**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 “[Uncovering Cybercrimes in Social Media through Natural Language Processing](https://drive.google.com/file/d/1rAJ1djjueV3NpZ7rvedARvSnqdkh2rxB/view?usp=drive_link)”. The approach presented in this research monitors suspicious activity in social networks using neural network (NN) based natural language processing (NLP), which enables the detection and stopping of connected cybercrimes.

**Summary:**

* The report describes a natural language processing (NLP) and data preprocessing method used to analyze tweets in a research study. First, the tweets were cleaned to eliminate hashtags, mentions, and URLs in order to guarantee the clarity of the resulting model. After that, text normalization was used to standardize letter casing, and emojis were transformed into textual equivalents in order to retain their meaning.
* The tweets were then translated into the embedding language, which was usually English. The last stage was vectorization, which included taking each tweet and turning it into a single vector that represented the average of all the word vectors in the tweet.
* A matrix of tweets and their cosine distances was produced as a result of this approach. In order to evaluate how well the similarity model ranked related tweets, a validation dataset was carefully selected.
* The model was assessed using metrics including discounted cumulative gain (DCG) and hits. DCG@K evaluated both relevance and order within this ranking, whereas Hits@K determined if a tweet was in the top K most similar tweets.
* Results from using this strategy to two protest scenarios in 2020: one in Colombia and one in the US were described in the study. The data source was Twitter, and the TAGS application was used to collect the data. TinfoLeak was used to extract more information after Google News Embedding and word2vec were utilized for the embedding procedure. Gephi was used to create graphs of social networks.
* A training dataset was created in one scenario using 1,105 preprocessed tweets that were gathered, translated, and utilized. Hits and DCG measures were used to assess the model's performance, proving the method's efficacy.
* In brief, the research showcases the effectiveness of a similarity model across multiple metrics by highlighting data preparation and natural language processing techniques used to Twitter data in two protest scenarios.
* The elbow approach was used to choose the best tweet clusters after a thorough cleaning, translation, and vectorization process.

**Outcome:**

The study used a thorough preprocessing and analysis method, effectively employing Hits and DCG metrics to rank the similarity of tweets. After using this method to analyze Twitter data from two protest situations, tweet clusters based on cosine distances were optimized using efficient data translation, cleansing, and vectorization procedures.

**References** *(with citation)*

[1] J. R. Sánchez et al., “Uncovering Cybercrimes in Social Media through Natural Language Processing,” Complexity, vol. 2021, pp. 1–15, Dec. 2021, doi: 10.1155/2021/7955637.

**Evaluation of Report**

**Evaluation summary with justification.**

This report outlines a comprehensive approach to analyze tweets in a research study using natural language processing (NLP) techniques. The process began with tweet cleaning to remove hashtags, mentions, and URLs, ensuring model clarity. Text normalization was then applied to standardize letter casing, and emojis were transformed into textual equivalents to maintain meaning. The tweets were subsequently translated into English, serving as the embedding language.

**The quality of the major result(s) with justification.**  
  
This research demonstrates the effectiveness of a similarity model using various metrics, highlighting the importance of data preparation and NLP techniques in analyzing Twitter data from two protest scenarios. The elbow method was used to optimize tweet clusters after extensive cleaning, translation, and vectorization procedures.

**The usefulness of the paper to the overall project.**   
  
The paper is useful to the overall project by providing an efficient data preprocessing method by cleansing twitter posts and ensuring the content posted on social media is free of suspicious content. The text normalization approach used in this paper was also phenomenal and the results of the same were positive.

**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 discusses an approach to analyze tweets in a research study using NLP techniques to segregate tweets by cleansing them and translating them to obtain meaningful content by text normalization.

**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.