Harnessing Bullying Traces to Enhance Bullying Participant Role Identification in Multi-Party Chats

Anaïs Ollagnier, Elena Cabrio, Serena Villata

Harnessing Bullying Traces to Enhance Bullying Participant Role Identification in Multi-Party Chats

Anaïs Ollagnier, Elena Cabrio, Serena Villata
Université Côte d'Azur, Inria, CNRS, I3S
930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France

Abstract

As online content continues to grow, so does the spread of online hate, especially on social media. Most research efforts conducted on the task of bullying participant role identification are directed towards social networks such as Twitter and Instagram. However, private instant messaging platforms and channels were pinpointed in recent studies as the most prominent grounds for cyberbullying, especially among teens. Since data collection from major social media platforms is strictly limited, very few studies have investigated this task in a multi-party setting. However, the recent release of resources mimicking online aggression situations that may occur among teens on private instant messaging platforms contributes to filling this gap. In this study, we introduce a full pipeline aiming at automating the identification of bullying participant roles (bully and victim) in multi-party chats. Leveraging pre-trained language models and different learning frameworks, we perform hateful content classification of exchanged messages according to a binary scheme (online hate or no online hate). Then, - from these bullying traces - bullying behavioural cues (repetition and intention to harm) are derived and formalised into a role scoring function. As a result, the proposed pipeline identifies the bully and the victim among chat participants. Evaluated against state-of-the-art methods, the proposed pipeline achieves better performances considering all the datasets and roles to predict. In addition, the error analysis confirms that deriving bullying behavioural cues is beneficial to the task of participant role identification.

INTRODUCTION

Nowadays, people increasingly use social media platforms, not only as their main source of information but also as a means to post content, sharing their feelings and opinions. However, such environment is raising concerns as there are more and more episodes of online hate and harassment on these platforms. *Online hate* (OH hereafter) is defined by Salminen et al. (2020) as "comments using language that contain either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity" – in other words, comments that are rude, disrespectful, and can result in negative online and offline consequences for the individual, community, and society at large. Over the past years, the NLP community has introduced a plethora of theories, methodologies, taxonomies and real-world applications aiming at preventing and curbing this phenomenon. Recently, several important surveys have been published summarising research efforts conducted to address the task of online hate detection (OHD hereafter) (Jahan and Oussalah, 2021; Yin and Zubiaga, 2021; Chinivar et al., 2023). OHD encompasses various sub-tasks including the classification of hateful content (e.g. distinguishing hate and non-hate speech, identifying hate targets or detecting types of hate) and more recently the task of participant role identification (PRI hereafter). Concerning the latest, PRI consists in identifying the different participant roles of involvement in the context of cyberbullying episodes (Ratnayaka et al., 2020), i.e., bully, victim, bystander, assistant, defender, *inter alia*. Currently, most research efforts conducted on this task are directed toward social networks such as Twitter and Instagram. 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). Due to social media privacy policies, very few studies have investigated the task of PRI in multi-party chats.

In this paper, we propose to contribute to filling this gap by introducing a pipeline aiming at automating the identification of participant roles in cyberbullying episodes occurring on multi-party chats, i.e., bully and victim. The proposed pipeline consists of two main tasks: (1) the identification of hateful content on exchanged messages, and (2) the assignation of a bullying role to chat participants. In detail, leveraging state-of-the-art pre-trained language models and different learning settings (mono-lingual, cross-lingual, few and zero shot learning) we perform hateful content classification distinguishing the presence or absence of OH in chat participants' contributions. Then, learning from these bullying traces we derive bullying behavioural cues allowing to distinguish bullies from victims among chat participants, i.e., repetition and intention to harm. The contribution of this work is twofold: (1) it introduces a full pipeline portable to multiple languages, and (2) it formalises bullying behavioural cues to reinforce OHD tasks. Evaluations of each task were conducted on both French and Italian using two datasets mimicking cyber aggression situations on private instant messaging platforms (Sprugnoli et al., 2018;

Ollagnier et al., 2022). Concerning the task (1), fine-tuned models relying on the *augmented cross-lingual* learning framework (models trained using both the French and Italian data) reach the best balance on both languages. Supported by post-hoc explanations, this learning setting demonstrates better robustness in facing different types of aggression and verbal abuse unlike *zero* and *few-shot* settings. Concerning the task (2), deriving bullying behavioural cues related to concepts of aggressive behaviours and from cyberbullying criteria (repetition and intention to harm) contributes in contrasting aggressive behaviours between bullies and victims.
NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

Related work

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). Among them, we have observed the prevalence of methodologies dedicated to social networks such as Twitter and Instagram. Many of them introduce various metrics exploiting user- (e.g., age, gender, location) (Chatzakou et al., 2017b; Tahmasbi and Rastegari, 2018; Balakrishnan et al., 2019) and social network-based features (e.g., number of followers, network centrality) (Chatzakou et al., 2017a; Kao et al., 2019). While reported performance of these approaches suggest that network-based features benefit in understanding participant roles of involvement in bullying episodes, their coverage is limited to social media channels exploiting a such structure and dynamic. To overcome this limitation, other studies have built beyond existing works to other social media including online forums (e.g. Reddit) and question asking and answering services (e.g. ASKfm). In this context, most of the proposed approaches have applied machine learning techniques using linguistic-based features ranging from shallow (surface form of the post) to deep-level (theoretical and descriptive linguistic analysis) as well as sentiment analysis (Kim et al., 2021). Among them, Ratnayaka et al. (2020) propose multiple role modelling relying on the pre-trained language model BERT aiming at distinguishing posts written by harasser/bystander assistant and victim/bystander defender. Recently, Jacobs, Hee, and Hoste (2022) investigate both supervised and ensemble learning relying on various pre-trained language models (RoBERTa, XLNet, BERTje, RobBERT) to build a multiclass classifier performing fine-grained role classification.
Concerning available resources, most of the proposed studies rely on “easy to access” data (Facebook, YouTube or Instagram) annotated using binary schemes (e.g. content or behaviours categories as abusive or not abusive) (Milosevic, Van Royen, and Davis, 2022). Whilst Van Hee et al. (2015) introduce fine-grained annotation guidelines to improve cyberbullying detection, few datasets were released using a more nuanced description of the type of abuse involved, or the severity of the case; or roles played by those involved. Concerning the latest, only Xu et al. (2012) and Jacobs, Hee, and Hoste (2022) use publicly available datasets providing fine-grained role labelling. Recently, two datasets collected through role-playing games aiming at mimicking cyber aggression situations on private instant messaging platforms in French and Italian were released (Sprugnoli et al., 2018; Ollagnier et al., 2022). These datasets provide a multi-level annotation scheme including the different participant roles, i.e., bullies, victims and bystanders.
Currently, methods addressing the task of PRI occurring on instant messaging platforms is only reported in Ollagnier et al. (2023). In this work, a pipeline aiming at predicting a unique role to chat participants is introduced. Despite promising results, this task remains challenging due to the ambivalence in the use of OH among participants (Haddock and Jimerson, 2017). Indeed, OH can be used with different intentions considering the role of participants in bullying episodes; it can be used as a means to cause harm (harassers), to defend (victims), to encourage (bystanders), and so on. In response to this limitation, we propose to explore OHD methods coupled to a scoring function relying on behavioural cues (i.e., repetition and intention to harm) to perform PRI in multi-party setting. As reported in Ollagnier et al. (2022), behavioural cues from OH usage emerge from statistical evidences; thus, we believe they can contribute to better understand different bullying roles. In addition, a recent study has proven the positive impact of including psychological features to perform PRI (Kao et al., 2019).

Participant Role Identification Pipeline

The proposed pipeline consists of two main tasks: (1) detecting OH in exchanged messages, and (2) assigning a unique role to chat participants. They are detailed below.

Online Hate Detection

OHD is mainly considered as a text classification problem (e.g. distinguishing hate and non-hate speech, identifying hate targets or detecting types of hate). The comprehensive survey conducted in Jahan and Oussalah (2021) reports a plethora of methods aiming at addressing this task, including dictionary look up, bag-of-words and word embeddings. Among existing methods, pre-trained models based on Transformer mechanism have allowed for significant breakthroughs in most of the NLP tasks including OHD (Mozafari, Farahbakhsh, and Crespi, 2019; Gokhale et al., 2022). However, dealing with conversational data, especially in a multi-party setting, these methods have not been much investigated so far. As a part of this work, we explore state-of-the-art pre-trained models to perform OHD on exchanged messages according to a binary scheme (online hate or no online hate). Here, both mono and multi-lingual models are assessed using different learning settings. The lack of models and labelled corpora to address OHD in multi-party setting led us to investigate *mono* and *multi-lingual* learning as well as *zero-shot*, *few-shot* and *cross-lingual* frameworks. The proposed settings are both evaluated on the French and the Italian datasets, respectively introduced in Ollagnier et al. (2022) and Sprugnoli et al. (2018).

The investigated learning settings are: 1) *zero-shot, cross-lingual*, i.e., training on one language and testing on unseen languages; 2) *mono-lingual*, i.e., training and testing on the same language; 3) *few-shot, cross-lingual*, i.e., training on one language and a small percentage of samples from the test language and testing on the test language; 4) *augmented cross-lingual*, i.e., training on several languages and testing on a language included in the training. In detail, as multilingual models the XLM Roberta base model from Conneau et al. (2020) (XLM-Base) and multilingual BERT base (Devlin et al., 2019) (mBERT) are used. As mono-lingual models we use respectively for French and Italian CamemBERT base (Martin et al., 2020) and ITA-Base [1]. In addition, we use the mono-lingual version of DeHateBert (Aluru et al., 2020).

Participant Role Assignment

There is extensive research on understanding individual behaviours and interactions during cyberbullying in various fields, such as psychology and sociology. Some scholars have focused on conceptualising cyberbullying against other hurtful online peer behaviours, while other studies have investigated practices regarding the role of involvement in bullying episodes. Patchin and Hinduja (2015) have defined cyberbullying criteria identifying repetition, intent, harm and imbalance of power as the core elements of bullying and cyberbullying. In Volk, Andrews, and Dane (2022), subcategories of aggressive behaviours are pinpointed as common means used regarding participant roles in bullying episodes; while proactive aggression (instrumental and planned aggression) is significantly related to bullies' practices reactive aggression (reactions after provocation) tends to be associated with victim behaviours.

At this step, we have leveraged these different findings to establish a role scoring function aiming at identifying bullies and victims in multi-party chats. Unlike the work presented in Ollagnier et al. (2023) which derives cues from a PRI classifier, the proposed scoring function relies on behavioural cues based on bullying traces identified using the aforementioned OHD classification strategies. Combining concepts from subcategories of aggressive behaviours and from cyberbullying criteria, we establish two measurements aiming at distinguishing bullies from victims among chat participants. Considering individual behaviours and interactions in chats, measures are defined as follows:

- repetition: the more hurtful comments are attributed to a participant, the more it is considered as being proactive. Here, all exchanged messages in a chat are considered, which allows contrasting with participants reacting after provocation (reactive aggression). An individual ratio is computed for each participant, a high ratio being associated with bully behavioural cues. Conversely, a low one corresponds to victim behaviours experiencing bullying.
- intention to harm: intent and harm criteria are combined to reflect the will to cause harm. Focusing on individual interactions, exchanged messages containing OH and those having any are counted and compared for each participant. The more the interactions of a participant consist of hurtful comments, the more it tends to be the bully. A ratio is computed for each participant corresponding to the balance between passive and aggressive behaviours unveiling the intention to harm peers.

Formally, for a conversation c let U be a set of users, L the set of labels between $[0,1]$ and S 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.
- $Clf : [0,1]^L \rightarrow L$ 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.

Behavioural cues are computed as follows:

$$u \in U, repetition(u) = \frac{|\{s \in S : Clf(D(s)) == 1, s \in speech\_acts(u)\}|}{\sum_{u \in U} |speech\_acts(u)|} \tag{1}$$

$$u \in U, intention(u) = \frac{|\{s \in S : Clf(D(s)) == 1, s \in speech\_acts(u)\}|}{|speech\_acts(u)|} \tag{2}$$

Then, the metric $Role\_scoring : U \rightarrow [0,1]$ such that, for all $u \in U$:

$$Role\_scoring(u) = \frac{repetition(u) + intention(u)}{2}$$

For a conversation c, a user u getting the largest $Role\_scoring$ is identified as the bully, while the user u having the lowest one as the victim.

Data

To evaluate the proposed pipeline on both tasks of OHD and PRI, we use two datasets in French and Italian respectively introduced in Ollagnier et al. (2022) and Sprugnoli et al. (2018). These two datasets contain conversations collected through role-playing games aiming at mimicking cyber aggression situations on private instant messaging platforms. Both of them consist of conversations manually annotated using a multi-level annotation scheme. Table 1 presents a sample of a conversation extracted from the French dataset.

French This resource, named *CyberAgressionAdo-v1 dataset*[2], is composed of 19 conversations including 4 addressing the topic of homophobia, 7 the topic of obesity, 3 the topic of religion and 6 about ethnicity. The conversations provide annotations such as the participant roles, the presence of hate speech, the type of verbal abuse present in the message, and whether utterances use different humour figurative devices (e.g., sarcasm or irony). To evaluate the task of OHD, we have transposed the annotations corresponding to the different types of verbal abuse into a binary scheme (online hate or no online hate), i.e., exchanged messages containing any type of verbal abuse are considered hateful while

[1] https://huggingface.co/dbmdz/bert-base-italian-cased

[2] https://github.com/aollagnier/CyberAgressionAdo-v1

conversation	translation	role
- tg jules	- shut up jules	bully
- jules ftg	- Jules shut up	bully_support
- ok pour le pont	- ok for the bridge	victim_support
- vazy jette toi connard	- go jump off asshole	bully
- nan nan tinquiète brice pas de prob	- nah nah don't worry brice no prob	conciliator
- je vous mange tous	- I eat all of you	victim
- brice jette toi	- brice jump off	bully
- Brice bien jouer pour ton role lorsque tu as fait la boule dans indiana jones	- Brice well done for your role of the ball in indiana jones	bully_support
- nan c'est fort boyard	- nah it's fort boyard	conciliator
- oe on sait que tu mange beacoup	- yeah we all know that you eat a lot	bully
- vous ete pas genti.	- you are not nice.	victim

Table 1: An excerpt from a cyberbullying episode about obesity extract from the *CyberAgressionAdo-v1*

the others are annotated as non-hateful. Concerning the task of bullying participant role detection, performances are assessed using the provided labels corresponding to the participant roles, i.e., bully, victim, bystander-assistant, bystander-defender and conciliator.

Italian This dataset, hereinafter referred to *CyberViolenzaAdo-v1*[3], consists of 10 conversations including 4 about gendered division of sport practices and 2 conversations per cyber aggression situations addressing the topics of interference in others' businesses, lack of independence, parental intromission and web virality. The provided annotation layers correspond to: (i) the cyberbullying role of the message's author; (ii) the cyberbullying type of the expression; (iii) the presence of sarcasm in the expression; (iv) whether the expression containing insults is not really offensive but a joke. As carried out for the *CyberAgressionAdo-v1 dataset*, we have transposed the annotation corresponding to the layer (ii) (i.e., *bullying_entity*) into a binary scheme distinguishing the presence or absence of OH. Here, the labels `Threat or blackmail`, `General Insult`, `Body Shame`, `Sexism`, `Racism`, `Curse or Exclusion`, `Insult Attacking Relatives`, `Defamation` and `Sexual Harassment` are considered hateful while the others are annotated as non-hateful. To evaluate the task of PRI, we use the provided labels corresponding to the annotation layer (i), i.e., bully, victim, bystander-assistant, bystander-defender.

Experimental Methodology

Our experimental setup illustrates four aspects: (1) the performance of the different models and learning frameworks on the task of OHD on conversational data, (2) the qualitative evaluation section in which we use explainability methods to observe model limitations, (3) the performance of the best OHD classifier on the task of PRI, and (4) the error analysis of PRI predictions to assess the reliability of the proposed scoring function.

To evaluate the task of OHD, training sets are composed of messages extracted from the aforementioned datasets, i.e. 2,921 entries in French and 2,074 in Italian. Concerning pre-processing, all the messages are lower-cased and tokenized. In turn, they are respectively truncated/padded to a length of 100 in both languages. Then, they are encoded using the pre-trained model required according to the learning setting. Next, generated sentence vector-based features are used to fine-tune the given pre-trained model to perform OHD according to the used binary scheme. Here, we reproduce the same setting used to perform the task of OHD on Italian introduced in Nozza, Bianchi, and Attanasio (2022). Models are trained for 5 epochs and evaluated every 50 steps, and we select the best checkpoint considering the validation loss. For each dataset, the train consists of 70% of stratified sampling (2336 messages in French and 1451 in Italian) and both the validation and the testing sets represent 15% of the remaining, i.e., 293 messages for both sets in French and 311 messages in Italian. To fine-tune the pre-trained models, we run different experimental frameworks: (1) *mono-lingual* (MONO), in which we fine-tune mono-lingual pre-trained models on the source language and test on the same language; (2) *zero-shot, cross-lingual* (ZERO), on which we fine-tune the best MONO model on each language and test on the unseen language; (3) *few-shot, cross-lingual* (FEW), in which we fine-tune mono-lingual models either on French or Italian and add 20% of the training set of the test language, then testing on the test language accordingly; (4) *augmented cross-lingual* (AUG), in which we fine-tune multi-lingual models using both languages and test on each separately.

Concerning the task of PRI, we have computed the $Role\_scoring$ function on each conversation extracted from the datasets, namely, 19 chats in French and 10 in Italian. The best OHD fine-tuned model trained is used in this evaluation to generate predictions of the bully and the victim over conversations.

Results

OHD Results

Table 2 presents the results obtained on the task of OHD using different pre-trained models and learning frameworks on both the French and the Italian datasets. For both languages, the best performances are observed using the `MONO` and `AUG` settings. In detail, the `AUG` setting combined with `mBERT` and the `MONO` setting combined with `ITA-Base`

[3] `https://dhsite.fbk.eu/2018/09/whatsapp-dataset-on-cyberbullying/`

Setting		CamemBERT			ITA-Base			DeHateBert			XLM-Base			mBERT		
		F1-h	F1-nh	F1	F1-h	F1-nh	F1	F1-h	F1-nh	F1	F1-h	F1-nh	F1	F1-h	F1-nh	F1
FR	**MONO**	0.76	0.85	0.82	-	-	-	0.78	0.85	0.81	-	-	-	-	-	-
	ZERO	-	-	-	0.14	0.77	0.45	0.65	0.75	0.70	-	-	-	-	-	-
	FEW	-	-	-	0.57	0.77	0.67	0.69	0.81	0.75	-	-	-	-	-	-
	AUG	-	-	-	-	-	-	-	-	-	0.76	0.65	0.71	**0.85**	**0.90**	**0.88**
IT	**MONO**	-	-	-	**0.79**	**0.94**	**0.86**	0.67	0.91	0.79	-	-	-	-	-	-
	ZERO	0.21	0.80	0.50	-	-	-	0.46	0.82	0.64	-	-	-	-	-	-
	FEW	0.48	0.87	0.67	-	-	-	0.33	0.86	0.60	-	-	-	-	-	-
	AUG	-	-	-	-	-	-	-	-	-	0.64	0.90	0.77	0.73	0.92	0.82

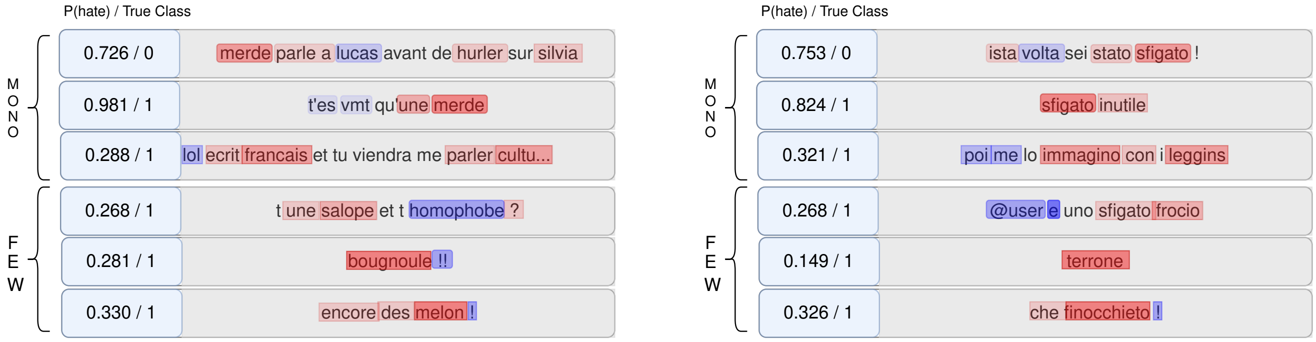
Table 2: Results obtained on the OHD task considering different learning settings. We report F1 score the for **h**ateful and **n**on-**h**ateful cases, and the macro-averaged F1 score.

are the first-ranked models, respectively in French and Italian. Overall, detecting the presence of OH performs better than identifying non-hateful comments, except with the `AUG` setting combined with `XML-base` in French. On average, the difference between the two classes is 8 points in French and 21 in Italian considering the two first-ranked models for each language. For both languages, the poorest performances are observed using the `ZERO` setting combined with `ITA-Base` and `CamemBERT`, respectively in French and Italian. For both languages, results obtained on the hateful class are improved using the `ZERO` setting combined with `DeHateBert`, especially in French. Such a finding suggests that either using data from social media or related to the task of OHD can be considered in the case of low-resource settings. However, it has to be evaluated empirically considering the targeted language; the gap between the F1-h scores of the two monolingual pre-trained models on both languages suggests that transferring knowledge from some languages to others can be confusing regarding the conceptualisation of meaning, idioms, and language structures, to name a few. Moreover, adding samples of the target resources to fine-tune models (the `FEW` setting) improves performances whatever the source of the data used for the original training.

Considering the first-ranked models, post-hoc explanation of the predictions exhibits two main issues: 1) the lack of diversity in hateful and non-hateful content in training and 2) the lack of high-level semantic and pragmatic knowledge in data representation. In both languages, most non-hateful content misclassification includes slur terms and taboo expressions not used in a pejorative way such as *merde* (literally *shit*), *putain* (*fuck*) or *tesoro un corno* (*sweetheart my ass*). As introduced in Technau (2018), there are different modes of use of slurs including hate speech (central use), other pejorative uses (mobbing, insulting), parasitic uses (banter, appropriation, comedy, youth language), neutral mentioning (academics, PC), and unaware uses. Jumping from one mode to another is particularly frequent in social interactions. As a part of cyber aggression situations, hate speech can be identified as the most central, albeit not the most frequent, mode of use. In the datasets, slur terms/taboo expressions are frequently exchanged among friends, e.g. between bullies and their bystanders. Conversely, most hateful content misclassification is due to the absence of slurs/taboo expressions; *La differenza tra @user e martedi ? Il martedì è grasso una sola volta all'anno* (EN: *The difference between @user and Tuesday? Tuesday is only Fat Tuesday once a year*), to go beyond the literal meaning it requires commonsense knowledge (the festival *martedí grasso*) and the bullying topical-focus (obesity). Hate can be conveyed in an insidious way using various victimisation strategies relying on verbal abuse (Hee et al., 2018; Ollagnier et al., 2022). Among them, `denigration` (harsh remarks that are meant to make the person feel bad about themselves) and `aggression-other` (content containing various types of abusive, derogatory language or insults deliberately harmful), respectively represent 12.6% and 11.5% of prediction mistakes. In Italian, 63.3% and 55.5% of messages identified as `Curse_or_Exclusion` (expressions of a wish that some form of adversity or misfortune) and `Threat_or_Blackmail` (expressions containing physical or psychological threats or indications of blackmail) are misclassified. Concerning the `ZERO` and `FEW` learning settings, the main limitations are related to language and/or cultural specificities which are not transferable between languages. Specific concepts/meanings can be expressed using words/expressions which do not have an equivalence in the source language, e.g. *tapette/pédale* to refer to homosexuals in French. Frequent mistakes are also observed when models have to deal with common (taboo) language-specific expressions, e.g. *cazzo* or *merde* respectively meaning *shit* in Italian and French. We have also observed different connotations between words/expressions according to the language leading to misclassification, e.g. *finocchietto* means literally fennel but having a pejorative connotation in Italian but not in French. Figure 1 presents examples of predictions with SHAP illustrating the aforementioned findings.

PRI Results

This evaluation was conducted using the `AUG` setting combined with `mBERT` which is the learning framework reaching the best balance considering all evaluation metrics on both languages. Table 3 presents the performances obtained by our pipeline against the baseline introduced in Ollagnier et al. (2023) on the task of PRI for the roles of `bully`



(a) Samples in French (b) Samples in Italian

Figure 1: Examples of predictions with SHAP (Lundberg and Lee, 2017) contributions on a color scale; color scale: darker red (blue) show higher (lower) contribution to the prediction. Translation available in the appendix.

Setting	FR		IT	
	F1-b	F1-v	F1-b	F1-v
Baseline	0.272	0.0	0.333	0.0
Our pipeline	0.479	0.538	0.333	0.714

Table 3: Results obtained on the PRI task. We report F1 score for the roles of **`bully`** and **`victim`**.

and `victim`. As we can observe, our pipeline significantly outperforms the baseline on most of the labels suggesting that bullying traces are pertinent features to determine participant roles of involvement in bullying. Moreover, our pipeline proves its ability to deal with short and noisy texts (6 tokens on average per message) which is beneficial considering the nature of social media data, especially in multiparty chats. In detail, the detection of `victim` performs better than `bully`, especially in Italian. This gap between the two languages is due to the presence in the French dataset of the role of `conciliator`, i.e., a friend of both the `bully` and the `victim` mediating the disagreement among active participants. For this role, we observe few occurrences of OH considering the whole French conversations, namely 0.013%; the percentage for the victims is 0.048% which suggests similarities in their behavioural practices in facing bullying. For the Italian dataset, reported misclassifications for victims and bullies are between victims and bystander-assistants and victim and bystander-defenders, respectively representing 100% and 85.7% of the model's mistakes. Concerning the French dataset, misclassifications are observed between bullies and bystander-assistants (75%) and between victims and conciliators (72.7%). Observations on the datasets have shown different aggressive tendencies and variations in bullying engagements suggesting that the bullying behavioural cues need to be nuanced to establish contrasts between aggressive behaviours, i.e., the types of aggression and verbal abuse as suggested in Ollagnier et al. (2022) and Sprugnoli et al. (2018). Considering the derived behavioural cues, they contribute to understanding the construction of aggressive behaviours regarding the role of involvement in bullying; for instance, bullies have a ratio of repetition and intention to harm of 0.084 and 0.566 on average against 0.025 and 0.376 for victims in Italian.

Conclusion and Future directions

This paper presents a pipeline aiming at identifying bullies and victims in chats. Consisting of two tasks (OHD and PRI), the proposed pipeline outperforms the existing baseline considering all the datasets and roles to predict.

Despite these promising results, training and testing are carried out on data "mimicking" cyber aggression situations occurring in chats. Whilst role-play data argue to be similar to naturalistic interactions, they can include biased data (e.g. personal student interactions not related to the scenario or extrapolated aggressive behaviours). The proposed pipeline should be tested in *real-world* settings to ensure its reliability. In this study, we have focused on identifying bullies and victims. However, other social roles can be involved and play a key role in bullying episodes such as bystanders and conciliators. All the roles should be considered to fully grasp engagement in bullying of each participant and thus better capture the construction of aggressive behaviours. Technically, we only consider one exchanged message in a row. However, due to the semi-asynchronous and "entangled" nature of the contributions by chat participants, the narrative chain can be held over several messages. Consider strategies aiming at reorganising the structure of events that unfold in the narrative should be investigated to validate the consistency of the obtained performances.

In future work, we plan to address the two latest limitations. We aim to leverage a more nuanced description of the type of OH in order to identify more participant roles and thus derive richer behavioural cues. Moreover, we intend to integrate contextual windows (aggregating multiple messages authored by the same participant) and pragmatic annotation to help in reorganising the narrative chaining.

References

Aluru, S. S.; Mathew, B.; Saha, P.; and Mukherjee, A. 2020. A deep dive into multilingual hate speech classification. In Dong, Y.; Ifrim, G.; Mladenic, D.; Saunders, C.; and Hoecke, S. V., eds., *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part V*, volume 12461 of *Lecture Notes in Computer Science*, 423–439. Springer.

Balakrishnan, V.; Khan, S.; Fernandez, T.; and Arabnia, H. R. 2019. Cyberbullying detection on twitter using big five and dark triad features. *Personality and Individual Differences* 141:252–257.

Bedrosova, M.; Machácková, 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. 2017a. 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.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Cristofaro, E. D.; Stringhini, G.; and Vakali, A. 2017b. Mean birds: Detecting aggression and bullying on twitter. In Fox, P.; McGuinness, D. L.; Poirier, L.; Boldi, P.; and Kinder-Kurlanda, K., eds., *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, 13–22. ACM.

Chinivar, S.; M.S., R.; J.S., A.; and K.R., V. 2023. Online offensive behaviour in socialmedia: Detection approaches, comprehensive review and future directions. *Entertainment Computing* 45:100544.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 8440–8451. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Gokhale, O.; Kane, A.; Patankar, S.; Chavan, T.; and Joshi, R. 2022. Spread love not hate: Undermining the importance of hateful pre-training for hate speech detection. *CoRR* abs/2210.04267.

Haddock, A. D., and Jimerson, S. R. 2017. An examination of differences in moral disengagement and empathy among bullying participant groups. *Journal of Relationships Research* 8.

Hee, C. V.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; Pauw, G. D.; Daelemans, W.; and Hoste, V. 2018. Automatic detection of cyberbullying in social media text. *CoRR* abs/1801.05617.

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.

Jahan, M. S., and Oussalah, M. 2021. A systematic review of hate speech automatic detection using natural language processing. *CoRR* abs/2106.00742.

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.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4765–4774.

Martin, L.; Müller, B.; Suárez, P. J. O.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. Camembert: a tasty french language model. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7203–7219. Association for Computational Linguistics.

Milosevic, T.; Van Royen, K.; and Davis, B. 2022. Artificial intelligence to address cyberbullying, harassment and abuse: New directions in the midst of complexity. In *International Journal of Bullying Prevention*, volume 4.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In Cherifi, H.; Gaito, S.; Mendes, J. F.; Moro, E.; and Rocha, L. M., eds., *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, 928–940. Springer.

Nozza, D.; Bianchi, F.; and Attanasio, G. 2022. HATE-ITA: Hate speech detection in Italian social media text. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 252–260. Seattle, Washington (Hybrid): Association for Computational Linguistics.

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 Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, 867–875. European Language Resources Association.

Ollagnier, A.; Cabrio, E.; Villata, S.; and Tonelli, S. 2023. Birdy: Bullying role detection in multi-party chats. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, February 7 - 14, 2023*. AAAI Press.

Patchin, J. W., and Hinduja, S. 2015. Measuring cyberbullying: Implications for research. *Aggression and Violent Behavior* 23:69–74.

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.

Salminen, J.; Hopf, M.; Chowdhury, S. A.; Jung, S.; Almerekhi, H.; and Jansen, B. J. 2020. Developing an online hate classifier for multiple social media platforms. *Hum. centric Comput. Inf. Sci.* 10:1.

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.

Tahmasbi, N., and Rastegari, E. 2018. A socio-contextual approach in automated detection of public cyberbullying on twitter. *ACM Trans. Soc. Comput.* 1(4):15:1–15:22.

Technau, B. 2018. Going beyond hate speech: The pragmatics of ethnic slur terms. *Lodz Papers in Pragmatics* 14(1):25–43.

Van Hee, C.; Verhoeven, B.; Lefever, E.; De Pauw, G.; Hoste, V.; and Daelemans, W. 2015. Guidelines for the fine-grained analysis of cyberbullying. *Language and Translation Technology Team-Ghent University* LT3 15-01.

Volk, A. A.; Andrews, N. C.; and Dane, A. V. 2022. Balance of power and adolescent aggression. *Psychology of violence* 12(1):31.

Xu, J.; Jun, K.; Zhu, X.; and Bellmore, A. 2012. Learning from bullying traces in social media. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, 656–666. The Association for Computational Linguistics.

Yin, W., and Zubiaga, A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.* 7:e598.

Examples Translation

We provide as literal as possible translations.

- FR: *merde parle a lucas avant de hurler sur silvia*
- EN: *shit talk to lucas before yelling at silvia*

- FR: *t'es vmt qu'une merde*
- EN: *you're such a piece of shit*

- FR: *lol ecrit francais et tu viendra me parler culture ...*
- EN: *lol written French and you will come and talk to me about culture ...*

- FR: *t une salope et t homophobe ?*
- EN: *you're a slut and you're homophobic ?*

- FR: *bougnoule !!*
- EN: *rag-head !!*

- FR: *encore des melon !*
- EN: *again rag-head!*, here *melon* means literally watermelon but refers to Middle-Eastern person in this context.

- IT: *ista volta sei stato sfigato !*
- EN: *what a bad luck !*

- IT: *sfigato inutile*
- EN: *useless loser*

- IT: *poi me lo immagino con i leggings*
- EN: *Then I imagine him wearing leggings*

- IT: *@user e uno sfigato frocio*
- EN: *@user is a loser*

- IT: *terrone*
- EN: *redneck*. literally, it means Southerner.

- IT: *che finocchieto !*
- EN: *what a faggot !*, here *finocchieto* means literally fennel but refers to homosexual person in this context.