BiRDy: Bullying Role Detection in Multi-Party Chats

Anaïs Ollagnier[1], **Elena Cabrio** [1], **Serena Villata** [1], **Sara Tonelli** [2]

[1] Université Côte d'Azur, Inria, CNRS, I3S
930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France
[2] Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy
{anais.ollagnier,elena.cabrio,serena.villata}@inria.fr, satonelli@fbk.eu

Abstract

Recent studies have highlighted that private instant messaging platforms and channels are major media of cyber aggression, especially among teens. Due to the private nature of the verbal exchanges on these media, few studies have addressed the task of hate speech detection in this context. Moreover, the recent release of resources mimicking online aggression situations that may occur among teens on private instant messaging platforms is encouraging the development of solutions aiming at dealing with diversity in digital harassment. In this study, we present BiRDy: a fully Web-based platform performing participant role detection in multi-party chats. Leveraging the pre-trained language model *mBERT* (multilingual BERT), we release fine-tuned models relying on various contextual window strategies to classify exchanged messages according to the role of involvement in cyberbullying of the authors. Integrating a role scoring function, the proposed pipeline predicts a unique role for each chat participant. In addition, detailed confidence scoring are displayed. Currently, BiRDy publicly releases models for French and Italian.

Introduction

Over the past few years, numerous studies about role detection on social platforms aiming at identifying malicious users in the context of cyberbullying episodes have been proposed (Rosa et al. 2019; Salawu, He, and Lumsden 2020; Kim et al. 2021). Whilst most of existing studies rely on network-based features (e.g., number of followers, network centrality) (Chatzakou et al. 2017; Kao et al. 2019), their coverage is limited to social media channels exploiting such structure and dynamic. To overcome this limitation, recent studies have focused on other social media including online forums (e.g. Reddit) and question answering services (e.g. ASKfm) investigating NLP-powered techniques. Among them, we have observed the prevalence of machine learning techniques relying on linguistic-based features ranging from shallow (surface form of the post) to deep-level (theoretical and descriptive linguistic analysis), as well as sentiment analysis (Ratnayaka et al. 2020; Kim et al. 2021; Jacobs, Hee, and Hoste 2022). Despite promising approaches which have contributed to better understand and curb online cyber aggression, most of them are limited to the scope of public bullying episodes (Alkomah and Ma 2022). However, private instant messaging platforms and channels have recently been pinpointed as one of the main platforms used to perpetrate bullying, especially among teens (Bedrosova et al. 2022). Since data collection from major social media platforms is strictly limited, very few studies have investigated the task of participant role detection in multi-party chats. To contribute filling this gap, two datasets were recently released (Sprugnoli et al. 2018; Ollagnier et al. 2022), which were collected through role-playing games mimicking cyber aggression situations on private instant messaging platforms. Both of them consist of conversations manually annotated using a multi-level annotation scheme including the different participant role of involvement in cyberbullying episodes. In this context, we introduce BiRDy[1] for Bullying Role Detection: a solution aiming at automating the identification of participant roles in cyberbullying occurring on multi-party chats. The proposed pipeline consists of two main tasks: (1) the labelling of participant role on exchanged messages, and (2) the assignation of a unique role to chat participants.

BiRDy Overview

Considering the nature of the considered datasets, all the conversations report cyber aggression situations in which each user is assigned to a unique role such as victim, bully, conciliator, bystanders assistant or defender. Using a conversation as input, the proposed pipeline attributes one of the roles to the identified users. In detail, fine-tuned *mBERT* models are trained to classify each exchanged message according to the role of the author. The fine-tuning methodology consists of exploring the use of contextual window by collecting for each message the n previous sentences authored by the same user. Due to the semi-asynchronous and "entangled" nature of the contributions by chat participants, using this strategy aims at reorganising the structure of events that unfold in the narrative. Once roles are predicted for the given conversation, a role scoring function is computed to assign a unique role to users by considering the probabilities established by the model. Formally, for a conversation c let U be a set of users, L the set of labels and S

[1] BiRDy is publicly available online: http://134.59.134.227/demo_prd/index.html

the set of all messages posted in c. We define the following functions:

- $speech\_acts : U \rightarrow 2^S$ which represents the whole messages posted by the user U.
- $D : S \rightarrow [0,1]^L$ which assigns for all users $u \in U$ and a sentence $s \in S$ a vector of probabilities.
- $Classifier : [0,1]^L \rightarrow L$ the function assigning a label to each vector. In this context, we consider $argmax$ to find the label with the largest predicted probability, named $P_{am}(m,l)$ for a pair of message and label.

We opted for the following metric $Role\_scoring : U \times L \rightarrow [0,1]$ such that, $\forall (u,l) \in U \times L$:

$$Role\_scoring(u,l) = \frac{\sum\limits_{\substack{m \in speech\_acts(u) \\ argmax(D(m))=l}} P_{am}(m,l)}{|speech\_acts(u)|} \quad (1)$$

In addition, to evaluate the ability of a given $Classifier$ to predict the appropriate role to users, we have established a confidence score. It is aiming at detailing for each label the degree of likelihood per user, such that, $\forall (u,l) \in U \times L$:

$$Confidence(u,l) = \frac{|U| \times Role\_scoring(u,l)}{\sum\limits_{u' \in U} Role\_scoring(u',l)} \quad (2)$$

Evaluation

Experiments were conducted considering different contextual windows on: (1) the task of participant role labelling on exchanged messages, and (2) role assignment to chat participants. For the experiment (1), training sets are composed of messages extracted from the aforementioned datasets, *i.e.* 2,921 entries in French and 2,114 in Italian. Several training sets are built, each corresponding to a contextual window ranging from 0 to 10. Concerning preprocessing, all the messages are lower-cased and tokenized. In turn, they are respectively truncated/padded to a length of 185 in French and 200 in Italian. Then, they are encoded using the *mBERT* base uncased model released by Hugging Face[2]. Next, generated sentence vector-based features are used to fine-tune *mBERT* to address the task of participant role detection. For both languages, the number of epochs is set to 6, the number of batches to 4, the learning rate to 2e-05 and the remaining parameters use default values. Figure 1 presents results obtained w.r.t. the weighted F-score. The best results are respectively obtained using a window of 9 in French (weighted F-score of 0.778) and of 10 in Italian (weighted F-score of 0.869). The procedure used to collect the scenarios, the type of cyber aggression topics addressed and the age of participants involved into role playing games are among factors explaining such differences in model performances between languages. Overall, reported results encourage in pursuing efforts in leveraging contextual and narrative discourse information in this context.

Concerning the experiment (2), we have computed the $Role\_scoring$ function on each conversation extracted from the datasets, namely, 19 chats in French and 10 in Italian. Fine-tuned models trained in the experiment (1) are used in this evaluation to generate predictions of the bullying participant roles over conversations. The preprocessing and the model settings are similar to the ones reported for each language. Figure 2 presents the average F-scores obtained for the detection of bullies and victims in chats. In detail, the identification of bullies in the Italian dataset is more accurate than for victims, it reaches twice a F-score of 1.0. Conversely, the identification of victims performs better in the French corpus than for bullies. Reported misclassifications for victims and bullies for the Italian dataset are between victims and bullies and vice-versa, respectively representing 11% and 34% of models' mistakes. Concerning the French dataset, misclassifications are observed between bullies and bystander assistants (31%) and between victims and conciliators (24%). Observations on the datasets have shown variations in bullying engagements suggesting that the role of some participants evolves/switches over the conversation. Overall, obtained results are promising and suggest that BiRDy can help in monitoring cyberbullying on private instant messaging platforms, channels or any media relying on multi-party setting.

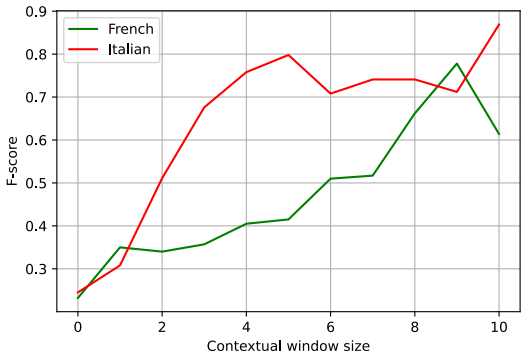

Figure 1: Results of the experiment (1) on the French and the Italian datasets.

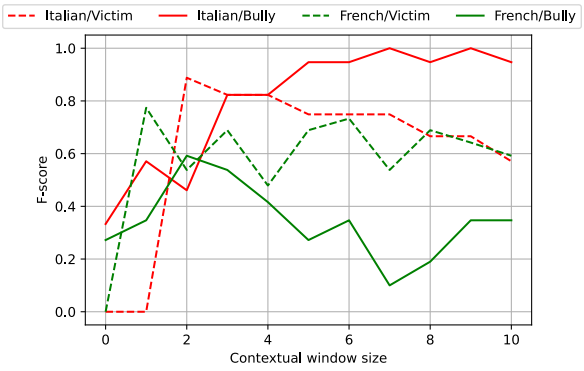

Figure 2: Results of the experiment (2) on the French and the Italian datasets for the detection of bullies and victims.

[2]https://huggingface.co/bert-base-multilingual-uncased

Acknowledgements

This work is funded under the IDEX UCA OTESIA, the UCA Academy 1 project (C870A021–D103–ACAD1_FIN_17_20Y) and the French government, through the 3IA Côte d'Azur Investments managed by the National Research Agency (ANR-19-P3IA-0002).

References

Alkomah, F.; and Ma, X. 2022. A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Inf.*, 13(6): 273.

Bedrosova, M.; Macháčková, H.; Serek, J.; Smahel, D.; and Blaya, C. 2022. The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech Republic, Poland, and Slovakia. *Comput. Hum. Behav.*, 126: 107013.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Cristofaro, E. D.; Stringhini, G.; and Vakali, A. 2017. Detecting Aggressors and Bullies on Twitter. In Barrett, R.; Cummings, R.; Agichtein, E.; and Gabrilovich, E., eds., *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, 767–768. ACM.

Jacobs, G.; Hee, C. V.; and Hoste, V. 2022. Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Nat. Lang. Eng.*, 28(2): 141–166.

Kao, H.; Yan, S.; Huang, D.; Bartley, N.; Hosseinmardi, H.; and Ferrara, E. 2019. Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection. In Amer-Yahia, S.; Mahdian, M.; Goel, A.; Houben, G.; Lerman, K.; McAuley, J. J.; Baeza-Yates, R.; and Zia, L., eds., *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 183–188. ACM.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and Choudhury, M. D. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2): 1–34.

Ollagnier, A.; Cabrio, E.; Villata, S.; and Blaya, C. 2022. CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game. In *Proceedings of the Language Resources and Evaluation Conference*, 867–875. Marseille, France: European Language Resources Association.

Ratnayaka, G.; Atapattu, T.; Herath, M.; Zhang, G.; and Falkner, K. 2020. Enhancing the Identification of Cyberbullying through Participant Roles. In Akiwowo, S.; Vidgen, B.; Prabhakaran, V.; and Waseem, Z., eds., *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, 89–94. Association for Computational Linguistics.

Rosa, H.; Pereira, N. S.; Ribeiro, R.; Ferreira, P. C.; Carvalho, J. P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A. M. V.; and Trancoso, I. 2019. Automatic cyberbullying detection: A systematic review. *Comput. Hum. Behav.*, 93: 333–345.

Salawu, S.; He, Y.; and Lumsden, J. 2020. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Trans. Affect. Comput.*, 11(1): 3–24.

Sprugnoli, R.; Menini, S.; Tonelli, S.; Oncini, F.; and Piras, E. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In Fiser, D.; Huang, R.; Prabhakaran, V.; Voigt, R.; Waseem, Z.; and Wernimont, J., eds., *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP*. ACL.