Discovery of Insights on Cybersecurity Education from Twitter Using Analytics

Dr. Azene Zenebe
Associate Professor, Management Information Systems

Bowie State University, Maryland
azenebe@bowiestate.edu

Prof. Tony Yorkman
Lecturer, Management Information Systems

Bowie State University, Maryland
tyorkman@bowiestate.edu

Abstract - There are enormous amount of data generated about various topic or organizations on several matters on cyberspace on a daily basis. The challenge is what and how to extract useful information and knowledge from such data. We used IBM® Analytics™ for the retrieval, extraction and analysis of social media contents on the topic Cybersecurity education. The contents were from Twitter, and the time frame selected was from January 1, 2015 to March 14, 2017. One thousand three hundred eighty seven (1,387) tweets that have both #cybersecurity and #education hashtags were retrieved, relevant data were extracted and analysis was performed using Watson Analytics. We are able to identify patterns and discover useful insights. The trends of the tweets, distribution of the geographic locations (countries and states) and gender of the authors of the tweets were presented. Furthermore, in order to understand the tone of the tweets, results of sentiment analysis were presented including overall sentiments and sentiments by gender as well as by states for USA. The discovered insights such as the several trends and sentiments towards Cybersecurity education can be used for policy and program development in Cybersecurity education, as well as recruitment and retention of students in Cybersecurity education.

Keywords

Twitter Analytics, Cybersecurity Education, Sentiment Analysis, Visual Analytics

1 INTRODUCTION

Social media analytics refers to the systematic and scientific ways to utilize the vast amount of content created by Web-based social media outlets, tools, and techniques. It is done using many analytic methods, including text mining, sentiment analysis, and social network analysis. Companies use it to get a better understanding of their customer base, and can gain financial and competitive advantages from doing so. Governments use it to track potential terrorist threats, which can lead to enhanced national security. Social scientists use it to get a better understanding of how communities and societies work, which can provide guidance on how to best manage these societies.

Closely related to Social media analytics is Text analytics. It is a concept that includes information retrieval (e.g. searching and identifying relevant documents) as well as information extraction, data and text mining, and Web mining. By contrast, text mining is primarily focused on discovering new and useful knowledge from textual data sources. The overarching goal for both text analytics and text mining is to turn unstructured textual data into actionable information through the application of natural language processing (NLP) and analytics. Text mining entails three tasks (Sharda, R., Delen, D., & Turban, E.,2015):

- Establish the Corpus: Collect and organize the domain-specific unstructured data
- Create the Term–Document (T-D) Matrix: Introduce structure to the corpus
- Extract Knowledge: Discover novel patterns from the T-D matrix.

One of the most useful applications of text mining is sentiment analysis. Sentiment analysis tries to answer the question, "What do people feel about a certain

topic?" by digging into opinions of many using a variety of automated tools. It is also known as opinion mining, subjectivity analysis, and appraisal extraction. Sentiment analysis process involves several steps. The first step is called sentiment detection, during which text data is differentiated between fact and opinion (objective vs. subjective). This is followed by negative-positive (N-P) polarity classification, where a subjective text item is classified on a bipolar range. Following this comes target identification (identifying the person, product, event, etc. that the sentiment is about). Finally come collection and aggregation, in which the overall sentiment for the document is calculated based on the calculations of sentiments of individual phrases and words from the first three steps (Sharda, R., Delen, D., & Turban, E., 2015).

In this paper, we use text analytics, also referred as analytics for social media, to analyze invaluable social media content, particularly tweets from the Twitter, and identify patterns and discover insights on Cybersecurity education including sentiments. The discovered insights are important, first for decision makers and then to the public to understand the sentiments and interest towards Cybersecurity education.

The paper is organized into five sections. Section 2 presents the background of the research, followed by the methodology in section 3. Section 4 presents the results and discussions, followed by the conclusion in sections 5.

2 BACKGROUND

There are enormous amounts of digital data generated about various topics, issues or organizations (e.g.) on their products, services and other issues on cyberspace including Facebook, Twitter, forums, reviews, video descriptions and comments, blogs, and news on daily basis. As a result, big data is generated in high volume, velocity and variety as well as with veracity. IBM estimates that every day we create 2.5 quintillion bytes (2.3 trillion giga bytes, i.e., about 10 million blu-ray discs) of data. The data come in different formats: structured and non-structured documents, images, audio and video. The data come with high frequency; per

minute, it is estimated that 204 million emails, 216 thousand Instagram posts, and 72 hours of video are added (IBM Big Data & Analytics Hub). As of June 2016, the numbers of monthly active twitter users worldwide are about 313 million; of which 67 million users are from USA.

It is important that such resources are mined to unlock meaning and provide patterns. Text analytics has been in use to unlock meaning from huge text resources. There is significant advance in text analytics in recent years. One of the leading systems is the IBM Watson Analytics, a product of the IBM research team. Developed in 2010, IBM Watson is a system designed to answer questions raised in human language. It employs text mining and a deep natural language processing (High, 2012). In 2011, in the 1st human-versus-machine match-up, the three Jeopardy Episodes during February 14-16 was presented. Watson did not have access to the INTERNET, but had access to 200 million pages of structured and unstructured content using 4TB storage. Watson out performed both the biggest money winner (Brad Rutter) and the record holder for the longest championship streak for 75 days (Ken Jennings) (Sharda, Delen, & Turban, 2015). Other competitive products are SAS analytics and Tableau.

There are rather limited studies that have explored the sentiment analysis and other analytics on several topics, products, services and organizations in the past based on Twitter data. Research by Camargo, Torres, Martinez, and Gomez (2016) describes a system that allows government planners to analyze citizens' perception of security to Bogota-Colombia. The proposed system uses big data technology to collect, process, index, store, analyze and visualize data from Twitter. Authors implemented a Java-based crawler component using the Twitter Streaming API. The component allows the researchers to collect a set of 2,476,426 tweets of Bogota in a period of 111 days (from August to December of 2015). The API allows the authors to query only for tweets that match with the string "Bogot´a" in the tweet field "place". On a related context, Bouazzi & Ohtuski (2016) proposed an approach that relies on writing patterns and special unigrams to classify 21,000 tweets into 7 different classes, each containing 3000 tweets. The authors work

suggests that instead of focusing on the binary ("positive and "negative") and ternary ("positive", "negative" and "neural") classification, it would be more interesting to study the opinion of users to go deeper in the classification to detect the sentiment behind post.

Twitter has been the focus of numerous recent studies, with a broad range of focus. For instance, Bian, et al. (2016) mined Twitter to understand the public's perception of the Internet of Things (IoT). Search keywords used to define the trend of the IoT were variations of the word "Internet of Things" (e.g., "IoT", and "InternetOfThings") as well as their hashtag versions (e.g., "#IoT" and "#InterentOfThings"). Researchers collected over 2.9 billion raw tweets, however only a fraction of the data (30, 454 tweets) was deemed relevant to the study. Through sentiment analysis using Linguistic Inquiry and Word Count (LIWC), the authors discovered that the public's perception on the Internet of Things is mostly positive for the period of 2009 to 2015.

Based on our search of the literature, there is no previous study that has explored the sentiment analysis and other trends on Cybersecurity education based on Twitter data. Furthermore, our research is unique as it used four-way sentiment analysis: "positive", "negative", "ambivalent", "neutral".

3 METHODS

The main research questions addressed in this study are: What are the temporal patterns of the tweets as indicator of public interest about Cybersecurity education?; What are the spatial patterns of tweets about cyber security education?; What are the sentiments of tweets about cyber security education?; Who are the influential authors of tweets?; and what are the demographics (such as gender) of authors of the tweets?

To analyze the content of tweets and generate dataset a qualitative method, i.e. NLP Natural Language Processing), is used. Quantitative method is used to analyze the dataset to determine trends and other statistical outcomes. There has been enormous progress in the field of business intelligence and analytics through the

application of AI techniques for NLP and machine learning, and visualization for knowledge discovery. Advanced analytic tools like IBM Watson and SAS analytics are products with these capabilities. We use IBM® Watson Analytics™ to collect and analyze the tweets as well as for visual analytics on the resulting dataset. The contents were from Twitter, and the time frame selected was from January 1, 2015 to March 14, 2017. Both hashtags: #Cybersecurity and #education were used. That is, tweets with both #Cybersecurity and #education hashtags were retrieved.

4 RESULT AND DISCUSSION

There were about 658,153 tweets with the #cybersecurity hashtag and 1,115,258 tweets with the #education hashtag. The IBM Watson Analytics retrieved and tagged all the 1,387 tweets posted in English that have both #cybersecurity and #education hashtags, which constitutes the Twitter dataset for this paper. The dataset has 28 columns including Author name, Author friend count, Author follower count, Tweet type, Retweet count, Posted from, Language, Author city, Author state, Author country, Author gender, Sentiment, and Year / Month / Day / Hour Posted. All counts are a snapshot at the time when the Tweets are originally collected.

4.1 Trends and Distributions

Figure 1 shows trend of tweets on the topic by dates. Overall the level of interest on Cybersecurity education was high and stable. It also shows a few periods of high picks in early July 2015 to early August 2015. That is, the graph shows a high number of tweets during July 6th 2015 to August 4th, 2015. This could be due to cyber-attacks or breaches that occurred during the period, or Cybersecurity related events. Some of the breaches include the Hacking Team breach that published more than 1 million emails from the Italian surveillance company, revealing its involvement with oppressive governments as well as multiple Flash zero-day vulnerabilities and Adobe exploits. As result, a list of Hacking Team's customers including military, police, federal and provincial governments were leaked in the

July 2015. Also in July 2015, the Impact team penetrated Ashley Madison's servers and published the information of all 37 million users online.

There are minor spikes on 24th of August 2016, 26th of September 2016, and 10th of November 2016. Overall after 4th of August 2015, the trend of tweets is declining and low. In 2015 there were 749 compared to the 519 tweets in 2016. This indicates discussion about Cybersecurity education was settled. A similar trend is also found using Google trends, see Figure 2.

The visualization, in Figure 3, examined from where the tweets were posted based upon the location information in the author's profile, which does not necessarily reflect the author's location when they created the tweets, indicated that majority were from United States (778 tweets) and from United Kingdom (118 tweets). The visualization also shows a broad distribution of tweets about the topic from 38 different countries.

Figure 4 shows the number of tweets within the United States, where about 40 states are represented. The highest number of the tweets came from DC (124), followed by New York (103), California (74), Texas (46), Louisiana (43) and Colorado (37). Figure 5 also shows of pick of user tweets occurs at 21 hours.

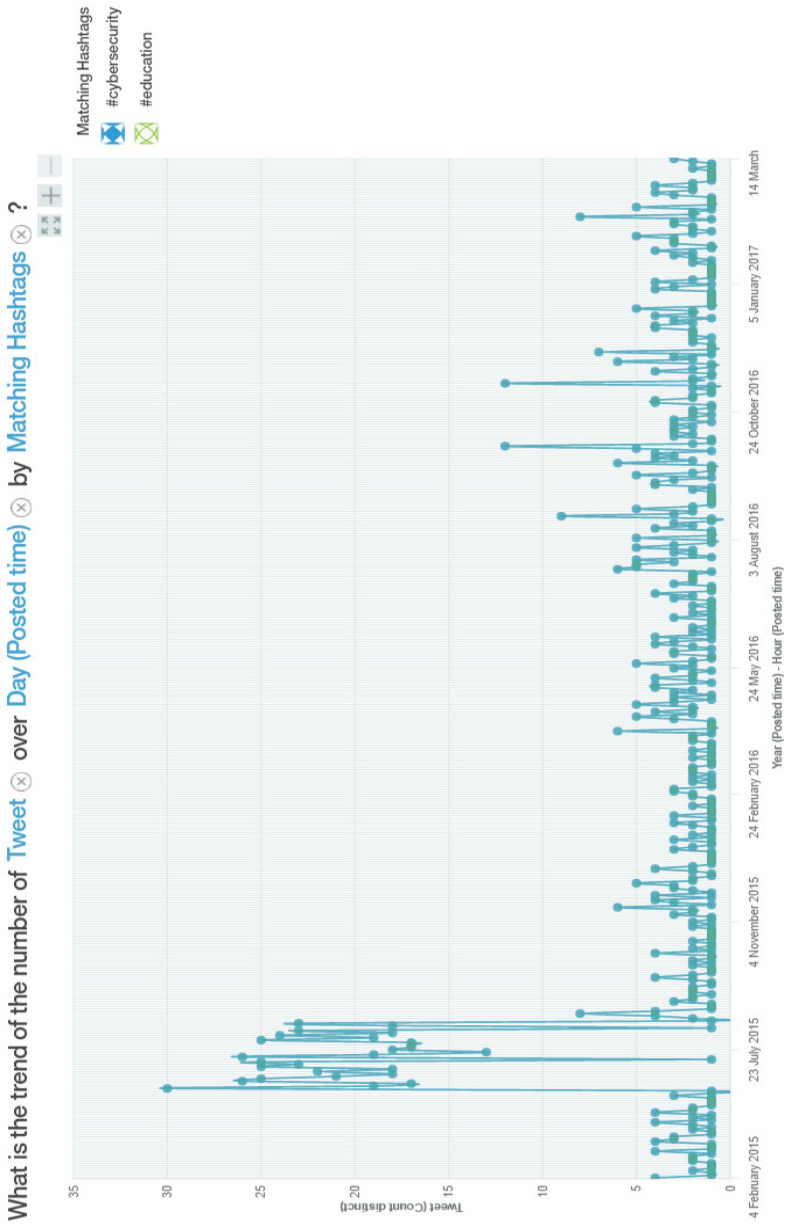

Figure 1. Trend of the number of tweets

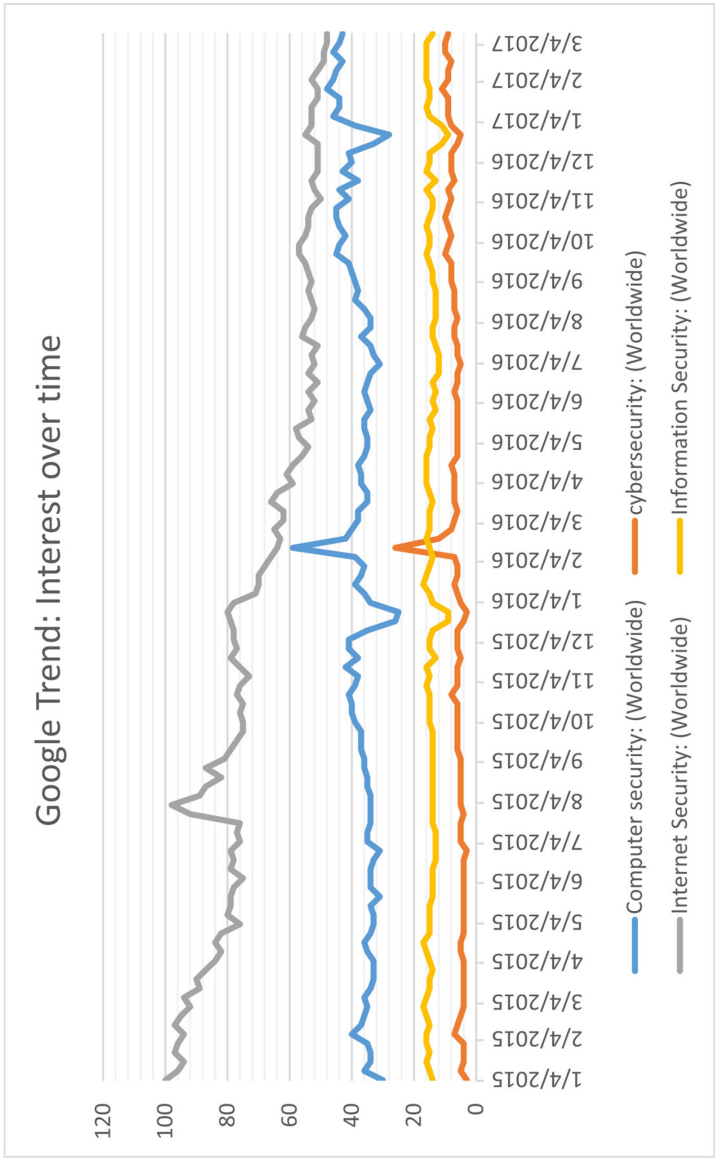

Figure 2. Trend of Interest over time (based on the number search on Google) (Source: https://trends.google.com/trends/)

Figure 3. Number of Tweets by Country

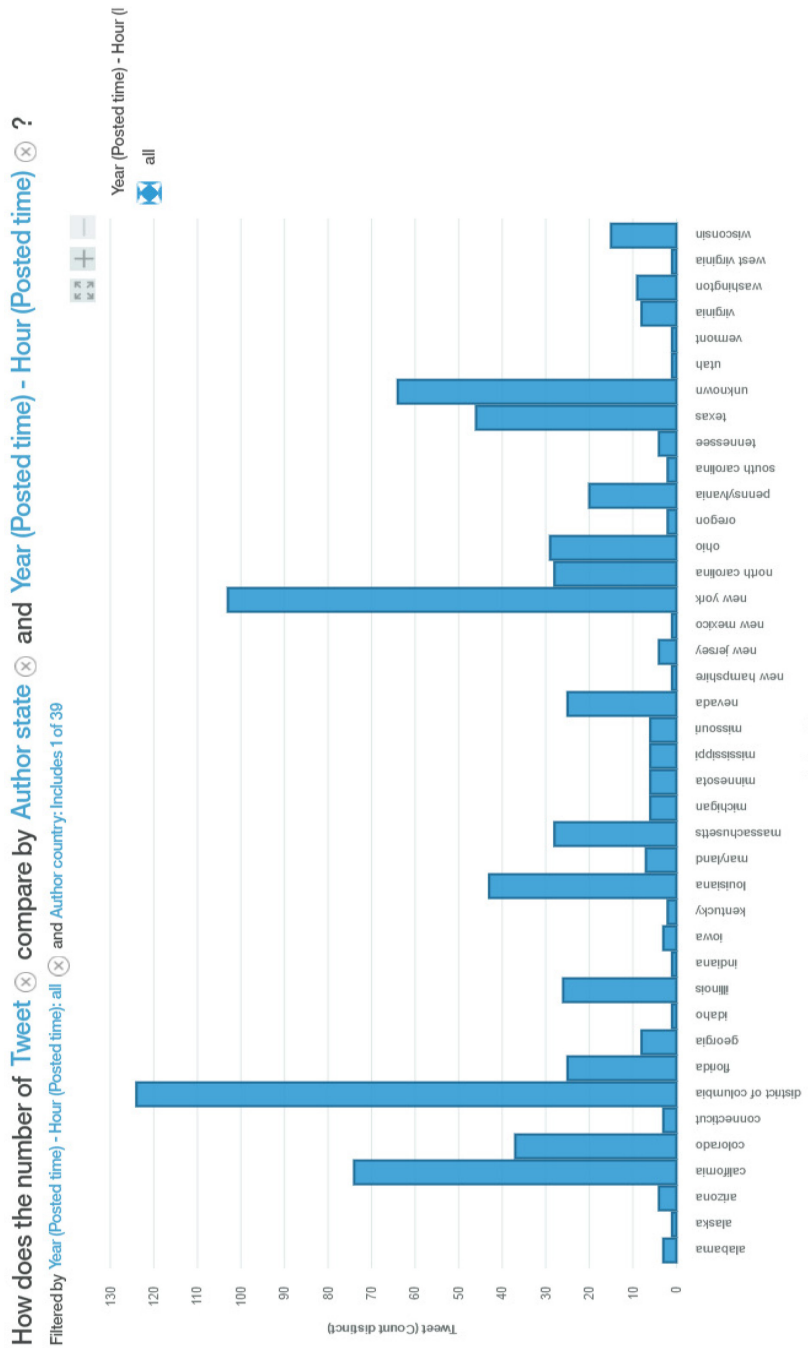

Figure 4. Number of Tweets by States in USA

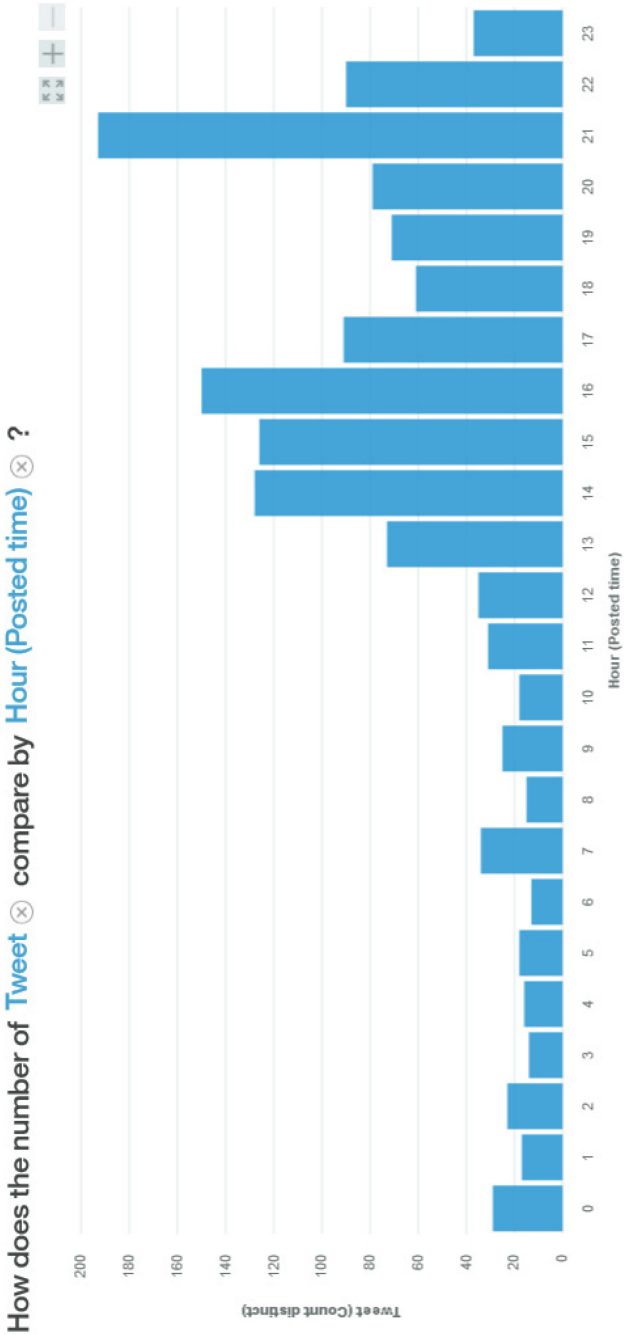

Figure 5. Pick hours of Tweets

4.2 Sentiment Analysis

Sentiment terms are words that measure the tone of a tweet. It indicates whether a tweet is positive, negative, ambivalent, or neutral. A tweet is categorized as ambivalent when it has the same number of positive and negative sentiment terms. A tweet is categorized as neutral when there are no sentiment terms that are detected in it. Table 1 presents the relative sentiment distribution by year. Overall the system identifies 44% positive, 3% negative, 51% neutral, and 1% ambivalent sentiments. Moreover, the positive sentiment is declining, and the neutral and negative sentiments are increasing during the three years.

	Positive (P)	Ambivalent	Neutral	Negative (N)	Unknown	Total
2015	481	4	246	10	8	749
%	64%	1%	33%	1%	1%	
2016	115	14	365	25		519
%	22%	3%	70%	5%		
2017	20	0	91	8		119
%	17%	0%	76%	7%		
Total	616	18	702	43	8	1387
%Total	44%	1%	51%	3%	1%	100%

Table 1. Sentiments Analysis by Year

There are more, about four times, tweets from males compared to females, see Table 2. There are no negative tweets from female compared to the 3% negative tweets from 81% of the male tweets. There is disparity by gender on opinions about Cybersecurity education and males dominated the conversions on social media as the profession is also male dominated. Figure 6 presents the sentiments by state for USA, where positive sentiments dominated in all states. Figure 7 indicates that there was more positive sentiment in 2015 compared to 2016.

	Positive (P)	Ambivalent	Neutral	Negative (N)	Unknown	Total
Female	34		69			103
% from (M +F)	6%		13%			19%
% Within F	33%		67%			
Male	212	5	195	16		428
% from (M +F)	40%	1%	37%	3%		81%
% within M	50%	1%	46%	4%		
M+F	246	5	264	16		531
Unknown	370	13	438	27	8	856
Total	616	18	702	43	8	1387
%Total	44%	1%	51%	3%	1%	100%

Table 2. Sentiments Analysis by Gender

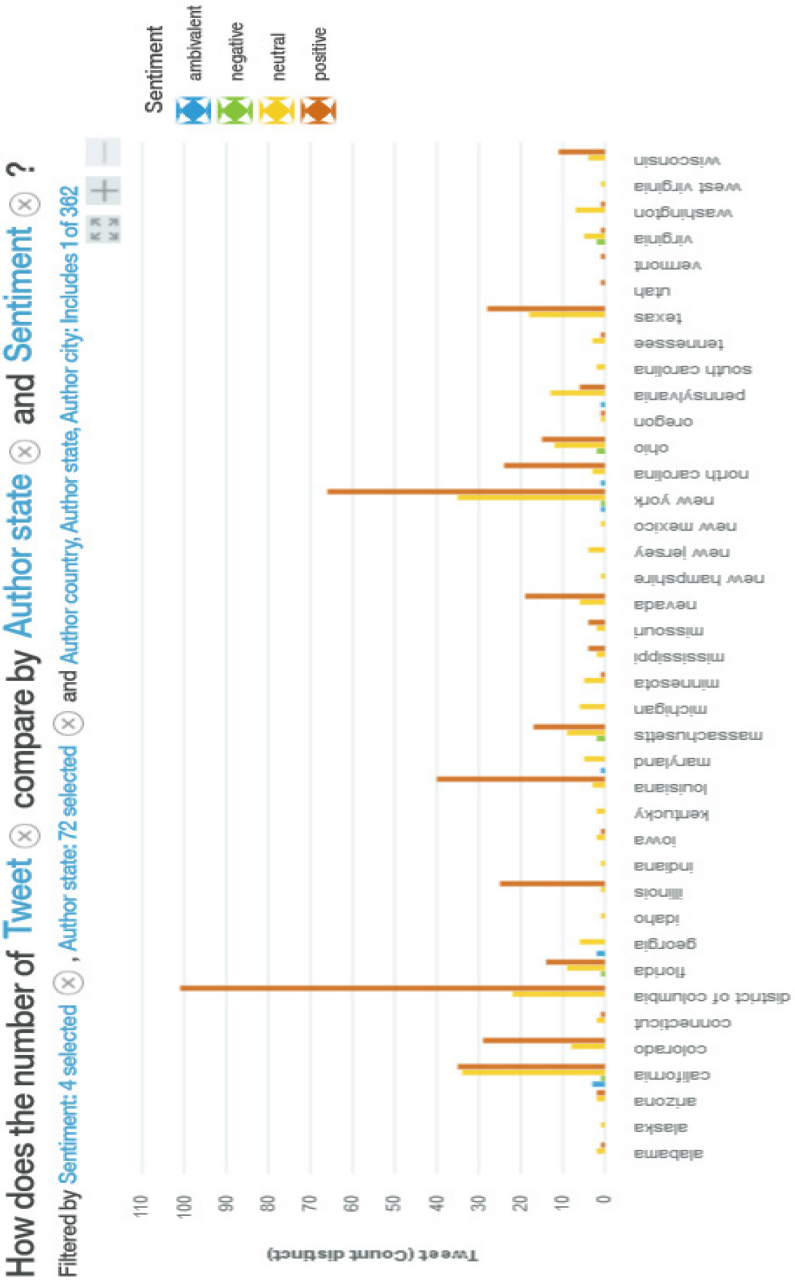

Figure 6. Sentiment of Tweets by States in USA

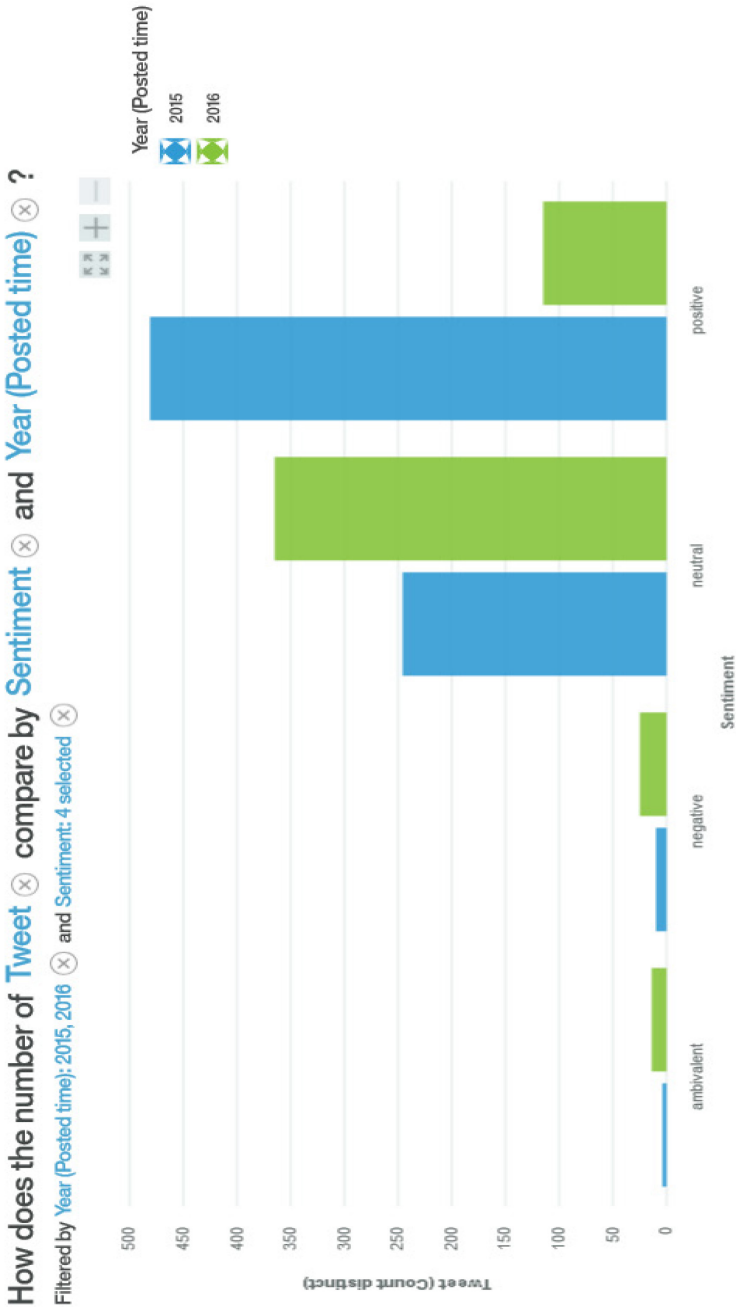

Figure 7. Sentiment of Tweets by Year

Figure 8 presents the word clouds of the author names we learned from the tweets. The author with the highest number of tweets is *Knol infos*, who is an ICT course instructor and IT-Security Thought Provoker

(https://twitter.com/knolinfos).

The author with the second highest number of tweets is *National Cyber League* (https://twitter.com/NatlCyberLeague). Other authors with high number of tweets include *Being Example* (https://twitter.com/BeingExample), and the *National Cyber Watch Center* (https://twitter.com/CyberWatchCtr).

Figure 8. Word cloud of Author names

5 CONCLUSION

Using Watson Analytics, we retrieve tweets, analyze them, and creates a dataset consists of 1,387 tweets. We also identify patterns and discover useful insights. Regarding the temporal patterns of tweets about Cybersecurity education, it has been steady recently. Regarding the spatial patterns of tweets about Cybersecurity education, most tweets come from United State. However, it unexpected to get a relatively low participation from California on Cybersecurity education compared to DC and New York, where California being the center for IT innovations – Silicon Valley.

Overall the sentiment was positive with respect to Cybersecurity education. Moreover, the positive sentiment is declining, and the neutral and negative sentiments are increasing during the three years. Furthermore, females did not show negative sentiments.

The discovered insights such as the several trends and sentiments towards Cybersecurity education can be used for policy and program development in Cybersecurity education, as well as recruitment and retention of students in Cybersecurity education.

We used free faculty version of Watson Analytics from IBM. This version comes with its limited feature and functionality including a random selection of 25,000 tweets per project. Another limitation is the possible inaccuracy of contents on social media due to short length (e.g. 140 characters for tweets), informality of the language, and credibility of sources. Future research will include topic modeling, predicative and perspective analytics on the social media data about the Cybersecurity education and related topics such as Cybersecurity job.

REFERENCES

[1] Bian J, Yoshigoe K, Hicks A, Yuan J, He Z, Xie M, et al. (2016) Mining Twitter to Assess the Public Perception of the "Internet of Things". PLoS ONE 11(7): e0158450. https://doi.org/10.1371/journal.pone.0158450

[2] Bouazizi, M., & Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. 2016 *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, 1-6. doi: 10.1109/ICC.2016.7511392

[3] Camargo, J., Torres, C. A., Martínez. O. H., & Gómez, F. (2016). A big data analytics system to analyze citizens' perception of security. *2016 IEEE International Smart Cities Conference (ISC2)*, Trento, 2016, 1-5. doi: 10.1109/ISC2.2016.7580846.

[4] High, R. (2012). The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. https://developer.ibm.com/watson/wp-content/uploads/sites/19/2013/11/The-Era-of-Cognitive-Systems-An-Inside-Look-at-IBM-Watson-and-How-it-Works1.pdf

[5] IBM Big Data and Analytics Hub. (n.d.). Extracting business value from the 4 V's of big data. Retrieved from http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data (accessed on 1/29/2017).

[6] Ribarsky, W., Wang, X., & Dou, W. (2014). Social Media Analytics for Competitive Advantage. *Computer & Graphics*, 38, 328-331.

[7] Sharda, R., Delen, D., & Turban, E. (2015). Business Intelligence and Analytics: Systems for Decision Support. Boston: Pearson.