Project 4: Natural Language Processing
Group 2

Chiara Iorio - s343732, Renato Mignone - s336973,
Claudia Sanna - s343470

The goal of this laboratory is to apply Natural Language Processing (NLP) techniques in the context of cybersecurity. Specifically, we will analyze a dataset consisting of SSH sessions, represented as sequences of SSH entities, such as commands, flags, parameters, and separators. We will perform dataset characterization, then we will analyze the impact of different tokenization strategies and experiment with various pre-trained Language Models, comparing their performance. Finally, we will use the fine-tuned model for inference to investigate cybersecurity threats and infer the intentions of potentially malicious users.

1 Dataset Characterization

2 Tokenization

3 Model training

4 Inference