Natural Language Processing in Cybersecurity

Ahmed Yasser Ibrahim

222110758

SE495

Introduction

Natural Language Processing (NLP) is a major field that has been recently introduced to cybersecurity. It enables machines to understand complex language similarly to humans, allowing them to process and analyze unstructured text. In this lab, we explore NLP techniques and build models for **social monitoring** using a dataset of tweets retrieved from Twitter.

Preprocessing

We loaded the dataset into **Google Colab**, removed **NaN values**, and truncated it due to the computational power required for handling the full dataset. For testing purposes, we limited our dataset to **10,000 entries** after preprocessing.

Data Cleaning

- **Removed URLs and special characters** as they hold little semantic value.
- **Tokenized** the text into words.
- **Lemmatized** words to retrieve their canonical forms.

Feature Extraction

We extracted two key NLP feature representations:

- **Bag-of-Words (BoW):** A vocabulary-based representation of text.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** A weighted representation that accounts for word importance.

Since the dataset was truncated, we **did not perform PCA**, as the computational requirements were manageable.

Model Training and Results

We trained three different models and evaluated their performance:

1. Logistic Regression

- Achieved an **accuracy of 83.6%**.
- Despite being a simple model, it performed well on our dataset.

2. Decision Tree

- Underperformed with an **accuracy of 74.9%**.
- Unlike previous labs, Decision Trees were **not well-suited** for this dataset.

3. Support Vector Machine (SVM)

- Required the **longest computation time**.

- Delivered the **best performance** with an **accuracy of 88%**.

Conclusion

This lab demonstrated how **NLP techniques** can be applied to cybersecurity by analyzing social media data. After **preprocessing and feature extraction**, we trained multiple models and observed that **SVM performed best**, followed by **Logistic Regression**, while **Decision Trees were ineffective** for this dataset.

Future Improvements:

- Exploring **deep learning models** (e.g., transformers).
- Optimizing **hyperparameters** for better performance.
- Expanding **dataset size** for better generalization.

These findings reinforce the importance of **feature engineering and model selection** in NLP-based cybersecurity applications.