A Worldwide Analysis of Cyber Security And Cyber Crime using Twitter

Kartikay Sharma, Siddharth Bhasin, Piyush Bharadwaj

Abstract—In the era of social media and the Internet, there has been an exponential increase in threats related to the privacy of user accounts and data. The confidentiality of personal data is compromised for various motives. This sudden increase in security threats has led to widespread problems. Our research is focused on analyzing the extent of cyber-attacks in various countries across the globe. We have proposed a novel approach for analyzing the tweets related to cyber-attacks and its surrounding fields. The analysis proves that Asian countries face more cyber security issues as compared to European countries. Further, it is also analyzed that developing countries like India are more prone to these issues as compare to developed countries like the United States or the United Kingdom.

Index Terms: cybercrime, cybersecurity, hacking, malware, security, sentiment analysis, machine learning, data science

I. INTRODUCTION

In the extensive domain of security, analysts need knowledge about the state of the world to make time-critical decisions. This knowledge is drawn from a variety of sources and then represented in a form that will allow further analysis and decision making. In this paper, we have analyzed tweets about cybersecurity and other related ones.
This system focuses on sentiment analysis on tweets, which are entries limited to 280 characters with the possible insertion of emoji. As of the fourth quarter of 2018, micro-blogging service averaged at 321 million monthly active users, hence it provides a remarkable environment to collect data in a close to real-time manner. Mining Twitter data for interpretations is one of the most common natural language processing tasks.
In our work, we propose a method to analyze a number of tweets on hashtags on cyber securities and perform sentiment analysis using VADER to comprehend the overall sentiment of a Twitter user. Along with the text of the tweet, other aspects are also considered like the username and location of the author of the tweet. We specifically concentrate on tweets in English due to the widespread nature of this language on the Web. Analyzed Tweets makes it simple to understand what people think (positively or negatively) in their tweet related to a certain keyword. Sentiment Analysis, or Opinion Mining, is a sub-field of Natural Language Processing (NLP) that tries to identify and extract opinions from within a given text. The purpose or intention of sentiment analysis is to measure the attitude, evaluations, and emotions of a speaker/writer based on the computational treatment of subjectivity in a text.
VADER or Valence Aware Dictionary and Sentiment Reasoner, defined as a lexicon and rule-based sentiment analysis tool, is specifically accustomed to the sentiments expressed on various social media platforms. VADER uses a mixture of a sentiment lexicon which is a list of lexical features (e.g., words) that are generally labelled according to their semantic orientation as positive, negative or neutral.
Data from surveys can be used for cybercrime measurements. Sentiment information can be used to compare various countries on the account of cybersecurity.
The fact that VADER generalizes more suitably across contexts than any of our standards makes it fit to be used for such estimates. However, when compared to worldly ML techniques, the lucidity of VADER carries several leads. To start with, it is both fast and computationally efficient without losing accuracy. Running directly from a regular common laptop machine with normal, fair specifications (e.g., 3GHz processor and 6GB RAM), a dataset that needs a fraction of a second to analyze with VADER can take hours while working more complicated models like SVM (if training is required). To say more, the lexicon and practices used by VADER are directly available. VADER is hence easily examined, understood or revised. By revealing both the lexicon plus rule-based model, VADER presents the inner functioning of the sentiment analysis system further convenient to a more comprehensive audience past the computer science community.

The paper is organized as follows: Section 2 gives the literature survey on papers related to cybercrime, cybersecurity, and related areas. Section 3 explains the proposed methodology. Section 4 provides the results obtained from the said approach. Finally, Section 5 concludes the work and talks about its future scope.

II. LITERATURE SURVEY

In paper [1], the author believes that social media today is an integral part of people's daily routines and the livelihood of some. He gives a method to measure consumer loyalty using the data gathered from Twitter. In paper [2], the author puts forward a position paper that reviews the developments in applying Data science for cyber security and cyber security for data science and then discusses its applications in Social Media. In paper [3], the author recommends that the traditional solutions along with the use of analytic models,

Revised Manuscript Received on 14 September, 2019.
Kartikay Sharma, Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India, kartikaysharma53@gmail.com
Siddharth Bhasin, Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India,
Piyush BharadwajNalini Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India,

machine learning and big data could be improved by providing relevant awareness to control or limit consequences of threats. In paper [4], a methodology for tracking social data that can trigger cyber-attacks is developed. Their main contribution lies in the monthly prediction of tweets with content related to security attacks and the incidents detected based on ℓ_1 regularization. In paper [5], the study examines the cybersecurity attitudes and the actual behavior over time, using the data collected from Twitter. They use the sentiment analysis and text mining techniques on original tweets related to cybersecurity collected at two different time periods. Upon completion of the research, they presented the analysis of the relationship between cybersecurity attitudes and behavior and how behaviors may be shaped by the attitudes. In paper [6], the analysis addressed the security specialists of machine learning techniques, applied to the detection of intrusion, malware, and spam. The goal was twofold: to assess the current maturity of these solutions and to identify their main limitations that prevent an immediate adoption of machine learning cyber detection schemes. The conclusion was based on an extensive review of the literature as well as on experiments performed on real enterprise systems and network traffic. In paper [7], a novel approach for sentiment analysis had been developed for extracting the opinions from a given data source. The proposed approach was applied to one of the biggest service industries in the world: the travel industry. With the application of this approach, an analysis of opinions and sentiments expressed on Twitter about TripAdvisor was done. In paper [8], the study is about the adversarial resilience of detection systems based on supervised machine learning models. A formal definition was provided for adversarial resilience while focusing on multisensory fusion systems. In paper [9], the author adopted a special semi supervision method to classify cybersecurity log into attack, unsure and no attack, by first, breaking the data into 3 cluster using Fuzzy K Mean (FKM), then manually labelling a small data (Analyst Intuition) and finally, training the neural network classifier Multi-Layer Perceptron (MLP), based on the manually labelled data. In paper [10], the author suggested that security attacks on the web have been perpetrated by hacker-activist organizations that aim to damage the web services in a specific context for which they are motivated. The author presented a sentiment analysis method on Twitter content. The method which was suggested by the author was based on the daily collection of tweets from the users; who use the platform as a means for expressing their views on relevant issues, and who use it to present content related to security attacks in the web. The information was turned into data that could be analyzed statistically to predict whether there is a possibility of an attack or not. The latter was done by analyzing the collective sentiment of users and groups of hacking activists in response to a global event. Survey paper [11], described a focused literature survey of the methods of machine learning and data mining for cyber analytics in the support of intrusion detection. In paper [12], the author suggested that, in order to secure vital personal and organizational systems, timely intelligence on cybersecurity threats and vulnerabilities is required. Intelligence about these threats is generally available in both overt and covert sources like the National Vulnerability Database, CERT alerts, blog posts, social media, and dark web resources. In paper [13], the author applied machine learning and sentiment analysis to cybersecurity, with the purpose of developing a method for detection of the cyber threats, which were often undetectable by traditional tools. Paper [14] depicted an engaged writing review of machine learning and information digging techniques for digital investigation with the help of interruption detection.

III. PROPOSED METHODOLOGY

The workflow given below is the depiction of methodology which was followed as part of this analysis

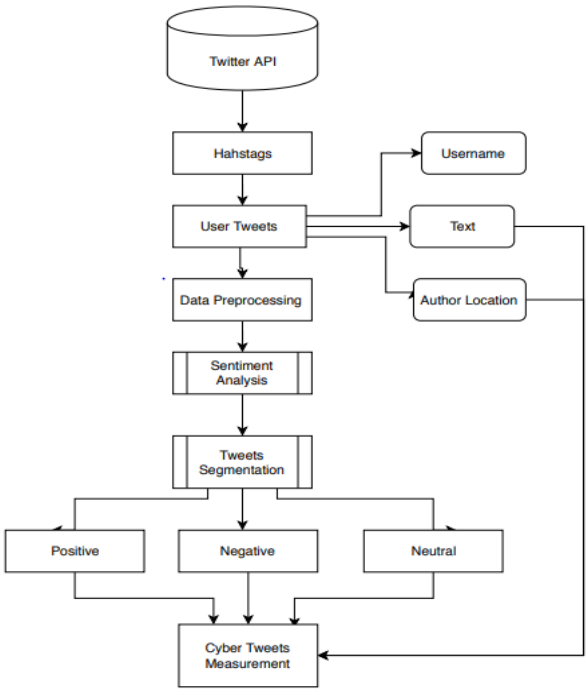

Fig. 1. Flowchart of the proposed work

Algorithm used

The algorithm used in the proposed methodology of this analysis is as follows:

1. Extract tweets using Python.
2. Extracting tweets for specified hashtags in the table provided.
3. Tweets extracted for 16 hashtags in seven regions, which are the US, India, Europe, Australia, Canada, Russia, and the UK.
4. Combine all tweets to make a single dataset containing Username, Author Location and the text of the tweet.
5. Pre-process the dataset by cleaning, parsing and tokenizing the tweets.
6. After pre-processing, each tweet is analyzed for Sentiment Analysis using VADER that is, Valence Aware Dictionary and Sentiment Reasoner.
7. The polarity of each tweet is known and added to the dataset.
8. It is then analyzed and various conclusions are drawn using the analysis.

Twitter Data Extraction

Twitter is a social networking service that allows its account holders to send and receive 280-character messages known as "tweets", follow other users, type communities around a trending topic (#hashtag), and forward tweets to 2018, the micro-blogging service averaged at 321 million monthly active users, therefore it provides a remarkable surrounding, to gather information in an exceedingly close to real-time manner. Mining Twitter data for insights is the one used in every of the foremost common natural language processing tasks. This system focuses on sentiment analysis on tweets. The workflow of the proposed methodology is shown in Fig 1.

authorloc	text	username
India	b"Dark Web Cryptocurrency Deals Likely in News Sites' Admin Hacks \n\	Mistral Solutions
India	b'RT @4orgexcellence: 7 Ways to achieve #InternetOfThings success!\n\	Anuj Saxena
India	b'Inviting #startups #SME for showcase their products or services, who a	Anuj Saxena
India	b'#Funding Calling Indian startups in #healthcare #learning and #Informa	Nitin Vinay
India	b'Technology and Banking goes hand in hand, kudos to #USAB for embra	Ratan Jyoti

Fig. 2. Extracted data

We have used the social network platform of Twitter to gather 60,000 original tweets associated with cybersecurity. Twitter users' posts are generated in a period of time with a high level of anonymity and thus can be freed from the biases identified by Peterson and Wilson (1992). We extracted the tweets with the Python engine (Spyder) using Twitter API. With Tweepy we were able to extract the user's name, their location and also the text of the tweet.

In order to determine the angle of the population of Twitter users towards cybersecurity, we got to know the communities and themes drawn by #Hashtags in Twitter associated with the subject of cybersecurity using the six shaping ICT security principles of confidentiality, authentication, and integrity of information, non-repudiation, access management, and accessibility. Some examples of these #hashtags are shown in Table 1.

TABLE I. HASHTAGS WITH THEIR COUNT

Hashtags	**Counts**
Cybersecurity	2525
Cyberattack	4538
Cyberthreat	813
Cyber	3725
Hackers	5409
Databreach	7935
Cyber risk	929
Hacking	2751
Spam	8194
Malware	10000
Cyberwar	862
Websecurity	984
Dataleak	84
InfoSec	1577
Information Security	2037

Data Preprocessing

Data pre-processing could be a very crucial step in data mining because it can have a grave impact on the results. An unprocessed dataset can cause wrong results and may ruin the analysis; therefore, it's necessary to pre-process the information before applying any data mining operation [5]. In data pre-processing we tend to remove the unwanted tags, web links and special symbols (@ # ^ * "/ : >, < \ |?), that may lead to wrong results. The whole task of data pre-processing is finished in an automatic fashion using an identical system developed for tweet assortment. Further, since an individual can send multiple tweets on Twitter, if these multiple tweets aren't removed then it may lead to a biased analysis. therefore, so as to form an unbiased analysis, we removed all the multiple tweets from an identical person.

Sentiment Analysis

authorloc	text	username	Sentiment	Polari
Canada	b Fileless infection ta…	CyberSecurity	neutral	0
Canada	b RT rtehrani Th…	NoSQL	negative	-1
Canada	b RT omvapt Hadoop coop …	Di993r	negative	-1
Canada	b RT WeAreConfian…	One Crypto News	positive	1
Canada	b Browser Extension St…	Jeff Multz	negative	-1
Canada	b Credential management i…	YourTechComp…	positive	0

Fig. 3. Preprocessed data with sentiment analysis

For Sentiment analysis, we moved to a sophisticated analytical tool for sentiment in informal postings. The results found with VADER were intriguing. VADER not only analyzes individual word sentiment but also makes an attempt to predict the normalized valence of positive or negative sentiment, supported by overall sentences, and accounting for factors like negation, punctuation or emoticon usage. It also provided a consistent analysis of SET comments that are usually written informally. The results are shown in Fig 3.

Fig 4 shows the values that VADER gives the polarity of the text and labels it as positive, negative or neutral, for positive sentiment the polarity should be greater than or equal to 0.05, for neutral sentiment the polarity should be between -0.05 and 0.05 and for negative sentiment the polarity should be less than or equal to -0.05.

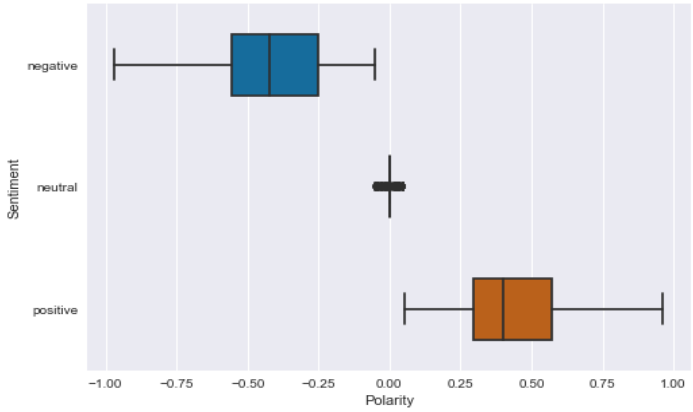

Fig. 4. Boxplot of polarity

IV. RESULT

The Pie chart depicts the hashtags used along with the percentage of tweets retrieved in the data set using that particular hashtag. With the highest number of tweets (10000) of hashtag Malware and the lowest number of tweets (84) of hashtag Data Leak.

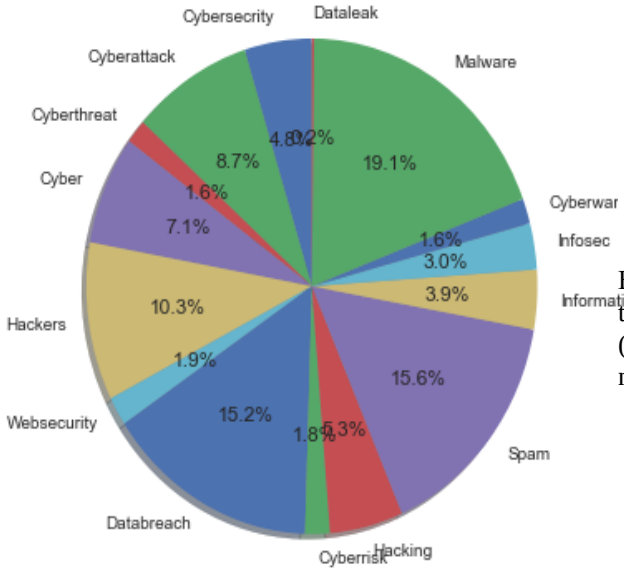

Fig. 5. Pie chart of the Hashtags

Figure 6 shows the top 10 countries in the dataset with their respective tweets. According to the graph, users in India tweeted the highest number of tweets followed by the US, the UK, Canada, France, Singapore, Israel, Brazil, Switzerland, and Italy.

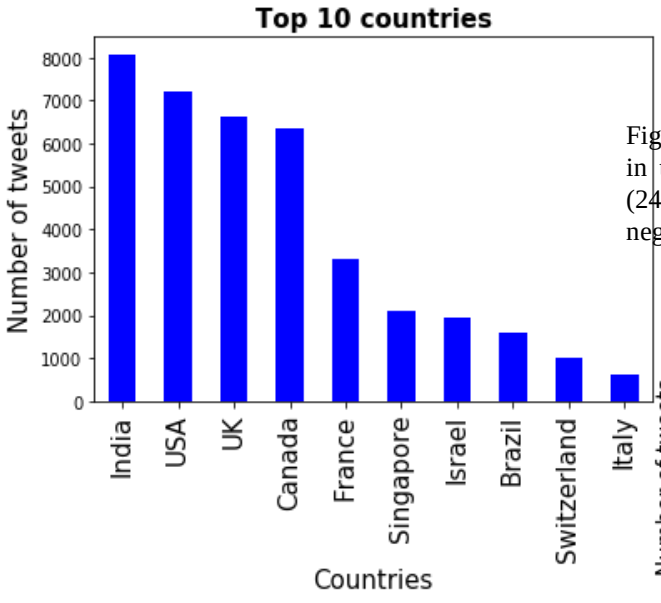

Fig. 6. Top 10 countries with the highest number of tweets

Figure 7 shows sentiment analysis of the total users in the dataset. With the highest number of positive tweets (24581) followed by the negative tweets (23569) and then the neutral tweets (17772).

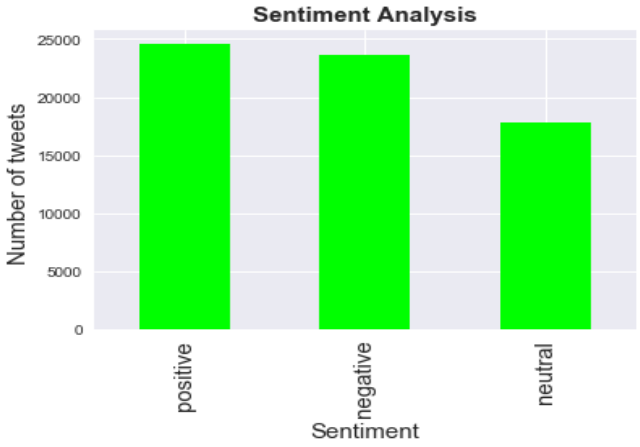

Fig. 7. Sentiment analysis of the dataset

Figure 8 shows the total user's sentiment analysis of India, in the dataset. With the highest number of negative tweets (2994) followed by the positive tweets (2972) and then the neutral tweets (2111).

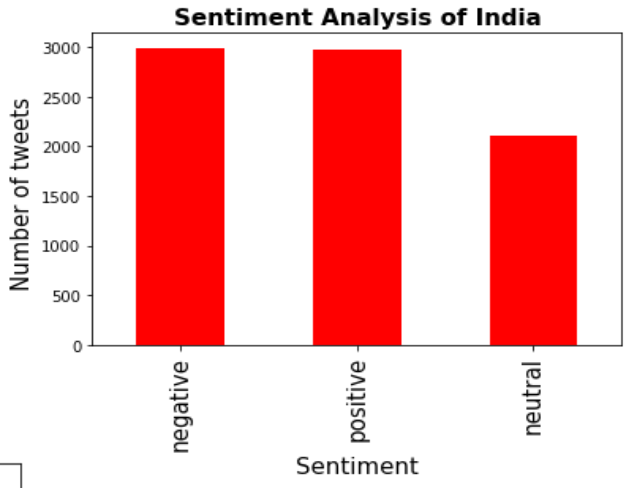

Fig. 8. Sentiment analysis of India

Figure 9 shows the total user's sentiment analysis of the US in the dataset. With the highest number of neutral tweets (2493) followed by positive tweets (2469) and then the negative tweets (2241).

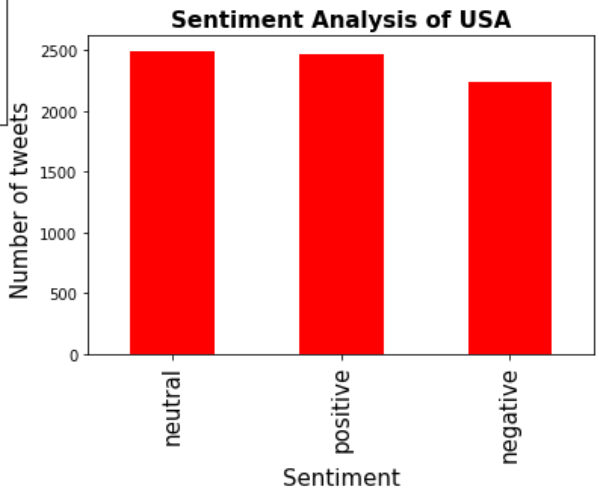

Fig. 9. Sentiment analysis of USA

Figure 10 shows the total user's sentiment analysis of the UK in the dataset. With the highest number of positive tweets (2582) followed by negative tweets (2391) and then the neutral tweets (1634).

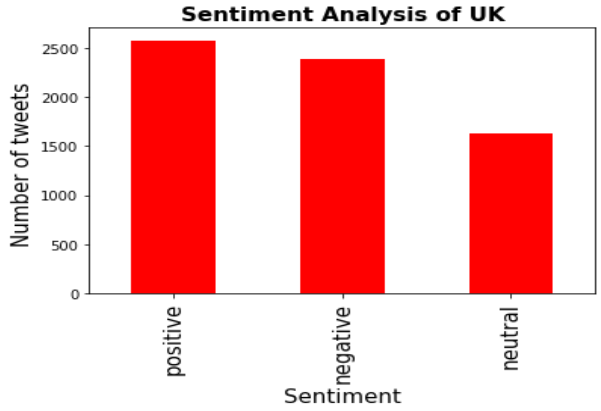

Fig. 10. Sentiment analysis of UK

Figure 11 shows the total user's sentiment analysis of Canada in the dataset. With the highest number of positive tweets (2444) followed by the negative tweets (2357) and then the neutral tweets (1559).

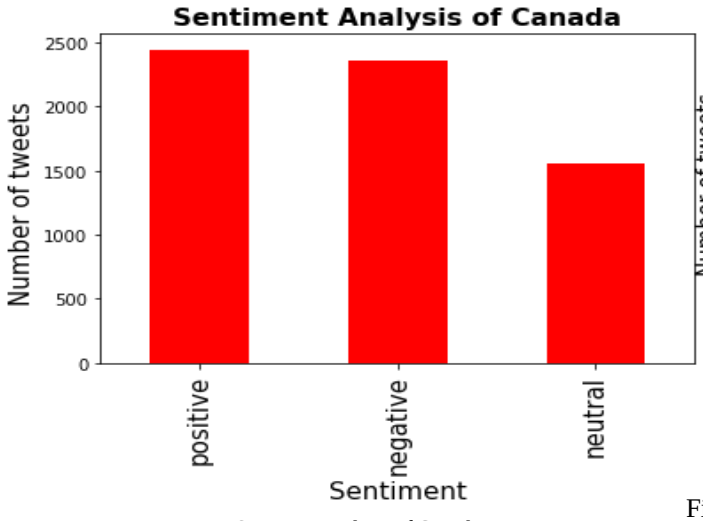

Fig. 11. Sentiment analysis of Canada

Figure 12 shows the total user's sentiment analysis of France in the dataset. With the highest number of positive tweets (1255) followed by the negative tweets (1109) and then the neutral tweets (935).

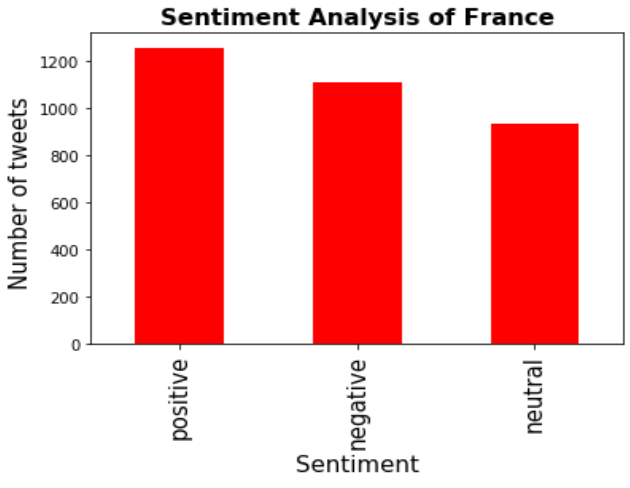

Fig. 12. Sentiment analysis of France

Figure 13 shows the total user's sentiment analysis of Singapore in the dataset. With the highest number of negative tweets (2444) followed by the positive tweets (728) and then neutral tweets (555).

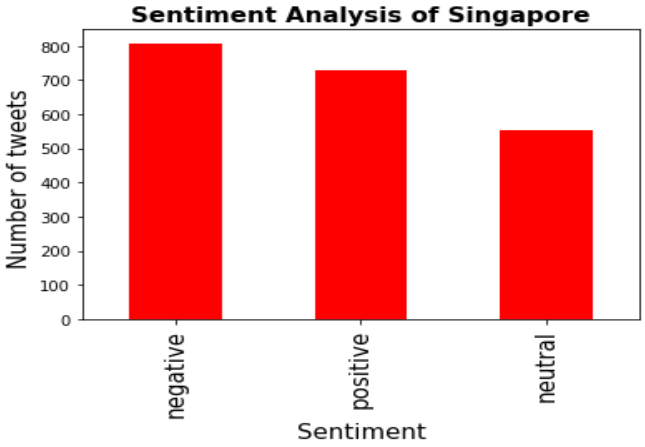

Fig. 13. Sentiment analysis of Singapore

Figure 14 shows the total user's sentiment analysis of Israel in the dataset. With the highest number of positive tweets (739) followed by the neutral tweets (642) and then the negative tweets (563).

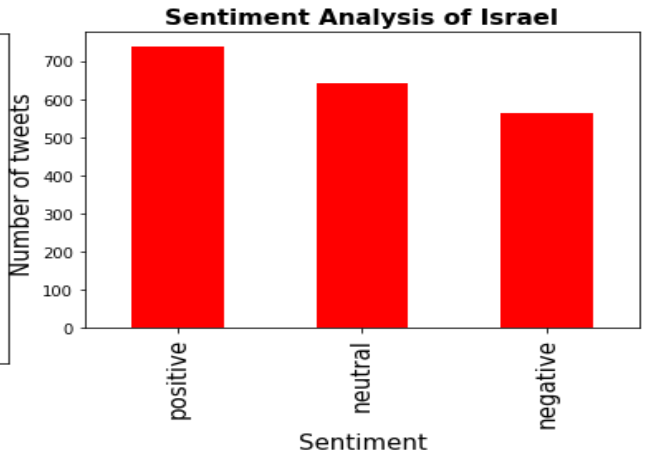

Fig. 14. Sentiment analysis of Israel

Figure 15 shows the total user's sentiment analysis of Brazil in the dataset. With the highest number of positive tweets (618) followed by the negative tweets (485) and then the neutral tweets (412).

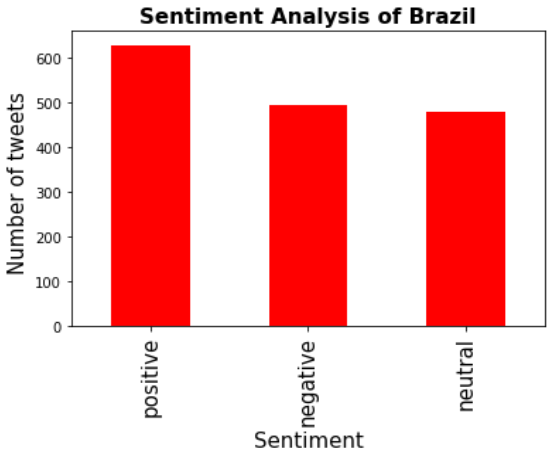

Fig. 15. Sentiment analysis of Brazil

Figure 16 shows the total user's sentiment analysis of Switzerland in the dataset. With the highest number of negative tweets (409) followed by the positive tweets (346) and then the neutral tweets (241).

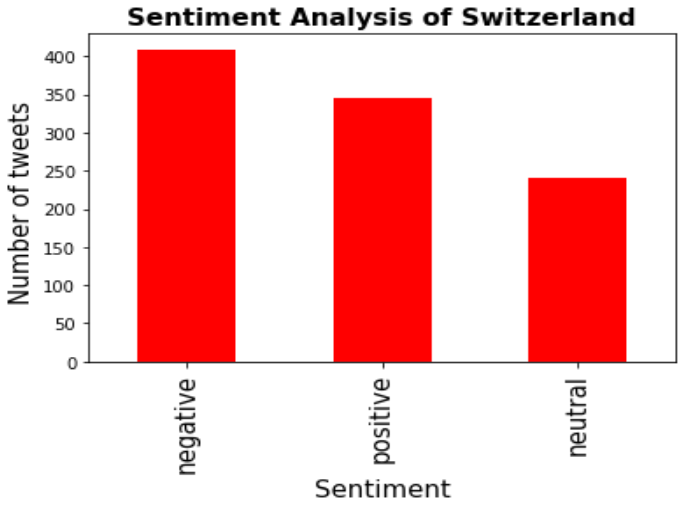

Fig. 16. Sentiment analysis of Switzerland

Figure 17 shows the total user's sentiment analysis of Italy in the dataset. With the highest number of negative tweets (223) followed by the positive tweets (211) and then the neutral tweets (174).

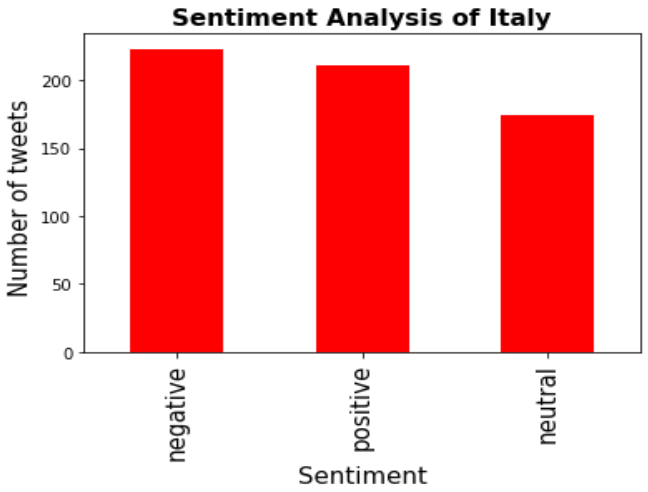

Fig. 17. Sentiment analysis of Italy

V. CONCLUSION

This paper analyze tweets associated with cyber-attacks and its close fields in various countries across the world. We employed a Social Sentiment sensor on Twitter. The research collects tweets for fifteen relevant cyber security related hash tags for a specific period and classifies them as negative, positive and neutral. The extracted data is analyzed for the cybercrime measurements. Sentiment information was used to compare various countries on the account of cyber security. The system has compared top 10 countries with the greatest number of tweets within the dataset, with India tweeted the maximum number of tweets followed by the USA, the UK, Canada, France, Singapore, Israel, Brazil, Switzerland, and Italy. These were then compared with the parameters; positive (POS), negative (NEG) and neutral (NEUTRAL). India was observed to have the most positive tweets (2972) followed by the UK with (2582), and then the US with (2469). The utmost negative tweets were additionally recorded against India with (2994), the UK with (2391) and then Canada with (2357). The analysis shows that Asian countries face additional cyber security problems as compared to European countries. Further, it also analyzed that developing countries like India are more viable to these problems as compared to developed countries just like the United States or the United Kingdom. Our future work would focus on improving the model to predict alternative real-life events like elections, education techniques, market trends and so on.

REFERENCES

1. Khan, R., & Urolagin, S. (2018). Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 9(6), 380-388.
2. Thuraisingham, B., Kantarcioglu, M., & Khan, L. (2018, May). Integrating Cyber Security and Data Science for Social Media: A Position Paper. In 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) (pp. 1163-1165). IEEE.
3. Foroughi, F., & Luksch, P. (2018). Data Science Methodology for Cybersecurity Projects. arXiv preprint arXiv:1803.04219.
4. Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J., & Sanchez, V. (2018). Social sentiment sensor in Twitter for predicting cyber-attacks using ℓ1 regularization. Sensors, 18(5), 1380.
5. Gupta, B., Sharma, S., & Chennamaneni, A. (2016). Twitter Sentiment Analysis: An Examination of Cybersecurity Attitudes and Behavior.
6. Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018, May). On the effectiveness of machine and deep learning for cyber security. In 2018 10th International Conference on Cyber Conflict (CyCon) (pp. 371-390). IEEE.
7. Bhardwaj, P., Gautam, S., & Pahwa, P. (2018). A novel approach to analyze the sentiments of tweets related to TripAdvisor. Journal of Information and Optimization Sciences, 39(2), 591-605.
8. Katzir, Z., & Elovici, Y. (2018). Quantifying the resilience of machine learning classifiers used for cyber security. Expert Systems with Applications, 92, 419-429.
9. Teoh, T. T., Zhang, Y., Nguwi, Y. Y., Elovici, Y., & Ng, W. L. (2017, July). Analyst intuition inspired high velocity big data analysis using PCA ranked fuzzy k-means clustering with multi-layer perceptron (MLP) to obviate cyber security risk. In 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 1790-1793). IEEE.
10. Hernández, A., Sanchez, V., Sánchez, G., Pérez, H., Olivares, J., Toscano, K., ... & Martinez, V. (2016, March). Security attack prediction based on user sentiment analysis of Twitter data. In 2016 IEEE international conference on industrial technology (ICIT) (pp. 610-617). IEEE.
11. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.
12. Mittal, S., Das, P. K., Mulwad, V., Joshi, A., & Finin, T. (2016, August). Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 860-867). IEEE Press.
13. Fink, E., Sharifi, M., & Carbonell, J. G. (2011, February). Application of machine learning and crowdsourcing to detection of cybersecurity threats. In Proceedings of the US Department of Homeland Security Science Conference–Fifth Annual University Network Summit, Washington, DC.
14. [14] Maloof, M. A. (Ed.). (2006). Machine learning and data mining for computer security: methods and applications. Springer Science & Business Media.

AUTHORS PROFILE

Kartikay Sharma, Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India. kartikaysharma53@gmail.com

Siddharth Bhasin , Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India siddharth88bhasin@gmail.com

Piyush Bharadwaj, Bhagwan Parshuram Institute of Technology, GGSIPU, Rohini, Delhi, India piyushb88@gmail.com