



# CYBER-BULLYING, HATE SPEECH, AND ONLINE SOCIAL BRIDGES DETECTION

Moustaka, V., Chatzakou, D., Founta, A.-M., Gogoglou, A., Papagiannopoulou, E., Terzidou, T., Vakali, A.

#### GEC 2019

Athens, 7 June 2019

### BUILDING A MALICIOUS BEHAVIOR DETECTION BROWSER ADD-ON

- Online social networks (OSNs) constitute a breeding ground for the spread of several risks and threats to privacy and security
  - affect quality of life regarding information security
  - civic participation
- ENCASE platform
  - leverages the latest advances in usable security and privacy
    - to design and implement a browser-based architecture for the protection of minors from malicious actors in OSNs



Fig. 1. The ENCASE architecture

#### **CYBER-BULLYING & HATE SPEECH DETECTION**

- Framework
  - bully and aggressive users' various attributes detection (i.e., user, text, and network based)
- Proposed Methodology
  - ML classification algorithms can efficiently detect users exhibiting bullying and aggressive behavior, with over 90% AUC
- Unified deep learning classifier
  - hate speech in OSNs
  - many types of available data and metadata were used
  - was tested in multiple Twitter datasets with high performance and gaming dataset in a plug and play fashion, showing the potential to easily generalize its use into other platforms



Fig. 2. Pipeline of abusive detection process



Fig. 3. Architecture for a high-accuracy hate speech classifier

### ONLINE SOCIAL BRIDGES DETECTION

- 40,000 suspended accounts from Twitter were used as seeds to collect their neighboring sub-graphs from a complete graph of 50 million users
- The connected components of the formulated sub-graphs were calculated using the Tarjan algorithm and their connectivity was measured using k-core decomposition
- Green component: strongly connected core, red: peripheral nodes, <u>black</u>: disconnected nodes (malicious)
- The largest connected group of the red component constitutes the "social bridges" - linking malicious to honest users



Fig. 4. Twitter graph and social bridges

## TOWARDS IDENTIFYING PREDATOR BEHAVIOR IN CHAT CONVERSATIONS

- Perverted Justice dataset: contains dialogues between predators and victims
- Analysis of predators' and the victims' posts
  - 2 different datasets were created
  - For each dataset, the posts per seduction case were grouped and each case's posts were concatenated, following ascending order based on the date / time information field
- Exploitation of both textual information and affect/sentiment scores
  - One-Class SVM model training and evaluation
    - Distinction between predators and victims / friendly conversations

Text Representation	Feature set	OC-SVM params (kernel-nu-gamma)	Precision on Predators	Recall on Predators	Precision on Victims	Recall on Victims
GloVe	vector+affects+#posts	sigmoid-0.5-0.001	<u>1.00</u>	<u>0.75</u>	<u>1.00</u>	<u>1.00</u>
-	affects+#posts	sigmoid-0.5-0.001	<u>1.00</u>	<u>0.75</u>	<u>1.00</u>	<u>1.00</u>
Tfldf	vector+affects+#posts	poly-0.5-0.001	<u>1.00</u>	0.50	<u>1.00</u>	0.51

#### Table 1. Precision and recall scores on predators and victims





This work was funded by the European Union's Horizon 2020 research and innovation program ENCASE under the Marie Skodowska-Curie grant agreement No 691025