Submission date: 13/10/2019
Resubmission date: 16/03/2020
Resubmission date: 07/06/2020
Camera ready submission date: 04/07/2020

1st round notification: 23/12/2019
2nd round notification: 18/05/2020
3rd round notification: 29/06/2020
Available online: 06/07/2020
Publication date: 31/07/2020

Section: extended version
Extended versions of papers presented at the VIII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2019)

Identificação de predadores sexuais brasileiros em conversas textuais na internet por meio de aprendizagem de máquina

Title: Identification of Brazilian sexual predators in textual conversations on the internet through machine learning

Leonardo Ferreira dos Santos[1], **Gustavo 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, a large number of children and adolescents have made use of social applications. Easy to access, these applications provide benefits and opportunities. However, at the same time, they expose users to different risks, including predatory sexual activity. Predatory sexual activity has several purposes, such as obtaining child pornography, extortion, and sexual abuse. The present work has three main objectives: (i) to create a data set of textual conversations containing a real predatory sexual activity for Brazilian Portuguese; (ii) to perform a statistical analysis in the data set created; (iii) to carry out an experimental evaluation considering the most popular machine learning algorithms in the research domain with the data set built. This evaluation regards F_1 measure as a basis. The results achieved with contributions (i) and (ii) enable new studies to focus on the problem of identifying sexual predators in textual conversations for Brazilian Portuguese. The results obtained with the contribution (iii) show that the Support Vector Machines behaved as the best of the considered algorithms, presenting a result of 89.87%.*

Keywords. *PAN-2012, Sexual predator Identification, Machine Learning, Convolutional Neural Networks, Support Vector Machine, Decision Tree, Naïve Bayes, Random Forests, Social Networks, Chats*

Resumo. *Nos dias de hoje um grande número de crianças e adolescentes tem usado aplicações sociais. De fácil acesso, essas aplicações promovem benefícios e oportunidades. No entanto, ao mesmo tempo, expõem os usuários à diferentes riscos, dentre os quais a atividade predatória sexual. A atividade predatória sexual possui diversas finalidades como a obtenção de pornografia infantil, a extorsão e o abuso sexual. O presente trabalho possui três objetivos principais: (i) criar um conjunto de dados de conversas textuais contendo atividade sexual predatória real para o português do Brasil; (ii) realizar uma análise estatística das conversas textuais presentes nesse conjunto de dados; (iii) realizar uma avaliação experimental considerando os algoritmos de aprendizado*

Cite as: Santos, Leonardo., & Guedes, Gustavo. (2020). Identification of Brazilian sexual predators in textual conversations on the internet through machine learning (Identificação de predadores sexuais brasileiros em conversas textuais na internet por meio de aprendizagem de máquina). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 13(4), 22-47. DOI: 10.5753/isys.2020.822.

de máquina mais populares no domínio da pesquisa com o conjunto de dados construído. Essa avaliação considera a medida de F_1 como base. Os resultados alcançados com as contribuições (i) e (ii) possibilitam que novos estudos possam se concentrar na problemática da identificação de predadores sexuais em conversas textuais para o português do Brasil. Os resultados obtidos com a contribuição (iii) evidenciam que as Máquinas de vetores de suporte obtiveram o melhor comportamento, apresentando um resultado de 89.87%.

Palavras-Chave. *Pedofilia, PAN-2012, Identificação de predador sexual, Aprendizado de máquina, Redes Neurais Convolucionais, Máquina de vetores de suporte, Árvore de decisão, Naïve Bayes, Florestas Aleatórias, Redes sociais, Conversas virtuais*

1. Introdução

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]. Essa preocupação é crescente na medida em que redes sociais *online* se tornam mais acessíveis à população, permitindo que, cada vez mais, crianças e adolescentes consigam utilizá-las sem grandes dificuldades. Cientes desse fato, os predadores sexuais se aproveitam da vulnerabilidade dessas crianças e adolescentes para realizarem atividades predatórias sexuais [Hernandez et al. 2018].

Conforme descrito na literatura, a atividade predatória sexual pode se manifestar de diversas maneiras [NCMEC 2017]: em 78% das atividades predatórias sexuais, o principal objetivo é a obtenção de conteúdo pornográfico da vítima por meio de conversas em redes sociais. Em 7% atividades predatórias sexuais, a vítima também sofre extorsão (*Sextortion*[1]), esta se caracterizando por pedidos de transferências financeiras, disponibilização de dados de cartão de crédito dos pais e pertences. Em 5% dos casos de atividade predatória sexual, os predadores buscam contato pessoal com a vítima, objetivando a realização do abuso sexual. Os 10% restantes apresentam outras finalidades que não se enquadram nessas principais razões.

Segundo dados da pesquisa *TIC Kids Online Brasil*, que estuda o uso da internet por crianças e adolescentes [Barbosa 2018], 85% das crianças e adolescentes com idades entre 9 e 17 anos possuem acesso à internet. Os dispositivos móveis são o único meio de acesso para 44% do total de crianças e adolescentes, cuja maioria pertence às classes A e B (76%). Ao serem questionadas sobre o que consideram um incômodo na internet, 10% do total de crianças e adolescentes mencionaram o risco de contato ou assédio de pessoas estranhas, indesejadas ou adultos desconhecidos. Esse valor é inferior à menção a outros riscos como: risco de conduta (32%), risco de exposição a conteúdo inadequado (28%), percepção de sexo (17%) e a percepção de violência (12%). A pouca ocorrência de menção ao risco de contato e a não alteração desse quadro ao longo dos anos conclui a existência de uma ameaça real para crianças e adolescentes [Barbosa 2018].

[1]O "Sextortion" pode ser compreendido como a realização de ameaças após obtenção de conteúdo com teor de nudez da vítima, de forma que seja enviado mais conteúdo pornográfico produzido pela vítima [Wolak et al. 2018].

Nesse cenário, uma preocupação legítima dos pais é garantir que os filhos não sejam expostos aos riscos oriundos da internet, ao mesmo tempo que sejam usufruídos todos os benefícios e oportunidades [Livingstone et al. 2017]. No entanto, intervenções parentais na seleção do conteúdo a ser acessado na internet apresentam um elevado grau de rejeição, devido ao fato de crianças e adolescentes almejarem encontrar na internet um ambiente para liberdade, além da possibilidade de demonstrar que são capazes de tomar boas decisões e aprender com os erros [Ghosh et al. 2018].

Diante do contexto apresentado, diversos estudos têm sido conduzidos com o objetivo de identificar predadores sexuais na internet. A maioria dos trabalhos encontrados na literatura foi viabilizada devido ao uso de conversas predatórias (em língua inglesa) disponibilizadas pela organização *Perverted-Justice* (PJ)[2]. Nessas conversas, agentes federais atuam na internet como crianças e adolescentes, de forma a servir de iscas para predadores sexuais convictos. Essa iniciativa permitiu o estudo de diversas características do predador sexual e, com isso, diferentes métodos de identificação foram propostos.

No entanto, uma limitação frequente nos trabalhos realizados no domínio da pesquisa é a carência de conversas que ocorreram entre predadores sexuais e vítimas (reais) de atividade predatória, sendo as vítimas crianças ou adolescentes. O principal impedimento para disponibilização dessas conversas é o teor sensível das informações presentes e por serem provas de processos que correm em sigilo na justiça. Nesse aspecto, o conjunto de dados fornecido pelo Ministério Público Federal de São Paulo (MPF-SP) em parceria com o Centro Universitário da Fundação Educacional Inaciana (FEI) [Andrijauskas et al. 2017] trouxe uma contribuição bastante relevante para o tema, dado que disponibilizou diversas conversas predatórias e, inclusive, foi utilizado como base para a construção do conjunto PAN-2012-BR [Santos and Guedes 2019]. No entanto, ambos os conjuntos consideram conversas transcritas de áudio para compor as conversas não-predatórias, o que não caracteriza um ambiente real de mensagens enviadas por texto. As conversas transmitidas por áudio omitem a riqueza linguística presente no vocabulário escrito em conversas virtuais na internet [Crystal 2002]. Além disso, a diferença de vocabulário empregado entre as diferentes fontes (transcrita de áudio e escrita) contribui para a redução da complexidade em uma tarefa de classificação, uma vez que existe uma menor sobreposição de termos empregados [Scott and Matwin 1998].

O presente trabalho estende a publicação realizada pelos mesmos autores no BraSNAM 2019 [Santos and Guedes 2019]. De forma a realizar essa extensão, dois tópicos foram considerados: (i) enriquecimento do conjunto de dados PAN-2012-BR por meio da aplicação do conceito de representatividade para a seleção de conversas não-predatórias; (ii) uso de técnicas de mineração de dados e algoritmos de aprendizado de máquina em uma avaliação experimental. Essa avaliação experimental tem como propósito apresentar um método para identificação de atividade predatória em conversas virtuais. Nesse cenário, as contribuições resultantes deste trabalho são:

- um novo conjunto de dados de conversas entre predadores sexuais e suas vítimas, denominado PREDADORES-BR, contendo apenas conversas provenientes de chats (tanto as predatórias como as não-predatórias). Nesse novo conjunto,

[2] `http://www.perverted-justice.com`

são consideradas todas as conversas *predatórias* presentes no conjunto de dados PAN-2012-BR. Para a composição das conversas *não-predatórias*, é apresentado um método de extração, transformação e seleção de chats em comunidades virtuais, baseado em um dos trabalhos desenvolvidos na competição PAN-2012 [Inches and Crestani 2012]. O método proposto no presente trabalho expande o desenvolvido na competição PAN-2012 ([Inches and Crestani 2012]) ao considerar a representatividade das conversas obtidas na criação do conjunto de dados. Isso é efetuado por meio da seleção dos tópicos definidos *a priori* no processo de coleta dos chats não-predatórios.

- uma análise estatística do conjunto de dados PREDADORES-BR; baseado no método proposto por [Sokolova and Bobicev 2018], são exploradas as características de mensagens e conversas, de forma a buscar uma melhor compreensão da distribuição dos dados entre as classes e explorar possíveis pontos de complexidade em uma posterior tarefa de classificação.
- avaliação experimental considerando técnicas de mineração de dados e algoritmos de aprendizado de máquina pertencentes a diferentes paradigmas e que correspondem ao estado da arte no domínio da pesquisa; nesse cenário, são considerados os seguintes paradigmas e algoritmos: (i) baseados em kernel: Máquina de vetores de suporte (SVM); (ii) probabilísticos: Naïve Bayes Multinomial (MNB); (iii) conexionistas: Redes neurais convolucionais (CNN); (iv): simbolistas: Árvores de decisão (DT) e Florestas aleatórias (RF). O objetivo é apresentar um resultado de base, evidenciando qual algoritmo de aprendizado de máquina apresenta o melhor desempenho no conjunto PREDADORES-BR, considerando a medida F_1. A partir desse resultado de base, novos trabalhos podem trazer contribuições para a detecção de atividade predatória sexual em textos na língua portuguesa do Brasil.

Vale ressaltar que os os códigos-fontes criados e os modelos gerados a partir da utilização do conjunto de dados PREDADORES-BR estão disponibilizados publicamente[3]. O objetivo é que esses modelos possam ser utilizados em aplicações reais para a detecção automática de potenciais predadores sexuais.

As demais seções desse trabalho estão dispostas da seguinte maneira: na seção 2, são levantados trabalhos semelhantes e relacionados ao tema abordado nesse estudo; na seção 3, é apresentado o conjunto de dados usado como base para a pesquisa; na seção 4, é apresentada a análise estatística do conjunto de dados PREDADORES-BR; na seção 5, é detalhada a avaliação experimental e os resultados preliminares são discutidos; e por fim, a seção 6 apresenta as conclusões, limitações e discussão sobre trabalhos futuros.

2. Trabalhos Relacionados

Ao longo das duas últimas décadas foram consideradas diversas abordagens para a identificação automática de atividade predatória sexual na internet. Em 2007, foi realizado um estudo piloto [Pendar 2007] considerando as conversas registradas no site PJ para a criação de um conjunto de dados. A tarefa de identificação dos predadores sexuais foi executada considerando o uso de unigramas, bigramas e trigramas para a geração

[3] `https://eic.cefet-rj.br/~lacafe/periodicos/`

dos vetores de características. São considerados os algoritmos de classificação *k-nearest neighbors* (kNN) e SVM nos experimentos. A aplicação do algoritmo kNN, considerando k = 30, obteve melhores resultados (F_1 = 94%), próximo dos resultados obtidos com o algoritmo SVM (F_1 = 90%).

Em [Villatoro-Tello et al. 2012], o problema de classificação de conversas predatórias é decomposto em duas partes. Na primeira parte os autores buscam descartar todas as conversas que não apresentam as características mais frequentes de uma atividade predatória sexual. Sendo assim, foram descartadas todas as conversas com as seguintes características: 1) Conversas com apenas um participante; 2) Conversas com menos de 6 mensagens por participante; 3) Conversas com sequências longas de caracteres, que não são compreensíveis como uma frase que faça parte de uma conversa. Esse filtro reduziu em 90% o volume de dados a ser considerado para a criação do modelo, o que permitiu otimizar o uso de recursos computacionais, assim como direcionar a análise para os cenários mais propícios ao aliciamento. A segunda parte, responsável pela identificação do predador sexual dentre os participantes de uma conversa, fez uso de uma Rede Neural Perceptron com Múltiplas Camadas e representação binária. O resultado atingido (F_1 = 87,3%, $F_{0.5}$ = 93,4%) conferiu aos autores a primeira colocação na competição PAN-2012.

No trabalho realizado por [Cano et al. 2014], a estratégia adotada para a identificação da atividade predatória sexual levou em consideração um dos modelos psicológicos que explicam como o aliciamento é realizado na internet [O'Connell 2003]. Para cada estágio do aliciamento previsto no modelo, foram identificados padrões por meio de características léxicas, sintáticas[4], psicolinguísticas, de polaridade dos sentimentos[5] contida nas mensagens enviadas, de padrões no conteúdo das mensagens e no discurso. Para a extração das características léxicas foi utilizada a representação *Bag of Words* (BoW). A ferramenta LIWC [Pennebaker et al. 2001] permitiu a extração de características psicolinguísticas. Para uma melhor seleção das características psicolinguísticas das sentenças, foi determinado o ganho de informação para cada uma das características possíveis em cada fase do aliciamento, sendo selecionadas as 5 características mais relevantes. Para a criação do conjunto de dados, foram consideradas 50 conversas predatórias disponibilizadas no site PJ e cada sentença contida nessas conversas passou por uma classificação perante as etapas de aliciamento do modelo psicológico citado. Os resultados obtidos com o algoritmo SVM se mostraram satisfatórios para os autores (F_1 = 85%), reconhecendo que características psicolinguísticas presentes no discurso contribuem para a identificação dos estágios de aliciamento.

Na primeira iniciativa documentada para estudo do comportamento predatório na comunicação em jogos *online* [Cheong and Jensen 2015], o jogo *MovieStarPlanet*, cujo público-alvo são crianças e adolescentes entre 8 e 15 anos, disponibilizou três conjuntos de dados contendo a comunicação escrita de diferentes jogadores: mensagens de *status*, comentários em vídeos e postagens em fóruns, além de conversas públicas e privadas realizadas dentro da plataforma. Em dois dos conjuntos, foram considerados apenas

[4] `https://nlp.stanford.edu/software/tagger.shtml`
[5] `http://sentistrength.wlv.ac.uk/`

conteúdos não predatórios e o terceiro conjunto de dados foi criado a partir de conversas realizadas por 59 predadores sexuais. Alguns dos desafios para a classificação das conversas nesse contexto foram o vocabulário usado pelo público, que apresenta um elevado uso de gírias, erros gramaticais e ortográficos, além de sequencias digitadas sem sentido. Também pode-se destacar a natureza de conversas realizadas nas plataformas, que podem ser similares às usadas pelos predadores: conversas sobre namoro, estar solteiro, procurar por um(a) ou estar apaixonado pelo(a) namorado(a), assim como assuntos relacionados à família. De forma a lidar com esse cenário, foi considerada uma combinação de características léxicas, sentimentais e comportamentais. Os resultados obtidos (F_1 = 78%, $F_{0.5}$ = 86%) foram considerados promissores, tendo em vista a natureza dos dados.

Em [Ebrahimi 2016], a identificação de atividade predatória sexual foi tratada como uma tarefa de detecção de anomalias. Nesse contexto, o algoritmo de aprendizado semi-supervisionado Máquina de vetores de suporte para uma única classe (OC-SVM) é aplicado em um conjunto de dados privado e disponibilizado pelo *Sûreté du Québec* (SQ) e no conjunto de dados da competição PAN-2012. O conjunto de dados SQ apresenta um total de 82 conversas, das quais 76 são predatórias. Os resultados mais significativos foram encontrados no conjunto de dados SQ (F_1 = 97,4%), no entanto os autores consideram o experimento realizado uma prova de conceito e apenas válido para comprovação da hipótese proposta no trabalho.

Uma estratégia semelhante à [Villatoro-Tello et al. 2012] é aplicada por [Cardei and Rebedea 2017] ao implementar um algoritmo em dois estágios. O primeiro estágio considera características textuais e comportamentais em conjunto com o algoritmo 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, o uso do algoritmo RF apresentou melhor desempenho quando comparado ao algoritmo 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 foram mais significantes na identificação do predador sexual. Os resultados encontrados (*acurácia* = 100%, *abrangência* = 81.8%, $F_{0.5}$ = 95.7%) superam os encontrados na competição PAN-2012.

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 uso de CNN se mostrou promissor (F_1 = 81.64%) quando comparado ao resultado obtido pelos algoritmos SVM (F_1 = 61.02%) e redes neurais Perceptron (NN-MLP) (F_1 = 79.96%). Um ponto que vale ressaltar nesse trabalho é que ele busca validar a aplicação de algoritmos para a classificação no problema de identificação de atividade predatória sexual em conversas na internet e não definir um novo estado da arte para o domínio da pesquisa.

Dentre os trabalhos relacionados, é possível enxergar uma lacuna no uso de algoritmos de aprendizado de máquina supervisionado que compõem o estado da arte

no domínio da pesquisa em conjunto com dados em português do Brasil. As conversas não-predatórias presentes no conjunto de dados PAN-2012-BR são transcrições de conversas de áudio. Conversas em áudio e escritas pela internet apresentam diferentes características [Crystal 2002], o que pode impactar o resultado de experimentos. Nesse cenário, o presente trabalho apresenta três contribuições: (i) criação do conjunto de dados PREDADORES-BR de acordo com o método descrito em [Inches and Crestani 2012], utilizado na competição PAN-2012, composto por conversas regulares ocorridas em comunidades virtuais e conversas com a presença de predadores sexuais; (ii) análise estatística do conjunto de dados PREDADORES-BR com base no método proposto em [Sokolova and Bobicev 2018]; (iii) avaliação experimental considerando os algoritmos que correspondem ao estado da arte em aprendizado de máquina no domínio da pesquisa.

3. Conjuntos de Dados PREDADORES-BR

A motivação para o enriquecimento do conjunto de dados PAN-2012-BR [Santos and Guedes 2019] se sustenta na necessidade de adicionar uma maior representatividade nos dados para o estudo. O conceito de representatividade na criação de um conjunto de dados remete a capacidade de contemplar a variabilidade de uma população-alvo em uma amostra [Biber 1993]. Para o presente trabalho, entende-se que as crianças e adolescentes que se comunicam pela internet em conversas virtuais no formato escrito, em português do Brasil, são a população-alvo. Sendo assim, existem três pontos de melhoria relacionados à representatividade dos dados, identificados após análise individual das conversas não-predatórias no conjunto de dados PAN-2012-BR:

- Ausência de conversas da categoria adulta (i.e., que podem conter termos sexuais): conversas predatórias apresentam a ocorrência de termos sexuais. As conversas não-predatórias presentes no conjunto de dados PAN-2012-BR não apresentam essa característica.
- Transcrição de mensagens de áudio: uma quantidade não informada de mensagens que compõem as conversas não-predatórias são oriundas de transcrições de mensagens de áudio [Andrijauskas et al. 2017]. Essa característica impacta diretamente a ocorrência da forma grafo-linguística difundida em textos de conversas virtuais, denominada internetês [Komesu and Tenani 2009].
- Vocabulário empregado: é possível observar que os assuntos das conversas não-predatórias presentes no conjunto de dados PAN-2012-BR não são compatíveis com a realidade de uma criança ou adolescente.

Na competição PAN-2012 [Inches and Crestani 2012], é proposto o uso de duas sub-categorias para conversas não-predatórias: 1) Conversas regulares, isto é, conversas pertencentes à diversas categorias, realizadas, originalmente, em formato textual e que não apresentam termos sexuais; e 2) Conversas pertencentes a categoria adulta. Para atingir tal objetivo, diferentes fontes de dados foram consideradas. Na Figura 1 é apresentada a metodologia para a criação do conjunto de dados PREDADORES-BR.

3.1. Conversas predatórias

Para o presente trabalho é considerado um total de 39 conversas predatórias. Estas conversas, antes sob sigilo de justiça, foram anonimizadas por meio de marcações específicas,

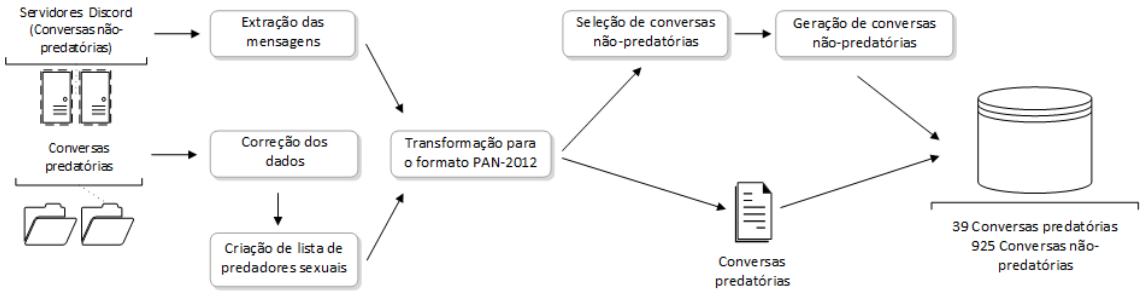


Figura 1. Etapas para criação do conjunto de dados PREDADORES-BR: 1) Extração de conversas não-predatórias de diferentes categorias oriundas de servidores de bate-papo Discord; 2) Acerto de dados em conversas predatórias e geração de lista de predadores sexuais; 3) Transformação das conversas extraídas para formato usado na competição "PAN-2012"; 4) Seleção de conversas não-predatórias considerando a representatividade de cada tópico; 5) Construção da base de conversas predatórias e não-predatórias PREDADORES-BR.

realizadas manualmente, de tal forma que a identidade dos participantes fosse preservada [Andrijauskas et al. 2017]. A Tabela 1 apresenta as marcações encontradas nas conversas predatórias.

Tabela 1. Marcações inseridas nas conversas 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

Após análise individual das conversas predatórias, 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. Esse erro foi ignorado, visto que não interfere nos objetivos propostos para o trabalho. Um outro ponto observado foi a presença de mais de uma codificação (ISO-8859-1 e UTF-8) nas conversas predatórias, o que poderia impactar o resultado de experimentos. Sendo assim, foi necessária a execução de algumas tarefas de correção na codificação e no formato dos dados:

- Conversão de todas as conversas para codificação UTF-8.
- Substituição dos caracteres "$>$" e "$<$" utilizados como delimitadores dos marcadores para preservação de identidade. Esses caracteres são considerados ilegais

para o uso dentro de elementos em um documento XML[6]. Para a substituição, foram considerados os caracteres "[" e "]".

De forma a viabilizar a avaliação dos algoritmos, fez-se necessário identificar o predador sexual dentre os participantes das conversas predatórias. A Figura 2 ilustra uma das conversas analisadas. Para atingir esse objetivo, cada conversa foi analisada e o *hash* de cada participante identificado como predador sexual foi preenchido em um arquivo de texto à parte. O arquivo resultante dessa análise, denominado "*pan12-br-sexual-predator-identification-training-corpus-predators.txt*" apresentou um total de 39 predadores sexuais.

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

Figura 2. Exemplo de conversa predatória presente no conjunto de dados PREDADORES-BR. Na conversa em questão é possível observar duas pessoas conversando: um predador sexual ("709916bfe16ef8cdd6102dc5453f302f") e um menor de idade ("13f27f55ef3622f4e987aac6a57b1ce8").

3.2. Conversas não-predatórias

A criação de conversas regulares (i.e., não-predatórias) foi realizada a partir da coleta de mensagens enviadas em comunidades virtuais hospedadas na plataforma Discord[7]. O Discord é uma plataforma constituída por servidores com o propósito de permitir a troca

[6] https://www.w3schools.com/xml/xml_syntax.asp
[7] https://www.discord.com/

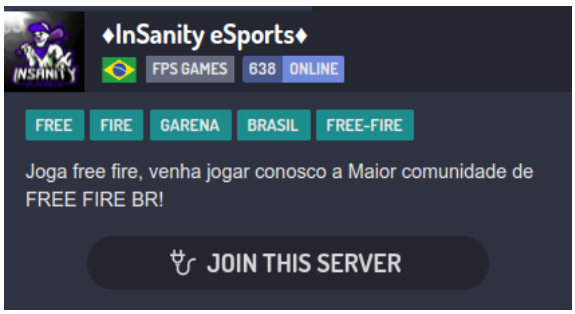


Figura 3. Exemplo de servidor Discord que abriga uma comunidade de jogos online.

de mensagens na forma de imagens, vídeos, voz ou texto. Um servidor Discord apresenta uma estrutura capaz de manter diversas salas de bate-papo, em que cada uma apresenta um tópico para direcionar as discussões. Inicialmente, a plataforma foi criada com o propósito de apoiar a comunidade de jogadores virtuais, entretanto tem sido expandida em grande escala e vem sendo utilizada para uma grande variedade de propósitos [Webb 2018].

O uso de um site especializado na indexação de servidores Discord foi usado como ferramenta de apoio[8] para a busca de servidores que pertencessem à diferentes categorias e contivessem textos em português do Brasil. As mensagens foram adquiridas por meio de uma ferramenta de código aberto[9]. Uma vez que a ferramenta é inicializada e autenticada em um servidor Discord, é possível selecionar os canais de bate-papo desejados e então realizar a extração de todo o histórico de mensagens enviadas nesses canais. Para o presente trabalho, a categoria das conversas de um servidor foi definida com base na descrição dos servidores e o conjunto de *tags* as quais o servidor Discord se encontra classificado na ferramenta de apoio. A figura 3 ilustra um dos servidores encontrados por meio da ferramenta.

No total, foram considerados 11 servidores e 5 categorias distintas: Jogos, Política, Tecnologia, Estudos e Adulto. A Tabela 2 apresenta as características dos dados obtidos ao final do processo de extração e transformação das mensagens em conversas não-predatórias.

Tabela 2. Características dos dados extraídos de servidores de bate-papo Discord e organizados por categoria.

Característica	Jogos	Política	Tecnologia	Estudos	Adulto
Servidores	2	3	3	1	2
Mensagens	882.800	594.986	19.621	523.247	1.341.813
Conversas	81.353	31.455	4.514	31.478	14.832

O método proposto em [Inches and Crestani 2012] e escolhido como base para a

[8] https://disboard.org/
[9] https://github.com/Tyrrrz/DiscordChatExporter

```
<conversation id="52">
  <message line="1">
    <author>a38386b66f95357fc5996b21fe1785bb</author>
    <time>2019-07-02 00:46:00</time>
    <text>Sim, amanhã faço isso, agora sou obrigado a dormir, por causa da minha mãe e o
        toque de recolher dela</text>
  </message>
  <message line="2">
    <author>e915385e151a03b29b5e797cb8d46847</author>
    <time>2019-07-02 00:47:00</time>
    <text>ss</text>
  </message>
  <message line="3">
    <author>e915385e151a03b29b5e797cb8d46847</author>
    <time>2019-07-02 00:47:00</time>
    <text>entendo</text>
  </message>
  <message line="4">
    <author>e915385e151a03b29b5e797cb8d46847</author>
    <time>2019-07-02 00:47:00</time>
    <text>Boa noite ae</text>
  </message>
</conversation>
```

Figura 4. Exemplo de conversa não-predatória presente no conjunto de dados PREDADORES-BR.

construção do conjunto de dados PREDADORES-BR sugere que a presença de conversas predatórias represente aproximadamente 4% do total de conversas. Tal desbalanceamento é justificado pelo fato de que as conversas predatórias são raras. A principal motivação em construir um conjunto de dados com essas características se justifica na possibilidade de fomentar diferentes campos de pesquisa. Também é sugerido que, para a seleção de conversas não-predatórias, sejam consideradas todas as conversas com até 150 mensagens. Após análise de todas as conversas extraídas dos servidores de bate-papo Discord, foi possível identificar um total de 163.632 conversas dentro do limite de mensagens estabelecido. A Figura 4 apresenta um exemplo de conversa não-predatória. Para o presente cenário, considerar todas as conversas aumentaria o desbalanceamento inicialmente proposto. Sendo assim, foi adicionada uma etapa para a seleção de conversas não-predatórias no método proposto.

Nesta etapa de seleção de conversas não-predatórias, a representatividade de cada categoria foi considerada por meio de uma amostragem estratificada em todas as conversas não-predatórias com até 150 mensagens, conforme recomendado na literatura [Biber 1993]. Para efetuar essa etapa, foi utilizada a biblioteca Scikit-learn [Pedregosa et al. 2011]. O conjunto de dados final após a etapa de seleção de mensagens não predatórias pode ser observado na Tabela 3. Esse conjunto de dados recebeu o

nome de PREDADORES-BR[10].

Tabela 3. Conjunto de dados PREDADORES-BR.

Classe	Conversas	Mensagens	Usuários únicos
Predatória	39	436	78
Não-Predatória	925	12.260	2.239

4. Análise estatística

Para uma melhor compreensão das conversas que compõem o conjunto de dados PREDADORES-BR, foi realizada uma análise estatística baseada na proposta de [Sokolova and Bobicev 2018]. Essa proposta considera a extração de medidas que permitam entender melhor a escala e diversidade dos dados. Isso possibilita uma melhor compreensão da complexidade dos dados, além de proporcionar comparações com outros conjuntos de dados.

Os resultados da análise são apresentados na Tabela 4. Observa-se uma diferença significante no volume de dados entre as classes disponíveis no conjunto de dados (i.e., predatória e não-predatória). Essa diferença é consequência do desbalanceamento proposto pelo método escolhido para criação do conjunto de dados. Também é possível observar que a quantidade de termos em conversas não-predatórias apresenta a média inferior ao desvio-padrão. Este comportamento se estende para a quantidade de termos por mensagem em ambas as classes de conversas. A ocorrência do fenômeno se deve ao fato dos termos não seguirem uma distribuição normal.

As conversas predatórias tendem a ser mais extensas e apresentar um volume maior de mensagens. Ao analisar as 436 mensagens enviadas em conversas com predadores sexuais, pode ser observado que 260 delas (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 predatória apresentou uma média pouco superior a 11 mensagens trocadas e um desvio padrão de 8,22. A maioria das conversas predatórias apresenta apenas 4 mensagens trocadas, porém é possível encontrar conversas contendo 41 mensagens.

O cálculo do percentual de termos raros (PTR) foi considerado para quantificar a riqueza do vocabulário do conjunto de dados. Entende-se por termos raros, aqueles que apresentam a frequência de uso abaixo de um limiar definido. Nesse contexto, o limiar pode ser definido de diferentes maneiras, dependendo do domínio. Por exemplo, um valor fixo de ocorrências em um vocabulário [Ponomareva and Thelwall 2012] ou a frequência inversa média das palavras [Sutskever et al. 2014]. O PTR pode ser definido conforme a

[10] `https://github.com/LaCAfe/PREDADORES-BR`

Tabela 4. Análise estatística do conjunto de dados PREDADORES-BR.

Característica	Classe predatória	Classe não-predatória	Total
Termos	2.370	49.704	52.074
Vocabulário	1.001	12.318	12.707
Número de mensagens	436	12.260	12.696
Termos por conversa (μ)	60,77	53,73	54,01
Termos por conversa (σ)	39,64	102,53	100,76
Termos por mensagem (μ)	5,44	4,05	4,10
Termos por mensagem (σ)	6,27	8,41	8,35
Hapax Legomena	690	8.132	8.366
Dis Legomena	137	1.731	1.796

Equação 1.

$$PTR = \frac{|\{t \in V | c(t) < 3\}|}{|V|} \tag{1}$$

Dado que V representa o vocabulário, $|V|$ é o tamanho do vocabulário, t o termo presente no vocabulário V e $c(t)$ corresponde ao número de ocorrências do termo t em um vocabulário V. No presente cenário, são considerados termos raros os *Hapax Legomena*[11] e *Dis Legomena*[12] [Sokolova and Bobicev 2018].

A alta ocorrência de termos raros em um corpus aumenta a complexidade em tarefas de classificação [Blitzer et al. 2006]. De forma a reduzir essa complexidade, a remoção de termos raros é, frequentemente, umas das ações realizadas no pré-processamento do corpus. Outro ponto que motiva a remoção é o entendimento de que termos raros podem ser descartados por não apresentarem impacto em tarefas de classificação [Yang and Pedersen 1997].

Ao considerar o vocabulário predatório (i.e., $|V| = 1.001$), 82,6% é considerado raro (*Hapax Legomena* + *Dis Legomena*). Por outro lado, as conversas e mensagens não-predatórias apresentam uma riqueza de vocabulário menor (80%). Por conta dos PTRs identificados no conjunto de dados, foram analisadas as sobreposições de termos raros e não-raros entre as classes predatórias e não-predatórias. Um alto percentual de termos sobrepostos também é um indicador a ser considerado ao avaliar a complexidade de uma tarefa de classificação [Scott and Matwin 1998]. Os resultados são apresentados na Tabela 5. É possível observar que, embora os percentuais de termos sobrepostos sejam baixos, a presença de termos raros em conversas predatórias é menor impactada com os efeitos de sobreposição (2%). Considerando os termos mais frequentes no conjunto de dados, é possível observar que dos 174 termos, 153 estão presentes no vocabulário de ambas as classes (87% dos termos). Esse fator contribui para o aumento da complexidade

[11]Hapax Legomena: termos que ocorrem somente uma única vez em todo o corpus.
[12]Dis Legomena: termos encontrados apenas duas vezes em todo o corpus.

em tarefas de classificação quando se considera a eleição de termos mais frequentes como estratégia de seleção de características para geração de um modelo.

Tabela 5. Sobreposição de termos presentes no conjunto de dados PREDADORES-BR.

Característica	Conversas Predatórias	Conversas Não-Predatórias	Sobreposição (% total)
Termos raros	827	9.863	165 (2%)
Demais termos (c(t)>2)	174	2.455	153 (5%)

5. Avaliação experimental

Esta seção descreve os experimentos executados com o conjunto de dados PREDADORES-BR. O objetivo inicial consiste em possibilitar que os resultados obtidos sirvam de base para a classificação de conversas predatórias na língua portuguesa do Brasil. A seção 5.1 descreve o ambiente de execução dos experimentos. A seção 5.2 discute o pré-processamento e seleção das conversas presentes no conjunto de dados. A seção 5.3 apresenta a configuração dos experimentos. A seção 5.4 descreve os algoritmos de aprendizado de máquina considerados para a execução dos experimentos. A seção 5.5 aborda a configuração utilizada para a seleção de hiperparâmetros. Por fim, a seção 5.6 discute os resultados encontrados.

5.1. Ambiente

Os experimentos foram realizados utilizando um computador com o processador Intel Core i9-9900k com 8 núcleos, 64GB de memória RAM DDR4 3200MHz e placa gráfica GeForce RTX 2080 Ti com 11Gb de memória. O Ubuntu 18.04 AMD64 foi a distribuição Linux usada no ambiente. Para a realização do experimento com a CNN foi considerada a biblioteca conText v4[13] [Johnson and Zhang 2015].

Os experimentos com os algoritmos SVM, MNB, DT e RF e a preparação dos dados de entrada para a CNN foram codificados com o auxílio da aplicação Web Jupyter Notebook 6.0.1 [Kluyver et al. 2016] e a biblioteca Scikit-learn 0.21.3. A linguagem Python (versão 3.7.4) foi considerada como padrão para toda a codificação realizada.

5.2. Pré-processamento e seleção de conversas

Conforme discutido na seção 3, o conjunto de dados denominado PREDADORES-BR apresenta um desbalanceamento em virtude do método proposto para a sua criação. Sendo assim, para a avaliação dos algoritmos, a primeira etapa da metodologia proposta consistiu no balanceamento das conversas não-predatórias por meio da técnica de amostragem sem substituição. Com isso, foram selecionadas 39 conversas não-predatórias. Englobando as 39 conversas predatórias, ao todo são consideradas 78 conversas.

[13] `http://riejohnson.com/cnn_download.html`

As conversas predatórias disponibilizadas pelo FEI e MPF-SP, conforme detalhadas na seção anterior, sofreram alterações de forma a preservar a identidade dos participantes. Sendo assim, a primeira etapa do pré-processamento buscou aplicar, por meio do uso de expressões regulares, as mesmas marcações nas conversas não-predatórias. Não foi necessário aplicar as marcações relacionadas ao envio de mensagens de áudio (“[audio]”) por conta de uma limitação do aplicativo Discord, que somente permite o envio de mensagens de áudio para mensagens particulares (i.e., entre duas pessoas) e não em canais públicos.

Em um momento posterior, buscou-se reduzir a quantidade de termos. Nesse cenário, foram removidos: *stopwords*, acentuação e pontuação. A conversão do texto para minúsculo também foi realizada. Por fim, foi realizada a conversão de todos os números presentes nas conversas para palavras. A motivação pela mudança remete ao comportamento identificado nas conversas predatórias em que vítimas informam suas idades, não somente por números, mas também por palavras.

5.3. Configuração dos experimentos

A criação dos modelos foi planejada considerando duas formas distintas de avaliação. Em um primeiro momento, a técnica de validação simples (Em inglês: *Holdout*) estratificada separando o conjunto de dados em 3 subconjuntos: treinamento (60%), validação (20%) e teste (20%). A escolha pelos 3 subconjuntos se fundamenta em [Bishop 2006] que estabelece que, para conjuntos de dados com tamanho limitado, o comportamento de sobreajuste pode ocorrer quando considerados apenas os subconjuntos de treinamento e validação. Outro ponto que motiva o uso da técnica neste cenário é o viés na acurácia dos resultados, uma vez que ocorreria a seleção dos hiperparâmetros durante o processo de validação cruzada [Varma and Simon 2006]. Sendo assim, a técnica de validação simples estratificada tem como propósito eleger os hiperparâmetros sem que ocorra sobreajuste durante a eleição.

Após a seleção dos hiperparâmetros, é aplicado o processo de validação cruzada estratificada em grupos. Neste contexto, foram considerados 5 grupos, destes, 1 grupo selecionado para teste e os demais para treinamento. Para a análise do desempenho dos modelos são consideradas as medidas de avaliação: acurácia, F_1, precisão e abrangência.

5.4. Algoritmos de aprendizado de máquina

Os experimentos consideraram a utilização de 5 algoritmos de aprendizado de máquina: Máquina de vetores de suporte (SVM), Naïve Bayes Multinomial (MNB), Árvores de decisão (DT), Florestas aleatórias (RF) e Redes neurais convolucionais (CNN). O esquema de ponderação de termos TF-IDF (*Term Frequency - Inverse Document Frequency*) foi utilizado para a escolha das características mais relevantes do conjunto de dados nos algoritmos SVM, MNB, DT e RF. A CNN fez uso da seleção dos N termos mais frequentes no conjunto de dados. O uso de termos mais frequentes como método de seleção de características se justifica na proposta de especialização de CNN considerada para uso [Johnson and Zhang 2015]. O objetivo dessa avaliação experimental é compreender como algoritmos reconhecidos na área de aprendizado de máquina se comportam na identificação da atividade predatória sexual na língua portuguesa do Brasil.

5.5. Seleção de Hiperparâmetros

A seleção dos hiperparâmetros dos algoritmos SVM, MNB, DT e RF foi feita uma busca extensiva com o auxílio do módulo GridSearchCV presente na biblioteca Scikit-learn e o conjunto de treinamento. Para cada algoritmo foi elaborada uma lista de hiperparâmetros específicos e faixas de valores para a realização da busca. Para o algoritmo SVM, foram selecionados 3 hiperparâmetros específicos: C = (0.1, 1, 10, 100, 1000), γ = (1, 0.1, 0.01, 0.001, 0.0001) e função de Kernel *k* = ("Linear", "Função de Base Radial", "Polinomial", "Sigmoidal"). O Algoritmo MNB explorou diferentes valores para o hiperparâmetro α = (1, 0.1, 0.01, 0.001, 0). O algoritmo DT explorou os seguintes critérios para medição da qualidade da divisão dos nós da árvore: *Índice de Gini* e *Ganho de Informação*. Além destes, também foram avaliadas árvores de diferentes profundidades, entre 3 e 50 níveis. O algoritmo RF explorou diferentes quantidades de árvores de decisão (5, 10, 25, 50, 100), assim como a profundidade máxima dessas árvores (2, 5, 6, 7, 8, 9, 10, 12, 15, 20, 30, 50).

Além dos hiperparâmetros específicos aplicados em cada algoritmo, para o esquema de ponderação de termos TF-IDF foram considerados os seguintes parâmetros e faixas de valores para exploração: mínima frequência do termo em um documento (0.01, 0.05, 0.1, 0.2), máxima frequência do termo em um documento (0.25, 0.30, 0.5, 0.75, 1.0), máximo de características a serem consideradas (Todos, 1000, 5000, 10000, 15000) e, considerando a aplicação do modelo de n-gramas para a geração de características (1G, 1G-2G, 1G-3G). No módulo GridSearchCV é possível definir uma medida "alvo", ou seja, ele retornará os hiperparâmetros que atingiram o maior valor na medida escolhida. Para a avaliação do modelo foi escolhida a medida F_1.

Diferente dos algoritmos de aprendizado de máquina apresentados anteriormente, que fizeram uso do módulo GridSearchCV para a exploração dos hiperparâmetros, a biblioteca conText v4, usada nos experimentos com CNN não apresenta funcionalidade similar, sendo assim foi necessário a implementação de uma função que permitisse realizar os experimentos. No entanto, para todos os experimentos, alguns hiperparâmetros foram considerados *default* para a eleição: 100 épocas para treinamento do modelo com a redução da taxa de aprendizado inicial a partir das épocas 80 (-90%) e 90 (-99%). A Tabela 6 apresenta os hiperparâmetros e faixa de valores.

5.6. Resultados

Esta seção tem por objetivo avaliar os experimentos realizados com os algoritmos de classificação selecionados. Na subseção 5.6.1 é detalhado o processo de seleção de hiperparâmetros; a subseção 5.6.2 aplica os hiperparâmetros selecionados no processo de validação cruzada estratificada.

5.6.1. Validação Simples

Ao analisar os resultados obtidos com o módulo GridSearchCV, é possível evidenciar alguns comportamentos. O algoritmo SVM apresentou resultados mais significativos com as funções de Kernel *k* não lineares. O resultado mais expressivo considerou *k* = "Sigmoidal". Ao avaliar o parâmetro C, responsável por influenciar diretamente na assertividade

Tabela 6. Hiperparâmetros considerados para a eleição de hiperparâmetros com a CNN.

Hiperparâmetro	Faixa de Valores
Camada de entrada (neurônios)	100, 500, 1.000
Regularização L2	10^{-3}, 10^{-4}, 10^{-5}
Dropout	0%, 10%, 50%
Termos mais frequentes	1.000, 5.000, 10.000
Taxa de aprendizado (TA)	0, 01, 0, 02, 0, 05
Otimizador	SGD, Rmsprop
momentum	0.5 , 0.9

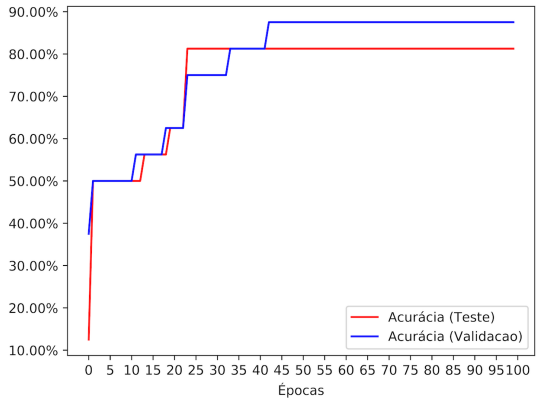
da classificação, é possível identificar o valor 0.1 como melhor resultado. Uma possível interpretação desse valor sugere que foi possível alcançar um fronteira de decisão bem definida entre as conversas predatórias e não-predatórias com uma maior capacidade de generalização. Outro parâmetro considerado para análise foi o γ. Este parâmetro influencia diretamente o comportamento de funções de kernel não lineares. O valor eleito (1) implica em considerar conversas predatórias e não-predatórias que possuam uma maior proximidade no espaço em detrimento das demais. Um possível efeito colateral desse valor é uma maior ocorrência de Falso-Positivos (FP) e Falso-Negativos (FN) ao se submeter novos dados para validação. Para a criação do esquema de ponderação de pesos TF-IDF, foi considerado um total de 10.000 termos. A partir desse total de termos, foram selecionados todos os termos com ocorrência mínima de 10% e máxima de 75% nas conversas presentes no conjunto de dados.

O algoritmo MNB, por sua vez, apresentou comportamento diferente para a seleção de características quando comparado ao algoritmo SVM. O hiperparâmetro α, responsável por regularizar as estimativas de probabilidade é considerado (0.1). Nesse caso, nenhuma probabilidade de ocorrência de um termo será zero. Com isso, os melhores resultados (considerando a medida F_1) fizeram uso de apenas 1.000 termos - unigramas somente, considerando os termos com ocorrência máxima de 50% em todas as conversas para a geração do esquema de ponderação de termos TF-IDF.

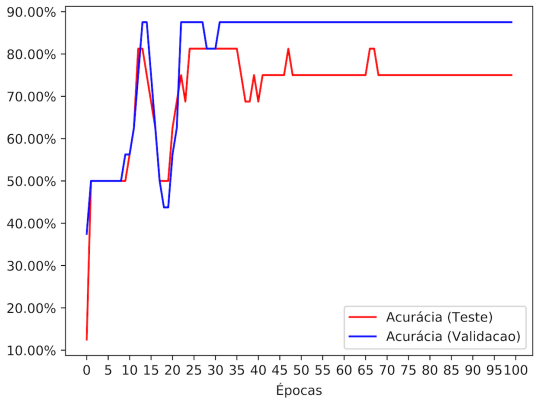
Para o algoritmo DT, o critério para a medição da qualidade da divisão dos nós da árvore não apresentou grande influência na decisão do modelo, embora a medida *Índice de Gini* tenha sido selecionada como a melhor recomendada para os dados. Um ponto importante que se destaca nesse algoritmo é o uso de termos de baixa frequência para compor a árvore de decisão. Embora tenha sido considerado um total de 15.000 termos, os melhores resultados foram obtidos com os termos presentes entre 10% e 25% do total de conversas do conjunto de dados.

As florestas aleatórias apresentaram uma configuração ótima com o total de 5 estimadores e a profundidade máxima igual à 12. Diferentemente do explorado com o algoritmo DT, para a criação do modelo de esquema de ponderação de pesos TF-IDF foram considerados um total de 10.000 termos. Destes, todos os termos com pelo menos

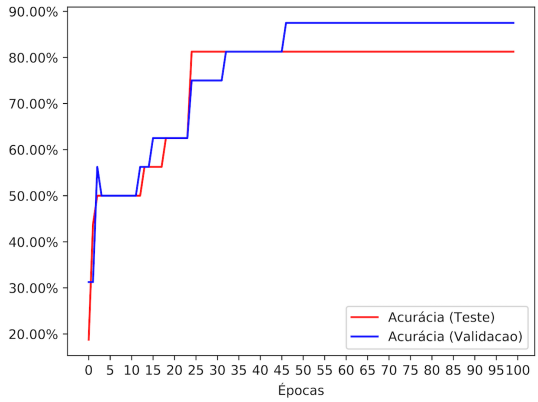
10% de frequência foram eleitos.



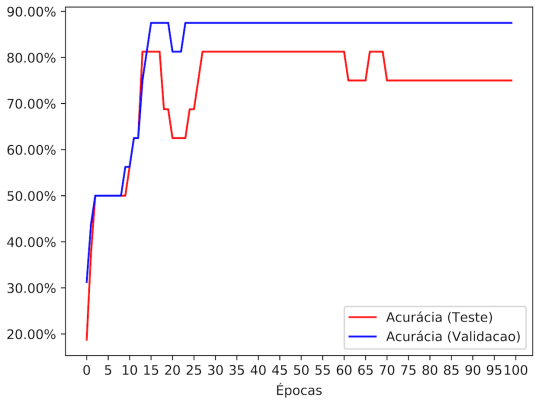
(a) *Dropout*: 10%; *momentum*: 0.5



(b) *Dropout*: 10%; *momentum*: 0.9



(c) *Dropout*: 50%; *momentum*: 0.5



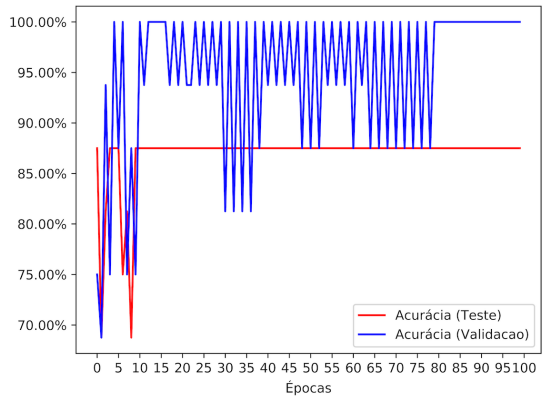
(d) *Dropout*: 50%; *momentum*: 0.9

Figura 5. Influência da variação da taxa de *Dropout* e do parâmetro *momentum* no treinamento de diferentes CNNs com o otimizador SGD e 1.000 termos mais frequentes.

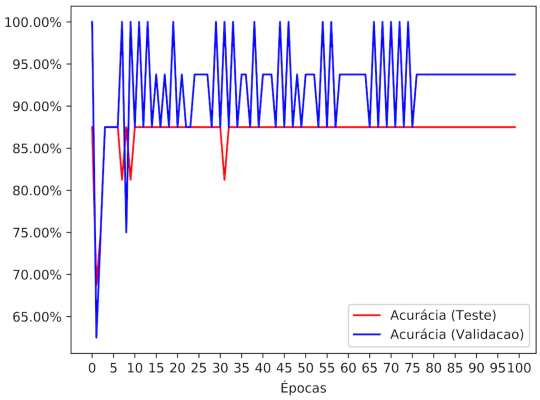
Ao avaliar o comportamento do treinamento das CNNs, foi possível observar que alguns hiperparâmetros não apresentaram influência significativa nos resultados. Nesse cenário, podemos citar a quantidade de neurônios da camada de entrada, o número de termos mais frequentes e a aplicação da regularização L2. No entanto, foi possível observar comportamentos distintos de acordo com o otimizador escolhido.

O otimizador SGD sofreu influência direta da taxa de *Dropout*, assim como dos parâmetros TA e *momentum*. A figura 5 ilustra o fenômeno. Todos os experimentos que consideraram uma TA menor que 0.05 não foram capazes de identificar nenhuma conversa predatória (F_1 = 0%). Ao considerar TA igual a 0.05, é possível observar como o parâmetro *momentum* e a taxa de *Dropout* atuam de forma a evitar um cenário de sobreajuste. O parâmetro *momentum*, responsável por acelerar o aprendizado, permitiu que os modelos (b) e (d) atingissem aproximadamente 90% de acurácia em aproximadamente 15 épocas. Os modelos (a) e (c), com a redução do *momentum* para 0.5, também atingiram o mesmo nível de acurácia, porém foi necessário um total de 45 épocas. Também é possível observar que, ao elevar a taxa de *Dropout* de 0.1 para 0.5 em conjunto com o parâmetro

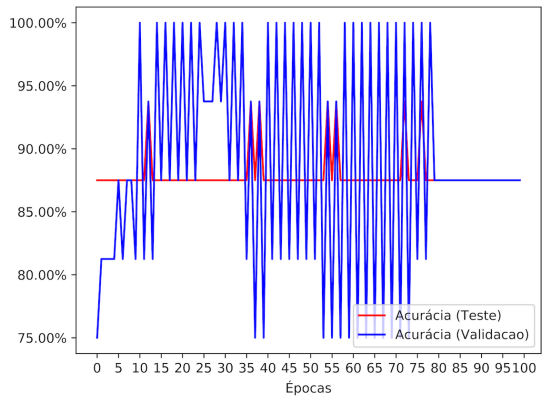
momentum igual à 0.9, o modelo (d) apresentou maior capacidade de generalização ao longo das épocas, atingindo acurácia máxima em pouco mais de 30 épocas.



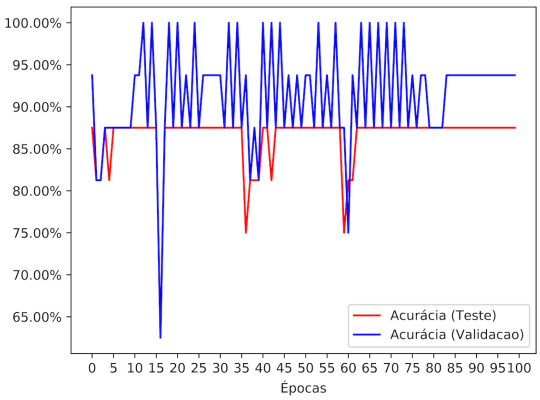
(a) TA: 0.01; *Dropout*: 10%



(b) TA: 0.01; *Dropout*: 50%



(c) TA: 0.05; *Dropout*: 10%



(d) TA: 0.05; *Dropout*: 50%

Figura 6. Influência da taxa de *Dropout* e taxa de aprendizado no treinamento de diferentes CNNs com o otimizador Rmsprop e 1.000 termos mais frequentes.

Ao analisar os resultados com o otimizador Rmsprop, é possível observar uma variação na acurácia do conjunto de validação ao longo das épocas em todos os cenários, conforme ilustrado na Figura 6. Essa variação é reduzida a medida em que a taxa de *Dropout* é elevada. Esse comportamento fica mais evidente nos modelos (b) e (d). Também é possível observar que a taxa de aprendizado aplicada em conjunto com o otimizador SGD é considerada alta para o otimizador Rmsprop, sugerindo um cenário de sobreajuste nos modelos (c) e (d). Vale ressaltar que a redução da taxa de aprendizado aplicada a partir da época 80 apresentou influência direta na estabilização da acurácia do modelo ao longo das épocas restantes. Nesse cenário, o modelo (a) apresentou o resultado mais significativo, atingindo 100% de acurácia ao final de 100 épocas.

Também é possível observar que a elevação da taxa de *Dropout* contribui para a garantia de generalização do modelo, no entanto, o resultado mais significativo (F_1 = 100%)

Tabela 7. Hiperparâmetros selecionados por meio do processo de validação simples que apresentaram os resultados mais significativos ao considerar a medida F_1.

Modelo	Hiperparâmetros
SVM	C = 0.1, γ = 1, *k* = "Sigmoidal" T = 10.000, Mín-TF = 10%, Máx-TF = 75%
MNB	α = 0.1; T = 1.000; Máx-TF = 50%
DT	Índice de Gini, T = 15.000, Mín-TF = 10%, Máx-TF = 25%
RF	Est. = 5, Prof. = 12, T = 10.000, Mín-TF = 10%
CNN-SGD	T = 1.000, TA = 0.05, Dropout = 50%, *momentum* = 0.9
CNN-RMSPROP	T = 1.000, TA = 0.01, Dropout = 10%

foi obtido com 10% de taxa de *Dropout*. Quando a taxa de *Dropout* e elevada à 50%, o modelo final apresenta uma pequena degradação da acurácia por conta da ocorrência de 1 Falso-Positivo.

5.6.2. Validação cruzada estratificada

Após a seleção dos hiperparâmetros para cada algoritmo de aprendizado de máquina apresentados na subseção anterior e sumarizados na Tabela 7, foi realizado um processo de validação cruzada estratificada. A motivação para tal se concentra na possibilidade de melhor estimar o viés e variância do modelo treinado [Kohavi 1995]. A escolha da quantidade de grupos é uma uma decisão importante no processo de validação cruzada. Um numero de grupos menor que o adequado pode implicar em um viés maior do modelo, assim como um valor maior pode ampliar a variância do modelo [Weiss and Kulikowski 1991]. Um fator que contribui para a escolha da quantidade de grupos é o grau de representatividade presente em cada um. Após uma análise empírica considerando diferentes quantidades possíveis de grupos, concluiu-se que 5 grupos seria uma quantidade ideal, tendo em vista o volume de dados no conjunto de dados. Dessa maneira foi efetuado o processo de validação cruzada e estratificada considerando um total de 5 grupos. Os resultados são apresentados na Tabela 8.

Ao analisar os resultados, é possível observar resultados promissores. O modelo SVM apresentou o melhor resultado (F_1 = 89.87%), seguido do modelo MNB (F_1 = 88.98%). Este resultado é influenciado pela capacidade de ambos os modelos em classificar uma maior quantidade de conversas predatórias corretamente. Esse característica presente em ambos os modelos pode ser observada na medida de Abrangência. No entanto, cada modelo fez uso de uma quantidade diferente de termos significativa como pôde ser observado na subseção 5.6.1. Entende-se que os resultados obtidos com o modelo MNB consolidam a importância de termos raros na identificação de predadores sexuais e se apresentam como fonte de informação relevante a ser explorada quando o conjuntos de dados apresenta um tamanho limitado. Após análise empírica, é possível observar

Tabela 8. Resultados obtidos por meio do processo de validação cruzada e estratificada com 5 grupos e aplicação dos hiperparâmetros selecionados e apresentados na Tabela 7.

Modelo	Acurácia	Precisão	Abrangência	F_1
SVM	90.00%	90.10%	92.50%	89.87%
MNB	88.57%	87.32%	92.50%	88.98%
DT	70.71%	79.81%	58.93%	62.19%
RF	87.50%	91.28%	82.50%	85.77%
CNN-SGD	86.25%	86.76%	87.50%	86.36%
CNN-RMSPROP	89.46%	100.00%	76.42%	85.92%

que uma quantidade significativa de termos raros presentes em conversas predatórias são oriundas de erros de ortografia e predadores sexuais usando o vocabulário de crianças e adolescentes (i.e. “istudo”, ‘iscola”, “BRAXAR”, “tah”). Sendo assim, conclui-se que o modelo SVM apresentará uma maior capacidade de generalização perante o modelo MNB.

O uso de CNNs no presente trabalho não obteve o mesmo desempenho quando comparado ao trabalho que introduziu o uso de CNNs no domínio da pesquisa para a língua portuguesa do Brasil [Santos and Guedes 2019]. Chama atenção a degradação de todas as medidas de desempenho, em particular, a queda de aproximadamente 15% na medida F_1. Dentre os resultados com o conjunto de dados PREDADORES-BR, os modelos CNN-SGD e CNN-RMSPROP apresentaram desempenho inferior perante os modelos SVM e MNB.

Vale ressaltar o resultado de 100% de precisão no modelo CNN-RMSPROP, sendo este o único o modelo dentre todos os avaliados sem a ocorrência de FPs. Essa característica do modelo CNN-RMSPROP é considerada importante, visto que, no mundo real, por conta de todo o tempo desprendido e da mobilização de profissionais da lei para atuar na investigação de um caso de suspeita de atividade predatória, a assertividade na identificação de uma conversa predatória é priorizada, e essa característica se encontra refletida na medida Precisão [Inches and Crestani 2012]. Os modelos que fizeram uso de CNN exploraram os 1.000 termos mais frequentes no conjunto de dados. É possível observar que, com uma maior quantidade de dados, os modelos baseados em CNN conseguiram resultados promissores pela capacidade da rede de considerar a ordem das palavras ao longo de uma conversa para, então, submetê-las as camadas de convolução. Embora seja um resultado promissor, é possível observar uma queda na abrangência, ocasionada pela maior incidência de FNs. No total, aproximadamente 25% das conversas predatórias não foram identificadas corretamente.

O modelo DT não obteve resultados satisfatórios. A ocorrência expressiva de FNs, isto é, conversas predatórias erroneamente identificadas como não-predatórias, provocou uma queda na medida F_1. Uma possibilidade para ocorrência desse fenômeno está na alta variância presente na escolha dos termos para compor a árvore. Ao comparar os resul-

tados do modelo DT com o modelo RF, é possível identificar uma maior capacidade de generalização das florestas aleatórias [Ross 1997]. Embora o modelo RF tenha apresentado resultados similares, porém inferiores ao modelo CNN-RMSPROP, esses podem ser considerados significativos devido a grande interpretabilidade das regras definidas para a identificação.

6. Conclusão e trabalhos futuros

Os predadores sexuais são uma ameaça à crianças e adolescentes na internet brasileira. Com o objetivo de contribuir nesse cenário, o presente trabalho apresentou três contribuições: o conjunto de dados PREDADORES-BR, uma análise descritiva desse conjunto de dados e uma análise experimental empregando cinco algoritmos de aprendizado de máquina para reconhecer a atividade predatória sexual no conjunto PREDADORES-BR.

A construção do conjunto de dados PREDADORES-BR explorou o uso de conversas regulares sobre os tópicos populares entre adolescentes no Brasil. Compreende-se que existe um enriquecimento dos dados perante ao PAN-2012-BR uma vez que é considerado o conceito de representatividade. A obtenção de conversas regulares é considerada uma tarefa complexa, porque, normalmente, apresentam informações sensíveis ou restritas às pessoas envolvidas nas conversas particulares [Inches and Crestani 2012]. Nesse cenário, a plataforma Discord se apresenta como uma relevante fonte de conversas regulares para o estudo da comunicação de adolescentes e jovens adultos na internet brasileira. Por meio das conversas dos servidores foi possível encontrar uma grande variedade de assuntos conversados na internet, o que abre espaço para o uso do conjunto de dados em outros domínios de pesquisa.

A presença de uma maior representatividade nos dados e ocorrência de termos em comum para ambas as categorias de conversas apresentaram um impacto mais significativo nos experimentos realizados com o algoritmo DT. Os experimentos com os algoritmos SVM e MNB apresentaram os resultados mais significativos ao considerar apenas a medida F_1. O algoritmo CNN em conjunto com o otimizador Rmsprop apresentou a melhor precisão dentre todos os experimentos realizados, o que o posiciona como alternativa relevante para aplicação em situações do mundo real. O algoritmo RF apresentou resultados inferiores aos realizados com o algoritmo CNN e o otimizador Rmsprop. Por conta da grande interpretabilidade das regras geradas com algoritmo RF entende-se que elas podem contribuir para a identificação de características com um maior grau de abstração, como as características comportamentais.

Uma limitação presente no trabalho consiste em apenas explorar características textuais, isto é, os termos usados nas mensagens trocadas. Por conta do volume de conversas considerada para análise - 78 conversas - a presença de novos termos pode impactar a identificação correta de conversas predatórias.

Para trabalhos futuros, a fim de buscar uma maior capacidade de generalização do modelo, pretende-se evoluir o método proposto para considerar características de maior nível, como as características comportamentais apresentadas por predadores sexuais e vítimas em uma conversa. A efetividade da exploração de características comportamen-

tais está diretamente relacionada ao volume de conversas predatórias obtidas. Nesse contexto, considerando a raridade do dado, entende-se que a obtenção de uma maior quantidade de conversas predatórias é essencial para a evolução no domínio da pesquisa. O uso da ferramenta LIWC [Pennebaker et al. 2001] e o uso de características psicolinguísticas também são uma possibilidade a ser explorada.

A metodologia proposta fez uso de um único método para a seleção de características: o esquema de ponderação de pesos TF-IDF. Nesse cenário, pretende-se explorar outros métodos para a seleção de termos, como por exemplo o teste estatístico χ^2, Ganho de informação e Informação mútua.

7. Agradecimentos

Ao Centro Universitário da Fundação Educacional Inaciana, em particular, o Prof. Dr. Rodrigo Filev Maia e ao Ministério Público Federal de São Paulo, especialmente à chefe do departamento de Crimes Cibernéticos, Adriana Shimabukuro, pela cooperação na disponibilização dos dados para a pesquisa. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

[Andrijauskas et al. 2017] 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 2018] 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*.

[Biber 1993] Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

[Bishop 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[Blitzer et al. 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

[Cano et al. 2014] Cano, A. E., Fernandez, M., and Alani, H. (2014). Detecting child grooming behaviour patterns on social media. In *International conference on social informatics*, pages 412–427. Springer.

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

[Cheong and Jensen 2015] Cheong, Y.-G. and Jensen, A. K. (2015). Detecting predatory behavior in game chats. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):220–232.

[Crystal 2002] Crystal, D. (2002). Language and the internet. *IEEE Transactions on Professional Communication*, 45(2):142–144.

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

[Ebrahimi 2016] Ebrahimi, M. (2016). *Automatic Identification of Online Predators in Chat Logs by Anomaly Detection and Deep Learning*. PhD thesis, Concordia University.

[Ebrahimi et al. 2016] 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.

[Ghosh et al. 2018] Ghosh, A. K., Badillo-Urquiola, K., Guha, S., LaViola Jr, J. J., and Wisniewski, P. J. (2018). Safety vs. surveillance: what children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 124. ACM.

[Hernandez et al. 2018] Hernandez, S. C. L. S., Lacsina, A. C., Ylade, M. C., Aldaba, J., Lam, H. Y., Estacio Jr, L. R., and Lopez, A. L. (2018). sexual exploitation and abuse of children online in the philippines: A review of online news and articles. *Acta Medica Philippina*, 52(4):306.

[Inches and Crestani 2012] 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 and Zhang 2015] 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 et al. 2019] 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.

[Kluyver et al. 2016] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B., editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.

[Kohavi 1995] Kohavi, R. (1995). Wrappers for performance enhancement and oblivious decision graphs. Technical report, Carnegie-Mellon Univ. Pittsburgh PA Dept. of Computer Science.

[Komesu and Tenani 2009] Komesu, F. and Tenani, L. (2009). Considerações sobre o conceito de"internetês"nos estudos da linguagem. *Linguagem em (Dis) curso*, 9(3):621–643.

[Livingstone et al. 2017] Livingstone, S., Ólafsson, K., Helsper, E. J., Lupiáñez-Villanueva, F., Veltri, G. A., and Folkvord, F. (2017). Maximizing opportunities and minimizing risks for children online: The role of digital skills in emerging strategies of parental mediation. *Journal of Communication*, 67(1):82–105.

[NCMEC 2017] 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 et al. 2018] 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 2014] 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.

[O'Connell 2003] O'Connell, R. (2003). A typology of child cybersexploitation and online grooming practices. *Preston, UK: University of Central Lancashire*.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pendar 2007] 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.

[Pennebaker et al. 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

[Ponomareva and Thelwall 2012] Ponomareva, N. and Thelwall, M. (2012). Biographies or blenders: Which resource is best for cross-domain sentiment analysis? In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–499. Springer.

[Ross 1997] Ross, S. M. (1997). *Introduction to Probability Models*. Academic Press, San Diego, CA, USA, sixth edition.

[Santos and Guedes 2019] Santos, L. F. d. and Guedes, G. P. (2019). Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na internet. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 143–154, Porto Alegre, RS, Brasil. SBC.

[Scott and Matwin 1998] Scott, S. and Matwin, S. (1998). Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*.

[Sokolova and Bobicev 2018] Sokolova, M. and Bobicev, V. (2018). Corpus statistics in text classification of online data. *arXiv preprint arXiv:1803.06390*.

[Sutskever et al. 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Varma and Simon 2006] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

[Villatoro-Tello et al. 2012] 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.

[Webb 2018] Webb, K. (2018). The world's most popular video game chat app is now worth more than $2 billion, as it gears up to take on

the makers of 'fortnite'. https://www.businessinsider.com/discord-funding-2-billion-value-2018-12 (Acessado em 17 de fevereiro de 2020).

[Weiss and Kulikowski 1991] Weiss, S. M. and Kulikowski, C. A. (1991). Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.

[Wolak et al. 2018] Wolak, J., Finkelhor, D., Walsh, W., and Treitman, L. (2018). Sextortion of minors: Characteristics and dynamics. *Journal of Adolescent Health*, 62(1):72–79.

[Yang and Pedersen 1997] Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. pages 412–420.