the picture of post-9/11 terrorism in america," 12-Jun-2016.
[5] M. Chong, "Sentiment analysis and topic extraction of the twitter network of #prayforparis," 2016.
[6] . S. T. M. Kumar, A., "Sentiment analysis on twitter," 2012.
[7] X. B. V. I. R. O. P. R. Agarwal, A., "Sentiment analysis of twitter data," 2011.
[8] C. N. Lau, J.H. and T. Baldwin, "On-line trend analysis with topic models: #twitter trends detection topic model online," 2012.
[9] J. W. J. H. E. L. H. Y. W. Zhao, J. Jiang and X. Li., "Comparing twitter and traditional media using topic models," p. 338349, 2011.
[10] D. B. Hong, L., "Empirical study of topic modeling in twitter," 2010.
[11] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.
[12] B. Grn and K. Hornik, "topicmodels: An r package for fitting topic models," *Journal of Statistical Software, Articles*, vol. 40, no. 13, pp. 1–30, 2011.
[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
[14] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy, *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*, pp. 391–402. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
[15] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," *Neurocomputing*, vol. 72, no. 7, pp. 1775 – 1781, 2009. Advances in Machine Learning and Computational Intelligence.
[16] R. Deveaud, E. Sanjuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval."
[17] F. Å. Nielsen, "Afinn," mar 2011.
[18] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.
[19] J. Silge and D. Robinson, "tidytext: Text mining and analysis using tidy data principles in r," *JOSS*, vol. 1, no. 3, 2016.
[20] M. Bergin, "One year later: a timeline of the pulse nightclub shooting and its aftermath," 12-June-2017.