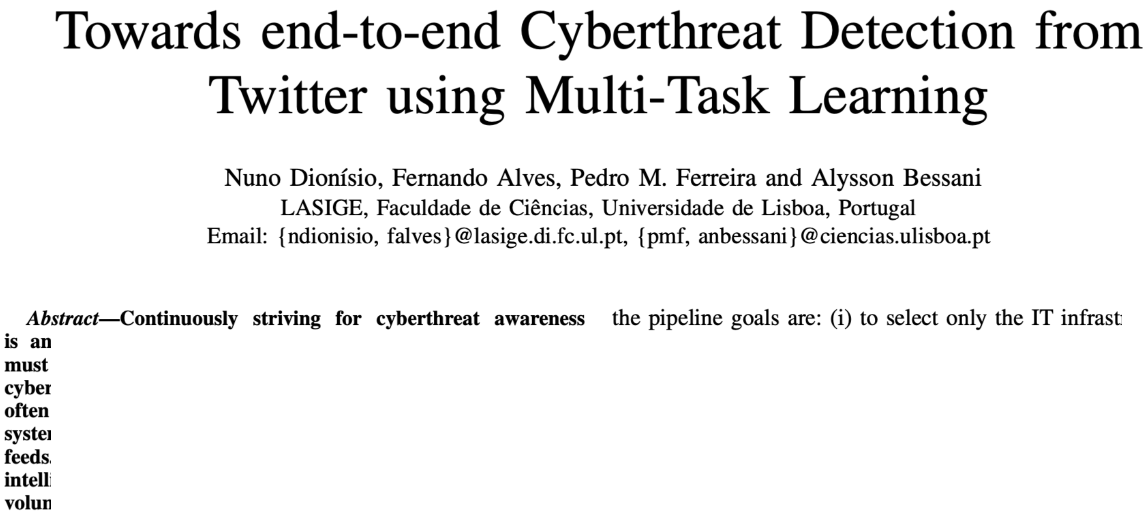
Evaluation of LLM Chatbots for OSINT-based Cyber Threat Awareness

https://arxiv.org/abs/2401.15127

Samaneh Shafee, Alysson Bessani, Pedro M. Ferreira

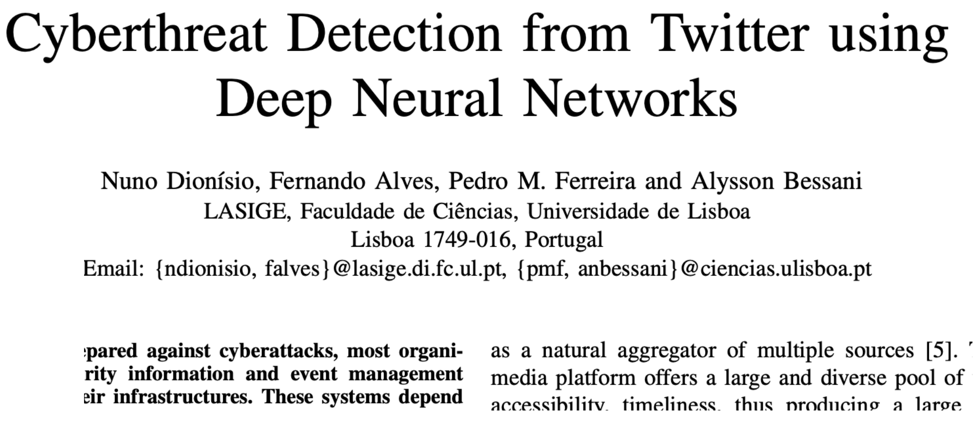


Towards end-to-end Cyberthreat Detection from Twitter using Multi-Task Learning

Nuno Dionísio, Fernando Alves, Pedro M. Ferreira and Alysson Bessani
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
Email: {ndionisio, falves}@lasige.di.fc.ul.pt, {pmf, anbessani}@ciencias.ulisboa.pt

***Abstract*—Continuously striving for cyberthreat awareness is an must cyber often syster feeds. intelli volun aggre strea**

the pipeline goals are: (i) to select only the IT infrast



Cyberthreat Detection from Twitter using Deep Neural Networks

Nuno Dionísio, Fernando Alves, Pedro M. Ferreira and Alysson Bessani
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Lisboa 1749-016, Portugal
Email: {ndionisio, falves}@lasige.di.fc.ul.pt, {pmf, anbessani}@ciencias.ulisboa.pt

pared against cyberattacks, most organi- rity information and event management eir infrastructures. These systems depend

as a natural aggregator of multiple sources [5]. media platform offers a large and diverse pool of accessibility, timeliness, thus producing a large

Follow the blue bird: A study on threat data published on Twitter⋆

Fernando Alves[1], Ambrose Andongabo[2], Ilir Gashi[2], Pedro M. Ferreira[1], and Alysson Bessani[1]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Centre for Software Reliability, City, University of London, UK

Abstract. Open Source Intelligence (OSINT) has taken the interest of cybersecurity practitioners due to its completeness and timeliness. In particular, Twitter has proven to be a discussion hub regarding the latest vulnerabilities and exploits. In this paper, we present a study comparing vulnerability databases between themselves and against Twitter. Al

Information Systems 95 (2021) 101586

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

ELSEVIER

Processing tweets for cybersecurity threat awareness

Fernando Alves *, Aurélien Bettini, Pedro M. Ferreira, Alysson Bessani
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

ARTICLE INFO

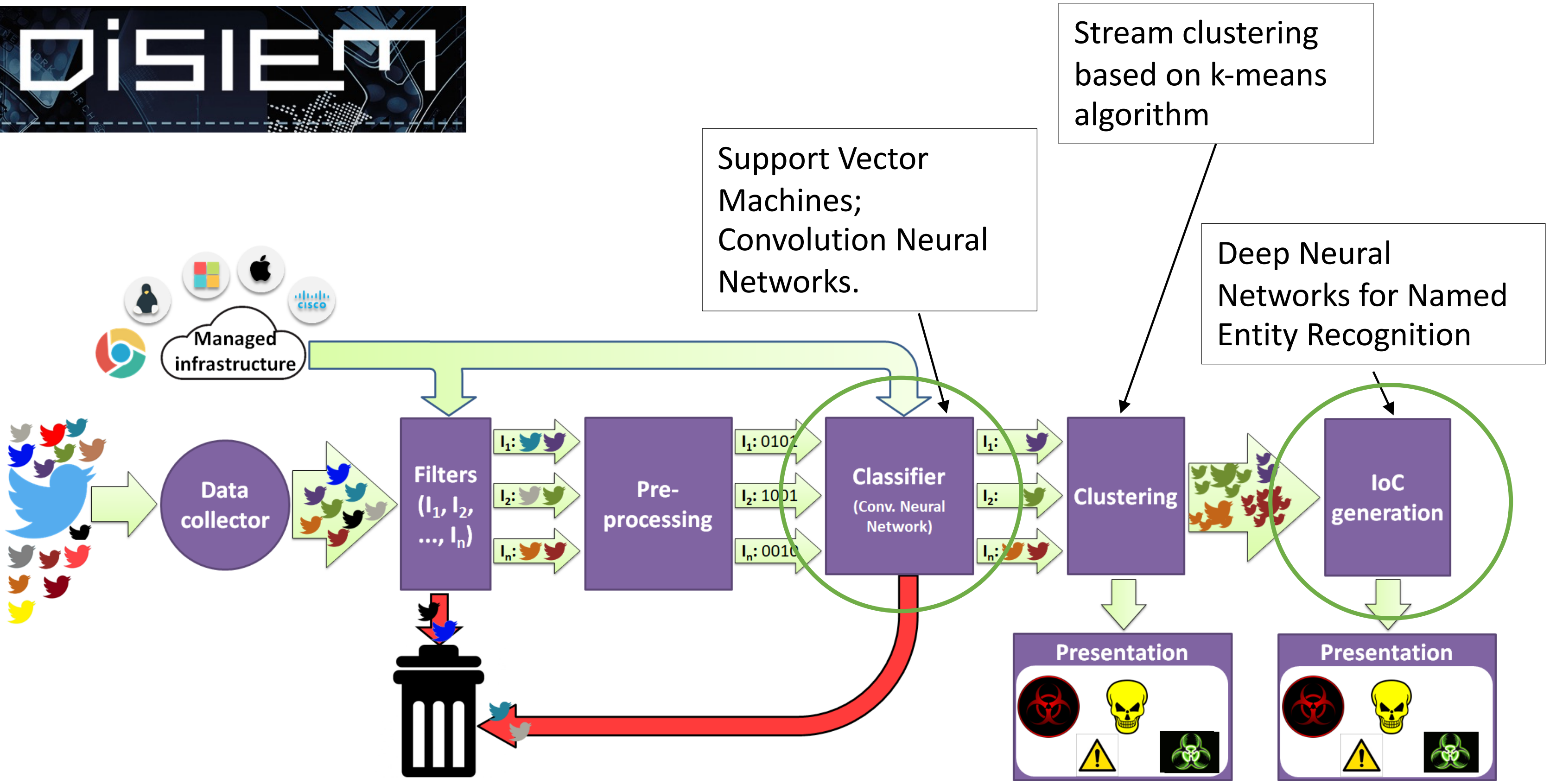
Article history:
Received 20 April 2020
Received in revised form 24 June 2020
Accepted 27 June 2020
Available online 4 July 2020
Recommended by Dennis Shasha

Keywords:
Threat intelligence
Threat discovery

ABSTRACT

Receiving timely and relevant security information is crucial for maint IT infrastructure. This information can be extracted from Open Sourc users, security organisations, and researchers. In particular, Twitter ha obtaining cutting-edge information about many subjects, including c SYNAPSE, a Twitter-based streaming threat monitor that generates a c the threat landscape related to a monitored infrastructure. SYNAPSE is kind of cybersecurity events and summarise them for the convenienc processing pipeline is composed of filtering, feature extraction, bin clustering strategy, and generation of Indicators of Compromise (

OSINT Processing Pipeline



Stream clustering based on k-means algorithm

Support Vector Machines; Convolution Neural Networks.

Deep Neural Networks for Named Entity Recognition

Managed infrastructure

Data collector

Filters (I_1, I_2, …, I_n)

I_1:

I_2:

I_n:

Pre-processing

I_1: 0101

I_2: 1001

I_n: 0010

Classifier (Conv. Neural Network)

I_1:

I_2:

I_n:

Clustering

IoC generation

Presentation

Presentation

The Question

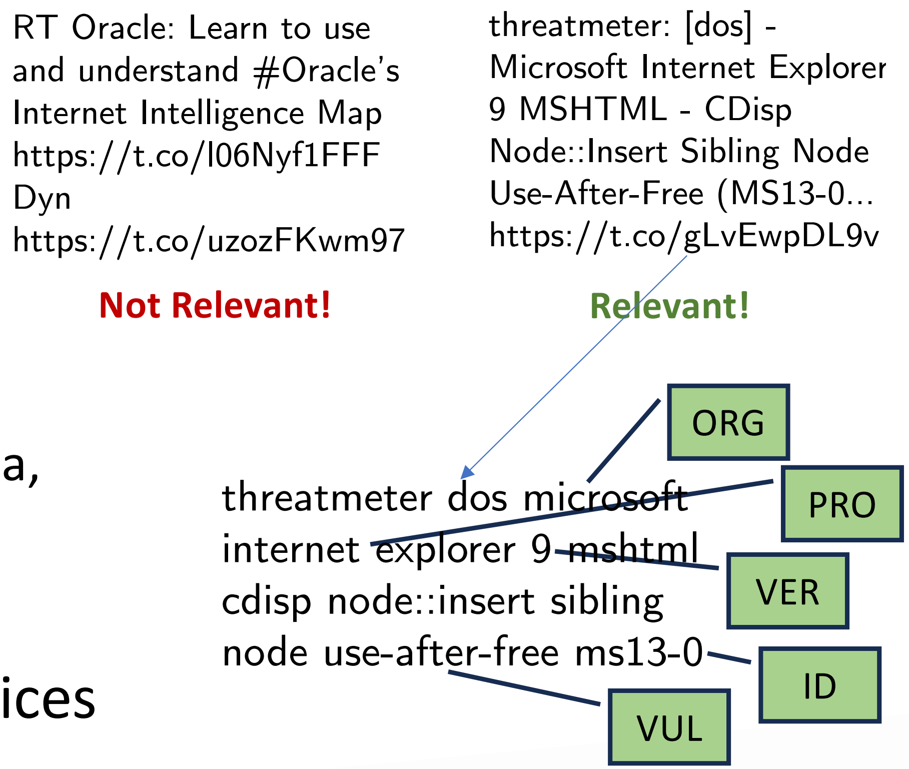
- *Given the success of LLM Chatbots, can we replace parts of this pipeline (Classification and IoC generation) by one of them?*

- Why?
  - Industry offers similar services (e.g., Microsoft copilot for Security)
  - They are very popular, so why not include them in automation pipelines?
  - Special-purpose models require updates and retraining

- There are similar research efforts for different tasks in other domains

Experiments

- Evaluated Chatbots
  - Commercial: ChatGPT
  - Open source: GPT4all, Dolly, Stanford Alpaca, Alpaca-LoRA, Falcon, and Vicuna
- Dataset: 38281 annotated tweets
- Followed prompt engineering best practices
- Different tests: ordered, shuttled, and isolated questions



RT Oracle: Learn to use and understand #Oracle's Internet Intelligence Map https://t.co/l06Nyf1FFF Dyn https://t.co/uzozFKwm97

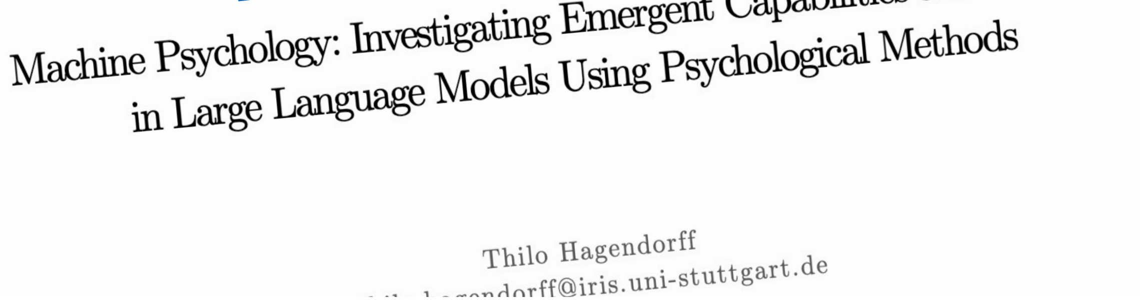
Not Relevant!

threatmeter: [dos] - Microsoft Internet Explorer 9 MSHTML - CDisp Node::Insert Sibling Node Use-After-Free (MS13-0... https://t.co/gLvEwpDL9v

Relevant!

threatmeter dos microsoft internet explorer 9 mshtml cdisp node::insert sibling node use-after-free ms13-0

https://arxiv.org/abs/2303.13988



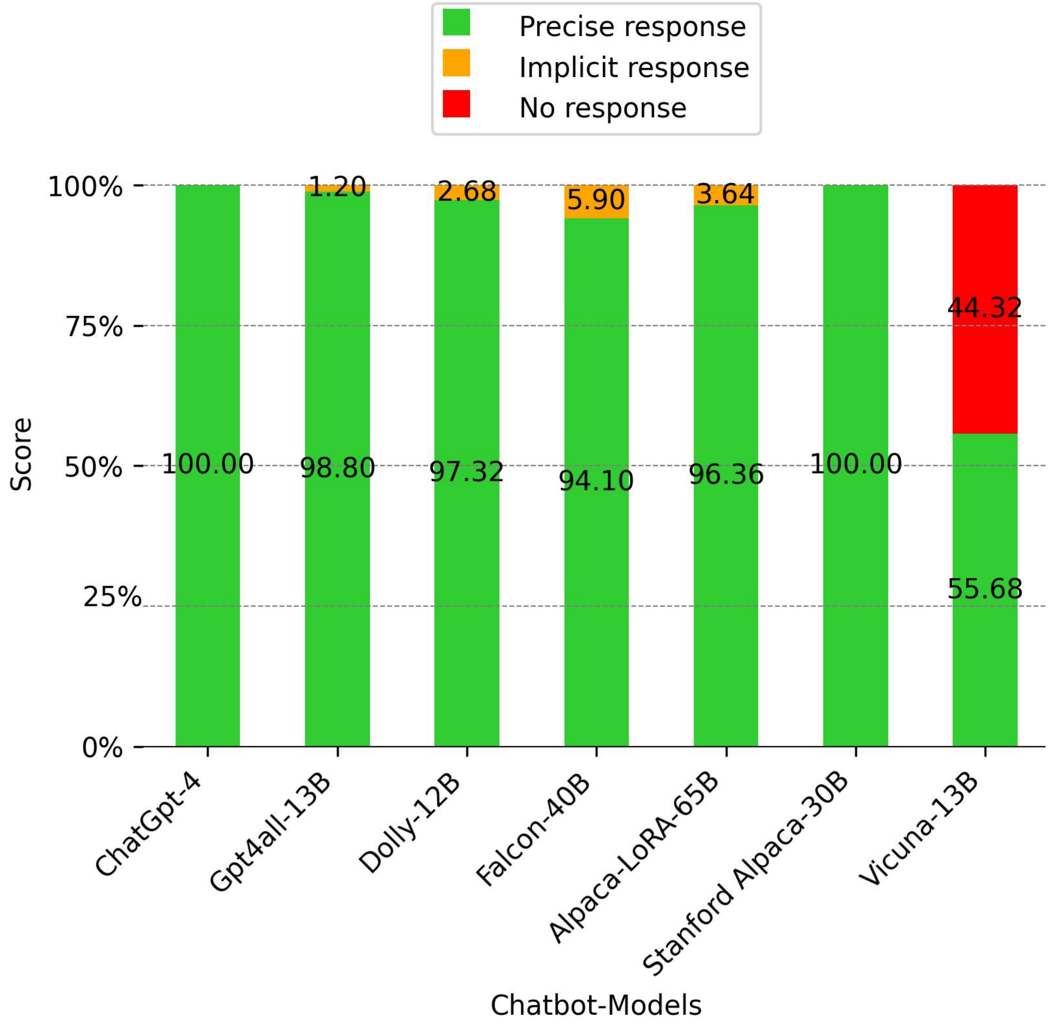
Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods

Thilo Hagendorff

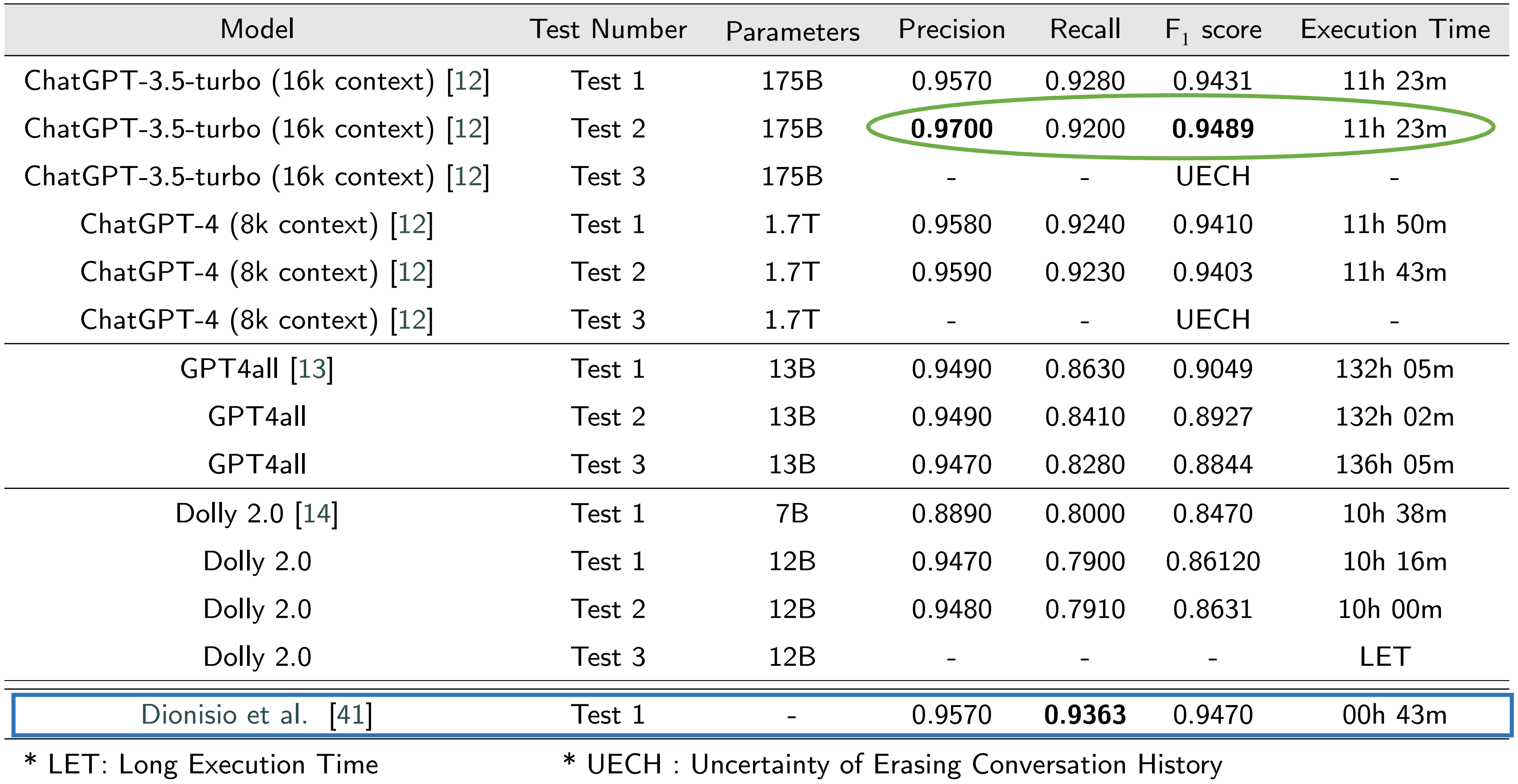
hagendorff@iris.uni-stuttgart.de

Chatbots “Failures”

- We ask, they answer. E.g.:
  - “Is the sentence ‘threatmeter dos microsoft internet explorer 9 mshtml cdisp node::insert sibling node use-after-free ms13-0’ related to cybersecurity? Just answer yes or no.”
- Wrong answers are expected, but the answer might not be clear



Classification



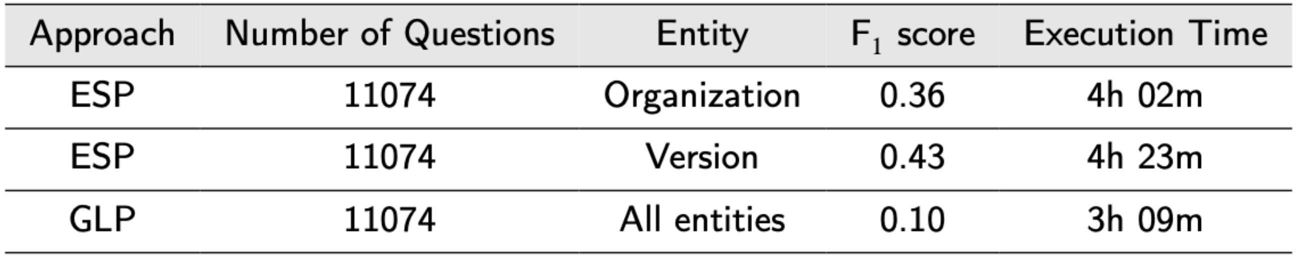
Model	Test Number	Parameters	Precision	Recall	F_1 score	Execution Time
ChatGPT-3.5-turbo (16k context) [12]	Test 1	175B	0.9570	0.9280	0.9431	11h 23m
ChatGPT-3.5-turbo (16k context) [12]	Test 2	175B	**0.9700**	0.9200	**0.9489**	11h 23m
ChatGPT-3.5-turbo (16k context) [12]	Test 3	175B	-	-	UECH	-
ChatGPT-4 (8k context) [12]	Test 1	1.7T	0.9580	0.9240	0.9410	11h 50m
ChatGPT-4 (8k context) [12]	Test 2	1.7T	0.9590	0.9230	0.9403	11h 43m
ChatGPT-4 (8k context) [12]	Test 3	1.7T	-	-	UECH	-
GPT4all [13]	Test 1	13B	0.9490	0.8630	0.9049	132h 05m
GPT4all	Test 2	13B	0.9490	0.8410	0.8927	132h 02m
GPT4all	Test 3	13B	0.9470	0.8280	0.8844	136h 05m
Dolly 2.0 [14]	Test 1	7B	0.8890	0.8000	0.8470	10h 38m
Dolly 2.0	Test 1	12B	0.9470	0.7900	0.86120	10h 16m
Dolly 2.0	Test 2	12B	0.9480	0.7910	0.8631	10h 00m
Dolly 2.0	Test 3	12B	-	-	-	LET
Dionisio et al. [41]	Test 1	-	0.9570	**0.9363**	0.9470	00h 43m

* LET: Long Execution Time * UECH : Uncertainty of Erasing Conversation History

REF.

Named Entity Recognition

- The way the question is asked is very important. Our method:
  - Find the name of **organizations|product** versions in the following sentence: ‘**TWEET**’. Give the shortest answer, and only use sentence segments in your response.
- ChatGPT4 results:



Approach	Number of Questions	Entity	F_1 score	Execution Time
ESP	11074	Organization	0.36	4h 02m
ESP	11074	Version	0.43	4h 23m
GLP	11074	All entities	0.10	3h 09m

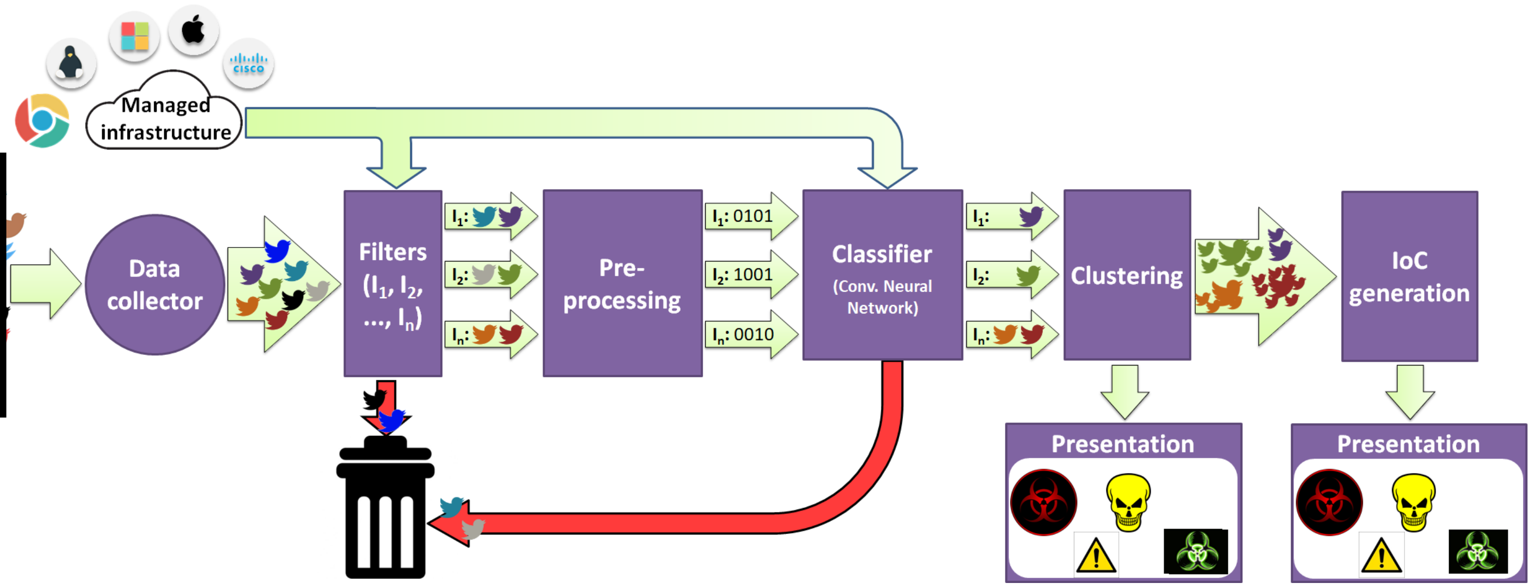
State of the art reports 0.94.

Main Takeaways

- LLM chatbots can do classification very well
  - They go slightly better than state-of-the-art deep learning models trained specifically for the task
  - Took more than 16x more time even running in better machines
- LLM chatbots cannot solve named entity recognition
  - Results are quite far from state-of-the-art
  - Also took a lot of time to process the queries
- This confirms what was observed in similar works for other domains

What’s next?

Random X accounts
Chan forums and
Other open spaces;
Dark Web forums.



New fault model:
- Misinformation
- Disinformation
  * LLM-generated disinformation

Questions?

- Alysson Bessani
  - anbessani@fc.ul.pt
  - www.di.fc.ul.pt/~bessani

This work was financially supported by the FCT through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020) and by the EU H2020 through the SATO project (957128).