Named Entity Recognition in Digital Forensics: An Exploration of Real-World Entity Extraction

Degree programme :

An ever-growing load of data constantly challenges Cybercrime Investigations. The data's unsorted and unstructured nature makes it additionally difficult to analyze. Named Entity Recognition can help analyze vast amounts of data and find Named Entities relevant to cybercrime investigations. Using real-world data to train Named Entity Recognition models has shown promising results, potentially leading to further research.

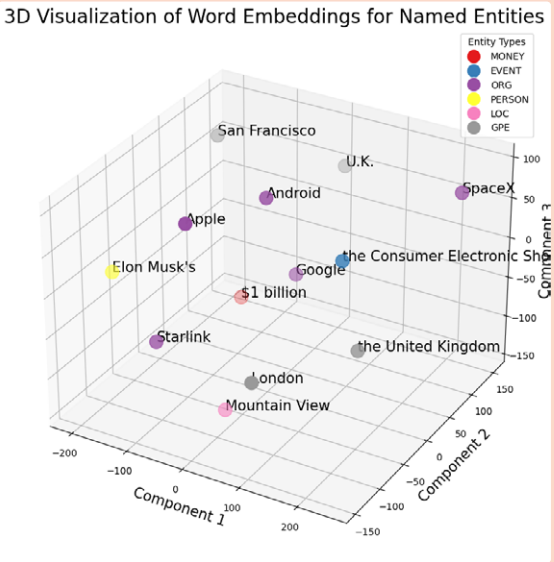
WARNING: Text overflow

Some text is too long and could not be fully displayed. The frame containing this text is indicated by a red background. Please reduce the length of the offending text and re-generate your abstract.

Overflow characters count: 1
Overflow content:

Challenges and opportunitie

Worldwide Cyber investigators
the rising volume and complexi
Digital Forensic and Cyber Inve
amounts of textual data, often f
or messenger apps like Telegra
in filtering out the relevant info
the textual data is riddled with
abbreviations, sometimes in di
complicates finding any data related to an investigation and forces the investigators to sift through the information manually. This is where Named Entity Recognition (NER) comes into the game. It is a technology that can extract Named Entities (NEs). NEs can be people's names, places, organizations, or quantities as monetary values. Applying this technology can speed up the investigation time substantially. The explorative research, implements classic and modern NER technologies, to augment the precision and speed up the analysis of cybercrime data. Ultimately creating a contemporary environment for cyber investigators.



ıta

Named
n
nologies
The data
ms like
lan or
Named
Entities of Interest (NEIs), specific groups like DDoS attack tools, botnets, and ransomware were defined. The NEIs were used to create specific datasets by annotating them after the groups. Prodigy has been used as a tool for annotating. The datasets were then used to train and measure the NER models from scratch. SpaCy, an open-source Natural Language Processing framework, was used for the training.

Findings and perspectives for practice

The NER models have been trained from scratch. The performance of the models has been measured using standard NLP metrics (precision, recall, F1-score). The results have shown that two of the NER models, which have been trained with a minimal amount of labels, have demonstrated surprisingly strong performance.

This research has furthermore led to insights into how labeled datasets created from real-world data affect the training of Named Entity Recognition models. Using real-word data for model training during the labelling process, allowed the NER models, not only get better when finding Entities (True Positives), but also to recognise when they found nothing (True Negatives).

The explorative work in the domain-overarching realm of Data Science, Cybersecurity, and Digital Forensics has proven to be an invaluable experience and made me quite passionate about NER. Gaining a deep insight was, on one side, very intense and, on the other side, very rewarding.

Charles Magri
MAS Data Science