**Individual In-depth Report**

**Member name:** [Avani Mundra](mailto:amudra@asu.edu)

**Evaluated by:** [Rahul Nayak](mailto:rrnayak@asu.edu)

**Date:** 09/03/2023

**Tasks Assigned:**

* Prepared questions related to data mining techniques to handle multimodal data that cover the topic holistically.
* Finding research papers based on the above topic
* Reviewing the paper ‘Cybercrime Profiling: Text mining techniques to detect and predict criminal activities in microblog posts’ in depth

**Summary:**

* The aim of the paper is to overcome the challenges of data sparseness and reduce semantic gap in text mining techniques to detect suspicious behavior in Twitter posts.
* The paper focuses on Twitter dataset and leverages the availability of ‘hashtags’ to classify the posts into criminal or non-criminal activities
* The concept of Cybercrime Profiling is discussed in detail. With the advent of Web 2.0, the internet has become vulnerable to criminals who take advantage of virtual communities to spread inappropriate and harmful content.
* Several data mining and text analytics techniques that were previously developed by distinguished authors have been discussed in this paper.
* The proposed approach in the paper to add semantics to data mining consists of three stages:
  + Text Corpus: This is an exhaustive dataset consisting of texts posted in twitter. It includes user profiles, tweets and relationships.
  + Corpus Processing : This stage involves removing stop words and performing stemming and analyzing hashtags. Twitter hashtags are used as indicators to detect events and trending topics, especially for identifying suspicious or illegal events. They help organize and search tweets by subject.
  + Classification Process: This stage employs the Normalized Compression Distance function to identify similarity between words present in the post and the ones present in the text corpus.
* The process begins with extracting hashtags, which are analyzed to detect potentially suspicious content.
* Each post is broken down into terms and automatically compared to suspicious terms. If a hashtag contains suspicious information, a threshold called "a" is used to determine the maximum allowable distance for term comparison. If a sentence contains two terms similar to those in the database, it is classified as suspicious post.
* The above implementation is proven using an example which contains three distinct suspicious words. The similarity function returns the value 0,thereby declaring the content suspicious.

**Outcome:**

* An example containing two suspicious words is used to prove the above approach. The similarity function returns 0, thereby declaring the content suspicious.
* It extends a previous solution for detecting suspicious profiles, emphasizing disambiguation and semantic enrichment, with plans to further improve execution time and precision rates in future work.

**References**

[16] S. Alami and O. Elbeqqali, “Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts,” *International Conference on Intelligent Systems: Theories and Applications (SITA)*, Oct. 2015, doi: 10.1109/sita.2015.7358435.

**Evaluation of Report**

**Evaluation summary with justification.**

The Summary of the report is consistent and crisp in addressing the Cybercrime Profiling: Text mining techniques to detect and predict criminal activities in microblog posts. It outlines the use of Twitter hashtags and a Normalized Compression Distance function to identify suspicious content, with a practical example demonstrating effectiveness.  
**The quality of the major result(s) with justification.**  
The quality of major results is justified mainly because it provides an exhaustive review of prior research, introducing Cyber Crime Profiling, and a practical Twitter example. The method effectively addresses social media data challenges, particularly on Twitter, supporting its importance in advancing online suspicious behavior detection.

**The usefulness of the paper to the overall project.**   
This paper significantly advances the field of detecting suspicious activities on social media, particularly Twitter. It introduces "Cybercrime Profiling" and addresses data sparsity and semantic challenges. By integrating previous techniques and demonstrating practical implementation, it enhances the effectiveness of detecting malicious content, contributing significantly to the overall research topic.

**Other comments**

The report provides valid explanations and presents the paper's core ideas in a comprehensible manner.

**Evaluation Approval  
  
Evaluation by:** [Rahul Nayak](mailto:rrnayak@asu.edu) **Date: 09/03/2023**

**Is the written report of the in-depth study complete with all the major result(s) of the paper(s)? If not, provide as many examples of the major result(s) missing in the written report as possible. (in bullet form). [Normally within 100 words]**

* Yes, The weekly member’s Report is complete in summarizing major results of the paper.
* It introduces the concept of cybercrime profiling and the summary of the paper has been written correctly in line with the topic chosen.

**Is each section of the guidelines sufficiently completed? If not, point out what is missing. [Normally within 40 words].**

Yes, each section has been sufficiently completed.

**Is the quality of this version of the written report satisfactory? If not, then why not? [Normally within 40 words]**

Yes, The quality of this version of the report is satisfactory.

**Approval.  
  
Approved by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu) **Date:** 09/03/2023 **Is the quality of this written in-depth study report and Evaluation report satisfactory? If not, then why not? (limit: 40 words)**

Yes, the quality of this written in-depth study report and evaluation report is satisfactory.