**Individual In-depth Report**

**Member name:** [Krupaben Kothadia](mailto:kkothadi@asu.edu)

**Evaluated by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu)

**Date: 10/14/2023**

**Tasks Assigned:**

* This report summarizes the paper “[Detecting malicious tweets in trending topics using a statistical analysis of language](https://drive.google.com/file/d/1pv0VpftrMVLTlXlObU3yEcF_XfkC8iGH/view?usp=drive_link)”. In order to identify possibly suspicious users, this study presents a natural language processing (NLP) technique that examines users' regular chats. By looking at odd behaviors that occur during their exchanges, the writers show user behavior.

**Summary:**

* By examining the user's communication habits, suspicious people on social networks can be identified. This study divides messages into three categories: links within messages, multimedia messages, and message topics.
* The key to identifying suspicious activity is the Message Topics. Discussions about certain topics, such as sports and TV series, are common. Abrupt changes in subject or irrelevant content may cause suspicion.
* On certain platforms, communications can be tagged with hashtags to help identify irregularities. Text, picture, audio, and video messages combined with other multimedia formats can indicate malevolent intent. Content that is harmful, particularly for blackmail, implies a phony user. Message links carry the risk of being dangerous. Personal information theft may result from clicking on misleading links.
* The study presents a novel method based on natural language processing (NLP). Text analysis is made possible by NLP, with a particular emphasis on word tokenization and prediction. While chain rule probabilities analyze word sequences, N-gram models evaluate word probabilities inside phrases.
* Messages containing designated "bad words" can be blocked by comparing sequences to training datasets. By calculating word probabilities, maximum likelihood estimate (MLE) improves the detection of malicious users.
* In conclusion, this research suggests a method for locating malevolent users in social networks. Through the use of natural language processing (NLP) algorithms and communication pattern analysis, the system can identify and block content that poses a risk to security.

**Outcome:**

In order to detect malicious users in social networks, this study presents a novel approach that employs natural language processing (NLP) tools to analyze communication patterns. Through message classification into discrete categories, N-gram models, and maximum likelihood estimate (MLE), the system improves security by identifying and obstructing questionable information, protecting users from potential risks.

**References** *(with citation)*

[1] J. Martínez-Romo and L. Araujo, “Detecting malicious tweets in trending topics using a statistical analysis of language,” Expert Systems With Applications, vol. 40, no. 8, pp. 2992–3000, Jun. 2013, doi: 10.1016/j.eswa.2012.12.015.

**Evaluation of Report**

**Evaluation summary with justification.**

The study introduces a novel method based on NLP. It enables text analysis through word tokenization, chain rule probabilities for word sequences, and N-gram models to evaluate word probabilities within phrases. The system can block messages containing designated "bad words" by comparing them to training datasets and calculate word probabilities through Maximum Likelihood Estimate (MLE) to improve malicious user detection.

**The quality of the major result(s) with justification.**  
  
This research proposes an innovative approach to identify malevolent users on social networks. By employing NLP tools, message categorization, N-gram models, and MLE, the system enhances security by recognizing and blocking suspicious content, safeguarding users from potential risks.

**The usefulness of the paper to the overall project.**   
  
This paper is useful to the overall project as it helps detect suspicious activities in social media by analyzing the content for the probabilities of “bad words” present in order to mitigate malicious activities.

**Other comments**

No Comments

**Evaluation Approval  
  
Evaluation by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu) **Date: 10/14/2023**

**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 written report of the in-depth study is complete with all the major results of the paper.
* The study explores the use of NLP technique to mitigate threats in social media by finding the probability of “bad words” in the social media posts

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

Yes. Each section of the guidelines are sufficiently completed. The outcome of this study contributes to the goal of the overall project.

**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 written report is satisfactory. The study of this reference paper is useful to the challenges which need to be addressed in this overall project.

**Approval.  
  
Approved by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu) **Date: 10/14/2023  
  
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.