Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na Internet

Leonardo Ferreira dos Santos[1], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracana, 229 - Rio de Janeiro - RJ - Brasil.

leonardo.santos@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *Nowadays, social applications represent one of the main threats to children and adolescents on the Internet. Among the various existent risks is the presence of sexual predators that seek, among the most diverse purposes, to obtain child pornographic content, to extort for financial purposes and, in more severe scenarios, the sexual abuse. The present work aims to identify Brazilian sexual predators through Convolutional Neural Networks. In order to achieve this goal, it is considered conversations coming from criminal evidence that recently became publicly available. Preliminary results consolidate the presented methodology as an alternative for tasks of binary classification of texts for the Portuguese language.*

Resumo. *Nos dias de hoje, aplicações sociais representam uma dos principais ameaças a crianças e adolescentes na Internet. Dentre os diversos riscos existentes destaca-se a presença de predadores sexuais que buscam, entre os mais diversos fins, obter conteúdo pornográfico infantil, extorquir com finalidades financeiras e, em cenários mais graves, o abuso sexual. O presente trabalho tem como objetivo identificar predadores sexuais brasileiros por meio de Redes Neurais Convolucionais. Para atingir esse objetivo, são consideradas conversas provenientes de provas criminais disponibilizadas recentemente ao público. Resultados preliminares consolidam a metodologia apresentada como uma alternativa para tarefas de classificação binária de textos para a língua portuguesa.*

1. Introdução

Hoje em dia, o uso constante de aplicações sociais na Internet representa um grande risco a crianças e adolescentes. Essas aplicações, acessadas principalmente por meio de *smartphones* e *tablets*, ampliam a exposição a riscos inerentes ao meio como, por exemplo, o *bullying*, conteúdo impróprio, *sexting*, comunicação com pessoas desconhecidas e encontros fora das redes sociais. Dentre os riscos destacados, o que levanta uma maior preocupação é a possibilidade de conhecer novas pessoas e a ocorrência de um abuso sexual.

Nesse contexto, predadores sexuais agem de forma a ludibriar crianças e adolescentes [Olson et al., 2007] com os mais diversos propósitos [NCMEC, 2017]: em 78% dos contatos, o principal objetivo é a obtenção de conteúdo pornográfico da vítima, enviado por aplicativos de conversas instantâneas. Em 7% dos contatos, além da obtenção de conteúdo pornográfico, também ocorre a extorsão, esta se caracterizando por pedidos

de transferências financeiras, disponibilização de dados de cartão de crédito dos pais e pertences. Por fim, 5% buscam diretamente o contato pessoalmente com a vítima com a finalidade de consumar o abuso sexual.

O perfil de acesso de crianças e adolescentes brasileiros à Internet reflete um cenário preocupante quanto aos riscos supracitados. Em 2017, 85% das crianças e adolescentes com idades entre 9 e 17 anos tinham acesso à Internet no Brasil [BARBOSA, 2018]. Destes, 44% usavam somente dispositivos móveis e são majoritariamente pertencentes às classes A e B (76%) [BARBOSA, 2018].

A possibilidade de contato entre um predador sexual e uma criança ou adolescente é uma preocupação global [Olowu, 2014; Dorasamy et al., 2018; Kloess et al., 2019], sendo o domínio de pesquisa de diversos trabalhos realizados nas últimas décadas e intensificada pela ocorrência da competição PAN-2012[1], que disponibilizou um conjunto de dados com conversas realizadas em várias comunidades virtuais. Dentre elas, conversas entre predadores sexuais convictos e agentes da organização *Perverted-Justice* (PJ)[2]. Esse conjunto de dados permitiu a identificação de diversas características do perfil do predador sexual, dentre elas: psicológicas [dos Santos and Guedes, 2018], comportamentais [Bogdanova et al., 2014] e textuais [Kontostathis et al., 2012], o que permitiu a criação de diversos algoritmos de aprendizado de máquina para a identificação automática de predadores sexuais na Internet.

Uma limitação recorrente nas pesquisas para identificação automática de predadores sexuais na Internet é a carência de dados reais, isto é, conversas que ocorreram entre predadores sexuais e vítimas, sendo as vítimas crianças ou adolescentes. Nesse cenário, o principal impedimento é o teor sensível dos dados e por serem provas de processos que correm em sigilo na justiça. Por não serem disponibilizados como artefatos resultantes da pesquisa, acabam prejudicando a reprodutibilidade e a sua evolução.

Para o presente trabalho, consideramos o conjunto de dados preliminar disponibilizado pelo Ministério Público Federal de São Paulo (MPF-SP) em parceria com o Centro Universitário da Fundação Educacional Inaciana (FEI) para realizar uma análise textual e reprodução da primeira proposta para detecção de predadores sexuais na Internet por meio de Redes Neurais Convolucionais (CNN) [Ebrahimi et al., 2016]. Ao melhor do nosso conhecimento, ainda não foram realizados trabalhos para identificação de predadores sexuais brasileiros por meio de conversas com crianças e adolescentes na Internet.

As demais seções estão dispostas da seguinte maneira: na seção 2, são levantados trabalhos semelhantes e relacionados ao tema; na seção 3, é apresentado o conjunto de dados; na seção 4 é discutida a comunicação entre predadores sexuais com crianças e adolescentes; na seção 5 é reproduzido o primeiro experimento considerando CNN para o domínio da pesquisa com o conjunto de dados apresentado e, por último, a seção 6 apresenta as conclusões, limitações e discussão sobre trabalhos futuros.

2. Trabalhos Relacionados

Uma abordagem típica para a identificação de predadores sexuais na Internet é considerá-la uma tarefa de classificação de textos [Villatoro-Tello et al., 2012]. Sendo assim, ao

[1] `https://pan.webis.de/clef12/pan12-web/author-identification.html`
[2] `https://www.perverted-justice.com`

longo das duas últimas décadas foram consideradas diversas abordagens. Em 2007, foi realizado um estudo piloto [Pendar, 2007] considerando as conversas registradas no site PJ para a criação de um conjunto de dados. Para a identificação dos predadores sexuais é aplicado *Bag of Words* (BOW) para geração das características, considerando unigramas, bigramas e trigramas. São considerados os algoritmos de classificação *k-nearest neighbors* (k-NN) e *Support Vector Machines* (SVM) nos experimentos. A aplicação do algoritmo k-NN considerando k = 30, obteve melhores resultados (F_1 = 94%), próximo dos resultados obtidos com o algoritmo SVM (F_1 = 90%).

No primeiro trabalho a fazer uso de características comportamentais que se tem conhecimento [Kontostathis, 2009], é considerado o uso da *Luring Communication Theory* (LCT) [Olson et al., 2007] e técnicas de mineração de texto. Foram consideradas 288 conversas extraídas do site PJ em que para cada etapa do processo de aliciamento da vítima descrito em LCT foi criado um léxico com os termos que melhor caracterizam a ação predatória. No primeiro experimento, foram consideradas 16 conversas, separando as frases enviadas por predadores e vítimas para então verificar a frequência de cada etapa da LCT e para a tarefa de classificação foi usado o algoritmo *J48*. Como resultado, foram obtidos resultados promissores (F_1 = 60%). No segundo experimento, foram consideradas 16 conversas extraídas do conjunto de dados da PJ e 16 conversas extraídas do ChatTracker [Bengel et al., 2004]. Após pré-processamento descrito no primeiro experimento, os dados foram submetidos a uma árvore de decisão *C4.5*, sendo obtidos resultados expressivos (F_1 = 93%).

Villatoro-Tello et al. [2012] apresentam uma abordagem para identificação de predadores sexuais composta por dois estágios. O primeiro estágio tem como objetivo identificar uma conversa predatória enquanto no segundo estágio buscou-se diferenciar o predador sexual dos demais participantes. Nenhum pré-processamento foi realizado, apenas uma filtragem das conversas seguindo alguns critérios que melhor caracterizam um cenário de tentativa de abuso sexual na Internet: 1) Conversas com apenas um participante; 2) Conversas com menos de 6 mensagens; 3) Conversas com sequencias longas de caracteres sem um significante aparente. Aplicando esses 3 critérios, foi possível reduzir em 90% a quantidade de conversas e usuários únicos. Para a classificação, foi considerada a criação de um modelo BOW e representação binária para determinar a relevância dos termos filtrados. No segundo estágio, foram empregados Redes Neurais Perceptron com Múltiplas Camadas (NN-MLP) e representação binária. Os resultados obtidos (*acurácia* = 98%, F_1 = 87%, $F_{0.5}$ = 93,4%) garantiram o primeiro lugar na competição PAN-2012.

A combinação de características psicolinguísticas e comportamentais tem apresentado resultados significativos. Bogdanova et al. [2014] define uma série de características de alto nível com base em 3 conjuntos de dados: conversas registradas no site PJ, conversas normais presentes no NPSChat [Forsythand and Martell, 2007] e conversas de teor sexual obtidas na Internet[3]. O trabalho usa o classificador SVM e exalta a diferença entre os resultados (*acurácia* = 97%) quando comparado com os resultados que usam apenas características de baixo nível (*acurácia* = 64%).

Recentemente, Cardei and Rebedea [2017] aplica estratégia similar à Villatoro-Tello et al. [2012] adotando como estratégia um classificador em dois estágios. O pri-

[3] `http://web.archive.org/web/20040728084602/http://www.geocities.com/urgrl21f/`

meiro estágio considera características textuais e comportamentais em conjunto com o classificador SVM para identificar conversas suspeitas de serem predatórias. O segundo estágio considera o mesmo conjunto de características previamente citado, porém somente as conversas que foram marcadas como suspeitas. Por conta do número reduzido de características no segundo estágio, *Random Forests* (RD) apresentaram melhor desempenho que o classificador SVM. Após análise dos experimentos, foi possível identificar que as características textuais apresentaram maior relevância para a identificação de conversas suspeitas enquanto as características comportamentais para a identificação do predador sexual. Os resultados encontrados (*acurácia* = 100%,*abrangência* = 81.8%, $F_{0.5}$ = 95.7%) superam os encontrados na competição PAN-2012.

Por fim, recentemente foi realizada uma revisão de literatura em Ngejane et al. [2018]. Nessa revisão, foram considerados todos os trabalhos com melhores resultados, considerando as métricas *acurácia* e F_1. Nesse contexto, destaca-se o primeiro trabalho fazendo uso de CNN [Ebrahimi et al., 2016], inspirado em uma abordagem que busca considerar a ordem das palavras para a tarefa de classificação [Johnson and Zhang, 2015].

O presente trabalho apresenta duas principais contribuições: a disponibilização do conjunto de dados na língua portuguesa no formato usado na competição PAN-2012 e reprodução da metodologia proposta em Ebrahimi et al. [2016] em experimentos considerando o conjunto de dados.

3. Conjuntos de Dados

Para o presente trabalho, foram analisadas conversas com predadores sexuais ocorridas no Brasil. A disponibilização dos dados[4] foi possível devido uma parceria entre MPF-SP e FEI [Andrijauskas et al., 2017]. A Tabela 1 apresenta as características do conjunto de dados.

Tabela 1. Características do conjunto de dados.

Classe	# Conversas	# Usuários únicos
Predatória	*39*	*78*
Não predatória	*137*	*274*

A fim de viabilizar a divulgação dos dados para o público, é observado que os dados passaram por pré-processamento visando omitir qualquer informação que permita identificar tanto o predador sexual quando a vítima. No lugar, foram inseridas marcações específicas para caracterizar o teor da informação trocada. Desta forma, a Tabela 2 apresenta todas as marcações presentes no conjunto de dados.

Conforme observado na Figura 1, os dados estão apresentados de forma similar ao formato usado na competição PAN-2012 [Inches and Crestani, 2012]. No entanto, algumas informações não foram disponibilizadas como, por exemplo, o horário de envio das mensagens e o arquivo com a lista de identificadores dos predadores sexuais presentes nas mensagens. Nas próximas seções são apresentados o pré-processamento e a análise feita para criação do arquivo de identificadores.

[4] `https://github.com/Andrijauskas/Datasets-Conversas`

Tabela 2. Marcações para preservação de identidade de predadores sexuais e vítimas.

Marcação	Teor da informação
>audio<	Mensagens de áudio enviadas e recebidas
>emoticon<	Emojis somente (Emoticons textuais foram mantidos)
>foto<	Imagens enviadas e recebidas
>local<	Nomes de Cidade, Estado, País ou Nacionalidade
>nome<	Nomes ou apelidos que caracterizem alguma das partes
>telefone<	Números telefônicos

```
<conversa id="1">
  <linha num="1">
    <autor>709916bfe16ef8cdd6102dc5453f302f</autor>
    <mensagem>Voce deita com migo na cama pelada</mensagem>
  </linha>
  <linha num="2">
    <autor>13f27f55ef3622f4e987aac6a57b1ce8</autor>
    <mensagem>Nao posso durmi ai</mensagem>
  </linha>
  <linha num="3">
    <autor>709916bfe16ef8cdd6102dc5453f302f</autor>
    <mensagem>Nao e pra dumir e so fica com migo ate as 3 horas da tardi</mensagem>
  </linha>
  <linha num="4">
    <autor>709916bfe16ef8cdd6102dc5453f302f</autor>
    <mensagem>Ai eu levo voce em bora</mensagem>
  </linha>
  <linha num="5">
    <autor>13f27f55ef3622f4e987aac6a57b1ce8</autor>
    <mensagem>Tah</mensagem>
  </linha>
</conversa>
```

Figura 1. Exemplo de conversa culpada disponibilizada pelo MPF-SP e FEI.

3.1. Pré-processamento

Entende-se que o formato proposto, estudado e difundido desde 2012, facilitará estudos comparativos assim como a validação de modelos previamente propostos. Portanto, o primeiro esforço de pré-processamento no presente trabalho foi a adequação do leiaute dos arquivos, criando arquivos análogos aos disponibilizados na competição PAN-2012. Uma vez que as conversas registradas não apresentam os horários de envio, o atributo *time* conforme demonstrado na Figura 2 não foi incluído.

De forma a permitir a construção de conjuntos de treinamento e teste, faz-se necessário identificar quais conversas apresentam atividade predatória ou teor predatório. Após análise individual das conversas foi possível identificar um erro de imputação dos dados em uma das conversas ($id = 2$). A conversa erroneamente apresenta dois predadores sexuais e uma vítima, porém o segundo predador sexual não apresenta relação com o contexto da discussão. No entanto, os dados foram preservados uma vez que o objetivo do presente trabalho não é impactado. O arquivo resultante "*pan12-br-sexual-predator-identification-training-corpus-predators.txt*" apresentou um total de 39 predadores e segue estritamente o formato proposto na competição PAN-2012.

Durante análise exploratória dos dados foram identificados dois pontos para correção: 1) os dados disponibilizados apresentavam mais de uma codificação (ISO-8859-1 e UTF-8), o que poderia impactar a validação de resultados de experimentos em ambi-

```xml
<conversation id="1d0d6eb4815de5e2b27d0c396abf9dc7">
  <message line="1">
    <author>634f0ee018e70d40d1db4a4bf3a2d35d</author>
    <time>02:55</time>
    <text>hi</text>
  </message>
  <message line="2">
    <author>898d2f30e39b4fc143ebdf8c0b5c6a92</author>
    <time>02:55</time>
    <text>asl</text>
  </message>
  <message line="3">
    <author>634f0ee018e70d40d1db4a4bf3a2d35d</author>
    <time>02:55</time>
    <text>m or f</text>
  </message>
  <message line="4">
    <author>898d2f30e39b4fc143ebdf8c0b5c6a92</author>
    <time>02:55</time>
    <text>m</text>
  </message>
</conversation>
```

Figura 2. Exemplo de conversa culpada disponibilizada no conjunto de dados PAN-2012.

ente Unix. Como medida, todos os conjuntos de dados foram convertidos para UTF-8; 2) os arquivos disponibilizados com a extensão *xml* não apresentam um formato válido, o que pode impedir o funcionamento de algumas bibliotecas de manipulação de arquivos XML. Foram encontrados dois pontos para correção: a) a adição de declaração da versão da especificação XML usada e encoding (UTF-8); b) a substituição do uso de caracteres "$>$" e "$<$", considerados ilegais[5], como delimitadores dos marcadores para preservação de identidade por "[" e "]". O conjunto de dados resultante do pré-processamento foi denominado "PAN-2012-BR".

4. Análise das conversas culpadas

De acordo com os dados disponibilizados, uma conversa é considerada culpada quando é constatada a presença de predadores sexuais conversando com outras pessoas, sendo essas menores de idade ou não. Dentre as 436 mensagens enviadas em conversas com predadores sexuais, 260 (59,63%) tiveram o predador sexual como remetente. As demais mensagens (176 - 40,37%) foram enviadas majoritariamente por crianças e adolescentes, porém, também foi possível encontrar ocorrências de mensagens produzidas por pessoas se passando por vítimas (e.g., pai se passando pela criança) e pessoas do círculo de amizades do predador. Cada conversa apresentou em média pouco mais de 11 mensagens trocadas. Embora a maioria das conversas apresente 4 mensagens trocadas, são encontradas conversas contendo até 41 mensagens.

A investigação teve como propósito identificar, dentre as marcações realizadas nos dados originais, quais foram mais usadas e quem mais as explorou. Dentre as 39 conversas culpadas não foram encontradas ocorrências de envio de mensagens de áudio. Na Tabela 3 é apresentada a quantidade de ocorrências de cada uma das marcações para predadores sexuais ou crianças e adolescentes.

Analisando as menções a nomes em conversas, estes foram usados com o

[5]https://www.w3schools.com/xml/xml_syntax.asp

Tabela 3. Ocorrência de marcações em conversas culpadas.

Marcação	Predadores Sexuais	Crianças e Adolescentes
>audio<	0	0
>emoticon<	19	5
>foto<	0	1
>local<	7	1
>nome<	12	0
>telefone<	0	0

propósito, primariamente, de obter intimidade com a vítima e passar segurança (e.g., *"oi seja bem vinda a meu face oficial sou eu [nome] vc tem que idade ?"*) nas fases iniciais, quando se desenvolve uma falsa relação enganosa com a vítima. Também foi usado em tom intimidatório, quando já existe uma intimidade desenvolvida. O acentuado uso de emoticons por parte do predador sexual também caracteriza a ação de dessensibilização do teor conversado, buscando amenizar o impacto negativo nas conversas, assim como ludibriar a vítima (*"Só uma pergunta so para sabe se posso dar em cima [emoticon]"*).

As menções à locais foram encontradas em vários momentos da comunicação do predador sexual, desde conversas introdutórias (*"Tc de [local] mesmo linda?"*) até estágios avançados do abuso (*"Vem até mim to com medo de ir ai Tem muita gente ai Frente a [local]"*). Um ponto importante é que não foram identificados envios de áudios e fotos por parte do predador sexual. Em algumas conversas o predador sexual se dispõe a enviar fotos para a vítima, porém não foram observadas ocorrências do envio de fato. Esse comportamento reflete uma característica dos predadores sexuais que é o uso da anonimidade para a realização do crime. O comportamento encontrado na comunicação dos predadores sexuais com as vítimas, identificado por meio dos marcadores, corrobora a teoria fundamentada em Olson et al. [2007].

5. Usando CNN para identificação de Predadores Sexuais na Internet

O conjunto de dados da competição PAN-2012 apresenta a distribuição descrita na Tabela 4. A fim de manter a mesma proporção, foi realizada a mesma distribuição dos dados para os conjuntos de treinamento (30%) e teste (70%), conforme descrito na Tabela 5.

Tabela 4. Características do conjunto de dados da competição PAN-2012

Subconjunto	# Conversas	# Usuários únicos	# Predadores Sexuais
Treinamento	*66.927*	*97.695*	*142*
Teste	*155.128*	*218.716*	*254*

Tabela 5. Características do conjunto de dados "PAN-2012-BR"

Subconjunto	# Conversas	# Usuários únicos	# Predadores Sexuais
Treinamento	*52*	*104*	*11*
Teste	*124*	*248*	*28*

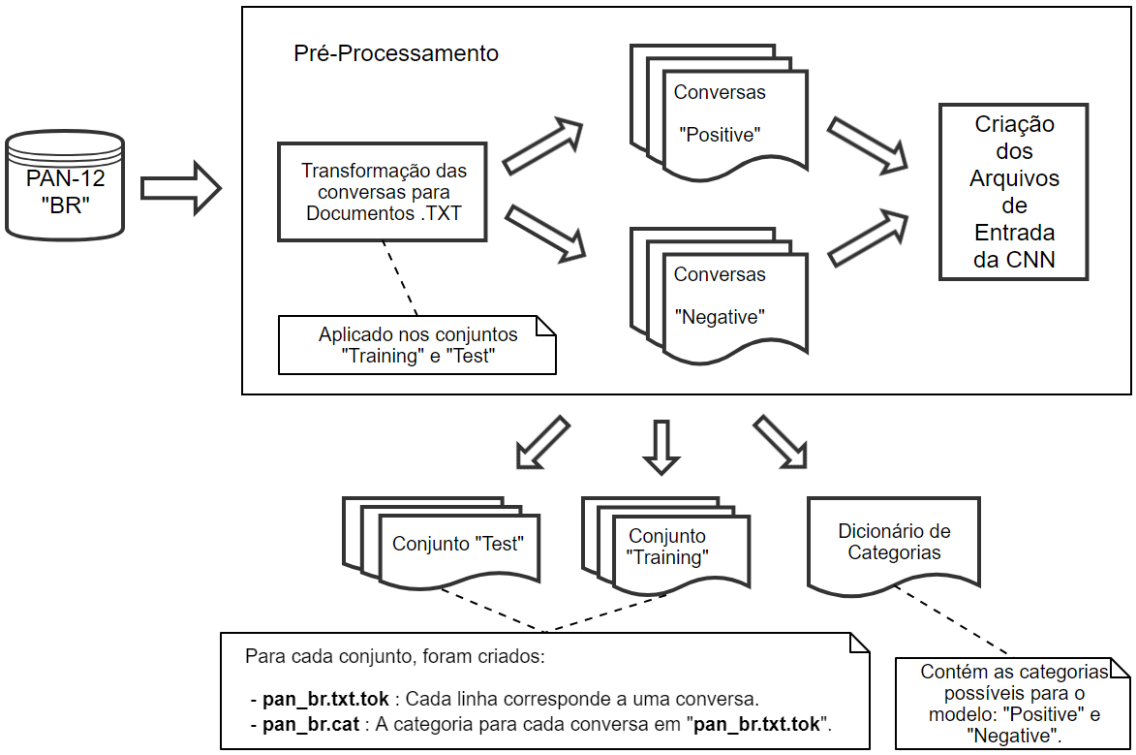

Figura 3. Preparação dos dados para posterior uso na biblioteca conText.

O primeiro experimento consiste na reprodução da arquitetura da CNN especificada em Ebrahimi et al. [2016] no conjunto de dados recém criado "PAN-2012-BR". Para a reprodução do experimento foi considerada a biblioteca conText[6], que implementa a especialização do algoritmo CNN para considerar a ordem das palavras [Johnson and Zhang, 2015].Os *scripts* contendo os parâmetros para montagem e configuração da CNN foram disponibilizados pelo autor. Para a execução apropriada do experimento foram necessárias tarefas de pré-processamento específicas, conforme pode ser observado na Figura 3.

Para o treinamento e teste da CNN foram utilizados os mesmos hiperparâmetros do trabalho considerado como referência [Ebrahimi et al., 2016]: 2000 neurônios para a camada de entrada, 3 regiões, regularização L2 = 10^{-4}, taxa de *Dropout* = $0,5$, função de perda quadrática, função de ativação *ReLU* e Kernel Linear. Conforme detalhado previamente no pré-processamento dos dados para posterior uso na biblioteca conText, cada conversa foi considerada uma característica e todas foram submetidas à diminuição de *case* e convertidas para codificação UTF-8. Na fase de treinamento, foram consideradas as 5.000 palavras com maior ocorrência nas conversas, 100 épocas e decaimento de peso. Nos demais hiperparâmetros, foram considerados os valores padrão da biblioteca usada. A arquitetura proposta pelo autor pode ser observada na Figura 4.

Realizamos três testes no conjunto de dados "PAN-2012-BR" considerando os hiperparâmetros citados acima: o primeiro teste, considerou a mesma proporção de dados do conjunto de dados PAN-2012 (30% para treinamento e 70% para testes). No segundo

[6]http://riejohnson.com/cnn_download.html

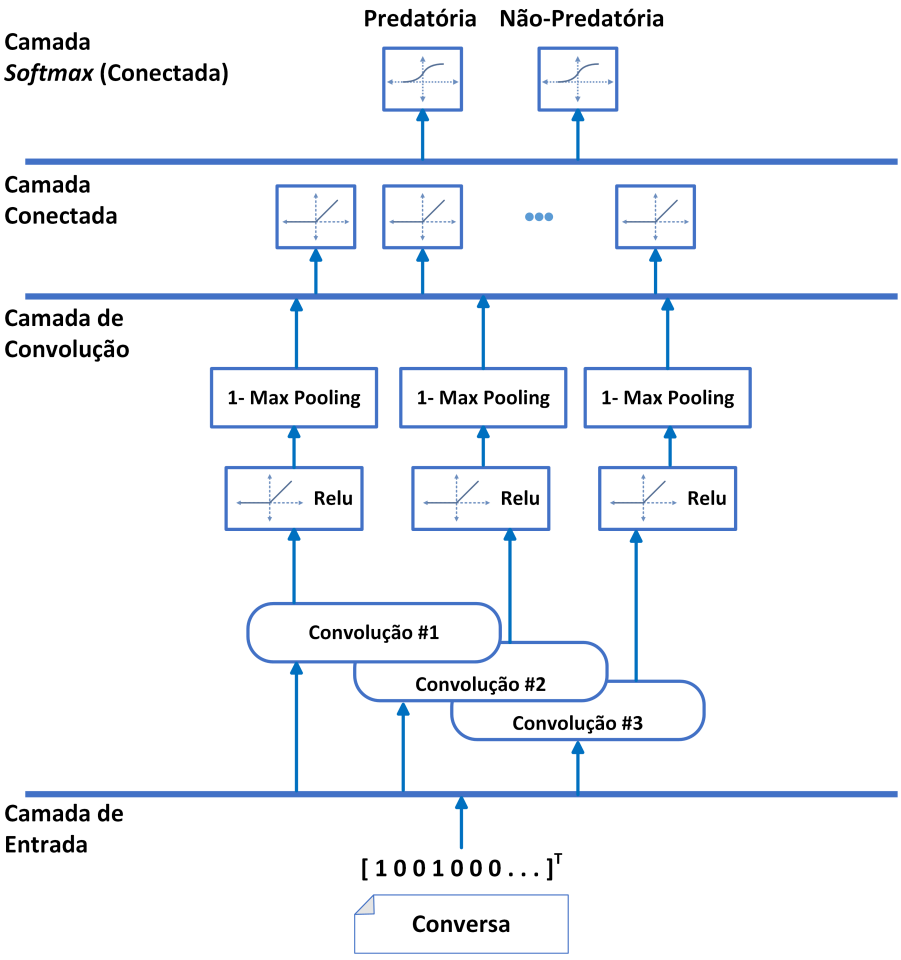

Figura 4. Adaptação para arquitetura de CNN reproduzida nos experimentos presente em Ebrahimi et al. [2016]. Autor: Próprio.

teste, invertemos a proporção dos conjuntos (70% para treinamento e 30% para testes). Por fim, no terceiro teste aplicamos o algoritmo K-Fold estratificado (K = 5). Os resultados podem ser observados na Tabela 6:

Tabela 6. Resultados do primeiro experimento com o "PAN-2012-BR" considerando os hiperparâmetros utilizados no trabalho proposto por Ebrahimi et al. [2016].

Modelo	Acurácia	Precisão	Abrangência	F_1	$F_{0.5}$
CNN (30%/70%)	0.99	1.0	0.98	0.99	0.99
CNN (70%/30%)	1.0	1.0	1.0	1.0	1.0
CNN (5-Fold)	0.99	1.0	0.99	0.99	0.99

Por conta dos resultados obtidos no experimento, resolvemos explorar a alteração de dois hiperparâmetros do modelo: o algoritmo de otimização que, por padrão na biblioteca conText, considera o Gradiente Descendente Estocástico (SGD) e o número de palavras no vocabulário do conjunto de treinamento. Os resultados podem ser observados nas Tabelas 7 e 8.

Tabela 7. Resultados do segundo experimento considerando as proporções (30%/70%) e (70%/30%) para os conjuntos de treinamento e teste. É considerado o otimizador SGD e diferentes tamanhos de vocabulário para treinamento.

Modelo	Acurácia	Precisão	Abrangência	F_1	$F_{0.5}$
CNN (30%/70%), vocab=1000	1.0	1.0	1.0	1.0	1.0
CNN (70%/30%), vocab=1000	1.0	1.0	1.0	1.0	1.0
CNN (30%/70%), vocab=10000	0.99	1.0	0.98	0.99	0.99
CNN (70%/30%), vocab=10000	1.0	1.0	1.0	1.0	1.0
CNN (30%/70%), vocab=15000	0.99	1.0	0.98	0.99	0.99
CNN (70%/30%), vocab=15000	1.0	1.0	1.0	1.0	1.0

Tabela 8. Resultados do segundo experimento considerando as proporções (30%/70%) e (70%/30%) para os conjuntos de treinamento e teste. É considerado o otimizador RMSProp e diferentes tamanhos de vocabulário para treinamento.

Modelo	Acurácia	Precisão	Abrangência	F_1	$F_{0.5}$
CNN (30%/70%), vocab=1000	0.99	0.98	1.0	0.99	0.99
CNN (70%/30%), vocab=1000	0.92	0.91	1.0	0.95	0.92
CNN (30%/70%), vocab=5000	0.99	1.0	0.98	0.99	0.99
CNN (70%/30%), vocab=5000	0.92	0.91	1.0	0.95	0.92
CNN (30%/70%), vocab=10000	0.99	0.98	1.0	0.99	0.99
CNN (70%/30%), vocab=10000	0.96	0.95	1.0	0.97	0.96
CNN (30%/70%), vocab=15000	0.99	0.98	1.0	0.99	0.99
CNN (70%/30%), vocab=15000	0.94	0.93	1.0	0.96	0.94

Durante a reprodução dos experimentos foi possível observar que, na maioria dos cenários, o modelo atingiu taxa de erro zero em até 50 épocas. A partir de então, o modelo perdeu capacidade de generalização embora tenha mantido ainda resultados altos nas métricas de classificação, o que, na nossa interpretação caracterizou um cenário de sobreajuste. Esse comportamento ficou mais evidente quando aplicado o algoritmo de otimização RMSProp considerando 70% dos dados para treinamento. Em alguns casos foi obtido 100% de acurácia em poucas épocas de treinamento, por exemplo, o modelo "CNN (70%/30%), vocab=5000" precisou de apenas 12 épocas para atingir 100% de acurácia. Um ponto importante é que o tamanho do vocabulário apresentou impacto discreto no uso do otimizador SGD.

6. Conclusão e trabalhos futuros

Predadores sexuais são uma ameaça a crianças e adolescentes na Internet brasileira. O presente trabalho apresentou a reprodução do primeiro experimento considerando CNN na identificação de predadores sexuais na Internet com dados oriundos de investigações do MPF-SP e liberados ao público. É o primeiro trabalho nesse contexto que se tem conhecimento com o uso de dados na língua portuguesa.

Resultados significativos obtidos nos experimentos consolidam o uso de CNN

como uma alternativa para a identificação de atividade predatória no Brasil. No entanto, a diversidade de vocabulário usado pelo predador, principalmente gírias e erros de ortografia, o tornam muito diferente do presente nas conversas não-predatórias. Com relação a esse ponto, conclui-se que o conjunto de dados apresenta uma baixa complexidade para a tarefa de classificação. Como citado na seção anterior, termos específicos usados em conversas predatórias começaram a ser encontrados após considerar um vocabulário com 15.000 termos. Considerando e a baixa complexidade citada previamente na classificação de um conjunto de dados desbalanceado, entende-se que os resultados obtidos se deram por conta da maior capacidade da rede de identificar corretamente uma mensagem não-predatória.

Um ponto para melhoria do conjunto de dados seria a adição de conversas predatórias. A principal motivação, além da necessidade de balanceamento, é ampliar o vocabulário de conversas predatórias. Entende-se que o regionalismo impacta diretamente na diversificação de gírias, portanto, o aumento de conversas predatórias teria impacto direto na capacidade de generalização do modelo. O aumento da diversidade de temas discutidos em conversas não-predatórias, cujo vocabulário apresente alguma similaridade com o usado em conversas predatórias, também é necessário. Por fim, é necessário alterar a arquitetura da CNN apresentada no experimento para melhor adaptação aos textos em português, visto que inicialmente essa arquitetura foi projetada para o conjunto de dados da competição PAN-2012.

Como oportunidade para trabalhos futuros, é considerado validar trabalhos que apresentaram resultados relevantes e candidatos ao estado da arte no domínio da pesquisa, com o principal propósito de identificar quais características melhor contribuem para a identificação da atividade predatória no Brasil.

7. Agradecimentos

Ao Centro Universitário da Fundação Educacional Inaciana, em particular, o Prof. Dr. Rodrigo Filev Maia e ao Ministério Público Federal, especialmente à chefe do departamento de Crimes Cibernéticos, Adriana Shimabukuro, pela cooperação na disponibilização dos dados para a pesquisa.

Referências

Andrijauskas, A., Shimabukuro, A., and Maia, R. F. (2017). Desenvolvimento de base de dados em língua portuguesa sobre crimes sexuais. *VII Simpósio de Iniciação Científica, Didática e de Ações Sociais da FEI*.

BARBOSA, A. F. (2018). Pesquisa sobre o uso da internet por crianças e adolescentes no brasil: Tic kids online brasil 2017. *São Paulo: Comitê Gestor da Internet no Brasil*.

Bengel, J., Gauch, S., Mittur, E., and Vijayaraghavan, R. (2004). Chattrack: Chat room topic detection using classification. In *ISI*.

Bogdanova, D., Rosso, P., and Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. *Computer speech & language*, 28(1):108–120.

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

Dorasamy, M., Jambulingam, M., and Vigian, T. (2018). Building a bright society with au courant parents: Combating online grooming.

dos Santos, L. F. and Guedes, G. P. (2018). Detecção de traços de narcisismo em conversas com predadores sexuais. In *7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*, volume 7. SBC.

Ebrahimi, M., Suen, C. Y., and Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, 18:33–49.

Forsythand, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

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

Johnson, R. and Zhang, T. (2015). Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112.

Kloess, J. A., Hamilton-Giachritsis, C. E., and Beech, A. R. (2019). Offense processes of online sexual grooming and abuse of children via internet communication platforms. *Sexual Abuse*, 31(1):73–96.

Kontostathis, A. (2009). Chatcoder: Toward the tracking and categorization of internet predators. In *Proceedings of Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009). SPARKS, NV. May 2009*. Citeseer.

Kontostathis, A., Garron, A., Reynolds, K., West, W., and Edwards, L. (2012). Identifying predators using chatcoder 2.0. In *CLEF (Online Working Notes/Labs/Workshop)*.

NCMEC (2017). The online enticement of children: An in-depth analysis of cybertipline reports. *National Center for Missing & Exploited Children Web site., https://missingkids-stage.adobecqms.net/ourwork/publications/exploitation/onlineenticement. Acessado em 16 de março de 2019*.

Ngejane, C., Mabuza-Hocquet, G., Eloff, J., and Lefophane, S. (2018). Mitigating online sexual grooming cybercrime on social media using machine learning: A desktop survey. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6. IEEE.

Olowu, D. (2014). Cyber-based obscenity and the sexual exploitation of children via the internet: Implications for africa. In *African Cyber Citizenship Conference 2014 (ACCC2014)*, page 115.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. 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. IEEE.

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