Analyzing Reactions to Major Events Using Topic Modeling of Twitter

Robert Turner
Mentors: Bret Hanlon, Fred Boehm
University of Wisconsin-Madison
Department of Statistics
Apr. 2016

Abstract

Each day millions of Tweets are sent from users worldwide concerning a variety of events and topics. Parsing through so much data can be difficult, but if successful would provide insight into public opinion and reactions. One possible avenue for this investigation is through the use of topic models. A topic model assumes that a Tweet would be created by a weighted selection from a variety of topics, followed by selecting major terms from these topics. In this way we can find clusters of co-occuring terms and determine the focus on people on Twitter. To examine this approach topic models were fit to the periods of time immediately before and after the April 2015 Earthquake in Nepal and the January 2015 Superbowl. We hoped to see topics concerning these events appear in the model and to see the models change over time. We discovered that topics concerning these events appeared shortly after both events. In the case of the Superbowl there were numerous topics, but these topics did not persist for long following the end of the game. Topics for the Superbowl were also seen shortly before the event started. For the Nepal Earthquake there was only one topic referencing the event, but this topic persisted throughout the entire time period. These results suggest that the topic modeling approach can be useful for detecting major events through Twitter and may also shed some light on public opinion.

Introduction

Topic models are multivariate models which represent a set of estimated probability distributions for words across a set number of topics. To create a document from this model, a probability distribution for each document across the topics is created, then words are chosen from each topic to create the final document (1,4,10,11). A document can be any body of text such as an author's body of work or text mined from popular websites. These topics represent blocks of co-occuring words, and ideally represent a summary of the major topics of interest or themes of a body of work. This allows for the examination of a large corpus of text. In this paper text downloaded from Twitter was analyzed using this topic model framework to determine the major points of interest discussed on Twitter. As more than 500 million tweets are sent daily, summarizing Twitter could provide an insight into what the average citizen is interested in.

Due to the huge volume of tweets sent daily, it can be very difficult to access and handle a significant portion of the Tweets sent within a given time period. The least expensive method of downloading Tweets is using Twitter's Streaming API, which promises a maximum of 1% of all Tweets for a given time period. Twitter also allows access to a "Firehose" feed which allows access to all Tweets sent but at a high price. While this small sample may seem unable to accurately represent all of Twitter, previous research has shown that not only does the Streaming API show good performance when detecting popular hashtags and keywords, it also allows access to far more Tweets than advertised, providing an average of 43.5% of Tweets when compared to the Firehose feed (2). While the Streaming API shows some signs of non-random filtering when compared to random samples of the same size created from the Firehose feed, these differences were small when the Tweet coverage was high.

With access to a significant portion of Tweets and the ability to discover the major topics of interest in a body of text, we endeavored to use topic modeling to analyze Tweets over time and see how the model changes in reaction to major events. In this paper we model the Tweets following the April 2015 earthquake in Nepal, and the 2015 Super Bowl. The earthquake was chosen as it represents a well known disaster, and was followed by people worldwide. We expected that topic models fit to Tweets sent after this event would

have fewer topics as the Nepal Eartquake eclipses other discussion. The Super Bowl is another popular topic on Twitter, but is popular due to its entertainment value rather than its severity. Models fit to Tweets sent after the Super Bowl were expected to have more varied topics, as the game has many unique facets to be discussed. The players, the game itself, the commercials, and discussion about the fans could potentially all generate their own topic. The time surrounding these events were broken up into small intervals and a topic model was fit to each interval to determine how Twitter reacted to these major events.

Methods

Dataset To access the Twitter Streaming API in R the streamR package created by Pablo Barbera was used (3). This package allowed us to download and use Tweets directly from the API within R. As shown before these Tweets provide close to random sample of Tweets sent within a time period and should perform well when performing model fitting. Tweets concerning the Nepal Earthquake were taken from 4-24-2015 at 18:11 GMT to 4-28-2015 at 6:11 GMT. Tweets concerning the Superbowl were taken from 2-1-2015 at 5:30 GMT to 2-7-2015 at 5:30 GMT. Tweets were downloaded in blocks, restarting the download every 5 minutes to provide robustness against potential server issues. Only Tweets originating from the US and sent in English were considered. Following the download the Tweets were cleaned of emoticons, leftover unicode, links, hashtags, and other garbage data which would affect later model fitting. This cleaning code is provided in the parseTweetFiles package on the referenced Github (appendix).

Model Fitting To analyze the data we fit a Bayesian topic model as described in Taddy (2012) (1). In this case a K-topic model assumes that each Tweet x containing n characters follows a multinomial distribution, or, $X \sim \text{MN}(\omega_1\theta_1 + ... + \omega_k\theta_k)$ where each ω_i is a weight specifying how much of topic i is represented in the document and each θ_i is a probability vector specifying how likely each word is to appear in a topic. The MAP estimates of θ and ω are used for this model specification. When topics are well defined, the weights can be used to determine the subject or subjects of a Tweet. To fit this model we used the maptpx R package referenced from Taddy (2012). This package fits topic models using independent Bayesian priors for θ and ω and finds the MAP estimation for the model parameters. While there are other approaches such as the Variational Bayes approach(4) or an MCMC model(11), the maptpx package fitting runs faster than other R packages and also allows us to compare models with different numbers of topics. It does this by calculating the Bayes Factor of each K-topic model when compared to the null model (1 topic). The model with the largest Bayes Factor is considered the best model for a given corpus. The VB or MCMC model would require a different method of choosing K.

To analyze the change of Tweets over time we first broke the full dataset of Tweets for a given event into multiple time intervals, each interval being considered as a separate corpus. The length of each interval was set at 1.5 hours. This length was chosen as larger intervals resulted in occassional integer overflow problems during model fitting, and smaller intervals were thought to show little differences from one another. Topic models ranging from 5 to 55 topics were fit to each corpus and the best model was found. This range was chosen as topic models with less than 5 topics were thought to be too restrictive and models with more were thought to be too difficult to parse when fitting multiple models. This resulted in a number of topic models with varying numbers of topics for each time interval. The models were then visualized using the LDAvis R package (5). This package displays each topic on a PCA plot of the first two principal components of the word-topic distributions for a given model(θ). This visualization also displays the primary words of a topic, and can be adjusted to display words than appear frequently within a topic or words which appear infrequently in every other topic. This visualization, and the Bayes Factors for each model fit were analyzed as the results of this study.

Results

The algorithm chose K topic models with K ranging from 7-25 for each interval of Tweets. The Bayes Factors for models on each time period showed the expected pattern; an increase as K increases until a peak, then a decline as K increases further. Models fit to Tweets sent immediately after each event show key words for

each event. Topics referencing the Nepal disaster were seen for the entire period of time analyzed in this paper, while topics referencing the Superbowl were less common after February 2nd. Nepal keywords were among the top 30 most common words seen during each time interval after the event. Superbowl keywords were among the top 30 words during the event, but were only common during the hours after the event. For the time interval of tweets taken during the first half of the Superbowl, numerous topics referencing the event were scene. Conversely, there was only one topic referencing the Nepal Earthquake during the time period analyzed. The visualizations generated by this project can be found on the referenced Github page (appendix).

Discussion

The topic model fitting procedure was successful in detecting these two events taking place. Topic models fit to Tweets sent after the events included topics with key words referencing both events. Breaking the Tweets into smaller intervals would allow for an examination of how quickly Twitter users respond to these events. These topics persisted longer for the Nepal Earthquake than for the Superbowl, suggesting that the disaster attracted more discussion on Twitter than the Superbowl. Though, as discussed in a similar paper on event detection the number of topics discussed as a result of the Superbowl is much higher than for other events (6). While Tweets referencing the Nepal Earthquake will almost always include the terms "nepal" and "earthquake", Tweets about the Superbowl can focus on the game, the players, the coaches, or even the commercials. This dividing of focus may result in these topics becoming more difficult to detect as interest in the Superbowl wanes. The differences in the number of topic models weakly fits the activity level on Twitter (as measured in Longley 2014), suggesting that more discussion not only generates more Tweets, it also generates more topics of discussion.

While notable keywords pertaining the events in question were found in the topic models, well-defined topics were hard to determine. As the number of topics fit and the contents of these topics seemed fairly consistent, perhaps removing the most popular terms in the dataset would allow a closer examination of reactions to these specific events as opposed to the average sentiment of Twitter. Similarly, removing swear words and slang which seem to be evenly and randomly distributed among the topics may also result in more informative topics. The lack of well-defined topics may also be due to the smaller time periods used when fitting topic models. While the smaller interval size was necessary due to integer overflow issues, the overflow was tied to the number of documents fit, not the word count. Compressing a number of Tweets for a given time period into a larger document then fitting on a larger time interval across these documents may fix the overflow problem and provide a better fit. This is reminiscent of the method of aggregating Tweets by users recommended by Hong & Davidson (2010), which showed that by creating larger documents by combining Tweets from the same users topic model performance can be improved (7).

In addition to altering the dataset, different topic models could be fit to this data. Wallach 2009 (8) provides a number of suggestions for determining the number of topics in a topic model, which would allow for this paper to be repeated using alternative topic models, potentially creating more intuitive topics. Comparing or combining the results of multiple different topic models or topic models fit on different document collections of Tweets could also produce better results. However, despite these concerns this simple topic modeling procedure was able to detect these major events occuring and also react to the change in interest over time. The automated selection for the number of topics also reveals a trend to the number of topics in a model. More active time periods on Twitter showed a larger number of topics in addition to more Tweets. Despite the possible avenues for improvement this study shows that topic modeling can be an effective method of event detection using Twitter and can also model the response to these events over time.

References

1) Taddy, M. (2012). On Estimation and Selection for Topic Models. Journal of Machine Learning Research.
2) Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ArXiv.org.
3) Pablo Barbera (2015). streamR: Access to Twitter Streaming API via R. R package version 0.3.2.

4) Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research.
5) Carson Sievert and Kenny Shirley (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces.
6) Buntain, C., Lin, J., & Golbeck, J. (2015). Learning to Discover Key Moments in Social Media Streams. Arxiv.
7) Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics - SOMA '10.
8) Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09
9) Adnan, M., Longley, P. A., & Khan, S. M. (2014). Social dynamics of Twitter usage in London, Paris, and New York City. First Monday, 19(5).
10) Jockers, M. L. (2014). Text analysis with R for students of literature. Cham: Springer-Verlag.
11) Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Supplement 1), 5228-5235. doi:10.1073/pnas.0307752101

Appendix

Github Paper Page: https://github.com/rturn/Topic-Modeling-Twitter
Github Package link: https://github.com/rturner/parseTweetFiles