Finding Evidence Of The Sexual Predators Behavior

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Sexual predator identification is a critical problem given that the majority of cases of sexually assaulted children have agreed voluntarily to meet with their abuser [10]. Traditionally, a term that is used to describe malicious actions with a potential aim of sexual exploitation or emotional connection with a child is referred to as "*Child Grooming*" or "*Grooming Attack*" [6]. This attack is defined by [4] as "*a communication process by which a perpetrator applies affinity seeking strategies, while simultaneously engaging in sexual desensitization and information acquisition about targeted victims in order to develop relationships that result in need fulfillment*" (*e.g.* physical sexual molestation). Clearly, the detection of a malicious predatory behavior against a child could reduce the number of abused children.

Given the difficulties involved in having access to useful data, *i.e.,* where real pedophiles are involved, nowadays the problem of sexual predator identification through pattern recognition techniques is still a challenging research area. The usual approach to catch sexual predators is by means of police officers or volunteers who behave as fake children in chat rooms and provoke sexual offenders to approach them[1]. Unfortunately, online sexual predators always outnumber the law enforcement officers and volunteers. Therefore, tools that can automatically detect and to evidence sexual predators in chat conversations (or at least serve as a support tool for officers) are highly needed. Recently, different research groups have proposed distinct approaches for anticipating the presence of a predator in a chat, i.e., deciding whether or not a conversation is suspicious, and if so, to point the predator [1, 2, 3, 7, 9]. However, an important aspect of the problem has been left behind, i.e., once the predator is identified, officers need to collect all the necessary evidence for sentencing a pedophile. The later is known as the identification of predatory behavior and implies to detect those lines (interventions within a conversation) that are distinctive of the predatory activities.

Accordingly, in this work we focus on the problem of detecting the predatory behavior. Our main proposal is focused on the representation of the chat interventions, thus we incorporate features that capture content, style, and contextual information. For performing our experiments, we used the only publicly available data set for sexual predator detection [5]. This data set was released in the context of the sexual predator identification task (SPI) at PAN-CLEF'12[2] and comprises a large number of chat conversations that include real sexual predators.

2 Proposed framework and initial experiments

For our performed experiments, we followed a traditional supervised machine learning framework. However, as we previously mentioned, we are mainly focus on proposing a suitable representation for the posed task, namely: content, stylistic, and behavioral features. Thus, for our initial set of experiments we used as content features a traditional *Bag-of-Words* with the 10K most frequent

[1] The American foundation, called Perverted Justice (PJ) (`http://www.perverted-justice.com/`), follows the above mentioned approach.

[2] `https://pan.webis.de/clef12/pan12-web/`

Table 1: Results obtained using three distinct families of features: *content*, *style*, and *behavioral*.

Representation		Classifiers performance NB			SVM			RF		
		P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BoW	1-gram	0.54	0.47	0.50	**0.75**	**0.48**	**0.59**	0.68	0.37	0.48
	2-gram	0.51	0.39	0.44	0.70	0.33	0.45	0.51	0.37	0.43
	3-gram	0.52	0.17	0.26	0.66	0.16	0.26	0.49	0.21	0.30
POS	1-gram	0.29	0.33	0.31	**0.50**	0.02	0.04	0.31	0.14	0.19
	2-gram	0.31	**0.41**	**0.36**	**0.50**	0.01	0.03	0.38	0.18	0.25
	3-gram	0.33	0.37	0.35	0.46	0.11	0.18	0.35	0.18	0.24
LIWC	—	0.30	**0.58**	0.39	**0.69**	0.09	0.16	0.62	0.37	**0.46**

Table 2: Examples of incriminatory and not incriminatory evidence found by our proposed method.

Incriminatory	**Not-incriminatory**
» i'd be so excited with u i'd probably cum just touchin u	» do i have anything to be jealous about?
» you like that I'd do nasty things to your young little body	» i cant beelieve that i am nervous abt tonmorrow
» i will wear condom for you	» If u were here we would not be worrying about internet either baby

features. As for the stylistic features, we considered as features the 36 POS tags contained in the TreeTagger[3] part-of-speech tagger. Finally, as contextual features we account the 68 LIWC [8] psychologically meaningful categories. The LIWC representation provides richer information regarding the words contained in a text, therefore gives context. For example, the word 'cried' matches with four word categories: sadness, negative emotion, overall affect, and a past tense verb.

For training our evidence detection model we used the test partition of the corpus described in [5][4]. In the test partition, a total of 3,737 conversations contain at least one sexual predator[5], and within these conversations, predators interventions are labeled as *incriminatory* or *not-incriminatory*. In order to perform our training, we firstly filtered the 3,737 conversations as done in [9], resulting in a total of 1,466 conversations containing full conversations between victims and a predators. Then, from the filtered version of the corpus we preserve the predator's interventions, giving a total of 59,410 interventions, where 6,395 (11%) are *incriminatory*, and 53,015 (89%) are *not-incriminatory*. As can be noticed, a highly unbalanced problem. Thus, to evaluate the classification performance (using three well know learning algorithms: Naive Bayes, Support Vector Machines and Random Forest) we used precision, recall and the F-score metric of the positive class (i.e., *incriminatory*), and for all experiments we employ a stratified 10 fold cross validation technique to compute the performance.

We observe from Table 1, the best performance ($F = 0.59$) is obtained by the SVM classifier when BoW (*content*) features are used, with $n = 1$ for the *n-gram* size. With respect to the *style* features, the best result was obtained when POS 2-grams are used as features with the NB classifier. As for the *contextual* features, we notice that is not possible to obtain a good performance in terms of F; however, the NB classifier obtains a very high recall level ($R = 0.58$). According to [5], having lot of relevant incriminatory lines, augments the possibility of finding good evidences towards a suspect. Thus, during SPI task at CLEF'12, organizers proposed using the F measure with the β factor equal to 3, hence emphasizing recall. Consequently, our best configuration so far is the one generated by the BoW (1-gram) representation with the SVM classifier, which obtains an $F_{(\beta=3)} = 0.4979$; outperforming the best result reported during CLEF'12 $F_{(\beta=3)} = 0.4762$. Table 2 shows a few examples of the type of evidence we are able to obtain with our proposed method.

As future work, we plan to evaluate fusion methods in order to exploit the best from every family of features. Additionally, we are interested in evaluating the performance of representing the information using word embedding strategies.

[3] `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[4] The training partition is not labeled with the incriminatory lines.

[5] The total number of conversation on the test partition is near 155K.

References

[1] S. G. Burdisso, M. Errecalde, and M. Montes-y Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019.

[2] C. Cardei and T. Rebedea. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, 23(4):589–616, 2017.

[3] H. J. Escalante, E. Villatoro-Tello, S. E. Garza, A. P. López-Monroy, M. Montes-y Gómez, and L. Villaseñor-Pineda. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111, 2017.

[4] C. Harms. Grooming: An operational definition and coding scheme. *Sex Offender Law Report*, 8(1):1–6, 2007.

[5] G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[6] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, 2008.

[7] A. P. López-Monroy, F. A. González, and T. Solorio. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, page 112909, 2019.

[8] J. W. Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

[9] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, and L. V. Pineda. A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178, 2012.

[10] J. Wolak, K. J. Mitchell, and D. Finkelhor. Online victimization of youth: Five years later. 2006.