Um modelo computacional para identificação de assédio em salas de bate-papo

João Paulo de Oliveira Ramos [1], Guilherme Alberto Wachs Lopes[2]
[1,2] Ciência da Computação, Centro Universitário da FEI
unifjramos@fei.edu.br
gwachs@fei.edu.br

Resumo: *É crescente a ocorrência de crimes de assédio infantil em Salas de Bate-Papo na internet. De acordo com o National Centre for Missing and Exploited Children (NCMEC), esse problema tem ocorrido entre 1 a cada 7 crianças expostas na rede* NCMEC [1]. *Na literatura, existem propostas, inclusive competições, para a detecção de conteúdo pedófilo em conversas textuais PAN2012 [2]. Esse trabalho faz o estudo de conversas de sala de bate papo com a intenção de identificar um se o locutor de uma mensagem pode ser classificado como pedofilo ou não.*

1. *Introdução*

É incontestável que a Internet trouxe novas formas de se comunicar. Voz por ip, e-mail, aplicativos móveis, fóruns, são algumas dessas inovações. Em particular, as salas de bate-papo online também têm tido seu período de destaque e, até hoje, milhares de salas são criadas por dia para atender os mais diversos públicos. Além disso, aplicativos de conversa, tais como Whatsapp, Messenger e Telegram, têm experimentado uma de suas maiores demandas. Contudo, um problema grave emerge dessa inovação: são os crimes online.

Atualmente, de acordo com relatórios do *National Centre for Missing and Exploited Children (NCMEC)***,** 1 em cada 7 crianças é aliciada ao sexo, 1 em cada 33 recebe ataques verbais para um encontro pessoal e 1 em cada 3 recebe material pornográfico sem solicitar NCMEC[1].

2. Trabalhos Relacionados

Na literatura, alguns trabalhos concretizaram a tentativa de alcançar a resolução desse problema. No trabalho realizado em GUPTA [3], foram analisados os conteúdos de conversas com conteúdo pedófilo de forma teórica, onde executou experimentos linguísticos para análise das conversas. Esse trabalho teve uma abordagem focada em entender mais sobre como as conversas de pedofilia avançam durante um ataque.

Um ponto importante é encontrar quais são as palavras chaves utilizadas por um pedófilo em sua abordagem, BOGDANOVA [4] utilizou Características Semânticas em sua análise, onde é indicado que o campo do sentimento deve ser mais explorado para uma detecção mais precisa dos textos.

3. Word2Vec

Word2Vec é um modelo para representação de palavras e suas semânticas em um espaço multidimensional MIKOLOV [5] . Nesse espaço, palavras com o mesmo significado ficam dispostas espacialmente próximas e palavras semanticamente diferente, em localizações opostas. A estratégia de representação de palavras na forma de um vetor também é chamada de *Word Embedding*. A Figura 1 ilustra a relação semântica entre quatro palavras. Note que a direção de "*man*" para "*woman*" é a mesma que de "*king*" para "*queen*".

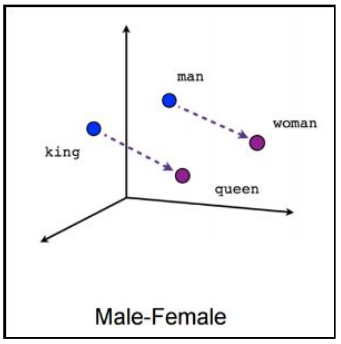

Figura 1 – Representação de contexto no espaço.

4. Metodologia

Para a execução desse projeto, foram utilizadas técnicas para tratamento e criação de um modelo semântico utilizando os conceitos do *Word2Vec*. Em seguida, foi desenvolvido um algoritmo classificador, que baseia-se nas distâncias das palavras no espaço *n*-dimensional. A Figura 2 apresenta as etapas da metodologia proposta.

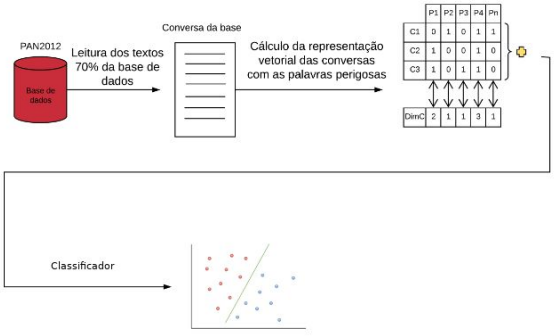

Figura 2 – Etapas da metodologia

4.1. Word2Vec

Após o treinamento do *Word2Vec*, as palavras são distribuídas no espaço, com seus contextos segmentos em um aglomerado de palavras.

Em seguida, é feita a classificação dos textos, que consiste na utilização das palavras já classificadas em

seus determinados contextos. Para a classificação dessas palavras, foram utilizados 2 sufixos "0"e "1", sendo "1" uma palavra encontrada no contexto de pedofilia e "0" uma palavra encontrada no contexto de não pedofilia.

4.2 Base de dados

A base de dados consiste de 60.000 conversas em português, extraídas de salas de bate papo na competição PAN2012, um desafio para a identificação de autores de textos com conteúdo pedófilo.

5. Resultados Parciais

Com as palavras treinadas, foi possível obter uma melhor visão de como as palavras foram distribuídas pelo modelo *Word2Vec* no espaço.

As Figuras 3 e 4 demonstram como o modelo se comportou na distribuição das palavras no espaço.

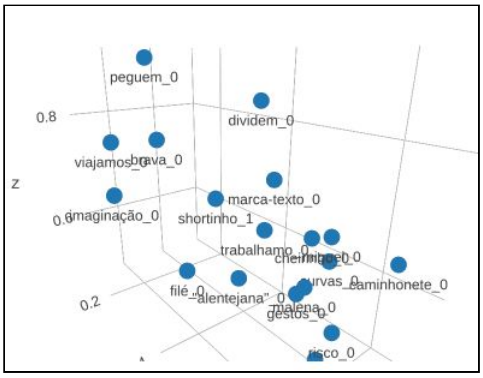

Figura 3 - 20 palavras que estão em torno da palavra "shortinho_1"

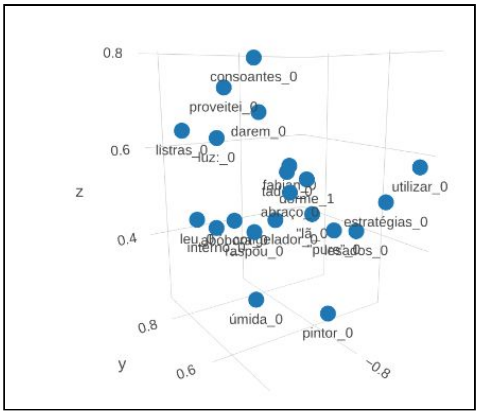

Figura 4 - 20 palavras que estão em torno da palavra "abraço_0"

Outra perspectiva gerada foi a distribuição das palavras utilizadas na etapa de treinamento.

As palavras em azul, são aquelas encontradas apenas no contexto de não pedofilia, as vermelhas as encontradas apenas no contexto de pedofilia e por final, as verdes, as quais são as palavras encontradas em ambos os contextos. A Figura 5 demonstra essa distribuição.

Até então, utilizando a métrica *F1 Score*, foi possível obter 46% de precisão na classificação. Possivelmente, o que motivou este resultado foi a falta na variação das amostras de treinamento e teste do modelo.

Os próximos passos desse projeto serão: a) variar as amostras de treinamento e b) testar e utilizar mais métricas de avaliação.

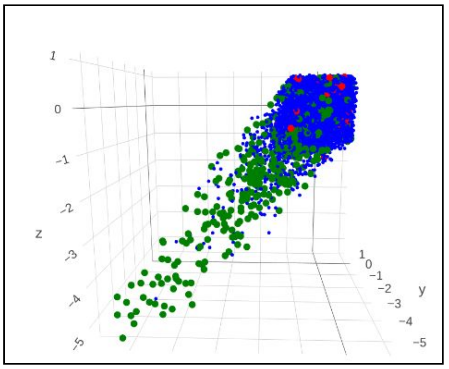

Figura 5 - Distribuição das palavras utilizadas para treinamento

6. Conclusão

Com o crescimento da utilização cada vez maior dos meios de comunicação digital, e muitas vezes essa utilização sendo feita por crianças e adolescentes, faz se necessário a elaboração de modelos que possam garantir a segurança desse indivíduo. Com os resultados obtidos até agora, é possível identificar que esse projeto está seguindo o caminho correto para o seu sucesso.

7. Referências

[1] NCMEC, National center for missing and exploited children, 2008, http://www.missingkids.com/missingkids/servlet/NewsEventServlet?LanguageCountry=en US&PageId=4303.

[2] PAN2012: Sexual Predator Identification. [S. l.], 2012. Disponível em: https://pan.webis.de/clef12/pan12-web/author-identification.html. Acesso em: 5 out. 2019.

[3] GUPTA, Aditi; KUMARAGURU, Ponnurangam; SUREKA, Ashish. Characterizing Pedophile Conversations on the Internet using Online Grooming. CoRR, [S. l.], 2012.

[4] BOGDANOVA, Dasha; ROSSO, Paolo; SOLORIO, Thamar. Exploring high-level features for detecting cyberpedophilia. Computer Speech & Language, [S. l.], v. 28, 2014.

[5] MIKOLOV, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Agradecimentos

À instituição Centro Universitário da FEI, por prover a iniciação científica, e ao meu orientador por todo suporte.

[1] Aluno de IC do Centro Universitário da FEI . Projeto PBIC082/18 com vigência de 08/18 à 07/19.