Aristotle University of Thessaloniki

Cyber-bullying , hate speech, and online social bridges detection

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

Abstract

Online social networks (OSNs) constitute a breeding ground for the spread of several risks and threats to privacy and security affecting life quality regarding information security and civic participation in data production [1].

Aiming at protection of minors from malicious actors in OSNs, a browser-based architecture* was designed and deployed, leveraging the latest advances in usable security and privacy.

Some of the methodologies that were used to develop the add-on, which aims to identify users' profiles for detection and prediction of malicious online behavior, are presented herein.

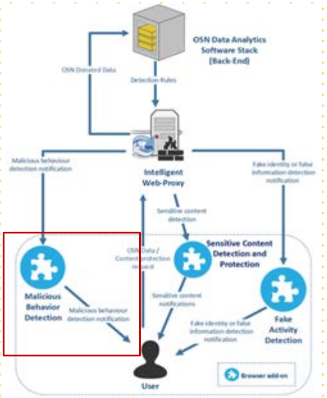

OSN Data Analytics Software Stack (Back-End)
OSN Donated Data
Detection Rules
Intelligent Web-Proxy
Malicious behaviour detection notification
Fake identity or false information detection notification
Sensitive content detection
OSN Data / Content protection request
Sensitive Content Detection and Protection
Malicious Behavior Detection
Malicious behaviour detection notification
Sensitive content notifications
Fake identity or false information detection notification
Fake Activity Detection
User
Browser add-on

Fig. 1. The ENCASE architecture

*https://encase.socialcomputing.eu/

Building a malicious behavior detection browser add-on

- A novel framework has been developed to detect bully and aggressive users various attributes, i.e., user, text, and network based
- The proposed methodology evaluated by a corpus of 1.6M tweets, showed that machine learning classification algorithms can efficiently detect users exhibiting bullying and aggressive behavior, with over 90% AUC [1]

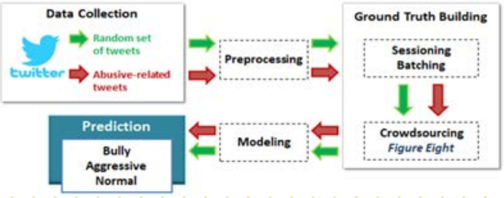


Fig. 2: Pipeline of abusive detection process

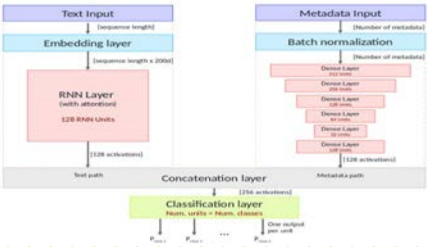


Fig 3. Architecture for a high-accuracy hate speech classifier

- A unified deep learning classifier has been proposed, aiming at predicting hate speech in OSNs [2]
- This classifier was tested in multiple Twitter datasets with high performance and one gaming dataset in a plug and play fashion, showing the potential to easily generalize its use into other platforms
- A methodology for annotating a large scale dataset of inappropriate speech was proposed on a 100k labeled Twitter dataset shared openly with communities [3]

- 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, black: disconnected nodes (malicious)
- The largest connected group of the red component constitutes the "social bridges" - linking malicious to honest users [4]

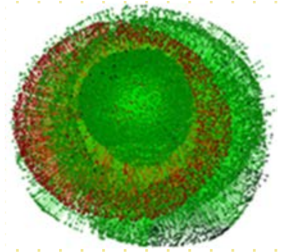
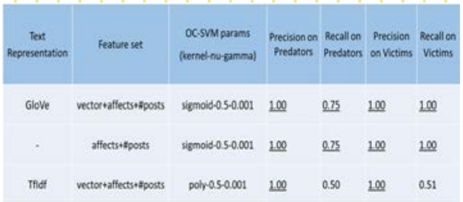


Fig. 4. Twitter graphs and social bridges

Table 1. Precision and recall scores on predators and victims



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	1.00	0.75	1.00	1.00
-	affects+#posts	sigmoid-0.5-0.001	1.00	0.75	1.00	1.00
Tfidf	vector+affects+#posts	poly-0.5-0.001	1.00	0.50	1.00	0.51

- The Perverted Justice dataset was exploited, with the purpose of identifying predator behavior in chat conversations
- The predators' and the victims' posts were analyzed separately
- Both textual information and affect/sentiment scores were exploited with the purpose of training and evaluating an One-Class SVM model which was used to distinguish between predators and victims/friendly conversations

References

[1] V., Moustaka, Z., Theodosiou, A., Vakali, A., Kounoudes, L.-G., Anthopoulos. (2019). Enhancing Social Networking in Smart Cities: Privacy and Security Borderlines. *Technological Forecasting & Social Change* 142, 285-300.

[2] D. Chatzakou, N., Kourtellis, J., Blackburn, E., De Cristofaro, G., Stringhini, A., Vakali. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, Troy, NY (USA).

[3] A.-M., Founta, D., Chatzakou, N., Kourtellis, J., Blackburn, A., Vakali, I., Leontiadis. (2019). A Unified Deep Learning Architecture for Abuse Detection. Accepted in *ACM Conference on Web Science (WebSci '19)*, Boston, USA.

[4] A.-M., Founta, C., Djouvas, D., Chatzakou, I., Leontiadis, J., Blackburn, G., Stringhini, A., Vakali, M., Sirivianos, N., Kourtellis. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the ICWSM '18*, Stanford, California.

[5] A., Gogoglou, Z., Theodosiou, T., Kounoudes, A., Vakali, Y., Manolopoulos. (2016). Early malicious activity discovery in microblogs by social bridges detection. *In 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Limassol, Cyprus.

Acknowledgments

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

Contact

Name: Athena Vakali
Email: avakali@csd.auth.gr
Organization: Department of Informatics (AUTH)
Website: https://datalab.csd.auth.gr/