American Journal of Management Science and Engineering
2019; 4(3): 49-55
http://www.sciencepublishinggroup.com/j/ajmse
doi: 10.11648/j.ajmse.20190403.12
ISSN: 2575-193X (Print); ISSN: 2575-1379 (Online)

SciencePG
Science Publishing Group

Social Media Text Data Visualization Modeling: A Timely Topic Score Technique

Zhenhuan Sui

Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio, USA

Email address:
suizhenhuan@gmail.com

To cite this article:
Zhenhuan Sui. Social Media Text Data Visualization Modeling: A Timely Topic Score Technique. *American Journal of Management Science and Engineering*. Vol. 4, No. 3, 2019, pp. 49-55. doi: 10.11648/j.ajmse.20190403.12

Received: April 7, 2019; **Accepted**: June 5, 2019; **Published**: July 26, 2019

Abstract: Due to the rapid growth of large size text data from Internet sources like Twitter, social media platforms have become the more popular sources to be utilized to extract information. The extracted text information is then further converted to number through a series of data transformation and then analyzed through text analytics models for decision-making problems. Among the text analytics models, one particular common and popular one is based on Latent Dirichlet Allocation (LDA), which is a topic model method with the topics being clusters of words in the documents associated with fitted multivariate statistical distributions. However, these models are often poor estimators of topic proportions. Hence, this paper proposes a timely topic score technique for social media text data visualization, which is based on a point system from topic models to support text signaling. This importance score system is intended to mitigate the weakness of topic models by employing the topic proportion outputs and assigning importance points to present text topic trends. The technique then generates visualization tools to show topic trends over the studied time period and then further facilitate decision-making problems. Finally, this paper studies two real-life case examples from Twitter text sources and illustrates the efficiency of the methodology.

Keywords: Text Analytics, Natural Language Processing, Cyber Security, Signaling, Pattern Detection, Social Media

1. Introduction

Previous studies have described some of the many possible uses of text analytics and social media [1]. Researchers have used social media as an observation source for timely decision making in project investment decision-making problems [2]. Some have also studied text in social media association and key word identification to track the speed with which information travel and the paths that the information takes [3]. Specifically for social media like Twitter, a method has been presented for predicting the spread of information in a social network using retweets as positive feedback and lack of retweets as negative feedback [1]. The number of retweets can be used as an important indicator in the prediction model for social events and changes. Studies used a Bayesian approach to develop a probabilistic model for the evolution of retweet counts [4, 5]. Their model successfully predicted the final total number of retweets through the time-series path of retweets. In this paper, Twitter feeds will also be the main sources of text information study and retweet counts are used as an indicator of importance and the proposed point-based “importance score” can be viewed as an approximate estimate of retweet counts.

Building on Latent Dirichlet Allocation (LDA), studies proposed subject matter expert refined topic (SMERT) for probabilistic clustering of texts to permit experts or users to edit the topics using knowledge about the system or their own needs [6, 7]. SMERT and LDA estimate the proportion of words in the overall corpus on each topic. As a special case of LDA, SMERT potentially incorporates “high-level” inputs from a subject matter expert to adjust the topics and clusters by zapping or boosting words in the topic definitions. Another study applied the SMERT model to course evaluation analysis [8]. Using Pareto charts, this method helped to screen out less effective feedback and allow researchers to focus on relevant and important information.

Topic models and SMERT have shown promise for creating intuitive summaries of bodies of text. But there are issues with estimation and in particular topic proportions are often poorly estimated and fail to capture what is new temporarily in the topic proportions. Therefore, this article proposes a visualization and point-base system designed to help users with sense-making of social media text data. The goals of this paper are to overcome the reported estimation issues from the SMERT models and demonstrate the value of the point system in relation to social media text data monitoring.

In this paper, we would explore the use of timely topic score technique for text visualization for improving sensemaking and monitoring. The specific case studies and examples that we use relate to improving the situation-awareness of system administrators in cyber security contexts. Cyber-security is a growing field of study due to the growing use of data collection and the use of newer internet enabled devices. Therefore, this paper will investigate through examples of the connection between cyber-security and social media, in particular Twitter, in addition to their individual importance.

The remainder of this paper is structured as follows. Section 2 will review LDA models and SMERT model. Section 3 proposes the point system associated a visualization method, which could aid in many text-related sense-making cases. In Section 4, we describe motivating examples relating to cyber vulnerabilities and then the proposed methods are illustrated through the two cyber-security case studies. Finally, we summarize our findings and suggest opportunities for future research.

2. Topic Models

In this section, we review the LDA model which is a probability distribution [9]. Note that virtually all text modeling methods begin with a natural language processing step in which text is transformed into numbers with irrelevant words removed and words "stemmed" (e.g., "jumping" and "jumps" both shorted to "jump") [10, 11, 12].

2.1. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) model is a statistical model that believes documents are random mixtures over latent topics and each topic is characterized by a distribution over the words. Assume that $w_{d,j}$ is the j^{th} word in d^{th} document with d = 1, …, D and j = 1, …, N_d, where D is the number of documents, and N_d is the number of words in the d^{th} document. Therefore, $w_{d,j} \in \{1, \ldots, W\}$, where W is the number of distinct words in all documents. The clusters or "topics" are defined by the estimated probabilities, $\hat{\phi}_{t,c}$, which signifies a randomly selected word in cluster t = 1, …, T (on that topic) achieving the specific value c = 1, …, W. Also, $\hat{\theta}_{d,t}$ represents the estimated probability that a randomly selected word in document d is assigned to cluster or topic t. The model variables $z_{d,j}$ are the cluster assignments for each word in each document, d = 1, …, D and j = 1, …, N_d. Then, the joint probability of the word $w_{d,j}$ and the parameters to be estimated, $(z_{d,j}, \theta_{d,t}, \phi_{t,c})$, is:

$$P\left(w_{d,j}, z_{d,j}, \theta_{d,t}, \phi_{t,c} \middle| N_d, \alpha, \beta, d = 1, \ldots, D, t = 1, \ldots, T, c = 1, \ldots, W\right)$$

$$= \left[\prod_{t=1}^{T} \frac{\Gamma(\sum_{c=1}^{W} \beta)}{\prod_{c=1}^{WC} \Gamma(\beta)} \prod_{c=1}^{W} \phi_{t,c}^{\beta-1}\right]\left[\prod_{d=1}^{D} \frac{\Gamma(\sum_{c=1}^{W} \alpha)}{\prod_{c=1}^{WC} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha-1}\right] \times \left[\prod_{d=1}^{D}\prod_{t=1}^{T} \theta_{d,t}^{n_t^{(d)}}\right] \times \left[\prod_{t=1}^{T}\prod_{c=1}^{W} \phi_{t,c}^{n_t^{(c)}}\right]$$

where $\Gamma(\ldots)$ is the gamma function and:

$$n_t^{(d)} = \sum_{j=1}^{N_d} \sum_{c'=1}^{W} I\left(z_{d,j} = t \,\&c = c'\right) \text{ and } n_t^{(c)} = \sum_{d=1}^{D} \sum_{j=1}^{N_d} I\left(z_{d,j} = t \,\&\, w_{d,j} = c\right) \quad (1)$$

and where $I(\ldots)$ is an indicator function giving 1 if the equalities hold and zero otherwise.

Note equation (1) is a simple representation of human speech in which words, $w_{d,j}$, and topic assignment, $z_{d,j}$, are both multinomial draws associated with the given topics. The probabilities $\phi_{t,c}$ that define the topics are also random with a hierarchical distribution. The estimates that are often used for these probabilities are Monte Carlo estimates for the posterior means of the Dirichlet distributed probabilities $\hat{\theta}_{d,t}$ and $\hat{\phi}_{t,c}$, produced by low values or diffuse prior parameters α and β.

To estimate the parameters in the LDA model in equation (1), "collapsed Gibbs" sampling is widely used. First the values of the topic assignments for each word $z_{d,j}$ are sampled uniformly [13, 14, 15]. Then, iteratively, multinomial samples are drawn for each topic assignment $z_{d,j}$ iterating through each document d and word j using the last iterations of all other assignments $z_{-(d,j)}$. The multinomial draw probabilities are

$$P\left(z_{d,j} = t \middle| d, j, z_{-(\boldsymbol{d},\boldsymbol{j})}, w_{d,j}\right) \propto \left(\frac{n_t^{(w_{d,j})} - I(z_{d,j}=t\ ) + \beta}{n_t^{(\cdot)} - I(z_{d,j}=t\ ) + W\beta}\right)\left(\frac{n_t^{(d)} - I(z_{d,j}=t\ ) + \alpha}{n_{\cdot}^{(d)} - 1 + T\alpha}\right) \quad (2)$$

where $n_t^{(w_{d,j})} = \sum_{d'=1}^{D} \sum_{j'=1}^{N_d} I\left(z_{d',j'} = t \,\&\, w_{d',j'} = w_{d,j}\right)$,
$n_t^{(\cdot)} = \sum_{d'=1}^{D} \sum_{j'=1}^{N_d} I\left(z_{d,j'} = t\ \right)$,
$n_t^{(d)} = \sum_{j=1}^{N_d} \sum_{c'=1}^{W} I\left(z_{d,j} = t \,\&c = c'\right)$, and
$n_{\cdot}^{(d)} = \sum_{t'=1}^{T} \sum_{j'=1}^{N_d} I\left(z_{d,j'} = t\right)$.

In words, each word is randomly assigned to a cluster with probabilities proportional to the counts for that word being assigned multiplied by the counts for that document being assigned. After M iterations, the last set of topic assignments generate the estimated posterior means using:

$$\hat{\phi}_{t,c} = \frac{n_t^{(c)} + \beta}{n_t^{(\cdot)} + W\beta} \quad (3)$$

And the posterior mean topic definitions using:

$$\hat{\theta}_{d,t} = \frac{n_t^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha} \quad (4)$$

Therefore, if words are assigned commonly to certain topics by the Gibbs sampling model, their frequency increases the posterior probability estimates both in the topic definitions $\hat{\phi}_{t,c}$ and the document probabilities $\hat{\theta}_{d,t}$.

2.2. Subject Matter Expert Refined Topic (SMERT) Model

In practice, not all of the distribution is relevant to the user and the topics can be represented by ordered lists of words which users often find interpretable. SMERT generalizes LDA in that it incorporates input from a Subject Matter Experts (SMEs) or ordinary users. The method derives the main topics with a body of documents and estimate what portion of the text corresponds to each topic. Extended from equations of LDA, SMERT has a distribution as equation (5). The distribution is fitted using collapsed Gibbs sampling which is a form of Markov Chain Monte Carlo. Collapsed Gibbs is an iterative process where the topic assignments and distribution are modified. The topic assignments converge to the samples from the new distribution and are then used for estimations for the topics and proportions.

$$P\left(w_{d,j}, z_{d,j}, \theta_{d,t}, \phi_{t,c} \middle| N_d, \alpha, \beta, d = 1, \dots, D, t = 1, \dots, T, c = 1, \dots, W\right)$$

$$= \left[\prod_{t=1}^{T} \frac{\Gamma(\sum_{c=1}^{W} \beta)}{\prod_{c=1}^{WC} \Gamma(\beta)} \prod_{c=1}^{W} \phi_{t,c}^{\beta-1}\right] \times \left[\prod_{d=1}^{D} \frac{\Gamma(\sum_{c=1}^{W} \alpha)}{\prod_{c=1}^{WC} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha-1}\right] \times \left[\prod_{d=1}^{D} \prod_{t=1}^{T} \theta_{d,t}^{n_t^{(d)}}\right] \times \left[\prod_{t=1}^{T} \prod_{c=1}^{W} \phi_{t,c}^{n_t^{(c)}}\right] \times \left[\prod_{t=1}^{T} \prod_{c=1}^{W} \binom{N_{t,c}}{x_{t,c}} \phi_{t,c}^{x_{t,c}} (1 - \phi_{t,c})^{N_{t,c}-x_{t,c}}\right] \quad (5)$$

What is new here is that $x_{t,c}$ and $N_{t,c}$ are collected from a boost and zap table, and $x_{t,c}$ is the successes out of $N_{t,c}$ Bernoulli trials for all topics $t = 1, \dots, T$ and words $c = 1, \dots, W$. They are the high-level input form SMEs.

3. Timely Topic Score Technique for Text Visualization

In this section, we define additional notations and assumptions for the proposed method. Consider a finite number of text document with L sentences and each sentence is signified as s_l, where $l = 1, \dots, L$.

Our method is based on the SMERT and LDA methods. In either case, the derived topics are denoted t_i, $\forall i \in I$, where I is the set of topic indices. Within each topic, the words are ordered as w_{ij}, $\forall j \in$ J and J is the set of word indices. P_{il} is the estimated mean posterior probability that sentence l falls in the topic t_i. A set of documents is called a corpus and q is the number of top words in each topic that are studied by the subject matter expert. The default is q = 10 words for each topic (top 10). Also, the predicted score or importance number is the *PS*.

Algorithm Outline:

Step 0. Select a corpus of text samples from the relevant time period.

Step 1. Run LDA on the corpus.

Step 2. Loop over each topic t_i, $\forall i \in I$.

Step 2.1 Loop over each word w_{ij}, $\forall j \in the\ first\ q\ words\ in\ the\ topic$, zap w_{ij} if w_{ij} does not make sense. otherwise, boost it. End loops.

Step 3. Run SMERT without sorting topics using the high-level boosts and zaps.

Step 4. Loop over each sentence s_l with property m, $V_{im} = \sum_{l \in m} P_{il}$.

Step 5. Loop over each topic $m \in M$, rank V_{im} from largest to smallest.

Step 6. Select *N* largest values of V_{im}, $V_{nm} = V_{im}$, where $n = 1, \dots, N$.

Step 7. For all $m \in M$ and $n = 1, \dots, N$,

Step 7. 1. If count of topic i, $C_i = 1$, assign predicted score $PS_{1im} = C_{1n}$, where $n = 1 \dots N$, $C_{11} > C_{12} > \cdots > C_{1N}$.

Step 7. 2. If count of topic i, $C_i = 2$, assign predicted score $PS_{2im} = C_{2n}$, where $n = 1 \dots N$, $C_{21} > C_{22} > \cdots > C_{2N}$.

Step 7. 3. If count of topic i, $C_i \neq 1$ or 2, assign predicted score $PS_{im} = 0$.

Step 8. $S_m = \sum_i (PS_{1im} + PS_{2im} + PS_{im} + PS)$, where PS is the constant predicted score for all $m \in M$. End loops.

Step 9. Plot $PS_{1im}, PS_{2im}, PS_{im}, PS$ in a column chart with short phrase extracts from the topic definitions as labels.

For *Step 4*, $V_{im} = \sum_{l \in m} P_{il}$ means to sum all of the probabilities for the same property. Here, the property includes examples like different months, years, or even days. For SMERT, normally 20 topics are selected as outputs. In Step 6, among the 20 topics, normally, $N = 5$, or the top 5 topics are selected in the method in most cases. In the method, predicted scores are normally either equal to predicted numbers or proportional to the predicted numbers.

4. Case Studies

To demonstrate the proposed technique, we use a case from 2014 relating to cyber security. During 2014, there were several major cyber vulnerabilities that became public knowledge. Most notably was the vulnerability commonly known as the Heartbleed. The Heartbleed vulnerability was made public knowledge on April 1, 2014. This vulnerability resulted from a lack of bounds in memory allocations for operating systems. The vulnerability and notification allowed for large amounts of information to be stolen from any susceptible computer. Upon this disclosure many hackers made use of the vulnerability before a patch could be created. As a result, the number of attacks on a large Midwest institution's computers increased by approximately 400% in the month of April as shown in *Figure 1*.

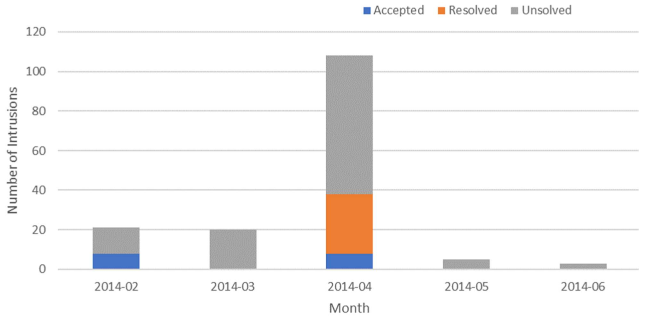

Figure 1. *Known computer intrusions for a large Midwest organization in 2014.*

No doubt, some system administrators at the Midwest organization knew about the Heartbleed vulnerability after the announcement but many did not. Yet, all observed the spike in attacks as detected using the intrusion detection system (IDS). The IDS generally intercepts only a fraction of all attacks so likely some were missed and all administrators needed to perceive the vulnerability and understand its cause. This is the objective of our proposed methods in this article, i.e., to improve situation awareness at all times by synthesizing Twitter feeds into an intuitive chart.

As another example, we consider the November 2014 attack on the Sony Corporation by, reportedly, North Korea. This was a well-publicized attack that received a large amount of media attention. These famous cyber events raise discussions on social media platforms. The proposed methods seek to identify and interpret the events in both cases.

In our research, 15 Twitter broadcasters were analyzed for the purposes of both studies (*Step 0*). These users were found by searching for the Twitter users who have a reputation for being cyber-security analysts. Also, a combination of individuals and organizations/groups were found to ensure there wasn't a bias based on the goal of the Twitter user. The Twitter sources (usernames) in the following examples are: Mathewjschwartz, Neilweinberg, Scotfinnie, Secureauth, Lennyzeltser, Dangoodin001, Dstrom, Securitywatch, Cyberwar, Jason_Healey, FireEye, Lancope, Varonis, DarkReading, RSAsecurity, and Mcafee_Labs.

4.1. Heartbleed Case Study

Figure 2 shows the total number of retweets for the first six months of 2014 which will be compared to the chart later that this new method generates. Notice that the retweet counts correlate with the known intrusion counts confirming that retweet counts often relate to important events.

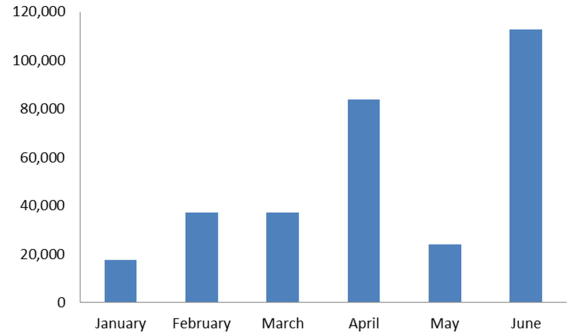

Figure 2. *Retweets for January to June 2014 with the Heartbleed announcement in April.*

Next, we applied the remaining steps in the proposed technique. Below are the topics that SMERT created based upon the tweets and zapping any unwanted words. Steps 1-3 involve applying SMERT. We zapped Heartbleed in February and March because we know from our expertise that there were no tweets about Heartbleed until April after it was publicly announced. The developed topics were then identified by words. Then, we manually translated the word lists into interpretable topics with the results in *Table 1*.

Table 1. *SMERT topics which were interpreted manually incorporating the highest frequency words.*

Number	Topics
1	Jason Healey and cyberwar, among others, tweeted with a moderate following tweeted in a few months about cyber security.
2	RSA security, among others, tweeted about its own products and an event called the archer summit with a small following.
3	Dangoodin001, among others, retweeted topics from many different months, without much following on Twitter.
4	Cyberwar, among others, tweeted about Eric Snowden and the NSA in multiple months
5	MacAfee Lab, Darkread, and dstrom, among others, tweeted about network security it multiple months
6	Dangoodin001 and Darkread, among others, tweeted about the Heartbleed with a moderate following on Twitter in particular during April.
7	Security watch and dangoodin001, among others, tweeted about apps and passwords with a moderate following on Twitter.
8	Lancop tweeted about its own company in particular during February and March with a low number of retweets.
9	Mathewjschwartz and Darkread, among others, tweeted about the target breach and information security with a low number of retweets.
10	Lennyzeltser and security watch, among others, retweeted topics and specifically at a neiljrubenk on Twitter.
11	Mathewjshwartz and dangoodin001, among others, tweeted with a high number of retweets in multiple months.
12	RSA security and Darkread, among others, tweeted about data security in multiple months
13	Fire eye, among others, tweeted about information security, malware, and threats with a moderate number of retweets particularly in April and May.
14	Varoni, among others, tweeted about information security and data privacy in multiple months.
15	Varoni, among others, tweeted about big data and security in multiple months with a low number of retweets.
16	Scotfinnie and security watch among others tweeted about Microsoft windows with a low number of retweets
17	Cyberwar and Dangoodin001 among others tweeted about thanking others in multiple months
18	Varoni and Darkread, among others, tweeted about social media and information security in multiple months.
19	McAfee Lab, among others, tweeted about security stories and particularly to Twitter users davemarcu and Slashdot in multiple months with a low number of retweets.
20	Darkread and Varoni, among others, tweeted about information security and Darkread in particular during April and June.

For both examples we use $N = 5$. Also, for steps 4-8, the predicted score is $PS = 10{,}000$. If the topics are unique among all the 6 months, a predicted score of 65,000, 52,000, 39,000, 26,000, 13,000 is assigned if the topic ranks No. 1 to 5 respectively. If the topics appear twice among all the 6 months, assign predicted score of 15,000, 12,000, 9,000, 6,000, and 3,000 if the topic ranks No. 1 to 5 respectively. *Figure 3* shows the predicted scores with a breakdown by topics (*Step 9*).

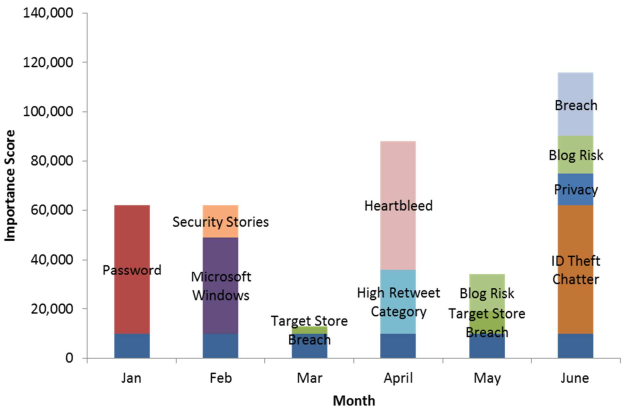

Figure 3. *Heartbleed example predicted score breakdown by topic.*

As can be seen in the figure above, April is characterized by the Heartbleed event and the high retweet category. This makes sense, as the Heartbleed event would cause a few particular announcements and updates to be highly retweeted. This characterizes April as a month which is abnormal and focuses on the Heartbleed event (as we now know is correct from the IDS data in *Figure 1*).

January and February both have slightly elevated PS and predicted retweet numbers as well. The January password focus and February story and system update focus may result in part from the Target store credit card theft in December of the previous year, and an increased focus on cyber security. The target theft involved many peoples credit card information being stolen and was a major event for many individuals who may not think of cyber security very often.

The month of June also had a large number of points associated. June seemed to have a large amount of discussion associated with breaches of security resulting in theft of personal information and privacy issues. However, this system did a good job of predicting the real retweet numbers. The month of April was clearly dominated by discussions of the Heartbleed vulnerability which is exactly what an IT professional would want to know about if they did not know already.

4.2. Shellshock and the Sony Hack Case Studies

Shellshock is a security bug in Unix Bash Shell. It was disclosed on September 24, 2014. Many web server deployments use Bash to process web requests. Therefore, the bug could cause potential vulnerability issues to execute arbitrary commands and allow attackers acquiring unauthorized access to hosts. This bug can be compared to the Heartbleed bug in severity as it could potentially compromise millions of unpatched hosts.

The data set is also from the 15 Twitter accounts as in the Heartbleed example, but the data in these two examples are from July to December in 2014.

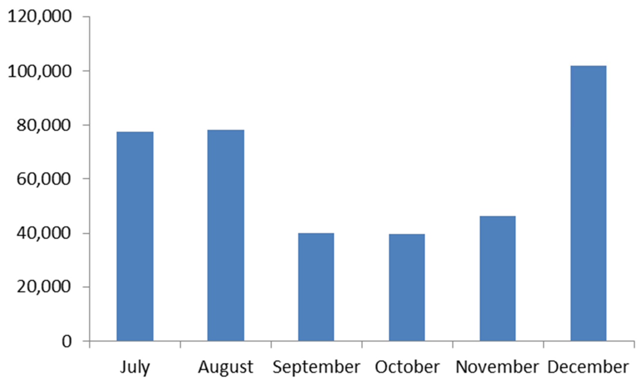

Figure 4. *Retweet numbers from the period involving Shockshell and the Sony hack.*

Figure 4 shows the total retweet number across the 16 accounts. Different from expectations about Shellshock and the Sony Hack events, the total retweet numbers in September and November are not very high compared to other months.

Table 2. *SMERT topics for the second case study during the Shockshell and Sony hack period.*

Number	Topics
1	Lancop tweeted about information security and cyber security for companies during July, August, October, and November with high number of retweets.
2	Lennyzelts, varoni and nealweinberg tweeted about new malware tool in October and December with high number of retweets.
3	Varoni tweeted about big data, information security, and data privacy in July and August with high number of retweets.
4	Mathewjschwartz, darkread and scotfinni tweeted about malware breach for Apple during September to November with high number of retweets.
5	Dangoodin001 and lennyzelts tweeted about year 2014 in August and December with high number of retweets.
6	Cyberwar and jasonhealei tweeted and retweeted about new things on internet during July, August and November with high number of retweets
7	Mathewjschwartz, cyberwar and dangoodin001 tweeted and retweeted about the Sony Hack during December with high number of retweets.
8	Dstrom, mathewjschwartz and cyberwar tweeted and retweeted about great reading and look during September and November with high number of retweets.
9	Secureauth tweeted and retweeted about security authenticity during September and October with high number of retweets.

Number	Topics
10	Securitywatch tweeted and retweeted about online ID security protection during October and November with high number of retweets.
11	Jasonhealei tweeted and retweeted about cyber attack and National Security Agency (NSA) during September and October with high number of retweets.
12	Securitywatch, dangoodin001 and mathewjschwartz tweeted and retweeted about apps on mobile device during July, August and November with high number of retweets.
13	Fireeye tweeted and retweeted about information security during July, August and October with high number of retweets.
14	Dangoodin001 tweeted about thank and questions during July, November and December with high number of retweets.
15	Darkread and dstrom tweeted about cloud data breach and security during July, October, and November with high number of retweets.
16	Rsasecur tweeted about blog, sharing security and RSA summit event during September and December with high number of retweets.
17	Cyberwar, darkread, and mathewjschwartz tweeted about the new bug Shellshock and potential attack during August to October with high number of retweets.
18	Rsasecur tweeted about cyber security threat detection in RSA during October and November with high number of retweets.
19	Mcafeelab tweeted about malware attack and new phishing threat report during July and December with high number of retweets.
20	Varoni tweeted about information security and password hack during July and August with high number of retweets.

After zapping the unwanted words, SMERT output 20 topics as in *Table 2*. Topics 7 is about the Sony Hack and Topic 17 is about Shellshock. In this example, the parameters and importance scores are assigned as the same values from the previous example. Then using the technique, *Figure 5* shows the predicted scores with a breakdown by topic for the Shellshock and the Sony Hack example.

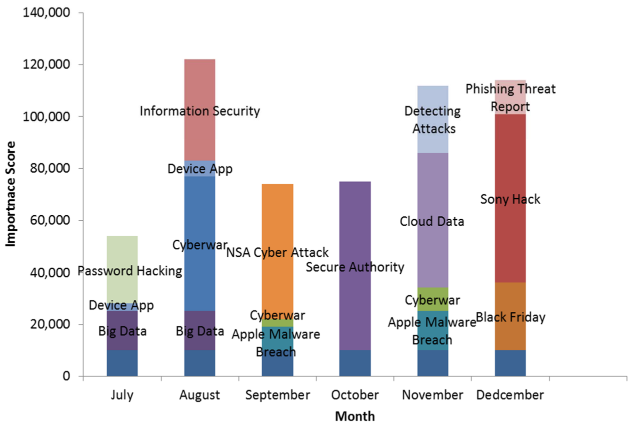

Figure 5. *Shellshock and the Sony Hack Predicted Scores Breakdown by Topics.*

In Figure 4, July, August and December have a higher total retweet number than other months. The reason that July and August have a higher total retweet number may be from discussion of the Unix Bash Shell on social media. Shellshock did not receive its name until September however. The reason that December has a higher total retweet number may be because the Sony Hack happened in late November. Although it aroused active discussions on social media in November, the total retweet does not react to this accident very sensitively due to the late time of the month. But the total retweet number of December behaves as we would expect.

Figure 5 shows that the predicted scores from the proposed technique are more sensitive to the social events than the real total retweet number. There is a peak in the August predicted score and the breakdown of topics for August has shown that the social media users have observed a new information security issue. As discussed in the previous paragraph, the bug was just not named as Shellshock yet. The Shellshock bug being referred to consistently over the time frame means it does not show up as clearly using this method however.

5. Conclusions and Future Research

In this article, we proposed a timely topic score technique for text visualization to aid in monitoring and sensemaking. We illustrated the application of the technique through two data sets related to cyber security. In the first case, the method explained the large uptick of cyber intrusions during the month of April 2014. The chart clearly shows that the uptick corresponded to the Heartbleed vulnerability. Similarly, for the second case study, the Shellshock vulnerability is also readily apparent. Another relevant occurrence (the Sony Hack) is clearly visible. In both case studies, the so-called "importance score" correlated highly with the number of retweets providing confirmation that the method generates relevant information. The technique leverages the explanatory capability of Twitter while simplifying the outputs into a single screen. This can potentially save reading streams from tens or hundreds of content generators.

Yet, a number of topics remain for future study. First, the technique can be compared with alternatives including methods based on more repeatable estimation procedures than collapsed Gibbs sampling. Second, the technique can be made more automatic. Instead of including manually generated labels in Step 9, auto generation can be investigated. Also, the technique based on the simpler LDA may be sufficient without human high-level data generation and the complications of SMERT. Third, the validation could be explored with simulated numerical examples and the related statistical properties can be evaluated. Finally, domains outside of cyber security can be studied. These might relate to sentiment analysis and the interests of populations relating to marketing or military conflicts.

References

[1] Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in Twitter. In *Workshop on computational social science and the wisdom of crowds, nips* (Vol. 104, No. 45, pp. 17599-601). Citeseer.

[2] Allen, T. T., Sui, Z., & Parker, N. L. (2017). Timely decision analysis enabled by efficient social media modeling. *Decision Analysis*, *14* (4), 250-260. https://doi.org/10.1287/deca.2017.0360.

[3] Yang, J., & Counts, S. (2010, May). Predicting the speed, scale, and range of information diffusion in Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*.

[4] Shah, D., & Zaman, T. (2010). Community detection in networks: The leader-follower algorithm. *stat*, *1050*, 2.

[5] Zaman, T., Fox, E. B., & Bradlow, E. T. (2014). A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, *8* (3), 1583-1611.

[6] Allen, T. T., & Xiong, H. (2012). Pareto charting using multifield freestyle text data applied to Toyota Camry user reviews. *Applied Stochastic Models in Business and Industry*, *28* (2), 152-163.

[7] Allen, T. T., Xiong, H., & Afful‐Dadzie, A. (2016). A directed topic model applied to call center improvement. *Applied Stochastic Models in Business and Industry*, *32* (1), 57-73.

[8] Allen, T. T., Vinson, S. M., Raqab, A., & Allam, Y. (2013). Using SMERT to Identify Actionable Topics in Student Feedback. *Integrated Systems Engineering Technical Report 2013*.

[9] Blei, D. M., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation Journal of Machine Learning Research (3).

[10] Allen, T. T., Sui, Z., & Akbari, K. (2018). Exploratory text data analysis for quality hypothesis generation. *Quality Engineering*, *30* (4), 701-712.

[11] Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

[12] Porter, M. F. (1980) An algorithm for suffix stripping. *Program*. 14 (3): 130-137.

[13] Teh, Y. W., Newman, D., & Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in neural information processing systems* (pp. 1353-1360).

[14] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101* (suppl 1), 5228-5235.

[15] Carpenter, B. (2010). Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling. *Rapport Technique*, *4*, 464.