Data Mining (Phase 2)

Team Name: Trio-Chargers

Team Members:

1. Sumanth Reddy Desireddy (Team Head) (sdesi4@unh.newhaven.edu)
2. Lakshmi Kara Gupta Chandulooru (lchan3@unh.newhaven.edu)
3. Uday kiran Karumburi Arumugam (ukaru1@unh.newhaven.edu)

Question:

Account security: Detecting spam or legit, unusual login behavior and unauthorized access attempts through Machine learning algorithms.

About:

Facebook is a platform used to generate and share content across a large network of users. The personal data of over 500 million Facebook users was posted in a low-level hacking forum. It includes phone numbers, full names, locations, email addresses, and biographical information. Security researchers say hackers could use the data to impersonate people and commit fraud. Facebook doesn't really secure your data but then again, you're putting it up for the world to see.

Literature Survey:

1. Christopher M. Hoadley, Heng Xu, Joey J. Lee and Mary Beth Rosson, "Privacy as information access and illusory control: The case of the Facebook News Feed privacy outcry," Electronic Commerce Research and Applications, Volume 9, Issue 1, January–February 2010, Pages 50–60, http://www.sciencedirect.com/science/article/pii/S1567422309000271

2. Bernhard Debatin, Jennette P. Lovejoy, Ann-Kathrin Horn M.A, Brittany N. Hughes, "Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences," Journal of Computer-Mediated Communication, 2009, pages 83–108, http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2009.01494.x/full

3. Stutzman, FD.,Gross, R. & Acquisti, A. (2013). Silent Listeners: The Evolution of Privacy and Disclosure on Facebook (2013). Journal of Privacy and Confidentiality, 4(2), https://ssrn.com/abstract=3305329

Dataset:

1. https://www.kaggle.com/datasets/khajahussainsk/facebook-spam-dataset
2. https://www.kaggle.com/datasets/sheenabatra/facebook-data

- Kaggle is a popular data-science competition website that provides free public datasets you can use to learn artificial intelligence (AI) and machine learning (ML).

About Datasets:

The dataset can be used for building machine learning models. To collect the dataset, Facebook API and Facebook Graph API are used and the data is collected from public profiles. There are 500 legit profiles and 100 spam profiles. The list of features is as follows with Label (0-legit, 1-spam).

1. Number of friends.
2. Number of the following.
3. Number of Community.
4. The age of the user account (in days).
5. Total number of posts shared.
6. Total number of URLs shared.
7. Total number of photos/videos shared.
8. Fraction of the posts containing URLs.
9. Fraction of the posts containing photos/videos.
10. Average number of comments per post.
11. Average number of likes per post.
12. Average number of tags in a post (Rate of tagging)
13. Average number of hashtags present in a post.

Various Data collection Methods to enhance Facebook security:

User Activity Monitoring: Facebook monitors user activity, looking for suspicious patterns or behaviors, such as unusual login locations or multiple failed login attempts.

Two-Factor Authentication (2FA): Users are encouraged to enable 2FA, which adds an extra layer of security by requiring a second form of verification, like a text message or authentication app, in addition to a password.

Device Recognition: Facebook collects data about the devices you use to access the platform. If a new device tries to log in, it may trigger additional security checks.

Algorithms Used for solving the research question:

Machine Learning Algorithms: Machine learning algorithms, such as clustering and classification, can be used to analyze patterns and detect anomalies in Facebook data. For example, you can use machine learning to identify suspicious or potentially malicious accounts based on their behavior and interactions.

Clustering: This is a data mining method used to place data elements in similar groups. Cluster is the procedure of dividing data objects into subclasses. Clustering quality depends on the way that we use. Clustering is also called data segmentation as large data groups are divided by their similarity.

To cluster your data, you'll follow these steps:

1. Prepare data.
2. Create similarity metric.
3. Run clustering algorithm.
4. Interpret results and adjust your clustering.

Prepare data:

As with any ML problem, you must normalize, scale, and transform feature data. While clustering, however, you must additionally ensure that the prepared data lets you accurately calculate the similarity between examples. The next sections discuss this consideration.

Create Similarity Metric

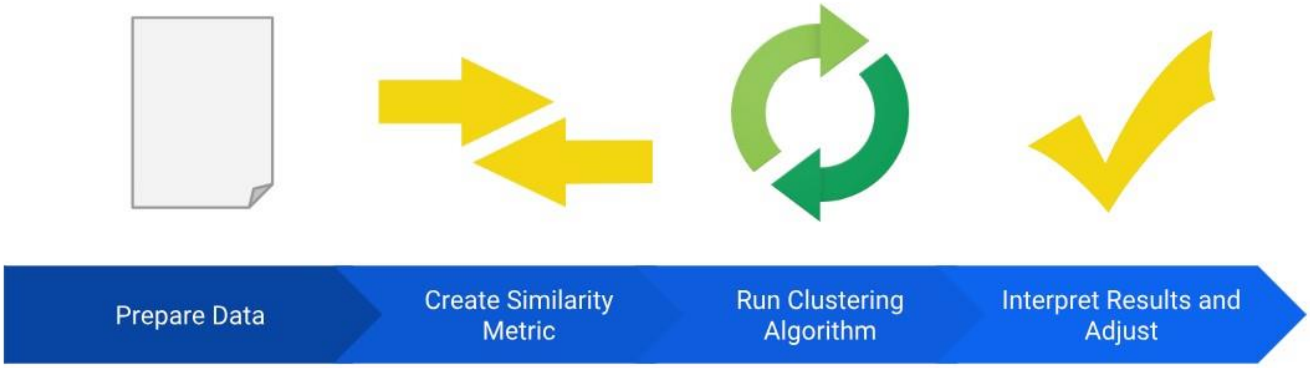
Before a clustering algorithm can group data, it needs to know how similar pairs of examples are. You quantify the similarity between examples by creating a similarity metric. Creating a similarity metric requires you to carefully understand your data and how to derive similarity from your features.

Run Clustering Algorithm

A clustering algorithm uses the similarity metric to cluster data. This course focuses on k-means.

Interpret Results and Adjust

Checking the quality of your clustering output is iterative and exploratory because clustering lacks “truth” that can verify the output.


Prepare Data
Create Similarity Metric
Run Clustering Algorithm
Interpret Results and Adjust