Text mining to segregate criminal activities profiles in microblogs (Review Paper).

Mr. Mahesh N. Rakheja, Mr. Pranjal Dhore, Mr. Siddhant Jaiswal

Department of Computer Science & Engineering
M.Tech. Jhulelal institute of Technology, Nagpur, Maharashtra, India

Abstract

The internet is a crucial source of information for almost everything. There was a time when people on the internet were into two categories, namely producer and consumer. The producer is the one who generates data, and the consumer is the one who uses data. Due to microblogs and social networks, every person on the internet becomes the producer and consumer. It leads to an increase in cybercrimes. There are many techniques to detect and predict cyber criminals profiles, But none of them proved successful. In this paper, we are going to introduce, Text mining approach based on hashtags to segregate criminal activities patterns that help us to detect cyber criminal's profiles in advance. It can lead to a reduction in cybercrimes. Microblogs have a restriction to 140 characters; It becomes difficult to express the views for the producer of content. Nowadays, people have started voicing their opinions on other websites and directly sharing the link of that website on microblogs. In this paper, We are also going to analyse the contents of websites posted on microblogs, which will increase in accuracy of profile segregation.

Keywords: Cyber Crime Profiling, Sentiment Analysis, Text mining, Websites, Microblogs, Text Analysis, Social Media, Suspicious Profile, NCD Normalised Compression Distance.

I. Introduction

The internet becomes a crucial source of information, attracting many criminals to perform criminalistic activities in the cyber world. As a leman have the power to generate and share the contents online, Every next person is launching websites. Many of the sites are serving a useful purpose, Whereas many are serving for illegal purposes. Law enforcement agencies are tracking such kinds of malicious web sites and banning them from the world of the internet. It increases lots of work and daily tasks.

Cyber users have made a shift from just a consumer to a producer (shown in fig. 1). In web 1.0, there was a specific set of people who used to generate the content, and the rest of the people used to consume the available information. A transformation from web 1.0 to 2.0 made no difference between consumer and producer. Any person can generate content, and any other person can contribute to it. Through social media, Wiki, Microblogs, and many more. This transformation increases the number of cybercrimes. It becomes difficult to trace cybercriminals, Based on the contents shared on the internet world.

Social media gives the freedom to express, communicate, and connect with people all over the world. It gives you the power to change other's viewpoints and convince people from your perspective. People can use energy positively and negatively. They use social media to achieve their illegal mindsets. Many cybercriminals portray themselves as legitimate and convince others to perform a specific task. Usually, The innocent become the victim of cybercrime.

There are many approaches to detect or predict the cybercriminal's profile either before the crime or after the crime using machine learning and deep learning. In our approach, we are suggesting the technique to use hashtags and URLs as a primary source of information to find the emotions of the post, which can further help in segregating the malicious profiles with legitimate profiles. This approach can help in reducing cybercrimes by predicting malicious patterns in advance.

There is a sudden increase in cyber crimes on social media platforms. Now, It becomes necessary to find priority solutions and methodology to deal with it. To achieve, We need a strong bond between the public sector (Law enforcement, etc.) and private sector (Facebook, Twitter, Google, etc.) to provide a robust reporting system. We need to observe every activity on social profiles including likes, comments, post, story, URL share, reply and all other events. This will help us to understand more about the emotions of the producer. Further help us to segregate cybercrime profiles.

This paper suggests an extension and improvement to an existing paper presented by Salim ALAMI [1]. In their article, they offered information about cybercrime profiling using text mining. They prefer to use hashtags to sort the profiles and use the similarity distance formula to detect and predict suspicious patterns. They decomposed the post into words and compared the words with existing names in the database using the similarity distance formula.

Our suggested approach is to overcome the problem of microblogs. Usually, microblogs restrict the size of contents in each post. Text with more than 140 characters is not allowed in microblogs. One hundred forty characters are not sufficient to express emotions; People started using websites to post contents and share the web link on microblogs. We are going to use the same technique (Similarity distance formula) to segregate cybercrime profiles, But we will also include the content specified on the shared website. Our suggestion will increase the accuracy of segregated cybercrime profiles as we are going to deal with more content.

This paper is organized as follows. In section II, We present an overview of cybercrime profiling. In part III, We offer the existing work by other authors. In section IV, We provide our view about the specific topic and in section V, Conclusion, and perspective about the subject.

II. Cybercrime Profiling

On the way to expand useful profiles of various cybercriminals, social media incorporates a massive quantity of data that needs to be used to improve the reporting of cybercrime. To apprehend the brand new tendencies of the cybercrime and additionally establishing the best solution so that it will be the muse stone to analyze and prosecute criminal activities, there may be an urgent want for cooperation and harmonization of public (e.g. law enforcement) and personal sectors (e.g. fb, Twitter, YouTube...) to encourage cybercrime reporting. Expertise the stairs in the method of committing crime, and know-how the situations that facilitate its fee, enables us to peer how we are able to interfere to frustrate crime" [2].

Criminal profiling is the system of Investigating and inspecting criminal behaviour if you want to assist discover the kind of individual chargeable for wrongdoing [3].

(Johnson, 2005), defined profiling as - An educated attempt to offer unique statistics as to the form of character who dedicated a sure crime. A profile based on traits patterns or factors of uniqueness that distinguishes sure individuals from the overall population [4].

To this point, all the countrywide safety groups rely on data and text mining strategies to come across and are expecting criminal sports, at the same time as records mining refers to the exploration and analysis of huge portions of information to discover meaningful patterns and guidelines. text mining, sometimes referred to as text statistics mining, is the process of analysing clearly happening textual content for the purposes of extracting and non-trivial styles or understanding from unstructured text [5].

S. Alami [1] proposed a solution in relation to applying records and text mining techniques to stumble on and predict the suspicious posted contents with the deduction of the suspicious behavior users had on the internet. The proposed answer is specially received by way of representing a chief project using techniques of textual content mining, based on the calculation of a similarity distance to come across suspicious posts, which is an effective way to examine the statistics published on the internet. The objective of this intelligence facts evaluation project is to use records mining to locate affiliation and discover relationships among suspect entities based on historical information, on the way to expect crook sports. Data mining is a powerful tool that enables crook investigators who may lack great training as facts analysts to explore big databases quickly and successfully. In this publication the authors haven't taken into consideration the disambiguation step of their proposed framework. In truth, imparting a powerful way to feature semantics to this shape of communique requires a disambiguation step, due to the fact many assets can be matched to the identical entity that result in synonymy and polysemy issues.

III. Related Works

Using text analytics to discover suspicious users in social media offers a vital project. There are numerous techniques to stumble on the means of expressions; many works were accomplished in this context showing numerous techniques for textual content analytics. The lexical matching suffers from many drawbacks; which includes ambiguity (polysemy and synonymy) and feasible lack of specificity(less "meaningful" ideas are identified). Short texts have the traits of sparsity, and are noisy due to their restricted length. whilst using the "bag of words" model to symbolize short textual content, contextual records is disregarded and therefore frequently leads to synonymy and polysemy troubles [6].

To conquer the hassle of data sparseness and the semantic gap in short text, numerous approaches have been proposed for adding semantics to textual content contained in tweets.

Meij [7] proposed a method to link n-grams to Wikipedia concepts primarily based on diverse features. Their technique is divided in steps; in the former they generate a ranked list of candidate ideas for each n-gram in a tweet by means of applying numerous types of functions. inside the later, they aim to enhance precision by making use of supervised machine learning.

Tang [8] have provided a framework for enriching brief textual content for clustering motive wherein they play multi- language information integration and feature reduction concurrently through matrix factorization strategies.

Mendes [9] proposed related Open Social alerts, a framework that consists of annotating tweets with facts from related data. Their approach is alternatively truthful and involves either searching up hashtag definitions or lexically matching strings to recognize (DBpedia) entities in tweets.

Banerjee [10] proposed a method to complement short texts representation with extra features from Wikipedia. This technique only used the titles of Wikipedia articles as additional external capabilities; it showed improvement within the accuracy of short texts clustering.

Liu [11] cognizance on Named Entity Recognition (NER) on tweets and use a semi-supervised learning framework to identify 4 sorts of entities.
Stephen Guo [12] proposed a structural SVM method to cope with the trouble of giving up-to-stop entity linking on Twitter. by way of thinking about mentioning detection and entity disambiguation together.

Hachey [13] determined that it's miles beneficial to divide the entity linking challenge into levels: seek and disambiguation. All through the former the system proposes a set of applicants for a named entity mentioned to be related to, which are then ranked by means of the disambiguation. They've also observed that much of the variant between NEL structures explained by the overall performance in their searchers, and the literature on named entity linking has targeted nearly solely on disambiguation.

Laniado [14] has explored the use of hashtags in Twitter and the relation to (Freebase) standards. clearly that hashtags are correct signs to stumble on activities and trending subjects. the use of guide annotations, they find that approximately half of the hashtags can be mapped to freebase ideas, maximum of them being named entities. In some cases, extra well known hashtags are mapped to principles. Assessors confirmed excessive agreement at the mission of mapping hashtags to principles. The authors make the belief that hashtags are mainly used to "ground" tweets, an assumption that we undertake in our work, enabling us to add semantics to tweets with the usage of hashtags definition.

IV. Review

As per the understanding of the utilization of text mining to detect and prevent cybercrimes by performing cybercrime profiling. We believe 140 characters are not sufficient to express a viewpoint. Hence many people are using websites. In order to improve accuracy, it is necessary to scrap the content of a shared website and perform text mining on it. This process will give us valuable inputs for cybercrime profiling. Now, based on emotions captured from microblog posts and shared URLs, We can segregate the cybercrime profiles. It will help to reduce cybercrimes.

V. Conclusion and Perspectives

Our suggested method is primarily based on the calculation of a similarity distance to stumble on and are expecting criminal activities in microblog posts. The reason for our technique is to decompose each publish and shared URLs in terms and examine them robotically to predefined suspicious terms database by way of using similarity distance calculation.

This paper offers an extension and improvement of the already proposed solution. The already proposed solution offers an idea of our worldwide research challenge together with an automated system for detecting suspicious profiles in the social media, via which we will find suspicious behavior and interests of users as well.

On this paper we're focused on along with a disambiguation step, due to the fact many resources can be matched to the same entity that result in synonymy and polysemy problems in an effort to upload semantics of exchanged statistics to pick out extra large suspicious profiles and additionally to improve the gadget in term of execution time.
For future work, we plan to enhance the machine in time period of execution time, developing new scoring strategies for disambiguation and the use of other knowledge sources a good way to enhance the precision charges.

Reference

[1] S. ALAMI, O. ELBEQQALI "Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts" IEEE, 20-21 Oct. 2015.

[2] Wilson, R. 2006, Understanding the Perpetration of Employee Computer Crime in the Organizational Context. Working paper no.04-2006.

[3] Turvey, B., (2002). Criminal Profiling, An Introduction To Behavioural Evidence. UK, Elsevier.

[4] Johnson, T. A., (2005). Forensic Crime Investigation. USA, CRC Press.

[5] Kanellis P., Kiountouzis E., Kolokotronis N., and Martakos D., (2006). Digital Crime and Forensic Science in Cyberspace, Idea Group Inc. (IGI), USA.

[6] J. Tang, X. Wang, H. Gao, X. Hu and H. Liu, "Enriching short text representation in microblog for clustering," Journal of Frontiers of Computer Science in China, pp. 88-101, 2012.

[7] E. Meij, W. Weerkamp and M. d. Rijke, "Adding Semantics to Microblog Posts," in WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining, 2012.

[8] X. W. H. G. X. H. H. L. J. TANG, "Enriching short text representation in microblog for clustering," springer, 2012.

[9] P. N. Mendes, A. Passant, P. Kapanipathi and A. P. Sheth, "Linked open social signals," WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 224-231, 2010.

[10] S. Banerjee, K. Ramanathan and A. Gupta, "Clustering short texts using Wikipedia," SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, p. 787–788, 2007.

[11] X. Liu, S. Zhang, F. Wei and M. Zhou, "Recognizing named entities in tweets," HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 359-367, 2011.

[12] G. Stephen, C. Ming-Wei and K. Emre, "To link or not to link? a study on end-to-end tweet entity linking," In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Atlanta, Georgia,June. Association for Computational Linguistics, p. 1020–1030, 2013.

[13] H. Ben, R. Will, N. Joel, H. Matthew and R. C. James, "Evaluating Entity Linking with Wikipedia," journal of Artificial Intelligence, vol. 194, p. 130–150, 2013.

[14] D. Laniado and P. Mika, "Making sense of twitter," ISWC'10 Proceedings of the 9th international semantic web conference on The semantic web, vol. Volume Part I, pp. 470-485, 2010.