**Individual In-depth Report**

**Member name:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu)

**Evaluated by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu)

**Date: 09/10/2023**

**Tasks Assigned:**

* Focus on my question to work upon as my area of research for this project.
* The question I chose was to do my research and find a system that maintains the balance between detecting suspicious activity in social media and safeguarding the user's privacy and maintaining data integrity.
* Review of [**Privacy-Preserving Data Mining: Why, How, and When**](https://drive.google.com/file/d/1JF6NG6eQJ1_Y9zLhEyODuRKGJUtQuPMf/view?usp=drive_link)

**Summary:**

* The main focus area of this paper is to find how technology from the security community can change data mining for the better, providing all its benefits while still maintaining privacy
* Data mining doesn’t inherently threaten privacy, and the paper focuses on two approaches that enable it without revealing private information: randomization and secure multiparty computation (SMC).
* The goal of data mining is to extract knowledge from data. Data mining is categorized into five tasks:
  + Exploratory data analysis (EDA)
  + Descriptive modeling
  + Predictive modeling: classification and regression
  + Discovering patterns and rules
  + Retrieval by content
* The first three tasks output models that essentially summarize the data in various ways.
* The last two find specific patterns, but they’re often generalized and don’t reflect particular data items. Since these models generally don’t contain individual data values, they don’t present an immediate threat to privacy.
* Most data mining applications operate under the assumption that all the data is available at a single central repository, called a data warehouse.
* This poses a huge privacy problem because violating only a single repository’s security exposes all the data. Whether a data warehouse is real or virtual is irrelevant: if the data mining algorithm can access the data, the opportunity exists for an attacker to get it, too.
* Distributed data mining algorithms minimize the exchange of data needed to develop globally valid models. Such algorithms can be performed in ways that provably preserve privacy and prevent disclosure of data beyond the original sources.
* The following approaches to prevent disclosure of data from data mining:
  + Data Perturbation - Modifying the data so that it no longer represents real individuals. Uses Randomization to modify data extracted from datasets.
  + Randomization - Produces random samples from the set of data matrices satisfying the already discovered patterns or models.
  + Secure multiparty computation - Based on the idea that every piece of private information is validly known to one or more parties.
  + Association rule mining - Finds interesting associations and relationships among large sets of data items. Used in SCM approach to demonstrate protocol.

**Outcome:**

* Although the benefits of data mining and analysis are considerable, no matter what the government says, they aren’t worth the price of individual privacy.
* Neither giving the government complete rights over your data nor banning data mining entirely is a wise or widely acceptable solution.
* If the SMC approach is used, the process of data mining doesn’t cause, or even increase the opportunity for, breaches of privacy because only selected parties have access to the data.
* However further research is required to increase multi-level security and increase accuracy in terms of efficiency.

**References** *(with citation)*  
  
[7] J. Vaidya, C. Clifton, “Privacy-preserving data mining: why, how, and when”, in *“IEEE Security & Privacy, Vol 2, Issue 6”*, USA, 2004, pp.19-27

**Evaluation of Report**

**Evaluation summary with justification.**

This report covers the major points given in the article regarding classifying suspicious content on social media and the methods or techniques that can be useful. This report also explains the steps necessary to complete this process proposed in the research.

**The quality of the major result(s) with justification.**  
Results from this research include an in-depth explanation of methods used in machine learning for classifying suspicious content, which are helpful and relevant to the goal of this project. Given in the research as well is an analysis between the randomization and secure multiparty computation (SMC) with results.

**The usefulness of the paper to the overall project.**   
This paper will provide the project with detailed explanations to methods used in machine learning to perform privacy preserving data mining techniques.

**Other comments**

None

**Evaluation Approval  
  
Evaluation by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu) **Date: 09/11/2023**

**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 member’s report is complete in summarizing major results of the paper.
* It introduces the concept of secure multiparty computation and data perturbation to perform secured data mining.

**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:** 09/11/2023 **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.