Uma Abordagem para Detecção Precoce de Predadores Sexuais em Conversas Virtuais

Marcelle R. Panzariello[1], **Geraldo Xexéo**[1,2]

[1] Prog. de Eng. de Sistemas e Computação / COPPE
Universidade Federal do Rio de Janeiro – Rio de Janeiro – RJ – Brasil

[2]Dep. de Ciência da Computação / IM
Universidade Federal do Rio de Janeiro – Rio de Janeiro – RJ – Brasil

{panzariello,xexeo}@cos.ufrj.br

Abstract. *This work proposes the development of three distinct strategies for early detection of sexual predators in internet conversations between two people. Conversations from the PAN 2012 database and algorithms such as SVM, Random Forest, KNN and BERT will be used. It is expected to overcome the state of the art and prove that the "Distinguish Predatory and General Conversations" strategy can achieve better results than the others.*

Resumo. *Este trabalho propõe o desenvolvimento de três estratégias distintas para detectar precocemente predadores sexuais em conversas realizadas na internet entre duas pessoas. Serão utilizadas conversas da base de dados do PAN 2012 e algoritmos como SVM, Floresta Aleatória, KNN e BERT. Espera-se superar o estado da arte e provar que a estratégia "Distinguir Conversas Predatórias e Gerais" pode conseguir melhores resultados que as demais.*

1. Introdução

Este trabalho desenvolve três estratégias para identificar precocemente predadores sexuais em conversas realizadas na internet entre duas pessoas: distinguir conversas predatórias e gerais; distinguir predador e vítima; e um classificador combinando as anteriores em sequência.

Como contribuição direta para a sociedade, buscamos apoiar ações de redução do número de casos de abuso sexual infantil. Segundo [Olson et al. 2007], o predador sexual executa certo padrão comportamental para abordar suas vítimas, antes de conseguir contato sexual. Assim, o objetivo é identificar o mais cedo possível que se trata de uma conversa envolvendo um pedófilo, para evitar que o contato sexual ocorra.

2. Fundamentação Teórica

Identificar predadores sexuais é um problema de classificação de textos [Pendar 2007]. Devido à natureza delicada do problema, dados de conversas entre predadores sexuais e vítimas, ou oficiais da lei atuando como vítima, são difíceis de serem encontrados [Pendar 2007]. Por conta disso, o site Perverted Justice (PJ) é uma das principais fontes de dados em problemas de identificação de predadores sexuais, pois contém conversas reais entre predadores e voluntários se passando por possíveis vítimas. Um dos primeiros

trabalhos da área, [Pendar 2007] conseguiu $F_1 = 94.3\%$, com o algoritmo KNN, ao identificar quem é a vítima e quem é o predador sexual em uma única conversa. Ele utilizou 701 conversas retiradas do site PJ.

Em 2012, a conferência CLEF lançou, na competição PAN, o desafio *Sexual Predador Identification* . Os dados disponibilizados e os resultados obtidos pelos competidores se tornaram referências na área [Inches and Crestani 2012], e são usados como base deste artigo.

[Villatoro-Tello et al. 2012] conseguiram o melhor resultado para a primeira tarefa ($F_{0.5}$ = 93.46%) do PAN 2012, com um classificador de dois estágios. O primeiro estágio classificava as conversas em suspeitas e gerais, e apenas as conversas suspeitas eram submetidas ao segundo classificador, que identificava quem é a vítima e quem é o predador. Esse modelo de classificador se tornou prática comum na literatura [Fauzi and Bours 2020].

Em 2017, [Escalante et al. 2017] propuseram uma nova abordagem, direcionada ao problema de detectar precocemente os predadores sexuais, conseguindo resultados satisfatórios utilizando *profile-based representations* e uma Rede Neural: com 50% das mensagens lidas o algoritmo foi capaz de obter quase 60% de F_1-*score*. Este foi, provavelmente, o primeiro trabalho nesta área.

Em 2019, [dos Santos and Guedes 2019] realizaram o possível primeiro trabalho de classificação utilizando dados verídicos de predadores sexuais brasileiros. Os autores utilizaram Redes Neurais Convolucionais para classificar conversas em culpadas e não culpadas.

Em um trabalho mais recente, [Kulsrud 2019] optou por utilizar a estratégia do classificador em dois estágios usando SVM para detecção precoce de predadores sexuais e realizou testes com diferentes tamanhos de conversas. Com 24 mensagens já foi possível obter $F_{0.5} = 80\%$.

Na literatura encontramos diversas abordagens para identificar predadores sexuais, mas poucos são os trabalhos que tratam da detecção precoce. O trabalho de [Kulsrud 2019] apresentou o desenvolvimento de três estratégias, semelhantes as propostas neste artigo, porém voltadas apenas à identificação predadores sexuais. Em seu trabalho, apenas a estratégia de dois classificadores foi utilizada para detecção precoce.

Este artigo propõe que as três estratégias sejam utilizadas para detecção precoce de predadores sexuais. Com isso, também acreditamos que a estratégia "Distinguir Conversas Predatórias e Gerais", que contém apenas um classificador, consiga obter melhores resultados, pois presumimos não haver necessidade de identificar qual autor é o predador sexual, uma vez que já sabemos que estamos tratando de uma conversa predatória.

3. Metodologia de Pesquisa

Essa é uma pesquisa de natureza aplicada, de abordagem quantitativa, com procedimentos técnicos experimentais. Foi iniciada com uma revisão narrativa da literatura, a partir de buscas nas bases de dados *Scopus*, IEEE e *Web of Science* entre janeiro de 2012 e abril de 2021. Usa, em seus experimentos, bases padronizadas na literatura, e compara seus resultados com o estado da arte da solução do problema.

4. Experimentos

Este trabalho propõe três estratégias para detectar precocemente predadores sexuais: I) distinguir conversas predatórias e gerais; II) distinguir predador e vítima; e III) combinar as duas estratégias anteriores em sequência. Como fonte de dados, estamos utilizando a base do PAN 2012, visto sua diversidade de conversas, grande quantidade de dados e boa aceitação na literatura.

O objetivo da primeira tarefa é identificar quais conversas são predatórias. Para isso, são consideradas conversas predatórias aquelas em que um dos autores é um predador sexual. A segunda estratégia consiste em identificar qual autor é o predador sexual e qual autor é a vítima, dentro de uma mesma conversa. Por fim, a terceira estratégia consiste em uma combinação das duas estratégias anteriores, onde apenas as conversas predatórias são passadas para o segundo classificador distinguir quem é o predador e quem é a vítima.

Em tarefas gerais de identificação de predadores sexuais, conversas inteiras são utilizadas em conjunto com algoritmos de classificação de textos para identificar se existe um predador sexual envolvido. Em detecção precoce de predadores sexuais, trabalhamos com partes das conversas para compreender se os resultados são satisfatórios.

Assim, estão sendo feitos experimentos com as três estratégias utilizando diversos tamanhos de conversas. Por exemplo, com as primeiras 5 mensagens das conversas, depois as 10 primeiras, e assim por adiante.

Embora a base não forneça rótulos para conversas predatórias, é possível obtê-los uma vez que são disponibilizados quem são os autores predadores sexuais. Assim, foram rotuladas como conversas predatórias as conversas em que um dos autores era um predador sexual.

Todos os experimentos utilizam *bag of words* e algoritmos *Naive Bayes*, *Random Forest*, KNN e SVM, e a estratégia *k-fold cross-validation*, onde $k = 10$. Também, na etapa de pré-processamento, foram excluídas todas as conversas que não envolviam apenas 2 autores. Esta estratégia já foi utilizada por [Kulsrud 2019] e é semelhante a utilizada por [Villatoro-Tello et al. 2012], que excluiu somente conversas que tinham apenas 1 participante.

Na Tabela 1 é possível observar alguns dos resultados obtidos para a estratégia "Distinguir Conversas Predatórias e Conversas Gerais" com apenas as 24 primeiras mensagens das conversas. Com o algoritmo SVM foi possível obter um resultado comparável ao resultado de [Kulsrud 2019].

Tabela 1. Resultados obtidos para a estratégia "Distinguir Conversas Predatórias e Conversas Gerais" utilizando apenas as 24 primeiras mensagens

Algoritmo	**Acurácia**	**Precisão**	**Recall**	F_1	$F_{0.5}$
Naive Bayes	0.9593	0.3154	0.6066	0.4148	0.3488
Random Forest	0.9785	0.9917	0.0956	0.1738	0.3416
KNN	0.9660	0.2870	0.2870	0.2870	0.2870
SVM	0.9934	0.8665	0.8539	0.8597	0.8637

Pretendemos investigar também novos algoritmos de Redes Neurais Profundas

ainda não usados na literatura neste contexto. Em especial, partiremos do algoritmo BERT, como referência inicial para as RNP, e investigaremos o uso de algoritmos posteriores.

5. Conclusão

Este artigo trata da detecção precoce de predadores sexuais, um tema relativamente novo e pouco explorado ainda. Embora existam dificuldades, como a falta de conversas verídicas públicas entre predadores sexuais e vítimas, esperamos com a base do PAN 2012, que contém dados do PJ, obter resultados que possam ser aplicados em bases de dados reais. Com o desenvolvimento das estratégias mencionadas ao longo do artigo, esperamos superar o estado da arte e provar que a estratégia "Distinguir Conversas Predatórias e Conversas Gerais" pode conseguir melhores resultados que as demais.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

dos Santos, L. and Guedes, G. (2019). Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na internet. In *Brazilian Workshop on Social Network Analysis and Mining*, pages 143–154.

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

Fauzi, M. A. and Bours, P. (2020). Ensemble method for sexual predators identification in online chats. In *2020 8th International Workshop on Biometrics and Forensics, IWBF 2020 - Proceedings*.

Inches, G. and Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. In *CLEF*.

Kulsrud, H. (2019). Detection of cyber grooming during an online conversation. Master's thesis, Norwegian University of Science and Technology.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. K. K. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241.

Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., y Gómez, M. M., and Villaseñor-Pineda, L. (2012). A two-step approach for effective detection of misbehaving users in chats. In *CLEF*.