CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

IDENTIFICAÇÃO AUTOMÁTICA DE ATIVIDADE PREDATÓRIA SEXUAL EM CONVERSAS VIRTUAIS NO BRASIL

Leonardo Ferreira dos Santos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Professor D.Sc. Gustavo Paiva Guedes e Silva

Rio de Janeiro,

Abril 2021

IDENTIFICAÇÃO AUTOMÁTICA DE ATIVIDADE PREDATÓRIA SEXUAL EM CONVERSAS VIRTUAIS NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Leonardo Ferreira dos Santos

Banca Examinadora:

Presidente, Professor D.Sc. Gustavo Paiva Guedes e Silva (CEFET/RJ) (Orientador(a))

Professor D.Sc. Eduardo Bezerra da Silva (CEFET/RJ)

Professor D.Sc. Eduardo Soares Ogasawara (CEFET/RJ)

Professor D.Sc. Ronaldo Ribeiro Goldschmidt (IME)

Rio de Janeiro,

Abril 2021

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S237 Santos, Leonardo Ferreira dos
Identificação automática de atividade predatória sexual em conversas virtuais no Brasil / Leonardo Ferreira dos Santos — 2021.
135f. : il. (algumas color.), enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2021.
Bibliografia : f. 124-135
Orientador: Gustavo Paiva Guedes e Silva

1. Assédio virtual. 2. Pedofilia. 3. Crime por computador – Prevenção. 4. Aprendizado de máquina. 5. Crime sexual contra as crianças. I. Silva, Gustavo Paiva Guedes e (Orient.). II. Título.

CDD 302.343

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Ao meu filho, Lucas, minha fortaleza.

AGRADECIMENTOS

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

Agradeço ao FEI, em particular, o Prof. Dr. Rodrigo Filev Maia e ao MPF-SP, especialmente à Adriana Shimabukuro, pela cooperação na disponibilização dos dados para a presente pesquisa.

Agradeço a minha família, pelo apoio incondicional.

Agradeço ao Prof. Dr. Gustavo Paiva Guedes e Silva, por toda a orientação, paciência, disponibilidade e amizade.

Agradeço aos corpos docentes do PPCIC e PPPRO, por todos os ensinamentos.

Agradeço ao LaCAfe, pelos aprendizados compartilhados.

Agradeço ao Prof. Renato Campos Mauro, por me apresentar o PPCIC e todos os aconselhamentos ao longo da minha trajetória profissional e acadêmica.

Agradeço ao Prof. Dr. José Gomes de Carvalho Júnior, pela amizade que perdura desde a graduação.

Agradeço ao Prof. Antônio Francisco da Silva Júnior, um grande amigo que me fez abraçar novos desafios profissionais.

Agradeço ao Dr. Marcelo Beckmann, pelo incentivo à continuidade acadêmica no *stricto sensu*.

Agradeço aos amigos Allan Oliveira, André Gordirro, Antero Neto, Daniel Mazzuca, Ellison Alves, João Pereira, Laudelino Lima e Victor Apocalypse pelos momentos de diversão proporcionados durante os períodos de descanso.

RESUMO

Identificação automática de atividade predatória sexual em conversas virtuais no Brasil

O uso da internet por crianças e adolescentes possibilita o acesso a um conjunto de oportunidades para o seu autodesenvolvimento. O acesso à informação, material educacional, entretenimento e socialização são algumas das oportunidades que podem ser usufruídas. O uso de redes sociais é um dos principais canais para socialização na internet. Por meio da criação de um perfil público no momento de ingresso à rede social, crianças e adolescentes podem criar conexões com outros perfis e estabelecer comunicação por meio de conversas virtuais. Predadores sexuais, por sua vez, fazem uso de redes sociais com o propósito de ludibriar essas crianças e adolescentes, estabelecendo uma relação enganosa para posterior execução de diversas atividades criminosas, como a obtenção de conteúdo pornográfico, a extorsão e a prática do abuso sexual. Nesse cenário, diversos estudos vêm se concentrando na identificação de predadores sexuais na internet. Embora seja um domínio de pesquisa amplamente explorado, não foram encontrados trabalhos que considerem o estudo de conversas virtuais realizadas na língua portuguesa do Brasil. Diante do problema exposto, a presente pesquisa tem como o principal objetivo propor um método que apresente resultados significativos para a identificação de atividade predatória em conversas textuais e virtuais em Português do Brasil. Para atingir esse objetivo, foi considerado como base de estudos um conjunto de 82 conversas predatórias anonimizadas e oriundas de provas criminais presentes em processos judiciais. Após a análise das conversas predatórias, um total de dezenove características textuais e comportamentais foram identificadas e consideradas para a criação de um método para detecção de atividade predatória em conversas textuais denominado MDAP. Para a validação do método, foi criado um conjunto de dados com características similares as da competição PAN-2012, utilizando como base as 82 conversas predatórias. Os resultados obtidos, quando comparados aos métodos candidatos ao estado da arte para o domínio da pesquisa, comprovam a eficiência do método MDAP para a identificação de atividade predatória em conversas textuais, se apresentando como uma alternativa para promoção de um ambiente virtual mais seguro para crianças e adolescentes.

Palavras-chave: pedofilia, conversas virtuais, aprendizado de máquina

ABSTRACT

Automatic identification of sexual predatory activities in virtual conversations in Brazil

The use of the internet by children and adolescents provides access to a set of opportunities for their self-development. Access to information, educational material, entertainment, and socialization are some of the options available. The use of social networks is one of the main channels for socializing on the internet. By creating a public profile when joining the social network, children and adolescents can develop connections with other people and establish communication through virtual conversations. Sexual predators, in turn, make use of social networks to deceive these children and adolescents, establish a deceptive relationship for subsequent execution of various criminal activities, such as obtaining pornographic content, extortion, and the practice of sexual abuse. In this scenario, several studies have focused on the identification of sexual predators on the internet. Although it is a widely explored research domain, no studies considered the task of virtual conversations conducted in Brazil's Portuguese language. Given the problem introduced, the present research has its primary objective to propose a method that offers significant results for identifying predatory activity on textual and virtual conversations in Brazilian Portuguese. A set of 82 anonymous predatory chats and criminal evidence present in judicial proceedings was considered the basis for the domain studies to understand sexual predatory activity in Brazil better. After the analysis of predatory conversations, a total of nineteen textual and behavioral characteristics identified served as the basis for creating the MDAP method. A data set with similar properties compared to the data set of the PAN-2012 competition validated the proposed method, using 82 predatory conversations as a basis. Compared to the state-of-the-art candidate methods for the research domain, the results obtained prove the efficiency of the MDAP method for identifying predatory activity in textual conversations, presenting itself as an alternative to promote a safer virtual environment for children and adolescents.

Keywords: pedophilia, virtual conversations, machine learning

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de uma conversa disponibilizada no conjunto de dados usado como referência durante a competição PAN-2012. 25

Figura 2 – Exemplo de problemas oriundos da presença de ruídos em um conjunto de dados. Autor: [García et al., 2015] 28

Figura 3 – Fatores condicionantes de degradação do desempenho dos algoritmos de aprendizado de máquina. Autor: [García et al., 2015] 29

Figura 4 – Exemplo de problema de classificação binário em duas dimensões. Autor: [Joachims, 2002] 34

Figura 5 – Exemplo de uma rede neural artificial Perceptron. 35

Figura 6 – Anatomia de um neurônio artificial. 36

Figura 7 – Exemplo de uma matriz de confusão. 39

Figura 8 – Etapas para criação do conjunto de dados PRED-2050-ALL. 53

Figura 9 – Exemplo de uma comunidade Discord sobre jogos em ambiente virtual. 55

Figura 10 – Distribuição das conversas predatórias no conjunto de dados PRED-2050-ALL de acordo com a quantidade de mensagens. 58

Figura 11 – Distribuição das conversas não predatórias no conjunto de dados PRED-2050-ALL de acordo com a quantidade de mensagens. 58

Figura 12 – Trechos de conversas predatórias em que pode ser observada a variação na escrita de termos. 64

Figura 13 – MDAP: Método de Detecção de Atividade Predatória. 65

Figura 14 – Comparação do desempenho dos cinco algoritmos de aprendizado de máquina após aplicação dos métodos MDAP e *baseline* para cada um dos valores *k* considerados. 90

Figura 15 – Resultados obtidos após a aplicação do teste não-paramétrico de Wilcoxon com os diferentes resultados obtidos por meio da aplicação do MDAP e o *baseline*. 92
Figura 16 – Média da importância das 50 principais características quanto à aplicação dos métodos MDAP e *baseline*. 96
Figura 17 – Duzentas características mais importantes após à aplicação do MDAP e agrupadas em linha com a seleção das Top-*k* características importantes. 97
Figura 18 – Resultados obtidos com a aplicação do MDAP e o *baseline* ao considerar todos os termos presentes nos conjuntos de dados “0” e “50”. 101
Figura 19 – Resultados obtidos com a aplicação do MDAP e o *baseline* quando considerado o vocabulário sem a presença de termos raros oriundos dos conjuntos de dados “0” e “50”. 103
Figura 20 – Matriz de confusão dos experimentos realizados com a aplicação do MDAP em conjunto com o algoritmo SVM no conjunto de dados “50”. 106
Figura 21 – Matriz de confusão dos experimentos realizados com a aplicação do MDAP em conjunto com o algoritmo SVM no conjunto de dados “0”. 107

LISTA DE TABELAS

Tabela 1 – Critérios de inclusão e exclusão considerados para a elaboração do mapa sistemático. 43

Tabela 2 – Resumo dos trabalhos relacionados. 50

Tabela 3 – Marcações inseridas inicialmente na conversas disponibilizadas pelo MPF-SP e FEI para preservação de identidade de predadores sexuais e vítimas. 54

Tabela 4 – Conjunto de dados PRED-2050-ALL. 56

Tabela 5 – Análise estatística descritiva do conjunto de dados PRED-2050-ALL. 57

Tabela 6 – Sobreposição de termos presentes no conjunto de dados PRED-2050-ALL. 60

Tabela 7 – Características exploradas na identificação da atividade predatória sexual pelo MDAP. 66

Tabela 8 – Léxicos criados a partir de fontes externas. 81

Tabela 9 – Exemplo de entradas presentes nos léxicos criados a partir de fontes externas. 81

Tabela 10 – Léxicos criados a partir de conversas predatórias (Origem interna). 83

Tabela 11 – Entradas presentes nos léxicos criados a partir de conversas predatórias (Origem interna). 83

Tabela 12 – Padrões textuais considerados para o mapeamento de CANs. 84

Tabela 13 – Resultados da aplicação do teste de Friedman em todos os experimentos realizados com a aplicação do método MDAP. 91

Tabela 14 – Resultados obtidos após a aplicação do teste *post-hoc* de Holm e $p = 0,05$. 93

Tabela 15 – Análise estatística dos conjunto de dados “0” e “50” após a aplicação do métodos MDAP e *baseline*. 99

Tabela 16 – Sobreposição de termos entre as classes de conversas presentes nos conjunto de dados “0” e “50”. 100
Tabela 17 – Desempenho dos algoritmos de aprendizado de máquina após a aplicação dos métodos MDAP e *baseline* ao considerar como limiar para a seleção de características todos os termos presentes nos conjuntos de dados. 102
Tabela 18 – Desempenho dos algoritmos de aprendizado de máquina após a aplicação dos métodos MDAP e *baseline* ao considerar como limiar para a seleção de características a remoção dos termos raros presentes nos conjuntos de dados. 103
Tabela 19 – Desempenho do algoritmo SVM após a aplicação dos métodos MDAP nos conjuntos de dados “0” e “50”. 105
Tabela 20 – Conversas classificadas incorretamente como pertencentes à classe “Predatória” (FP) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “50”. 109
Tabela 21 – Conversas classificadas incorretamente como pertencentes à classe “Predatória” (FP) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “0”. 110
Tabela 22 – Conversas predatórias classificadas incorretamente que apresentaram a medida TEM superior à 50% após a aplicação MDAP nos conjuntos de dados “0”ou “50”. 111
Tabela 23 – Resultados obtidos com a avaliação experimental. 121
Tabela 24 – Conversas classificadas incorretamente como pertencentes à classe “Não Predatória” (FN) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “50”. 122
Tabela 25 – Conversas classificadas incorretamente como pertencentes à classe “Não Predatória” (FN) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “0”. 123

LISTA DE ALGORITMOS

Algoritmo 1 – $mod\_mpcti(C_{xml}, DL)$ 68

Algoritmo 2 – $mod\_micp(C_d, concSet_{inc}, concSet_{exc}, DL_n, DP)$ 69

Algoritmo 3 – $apply\_concept\_for\_age\_verification(C_d, L_n)$ 71

Algoritmo 4 – $apply\_concept\_for\_photo\_interest(C_d, L_n, P_n)$ 72

Algoritmo 5 – $apply\_concept\_for\_location\_interest(C_d, L_n, P_n)$ 73

Algoritmo 6 – $mod\_mpctf(C_{micp})$ 75

Algoritmo 7 – $met\_mdap(DS_{xml}, concSet_{inc}, concSet_{exc}, DL, DP)$ 75

SUMÁRIO

Introdução **16**

1 Referencial Teórico **21**

1.1 O perfil do predador sexual na internet 21

1.1.1 As características da comunicação do predador sexual na internet 22

1.1.2 PAN-2012 23

1.2 Recuperação da Informação 25

1.3 Mineração de dados 27

1.3.1 Pré-processamento 27

1.3.2 Classificação 31

1.3.3 Avaliação e seleção de modelos de classificação 37

1.4 Considerações finais 41

2 Trabalhos Relacionados **42**

2.1 Mapa Sistemático 42

2.2 Identificação de atividade predatória sexual em conversas virtuais 44

2.3 Considerações finais 51

3 Conjunto de dados **52**

3.1 Motivação 52

3.2 Conversas predatórias 53

3.3 Conversas não predatórias 55

3.4 Análise estatística 56

3.4.1 Percentual de termos raros 58

3.5 Considerações finais 60

4 Metodologia **61**

4.1 MDAP: Método de Detecção de Atividade Predatória 61

4.1.1 MPCTI: Módulo de Padronização de Conteúdo Textual Inicial 67
4.1.2 MICP: Módulo de Identificação de Comportamento Predatório 68
4.1.3 MPCTF: Módulo de Padronização de Conteúdo Textual Final 74
4.2 MDAP: Proposta de implementação 75
4.3 Considerações finais 76

5 Avaliação Experimental 78
5.1 Léxicos 78
5.1.1 Origem externa 79
5.1.2 Origem interna 82
5.2 Configuração dos experimentos 83
5.2.1 Algoritmos de aprendizado de máquina 84
5.2.2 Critérios de avaliação 85
5.2.3 Método de validação 85
5.2.4 Definição do vocabulário e seleção de características 86
5.3 Resultados 87
5.3.1 Visão geral dos resultados 88
5.3.2 Análise dos conjuntos selecionados 98
5.3.3 Experimentos realizados 100
5.3.4 Discussão dos experimentos 103

6 Conclusão 115
6.1 Resumo das contribuições 116
6.2 Limitações 117
6.3 Trabalhos futuros 118

A Apêndice 121

Referências 121

Introdução

Recentemente, a pesquisa *TIC Kids Online Brasil* [BARBOSA, 2018], que estuda os riscos e as oportunidades presentes no comportamento de crianças e adolescentes brasileiros na internet, evidencia um cenário otimista quanto à adoção do uso da internet porém, ao mesmo tempo, traz alguns alertas. De acordo com essa pesquisa, 85% das crianças e adolescentes com idades entre 9 e 17 anos possuem acesso à internet. Vale destacar que, dentre essas crianças e adolescentes, 76% pertencem às classes A e B e 44% possuem os dispositivos móveis como o único meio de acesso à internet. Quando questionadas sobre o que consideram um incômodo na internet, crianças e adolescentes pouco mencionaram o risco de contato ou assédio de pessoas estranhas, indesejadas ou adultos desconhecidos (10% do total), quando comparado 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 últimos anos conclui a existência de uma ameaça real para crianças e adolescentes [BARBOSA, 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, visto que as crianças e os adolescentes almejam encontrar na internet um ambiente para terem liberdade e, além disso, querem mostrar que são capazes de tomar boas decisões e aprender com os erros [Ghosh et al., 2018].

A liberdade oferecida aos usuários de internet pode promover acesso a diferentes oportunidades, no entanto, também favorece uma exposição a variados riscos [van der Hof and Koops, 2011]. Dentre as oportunidades, pode-se destacar a possibilidade de conhecer novas pessoas e o desenvolvimento de novas amizades [Ellison and Boyd, 2013], comportamento considerado básico não somente para crianças e adolescentes mas para todos os seres humanos [Leary and Baumeister, 2000]. Em adição a isso, o acesso global à informação, à educação e ao entretenimento também contribuem para um maior uso do ambiente virtual [Hasebrink et al., 2009]. Uma preocupação, vale ressaltar,

é que as redes sociais não estão isentas dos mesmos riscos sociais que podem ocorrer a partir da interação com uma ou várias pessoas presencialmente [Clark et al., 2018]. Um fenômeno que atua como agravante nesse contexto é a distância psicológica promovida pela internet, que atua como um eventual inibidor do senso moral e social, aumentando a ocorrência de comportamentos não éticos [Crowell et al., 2008]. Nesse caso, um fator que contribui como iniciativa é o sentimento de anonimidade que a internet oferece [Soh et al., 2018]. Essa anonimidade pode se refletir em diversos níveis, desde o uso de um apelido em uma sala de bate-papo, como em construções de perfis falsos em redes sociais [Keipi, 2018]. Nesse cenário, destacam-se o *Bullying* virtual[1], *Sexting*[2] e atividades predatórias.

O uso da internet por crianças e adolescentes tornam o meio virtual um canal amplamente explorado por predadores sexuais [Hernandez et al., 2018]. Dessa forma, a possibilidade de contato entre um predador sexual e uma criança ou adolescente se torna uma preocupação global [Olowu, 2014; Dorasamy et al., 2018; Kloess et al., 2019]. Essa preocupação é crescente na medida em que as aplicações sociais *online* se tornam mais acessíveis à massa, permitindo cada vez mais que as crianças e os adolescentes consigam usá-las sem grandes dificuldades.

A atividade predatória sexual apresenta diversas finalidades, conforme identificadas na literatura [NCMEC, 2017]: em 78% das atividades predatórias sexuais, o principal objetivo é a obtenção de conteúdo pornográfico da vítima, enviado por meio de aplicativos de conversas instantâneas; em menor proporção, em 7% das atividades predatórias sexuais, além da obtenção de conteúdo pornográfico, a vítima também sofre sextorsão[3], esta normalmente caracterizada por pedidos de transferências financeiras, disponibilização de dados de cartão de crédito dos pais e outros pertences; por fim, em 5% dos casos de atividade predatória sexual busca-se o contato *in loco* entre o predador sexual e a vítima. A finalidade desse encontro é realização do abuso sexual; os 10% restantes apresentam outras finalidades que não se enquadram nessas principais razões para a realização do contato.

O impacto da atividade predatória sexual pode ser devastador para um menor de

[1] O *Bullying* virtual pode ser definido como uma ação agressiva e intencional realizada por uma pessoa ou grupo de pessoas ao longo do tempo, por meio de dispositivos com acesso à internet, contra uma vítima incapaz de se defender totalmente [Slonje and Smith, 2008].

[2] O “Sexting” é uma prática ilegal que consiste na prática de crianças e adolescentes tirarem fotos ou gravarem vídeos próprios com teor de nudez e, em seguida, compartilharem com outros que, por sua vez, divulgam em redes sociais na internet [McLaughlin, 2010].

[3] A sextorsão pode ser compreendida como a realização de ameaças após obtenção de conteúdo com teor de nudez da vítima, de maneira que a vítima acaba sendo forçada a produzir e enviar mais conteúdo pornográfico [Wolak et al., 2018].

idade. A ocorrência do abuso sexual destrói a espontaneidade e a liberdade da criança ou do adolescente, provocando um terror solitário [Bagley and King, 2003]. As vítimas de abuso sexual originados na internet tem maior probabilidade de desenvolver distúrbios mentais, tais como a depressão, estresse agudo, estresse pós-traumático, transtornos de conversão[4] podendo levar à tentativas de suicídio [Say et al., 2015].

Diante do cenário apresentado, diversos estudos têm sido conduzidos com o objetivo de identificar predadores sexuais na internet [Pendar, 2007; Morris, 2013; Liu et al., 2017]. A maioria dos trabalhos encontrados na literatura foi viabilizada por conta das conversas predatórias disponibilizadas pela organização *Perverted-Justice* (PJ)[5]. Essas conversas são concebidas por agentes federais que atuam na internet como crianças e adolescentes, de forma a servir de iscas para predadores sexuais convictos e, dessa maneira, permitem o estudo de diversas características de predadores sexuais [Cardei and Rebedea, 2017; Bogdanova et al., 2014]. Esses estudos viabilizaram o desenvolvimento de diferentes meios automáticos para a identificação de atividade predatória na internet.

No entanto, uma limitação frequente nos trabalhos realizados no domínio da pesquisa é a carência de dados reais, isto é, conversas que ocorreram entre predadores sexuais e vítimas reais, sendo as vítimas crianças ou adolescentes. Nesse cenário, o principal impedimento para a divulgação dos dados reais é o teor sensível dos dados, proveniente de provas criminais presentes em processos que correm em sigilo na justiça. Por não serem disponibilizados como artefatos resultantes da pesquisa, acabam prejudicando a reprodutibilidade e a sua evolução.

Ao considerar o domínio da pesquisa no cenário nacional, também observa-se uma ausência de trabalhos que busquem identificar a atividade predatória em conversas na língua portuguesa do Brasil. A única iniciativa observada no domínio da pesquisa, se encontra em Andrijauskas et al. [2017] com o desenvolvimento de uma base de dados com conversas predatórias. Esse desenvolvimento ocorreu após a criação de uma parceria entre o Ministério Público Federal de São Paulo (MPF-SP) e o Centro Universitário da Fundação Educacional Inaciana (FEI). Um ponto importante, vale ressaltar que o trabalho de Andrijauskas et al. [2017] não propõe um método para a identificação de conversas predatórias. Objetivando preencher a lacuna observada, o presente trabalho constrói um método para identificar a ocorrência de atividade predatória sexual em

[4] O transtorno de conversão pode ser definido como a apresentação de sintomas físicos, oriundos de estresse e outros conflitos, similares a um transtorno do sistema nervoso. [MSD, 2018]

[5] `https://www.perverted-justice.com`

conversas textuais na língua portuguesa do Brasil. O método se baseia na exploração de características textuais e comportamentais presentes na comunicação do predador sexual brasileiro e as vítimas. Um ponto importante é que, para o melhor do nosso conhecimento, este é o primeiro trabalho a fazer uso de um conjunto de dados com conversas predatórias reais na língua portuguesa do Brasil.

Dado esse contexto, é considerada a seguinte pergunta de pesquisa: é possível melhorar o desempenho de algoritmos de aprendizado de máquina utilizando características textuais e comportamentais, representadas por meio de conceitos de alto nível, para a identificação de atividade predatória em conversas virtuais na língua portuguesa do Brasil? Para responder a essa pergunta, parte-se do princípio que o desempenho de um modelo de classificação de textos (e.g., documentos) para um domínio específico pode variar devido às especificidades presentes na escrita de cada idioma, como por exemplo, regionalismos e gírias [Karan and Šnajder, 2018; Chen et al., 2018]. Desta forma, as contribuições resultantes desta pesquisa são:

- A criação de um método para a identificação de atividade predatória sexual, em conversas no formato de texto provenientes da internet.
- A criação de um conjunto de dados denominado PRED-2050-ALL, que contém: (a) conversas textuais entre predadores sexuais e suas vítimas; (b) conversas não predatórias. As conversas não predatórias são provenientes de uma colaboração efetuada entre o FEI e o MPF-SP. Para compor as conversas não predatórias do PRED-2050-ALL, foi criado um método baseado na competição PAN-2012 [Inches and Crestani, 2012]. Esse método instrumenta os passos para extração, transformação e seleção de conversas presentes em comunidades virtuais.
- O desenvolvimento de léxicos provenientes de diferentes fontes de dados para auxiliar a representação dos diferentes conceitos de alto nível presentes no domínio da pesquisa.

Algumas das contribuições acima encontram-se iniciadas nos presentes artigos ao longo do desenvolvimento do presente trabalho:

- Santos, L. 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).

- Santos, L. and Guedes, G. P. (2019). Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na internet. In: 8o. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2019).

- Santos, L. and Guedes, G. P. (2020). Identificação de predadores sexuais brasileiros em conversas textuais na internet por meio de aprendizagem de máquina. iSys - Revista Brasileira de Sistemas de Informação (Vol. 13 No. 4).

Além desta introdução, segue a organização do restante do trabalho. O capítulo 1, apresenta os conceitos relacionados ao domínio da pesquisa, as técnicas de mineração de dados, medidas de desempenho e os algoritmos de aprendizado de máquina aplicados nos experimentos; O capítulo 2 detalha todos os trabalhos pesquisados e considerados como base para a presente pesquisa; O capítulo 3 apresenta o conjunto de dados PRED-2050-ALL e suas características por meio de uma análise estatística. O capítulo 4 descreve a principal contribuição desse trabalho. O método proposto, denominado MDAP, é apresentado. No capítulo 5, a avaliação experimental é detalhada. Inicialmente é apresentada a configuração dos experimentos e, logo após, é realizada uma análise comparativa do MDAP com um *baseline* previamente definido. Nesse cenário, são aplicados testes de significância estatística com o propósito de melhor compreender o desempenho do MDAP após a realização de diferentes experimentos. Em seguida, são analisados e discutidos os resultados mais expressivos e as dificuldades encontradas. Por fim, no capítulo 6 são apresentadas as conclusões, as limitações identificadas ao longo da pesquisa e possibilidades de trabalhos futuros.

1- Referencial Teórico

O presente capítulo tem por objetivo apresentar os conceitos e técnicas aplicadas ao decorrer do desenvolvimento da presente pesquisa. Inicialmente, na Seção 1.1, é apresentado o perfil do predador sexual que atua na internet. Na descrição do perfil também são detalhadas as características frequentemente observadas na comunicação do predador sexual e suas vítimas, dado que essas características são de grande importância para a identificação do perfil predatório conforme apresentado ao longo dos demais capítulos. Na Seção 1.1.2 apresenta a proposta do evento PAN, ocorrida no ano de 2012, e a sua importância para o domínio da pesquisa. A Seção 1.2 introduz alguns conceitos relacionados à teoria da Recuperação da Informação aplicados na presente pesquisa. Na Seção 1.3 são discutidos diferentes conceitos relacionados à Mineração de Dados. Por fim, a Seção 1.4 apresenta as considerações finais.

1.1- O perfil do predador sexual na internet

O perfil do predador sexual que atua na internet tem sido amplamente estudado em todo o mundo. Normalmente, uma atividade predatória sexual pode ser caracterizada pelos seguintes pontos [Morris, 2013]: 1) Disparidade de idade entre as partes de uma conversa. Uma configuração típica de uma conversa predatória consiste de um adulto, isto é, uma pessoa responsável legalmente por seus atos e um menor de idade; 2) Intimidade inapropriada, ou seja, estímulo ou iniciativa a um teor de conversa que promova uma intimidade artificial e forçada. Uma vez essa intimidade estabelecida, é comum que o predador queira estender a relação recém-criada para um encontro *in loco*. A comunicação realizada normalmente apresenta algum teor sexual. É comum que inicialmente haja ocorrência de vocabulário moderado e, uma vez que se crie uma relação de confiança, as conversas comecem a apresentar maior frequência de elementos com teor sexual [Olson et al., 2007].

Em geral, os predadores sexuais são homens, solteiros e desempregados e, não

apresentam uma idade superior a 26 anos [Fortin et al., 2018; Babchishin et al., 2011] e apresentam traços de transtornos mentais [Sinnamon, 2017]. Sobre a escolha das vítimas, idade e gênero costumam ser os fatores decisórios, mas também são consideradas as que apresentam um perfil vulnerável e que aparentam ser alvos fáceis, como crianças ou adolescentes solitários e carentes de atenção, com sobrepeso e insatisfeitas com a sua própria imagem [Tener et al., 2015]. Assim, adolescentes do gênero feminino com idade entre 13 e 15 anos se apresentam como o grupo de maior risco [Tener et al., 2015]. Vale destacar que os predadores sexuais não costumam apresentar comportamento violento [Wolak et al., 2004] ou histórico criminal [Seto et al., 2011]. Por fim, sobre a questão comportamental, os predadores sexuais apresentam alto grau de empatia com as vítimas [Babchishin et al., 2011] e não são impulsivos [Wolak et al., 2009].

1.1.1 As características da comunicação do predador sexual na internet

Para os propósitos do domínio da pesquisa, uma característica é denominada predatória quando esta se encontra em um conjunto de padrões recorrentes na comunicação de um predador sexual e uma vítima em uma conversa textual. O uso de termos específicos em um determinado contexto e a realização de perguntas sobre determinados assuntos em conversas textuais são, por exemplo, alguns dos padrões na comunicação. Essas características predatórias podem ter origem em ambos os participantes de uma conversa textual, ou seja, tanto um predador sexual quanto uma vítima podem ser o criador do padrão presente na comunicação a ser considerado na interpretação da ocorrência de atividade predatória. Nessa perspectiva, diferentes características predatórias podem ser identificadas ao longo da realização de uma conversa textual.

Uma vez iniciada uma conversa predatória, é observado o envio de saudações à vítima [Lorenzo-Dus et al., 2016]. Este movimento se caracteriza, tal como em uma conversa regular, como um meio de ter a atenção e desenvolver uma abertura para iniciar a conversa. Em alguns cenários, a saudação é seguida de elogios [Black et al., 2015; Gunawan et al., 2016]. Os elogios atuam como estratégia de dissuasão para estabelecer uma conversa, assim como iniciar a dessensibilização da vítima, sendo esta uma das formas de se estabelecer uma falsa relação de confiança [Lorenzo-Dus and Izura, 2017].

Uma vez que ocorra uma demonstração de receptividade, geralmente inicia-se uma sequência de perguntas [Dhouioui et al., 2016] com o intuito de descobrir informações sobre a vítima como, por exemplo, o seu nome [Cardei and Rebedea, 2017]. Nesse contexto, também são observadas a troca de informações sobre a idade, visto que um dos principais objetivos do predador sexual é encontrar vítimas que sejam menores de idade [Gillam and Vartapetiance, 2012]. Além da idade, também são compartilhados os números de telefone [Winters et al., 2017]. Em cenários que existe a intenção do abuso sexual físico também é possível encontrar menções a locais [Olson et al., 2007].

Ao longo do processo de dessensibilização da vítima [Olson et al., 2007] é possível observar o uso de Emojis[1][Cano et al., 2014]. Também é observada a presença de questionamentos frequentes em conversas predatórias. Por exemplo, é possível observar questionamentos sobre a presença ou proximidade de pais e parentes [Hidalgo and Díaz, 2012] e o cômodo da casa em que a vítima se encontra no momento da conversa [Black et al., 2015]. Uma vez que o predador sexual se considera seguro a prosseguir com o abuso sexual virtual, é frequente a ocorrência de perguntas [Gunawan et al., 2016]. Dentre as perguntas realizadas, é comum ocorrer questionamentos sobre as peças de roupa que a vítima esteja usando no momento [Kontostathis et al., 2012].

Em estágios avançados da atividade predatória é possível identificar perguntas e discussões com alto teor adulto [Black et al., 2015] assim como o interesse na obtenção de conteúdo pornográfico da vítima [Cardei and Rebedea, 2017]. Também é observado que o predador sexual pode adotar um tom imperativo ao ordenar determinadas ações à vítima, uma vez que se observe a relutância por parte da vítima no atendimento de algum pedido feito [Bogdanova et al., 2014].

1.1.2 PAN-2012

A PAN é um evento colocado na conferência CLEF - um acrônimo para *Conference and Labs for Evaluation Forum* - composta por diferentes eventos científicos e disponibilização de tarefas no âmbito da análise forense e estilométrica de documentos encontrados na internet (no formato de uma competição). Nos últimos 10 anos, a

[1] https://unicode.org/emoji/charts/full-emoji-list.html

competição fomentou a pesquisa em uma grande variedade de temas como: identificação de autoria, caracterização de perfis, plagiarismo e identificação de atividades enganosas.

No que se refere à identificação de atividades enganosas, a PAN-2012 [Inches and Crestani, 2012] organizou uma competição com o objetivo de identificar predadores sexuais praticando aliciamento em conversas na internet. Para atingir tal objetivo, foi disponibilizado um grande conjunto de dados com o propósito de não somente atender à finalidade da competição mas servir como uma referência para o testes de diferentes abordagens oriundas das mais diversas áreas de pesquisa. A tarefa apresentava dois problemas para a solução:

1. Identificar predadores sexuais em conversas textuais na internet
2. Dada uma conversa predatória, identificar as mensagens enviadas por um predador sexual

O conjunto de dados criado para a competição considerou as características realísticas do domínio da pesquisa: pouca ocorrência de conversas predatórias, um grande número de conversas com alto teor adulto ou tópicos que apresentem alguma similaridade com a comunicação realizada pelo predador sexual e uma quantidade ainda maior de conversas com temas variados. A Figura 1 apresenta um exemplo de conversa predatória seguindo o formato definido pela competição PAN-2012. Para a obtenção de conversas predatórias, não foram consideradas conversas entre predadores e crianças ou adolescentes pois, historicamente, as agências policiais apresentam uma postura relutante e não colaborativa sobre o tópico [Inches and Crestani, 2012]. Diante deste cenário, foram consideradas as conversas disponibilizadas no site *Perverted-Justice.com* (PJ). No site, são apresentados os perfis de predadores sexuais identificados por meio de conversas com agentes federais voluntários que agem como crianças e adolescentes em salas de bate-papo. As informações obtidas sobre o criminoso são disponibilizadas para o público assim como as conversas registradas. Além do PJ, para a construção do conjunto de dados foram consideradas as seguintes fontes de dados:

1. Omegle[2]: Conversas abusivas e de teor sexual porém não consideradas predatórias.
2. Canais de IRC[3][4]: Conversas não predatórias sobre diversos tópicos.

[2] http://web.archive.org/web/20100710040611/http://omegle.inportb.com/
[3] http://web.archive.org/web/20080704032104/http://www.irclog.org/
[4] https://krijnhoetmer.nl/irc-logs/

```
<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 1 – Exemplo de uma conversa disponibilizada no conjunto de dados usado como referência durante a competição PAN-2012. A conversa, na língua inglesa, apresenta dois participantes anonimizados interagindo por meio de cumprimentos (e.g., “hi”) e questionamentos sobre idade, gênero e localização (i.e., “asl”, “m or f”).

1.2- Recuperação da Informação

Segundo Manning et al. [2008], o conceito de Recuperação da Informação (RI) pode ser descrito como a capacidade de encontrar um ou mais documentos relevantes, de acordo com um critério de busca definido, dentre uma coleção de documentos que apresentam uma natureza similar. Nesse contexto, Han et al. [2011] acrescenta que: (i) os dados presentes em documentos não são estruturados, isto é, são dados que apresentam uma determinada semântica, porém não dispõem de uma organização para que seja de compreensão imediata para uma máquina; (ii) os documentos são recuperados por meio de palavras-chave. De forma a permitir a recuperação de um ou vários documentos a partir de uma coleção de documentos, diferentes modelos de linguagem podem ser considerados. Um modelo de linguagem pode ser dependente de um contexto específico, no entanto, é importante dizer que o modelo de linguagem permite quantificar diferentes

incertezas presentes em aplicações de linguagem natural [Zhai, 2008].

O modelo de linguagem mais simples é o modelo baseado apenas em unigramas. Nesse modelo, todo o contexto é desconsiderado e cada termo presente no documento tratado de forma independente. Desta forma, a ordem dos termos não é considerada e a estrutura do documento é ignorada. Uma generalização desse modelo é o *Bag of words* em que o uso de n-gramas possibilita representar uma maior parte da informação presente em um texto, se apresentando como um modelo de linguagem eficiente para a recuperação de informações [Wang and Manning, 2012].

O modelo de linguagem *Bag of words* (BoW) é baseado no trabalho de Harris [1954], no qual assume-se a premissa de que a linguagem natural apresenta uma estrutura distribuível. A característica distribuível de uma linguagem vem do fato de que os termos não são encontrados de forma arbitrária próximos de determinadas classes de termos. Sendo assim, é possível identificar a frequência de ocorrência de um termo e co-ocorrências por meio da posição de demais termos.

O processo de busca dos top-*k* documentos relevantes em uma coleção de documentos apresenta dois requisitos: (i) um critério de busca; (ii) o uso de um modelo que permita a recuperação de informação. Com relação ao requisito (ii), o modelo de espaço vetorial é normalmente usado [Manning et al., 2008]. Dentre as diferentes medidas existentes para o ranqueamento de documentos, a medida TF-IDF, um acrônimo para *Term Frequency - Inverse Document Frequency*, é responsável por definir a importância dos termos em uma coleção de documentos, e comumente, é aplicada no contexto de RI [Han et al., 2011]. Nessa medida, a importância dos termos em um documento é definida por meio de um esquema de pesagem de termos que considera a combinação de duas funções estatísticas: TF e IDF.

A função TF é responsável por calcular o peso de um termo t em um determinado documento d presente em uma coleção de documentos $\mathcal{D}$. Para isso, é considerado o total de ocorrências do termo t em um determinado documento d. Por outro lado, a função IDF atua de forma a penalizar os termos mais comuns em uma coleção de documentos. A função considera uma base heurística em que os termos que aparecem poucas vezes nos documentos tendem a ser mais relevantes do que os termos que aparecem muitas vezes nos documentos de uma coleção [Gudivada et al., 2018]. Sendo assim, o cálculo do peso na função IDF é realizado por meio da razão entre $\mathcal{D}$ e o número de documentos na coleção que apresentam ao menos uma vez um determinado termo df_t. Por fim, o peso

final w de um termo t em um documento d é o produto das funções TF e IDF, conforme apresentado na Equação 1.

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{|\mathcal{D}|}{df_t}\right) \quad (1)$$

1.3- Mineração de dados

O processo de mineração de dados é um passo dentro da área de descoberta de conhecimento em bases de dados (Em inglês: Knowledge Discovery in Databases - KDD) que consiste na aplicação de diferentes técnicas de análise de dados e algoritmos para identificar e extrair padrões a partir dos dados disponíveis [Fayyad et al., 1996]. É um processo multidisciplinar, que envolve diferentes áreas de pesquisa, tais como a estatística, o aprendizado de máquina, o reconhecimento de padrões, a recuperação da informação, a inteligência artificial, dentre outras [Han et al., 2011]. A Seção encontra-se organizada da seguinte forma: a Subseção 1.3.1 comenta a motivação para o uso de técnicas de pré-processamento e técnicas tradicionais normalmente empregadas em textos. Na Subseção 1.3.2 é apresentado o conceito de classificação de dados e, em seguida, são detalhados diferentes algoritmos de aprendizado de máquina. Logo após, na Subseção 1.3.3 são introduzidas medidas e técnicas comuns para a interpretação e avaliação de desempenho de modelos de classificação.

1.3.1 Pré-processamento

A atividade de pré-processamento contempla um conjunto amplo de tarefas que objetivam preparar os dados para posteriores técnicas de mineração de dados. Desta forma, ao final da atividade, espera-se obter uma maior qualidade dos dados, isto é, a ausência de dados que sejam irrelevantes ou redundantes para o propósito da pesquisa, dados que adicionem ruído ou que não sejam confiáveis [García et al., 2015]. Para que esse objetivo seja alcançado, são consideradas diferentes tarefas tais como a limpeza,

a integração, a redução e a transformação dos dados [Han et al., 2011]. Importante ressaltar que para cada tarefa de preparação dos dados existem diferentes rotinas que podem ser aplicadas.

Dentre as tarefas de limpeza de dados, a remoção dos ruídos apresenta um destaque. Originados por variadas razões, frequentemente se originam após processos de integração de dados de distintas fontes [Han et al., 2011]. A presença de ruídos pode impactar diretamente o desempenho de diferentes algoritmos de aprendizado de máquina.

Ao considerar um cenário de aprendizado supervisionado entre duas classes, como por exemplo, o cenário descrito na competição PAN-2012, dois problemas são comumente encontrados: (i) a sobreposição de características entre as classes, isto é, ocorrência de exemplares de diferentes classes e que apresentam alta similaridade. Essa sobreposição eleva a complexidade de definição das fronteiras de decisão dos algoritmos de aprendizado de máquina por conta da não-linearidade inserida pela presença dos dados ruidosos [García et al., 2008]; (ii) a disjunção de pequenos grupos de exemplares pertencentes a uma classe dentre os exemplares de uma outra classe. A Figura 2 ilustra os problemas apresentados.

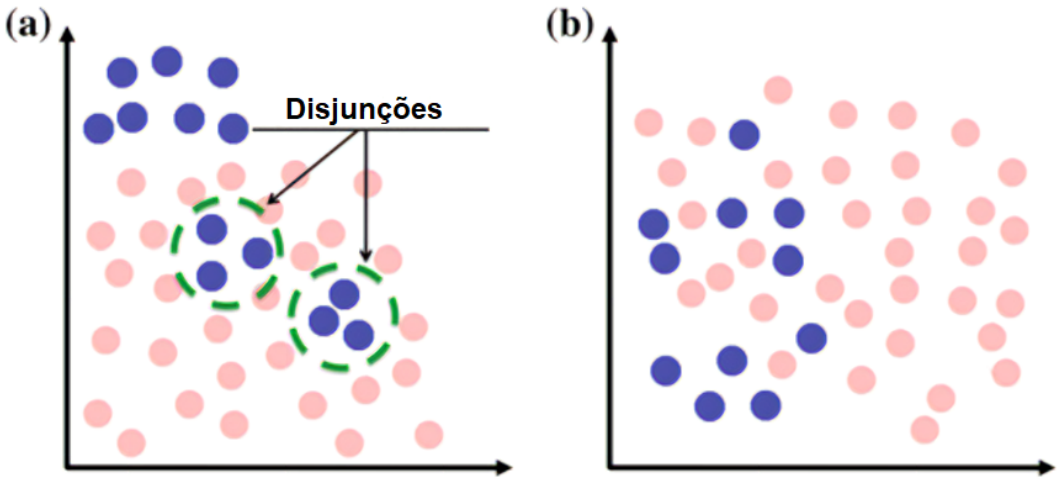


Figura 2 – Exemplo de problemas oriundos da presença de ruídos em um conjunto de dados. Na imagem, García et al. [2015] representa na Figura (a) o cenário em que pequenos grupos pertencentes a uma classe se encontram dentre os exemplares de uma outra classe enquanto a Figura (b) apresenta a sobreposição de exemplares. Autor: Adaptado de García et al. [2015].

Ainda nesse contexto, García et al. [2015] acrescenta que alguns outros problemas, tais como a presença de exemplares pertencentes a uma classe e muito próximas

à fronteira de decisão, assim como a presença de exemplares ruidosos distantes da fronteira de decisão, contribuem diretamente para a degradação do desempenho dos algoritmos de aprendizado de máquina. A Figura 3 ilustra uma intuição sobre os problemas apresentados.

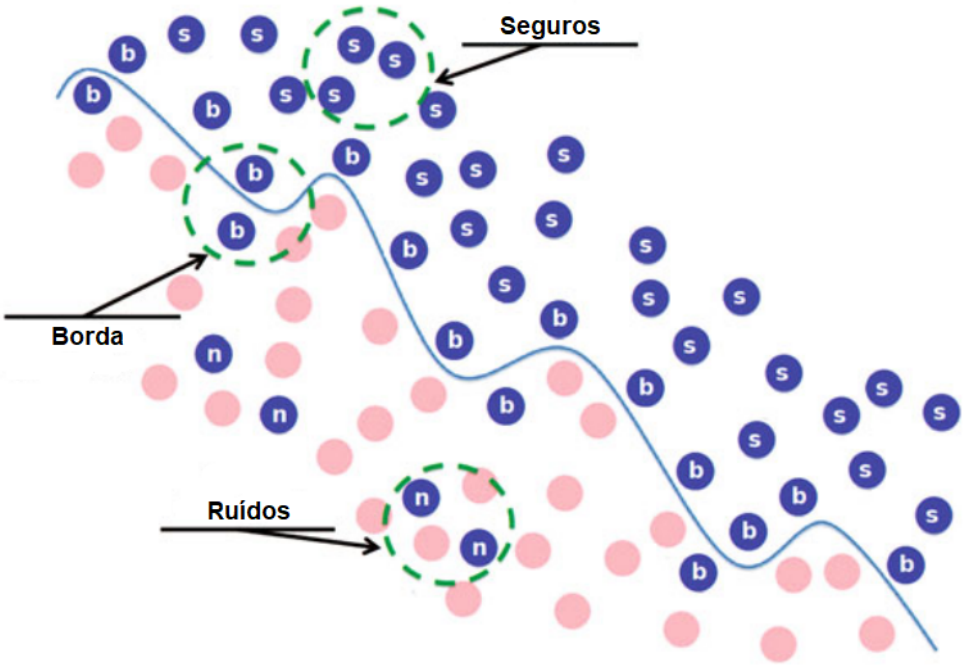


Figura 3 – Fatores condicionantes de degradação do desempenho dos algoritmos de aprendizado de máquina. São apresentadas três condições distintas para um exemplar presente em um conjunto de dados: Os exemplares seguros (nomeados como “s”), exemplares de borda (nomeados como “b”) que contribuem para o aumento da complexidade da tarefa de classificação e, por fim, os exemplares que contribuem para o aumento de ruído e subsequente degradação do desempenho do modelo gerado (nomeados como “n”). Autor: Adaptado de [García et al., 2015].

Ao considerar a aplicação de técnicas de limpeza em um conjunto de dados não estruturados, por exemplo, em uma coleção de documentos, o objetivo final da pesquisa define normalmente quais técnicas serão empregadas [Feldman et al., 2007]. Uma técnica normalmente aplicada em documentos em um primeiro instante é a *tokenização*. A *tokenização* consiste no aumento da granularidade do teor do documento de forma que cada *token* criado (a menor unidade que compõe o documento seja significativo).

Uma outra prática normalmente considerada é a remoção de palavras sem importância. As palavras sem importância são aquelas usadas com bastante frequência em uma língua e, por conta disso, não acrescentam informação relevante em um cenário de classificação de documentos [Kao and Poteet, 2007]. Exemplos comuns de palavras

sem importância são os artigos, preposições, pronomes, advérbios, etc. A remoção dessas palavras contribui não somente para a limpeza dos dados mas também para a sua redução do volume de dados a ser considerado nas tarefas subsequentes ao pré-processamento. Em adição a isso, a remoção de pontuações e caracteres especiais geralmente também compõem uma tarefa importante durante a limpeza de dados.

Construção de características

A tarefa de criação de características contempla um conjunto de técnicas frequentemente empregadas no pré-processamento quando há o desejo de reduzir a dimensionalidade dos dados [Han et al., 2011]. Essa técnica tem como objetivo extrair informações a respeitos dos dados de forma mais eficiente. Desta forma, é possível obter características com um maior poder discriminatório e com isso uma melhora do desempenho dos modelos em uma tarefa de classificação [Hu and Kibler, 1996].

Criar novas características requerem tanto um conhecimento estatístico quanto do domínio da pesquisa [Khurana et al., 2016]. A criação é realizada por meio de um processo iterativo em que a cada nova característica demanda uma avaliação do desempenho perante ao algoritmo de aprendizado de máquina escolhido. O uso de características criadas a partir do conhecimento do domínio da pesquisa permite uma captura mais eficiente de conceitos de alto nível quando comparado aos termos presentes nos documentos [Gabrilovich and Markovitch, 2005; Mayfield and Penstein-Rosé, 2010].

Uma dificuldade normalmente encontrada na criação de características baseadas no domínio da pesquisa é a necessidade do uso de fontes de dados digitais [Specia et al., 2009]. Nesse cenário, diferentes tipos de léxicos se apresentam como uma alternativa para a criação de características. Exemplos comuns são os de natureza psicolinguística, tais como o LIWC [Pennebaker et al., 2015] e dicionários tem sido aplicados em diferentes domínios de pesquisa [Akter and Aziz, 2016]. Em alguns cenários em que o uso de dicionários não atende plenamente a identificação de uma determinada característica do domínio da pesquisa é comum que sejam desenvolvidas rotinas específicas para a identificação e geração da característica desejada [Domingos, 2012; Patel et al., 2008].

1.3.2 Classificação

Segundo Aggarwal [2015], a classificação de dados é uma forma de análise em que é possível extrair modelos, denominados classificadores, com a capacidade de identificar diferentes grupos de dados. A tarefa de classificação de dados é um processo realizado em duas fases [Han et al., 2011]: a primeira fase é responsável pelo aprendizado e apresenta como produto a criação de um modelo de classificação. Para atingir esse objetivo, são empregados algoritmos de aprendizado de máquina e um conjunto de exemplares rotulados que formam o conjunto de dados de treinamento. Na segunda fase, novos exemplares, também rotulados, são submetidos ao modelo de classificação e, em seguida, é validada a capacidade de identificação do rótulo de cada exemplar.

Árvore de decisão

O algoritmo de árvore de decisão é alternativa normalmente considerada para a realização do aprendizado supervisionado. De natureza não-linear [Apté and Weiss, 1997], uma árvore de decisão é criada a partir dos dados disponíveis durante a fase de treinamento. Desta forma, todos os nós internos (não folha) presentes na árvore correspondem a uma característica presente no conjunto de dados. Em contrapartida, todos os nós folhas representam um atributo alvo dentre os presentes no conjunto de treinamento. Os nós da árvore de decisão são conectados por ramificações que interligam as características e definem a classificação de um determinado exemplar [Larose and Larose, 2014].

O uso de árvore de decisão permite a criação de modelos de classificação com uma maior interpretabilidade quando comparado a outros algoritmos de aprendizado de máquina [Kingsford and Salzberg, 2008]. Essa razão se deve a natureza do algoritmo em si, que se assemelha ao raciocínio humano, sendo composto por uma sucessão de testes simples em que, tipicamente, compara-se um atributo a um intervalo ou conjunto de valores possíveis [Kotsiantis, 2013].

Uma decisão importante na criação de uma árvore de decisão é a escolha das

características que irão compor os nós da árvore de decisão. Nesse contexto, os métodos mais tradicionais para a escolha de características são o Coeficiente de Gini [Hunt et al., 1966] e o Ganho de Informação [Breiman et al., 1984]. A árvore de decisão é um dos algoritmos de aprendizado mais populares atualmente [Kotu and Deshpande, 2019]. Dentre as implementações mais destaca-se a C4.5 [Quinlan, 1993] realizada em código aberto e desenvolvida na linguagem Java e a implementação J48, disponível por meio do software Weka [Witten et al., 2016].

Florestas Aleatórias

A floresta aleatória é um algoritmo de aprendizado de máquina supervisionado e não-linear, que consiste na combinação de classificadores com a estrutura de árvores de decisão. Cada árvore de decisão é treinada por entradas aleatórias distribuídas de forma uniforme [Breiman, 1996]. Após o treinamento de uma floresta aleatória, cada uma das árvores que compõem a floresta aleatória realizará a predição das novas amostras e então, por meio de uma votação, o resultado mais popular é considerado [Breiman, 2001].

Uma das vantagens do uso de florestas aleatórias é a sua capacidade efetiva de generalização. Por consequência, não estão sujeitas ao sobreajuste. Esta propriedade se justifica por conta da aplicação da Lei dos Grandes Números (LGN). A LGN é um teorema fundamental da lei da probabilidade que busca descrever o resultado de um experimento ao ser realizado mais de uma vez. Sendo assim, quanto mais experimentos forem realizados, mais próxima da probabilidade real será a probabilidade da média aritmética dos resultados [Ross, 1997].

Uma desvantagem presente na aplicação de Florestas Aleatórias é a dificuldade de interpretação dos resultados. Essa dificuldade ocorre devido a natureza do algoritmo, que pode empregar uma quantidade variada de árvores de decisão em que cada uma pode ser criada com diferentes configurações [Zhao et al., 2018]. Sendo assim, uma melhor compreensão do resultado obtido com o algoritmo requer uma análise individual das árvores de decisão que compõem o resultado obtido.

Naïve Bayes

O algoritmo de aprendizado de máquina Naïve Bayes é um classificador linear [Rennie, 2001] e probabilístico baseado no teorema de Bayes [Zhang, 2004]. O algoritmo é considerado “naïve” (Em português: ingênuo) pois assume que cada característica é independente, ou seja, que não existe a correlação entre as características que compõe o conjunto de dados. Por fim, é compreendido que cada característica contribui de forma igual para a identificação de uma determinada classe alvo [Aggarwal, 2015].

Uma vantagem da aplicação do classificador Naïve Bayes é a sua simplicidade para implementação, tolerante à ruídos e, ao considerar o desempenho computacional, é considerado um algoritmo rápido e com bom desempenho em grandes conjuntos de dados e com alta dimensionalidade, sendo recomendado para aplicações em tempo real [Misra and Li, 2019].

Uma variante comum do algoritmo Naïve Bayes, com aplicação na classificação de documentos, é a versão multinomial [Duda et al., 1973]. A ampla aplicação do algoritmo Naïve Bayes Multinomial (NBM) é motivada por considerar a frequência com que as palavras ocorrem nos documentos. Desta forma, o algoritmo NBM tem apresentado desempenho superior quando comparado às demais variantes e em diferentes domínios de pesquisa [Eyheramendy et al., 2003].

Máquina de vetores de suporte

A máquina de vetores de suporte (SVM) é um algoritmo de aprendizado supervisionado, com alta capacidade de reconhecimento de padrões e tem sido amplamente usado desde a sua publicação em diferentes domínios de pesquisa [Boser et al., 1992]. Desta forma, o SVM apresenta boa capacidade de adaptação na identificação de padrões lineares e não-lineares [Chang and Lin, 2011]. Essa característica é consequência da definição de um dos parâmetros presentes no algoritmo: a função de *kernel*.

A principal idéia para o algoritmo SVM consiste na representação dos dados do conjunto de treinamento como vetores de entrada, que podem pertencer à diferentes

classes em um espaço de características com alta dimensionalidade. Uma vez aplicada uma função de *kernel* nos vetores de entrada, é possível extrair um hiperplano que maximize a margem entre as classes presentes [Vapnik, 2013]. Os vetores responsáveis por definir a margem máxima do hiperplano resultante são chamados de "vetores de suporte". Na Figura 4 é apresentado o resultado de uma tarefa de classificação binária.

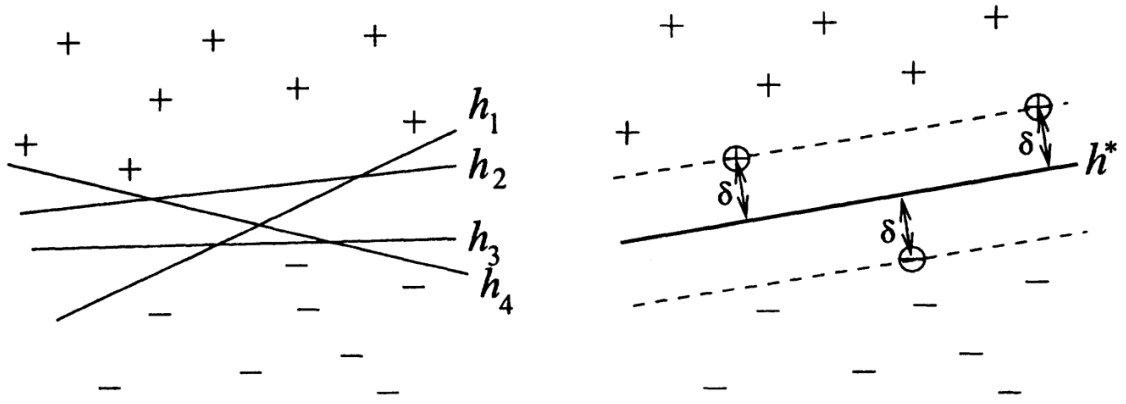


Figura 4 – Exemplo de problema de classificação binária em duas dimensões. Os exemplares pertencentes a classe "Positiva" são representados pelo caractere $+$ enquanto os exemplares da classe "Negativa" são representados pelo símbolo $-$. Esquerda: é possível identificar quatro diferentes hiperplanos $h_{1..4}$ de forma que os exemplares sejam separados sem erros. Direita: Após a aplicação do algoritmo SVM, é possível identificar o hiperplano otimizado h^* de tal forma que a margem δ entre os exemplares das diferentes classes esteja maximizada. Os exemplares marcados com círculos representam os "vetores de suporte". Autor: [Joachims, 2002].

O SVM apresenta algumas vantagens como a boa capacidade de generalização e baixa sensibilidade à ruídos oriundos da alta dimensionalidade dos dados [Hughes, 1968] e é capaz de prover melhores resultados de classificação. Segundo Joachims [1998], o SVM se apresenta como uma solução efetiva para problemas de classificação de textos. Embora a escolha do SVM apresente diferentes vantagens na sua escolha para uma tarefa de classificação, algumas características do SVM devem ser consideradas na escolha. Segundo Vapnik and Vapnik [1998], o número de exemplares considerados para treinamento influencia diretamente no tempo de treinamento e o consumo de recursos computacionais, o que, dependendo da configuração do ambiente de treinamento, pode tornar a tarefa de classificação proibitiva. Essa influência é decorrente da natureza do algoritmo SVM, em que o treinamento do modelo é tratado como um problema de otimização quadrática convexa [Aggarwal, 2015]. Uma possível alternativa para lidar com esse cenário (assim como realizar um melhor uso dos recursos computacionais) é o uso

de diferentes técnicas de seleção características de forma a reduzir a quantidade de características redundantes ou com menor poder discriminativo.

Redes neurais artificiais

A redes neurais artificiais surgiram a partir da concepção de que o cérebro humano processa as informações de forma totalmente diferente de um computador [Haykin et al., 2009]. Uma variante clássica das redes neurais artificiais são as redes *Perceptron* multicamadas (MLP). Essas redes apresentam algumas características: elas seguem uma única direção e são densamente conectadas, isto é, cada neurônio de uma determinada camada se comunica com os demais presentes da camada posterior. Uma MLP apresenta 3 tipos de camadas: uma camada de entrada, que será responsável por processar os elementos do padrão de ativação; as camadas ocultas que recebem os padrões processados da camada anterior, reprocessam e transmitem o resultado para a camada seguinte; por fim, uma camada de saída que disponibiliza o resultado final de todo o processamento realizado. A Figura 5 ilustra um exemplo de rede neural artificial MLP.

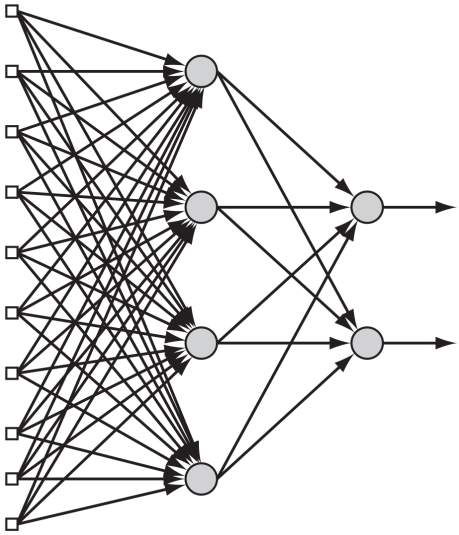


Figura 5 – Exemplo de uma rede neural artificial Perceptron. A rede apresenta quatro neurônios na camada de entrada e uma camada de saída com dois neurônios. Autor: Haykin et al. [2009].

Um neurônio artificial retrata uma abstração simplificada de um neurônio biológico e representa a estrutura básica para criação de uma rede neural artificial. A estrutura de

um neurônio k apresenta n entradas em que cada uma apresenta um peso w. Além das entradas, cada neurônio possui uma entrada adicional com viés b. Considerando $x_1 ... x_n$ os valores de entrada de neurônio artificial, é realizada a soma ponderada dos valores de entrada com os pesos previamente definidos e, por fim, a posterior adição de viés. Após o processamento, o resultado obtido é submetido a uma função de ativação φ em que é apresentada a saída $\hat{y}$. A Figura 6 apresenta a anatomia previamente descrita de um neurônio artificial.

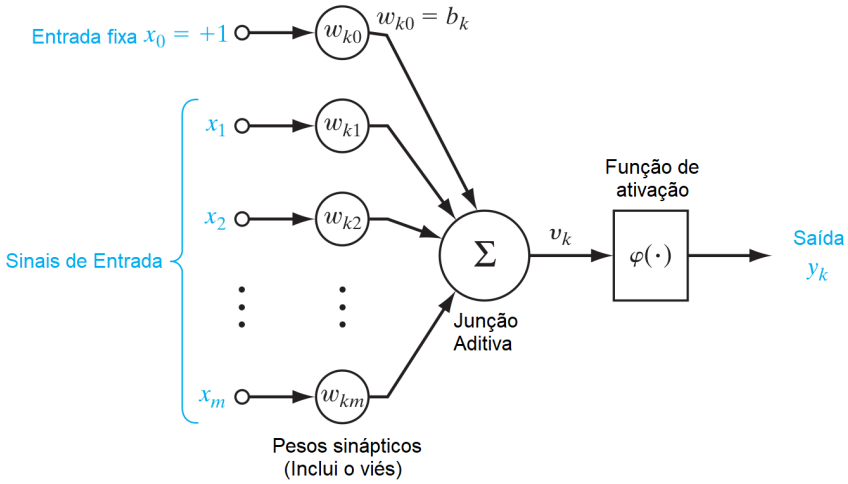


Figura 6 – Anatomia de um neurônio artificial. Autor: Adaptado de Haykin et al. [2009].

Ao longo das últimas décadas, diferentes funções de ativação tem sido testadas em redes neurais artificiais, de forma a descobrir a melhor aplicação para os variados domínios de pesquisa. Após extensa experimentação e pesquisa, atualmente a *Rectified Linear Unit* (ReLU) tem sido a principal recomendação [Fathi and Shoja, 2018]. A função de ativação ReLU é apresentada na Equação 2. Dado a entrada z, são retornados dois valores possíveis: (i) 0, quando a entrada z for um número negativo; (ii) z, nos demais casos em que a entrada for um número positivo.

$$\varphi(z) = max(0, z) \tag{2}$$

1.3.3 Avaliação e seleção de modelos de classificação

A seguir, são apresentadas as medidas de avaliação de desempenho consideradas nos trabalhos relacionados e nos experimentos. Medidas de avaliação são representadas de forma quantitativa e calculadas mediante fundamentação estatística. Desta forma, diferentes medidas de avaliação permitem a interpretação sob diferentes perspectivas.

Um ponto importante a ser considerado, uma vez que os dados são propensos a apresentarem uma alta taxa de ruído, é importante que sejam considerados algoritmos de aprendizado de máquina robustos [Verbaeten and Van Assche, 2003]. Um algoritmo de aprendizado de máquina pode ser considerado robusto ao apresentar a capacidade de criar modelos que sofram uma menor influência de ruídos e impacto oriundo de dados corrompidos [Huber and Ronchetti, 1981].

Validação cruzada e estratificada

O processo de validação cruzada foi inicialmente proposto em um estudo da área da psicologia [Larson, 1931] e tem como o principal objetivo identificar a capacidade de generalização de um determinado modelo treinado por meio de uma melhor estimativa do viés e da variância presente [Kohavi, 1995]. Em adição a isso, a processo de validação cruzada também contribui na identificação dos classificadores que apresentam melhores resultados para um determinado conjunto de dados [Refaeilzadeh et al., 2009].

Existem diferentes técnicas para a aplicação do processo de validação cruzada. Uma técnica popular, a validação cruzada e estratificada em *K* grupos, o conjunto de dados é dividido em K grupos de tamanho igual ou aproximado e de forma mutuamente exclusiva. Por utilizar uma abordagem estratificada, cada grupo apresentará a mesma proporção de documentos das diferentes classes que compõem o conjunto de dados original. Em cada rodada de validação, o classificador considerará *K-1* grupos como o conjunto de dados de treinamento e o grupo remanescente para o teste do modelo. Um ponto importante é que cada grupo é usado uma única vez como o conjunto de testes. Desta forma, ao fim das rodadas de validação, todo o conjunto de dados é considerado

para a realização dos testes assim como para o treinamento.

A escolha da quantidade dos grupos é 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 importante que contribui para a escolha da quantidade de grupos é o grau de representatividade possível de ser atingido em cada grupo. Em geral, $K = 5$ ou $K = 10$ são os valores mais comuns [Anguita et al., 2009].

Matriz de confusão

Uma ferramenta comum para avaliação de modelos de classificação é a matriz de confusão. No cenário de uma classificação binária, por exemplo, é possível identificar quantos exemplares foram corretamente previstos como pertencentes à classe positiva (verdadeiro positivo - TP) ou à classe negativa (verdadeiro negativo - TN) assim como também é possível identificar a quantidade de exemplares incorretamente classificados como pertencentes à classe positiva (falso positivo - FP) ou à classe negativa (falso negativo - FN). A Figura **7** ilustra uma possível representação de uma matriz de confusão para um cenário de classificação binária. Nela, são consideradas duas dimensões principais - Previsto e Real e são apresentados os resultados encontrados quanto a ocorrência das classes Positiva e Negativa.

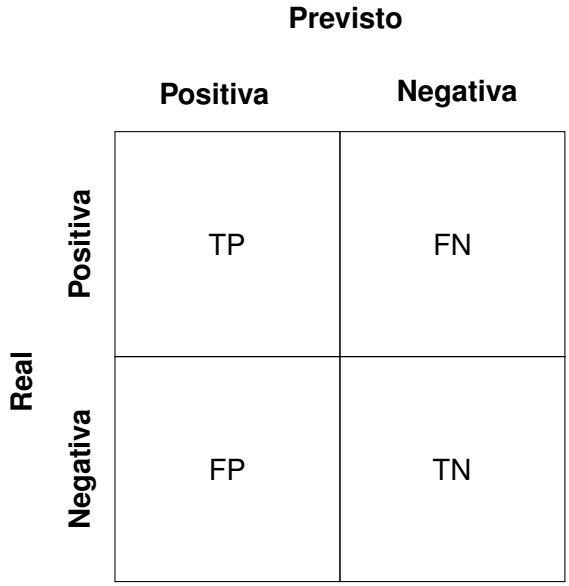


Figura 7 – Exemplo de uma matriz de confusão.

As principais medidas de desempenho consideradas atualmente em tarefas de classificação de textos são inspiradas na extração de informação inicialmente aplicada para a geração de textos sinônimos [Boyer and Lapalme, 1985]. As principais medidas consideradas (precisão, abrangência, medida F_β) valorizam a correta identificação de resultados da classe positiva enquanto a acurácia busca identificar a capacidade assertiva do modelo treinado. A seguir, são descritas as medidas de desempenho acima citadas.

Acurácia

A acurácia permite identificar a capacidade de acerto do modelo, ou seja, o total de predições corretas realizadas dentre o total de predições realizadas. A Equação 3 exemplifica o conceito apresentado.

$$Acurácia = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

Precisão

A medida de precisão representa a capacidade do modelo de classificação em identificar corretamente os exemplares pertencentes à classe positiva em um conjunto de dados dentre o total de exemplares identificados como positivos. A precisão é apresentada na Equação 4.

$$Precisão = \frac{tp}{tp + fp} \tag{4}$$

Abrangência

A abrangência representa a proporção com que um modelo de classificação é capaz de identificar corretamente os exemplares pertencentes à classe positiva em comparação a todos os exemplares positivos existentes. A abrangência pode ser compreendida conforme a Equação 5.

$$Abrangência = \frac{tp}{tp + fn} \tag{5}$$

Medida F_β

A medida F_β tem por objetivo medir o grau de eficiência na recuperação de ocorrências da classe positiva considerando o parâmetro de ponderação β e os resultados obtidos com medidas precisão e abrangência. O parâmetro β permite definir uma maior importância para uma medida. Por fim, uma vez definido β, é calculada a média harmônica e ponderada das medidas precisão e abrangência. A equação 6 define a explicação apresentada.

$$F_\beta = \frac{\left(\beta^2 + 1\right) tp}{\left(\beta^2 + 1\right) tp + \beta^2 fn + fp} \quad (6)$$

1.4- Considerações finais

Ao longo do presente capítulo foram apresentados os principais conceitos para uma melhor compreensão do método proposto para identificação de atividade predatória sexual em conversas textuais na internet, descrito no Capítulo 4 assim como o entendimento dos principais trabalhos relacionados a serem apresentados a seguir, no Capítulo 2. Inicialmente foram introduzidos diferentes aspectos do domínio da pesquisa, em particular, da atividade predatória sexual ocorrida na internet. Com isso, foram apresentados o perfil do predador sexual e as características comumente presentes em conversas predatórias. Após a introdução ao domínio da pesquisa, a competição PAN-2012 é apresentada.

Em seguida, foram detalhados diferentes conceitos que direcionam o RI, com destaque para o modelo de linguagem *Bag of Words* e a função TF-IDF. A apresentação do processo de mineração de dados considerou a discussão de diferentes técnicas de pré-processamento com destaque para as práticas mais tradicionais e amplamente usadas. Dentre elas, é dado um destaque para o uso da técnica de construção de características. Também foi introduzido o conceito de léxicos e como eles podem contribuir para a descoberta de informação em dados. Em conjunto a isso, o método denominado Ganho de informação como uma alternativa para a seleção dos termos mais relevantes em uma coleção de documentos. Por fim, foram descritos os diferentes algoritmos de aprendizado normalmente considerados para o domínio da pesquisa e as medidas para avaliar o desempenho de modelos de classificação.

2- Trabalhos Relacionados

O presente capítulo tem por finalidade apresentar todo o planejamento, a busca e os resultados da pesquisa realizada para o presente trabalho. Primeiramente, foram realizadas pesquisas em alguns dos principais canais de publicação (ScienceDirect, IEEE Xplore, ACM Digital Library e SpringerLink) e posteriormente no indexador Google Scholar. A proposta para expandir a busca e considerar o indexador Google Scholar se justifica em dois pontos: (i) busca de trabalhos presentes na literatura nacional; (ii) trabalhos realizados com base na competição PAN-2012 para identificação de predadores sexuais em conversas virtuais que não tenham sido indexados nos principais canais de publicação. Dessa maneira, esse capítulo se encontra estruturado da seguinte forma: na Seção 2.1 é apresentada a pesquisa realizada e criação de um mapa sistemático para a organização dos resultados, a Seção 2.2 apresenta os trabalhos que obtiveram os melhores resultados e o estado da arte na tarefa de classificação de conversas com conteúdo predatório sexual. Por fim, a Seção 2.3 são feitas algumas considerações a respeito dos principais trabalhos e a proposta da presente pesquisa.

2.1- Mapa Sistemático

A seguinte busca foi realizada nos principais canais de pesquisa: (“Sexual” AND (“Predator” OR “Offenders”) AND (“Identification” OR “Detection”) AND (“internet” OR “Online”) AND (“Conversation” OR “Chat”)). Após a realização das busca descrita foram encontrados um total de 203 trabalhos.

A criação do mapa sistemático exigiu a execução de dois estágios de análise. No primeiro estágio (ES1), é realizada a leitura do título e resumo/*abstract*. O segundo estágio (ES2) realiza uma leitura completa do trabalho. A avaliação de um trabalho no mapa sistemático leva em consideração um conjunto de critérios de inclusão (CI) e critérios de exclusão (CE). A tabela 1 detalha os critérios aplicáveis na avaliação preliminar de um trabalho.

Tabela 1 – Critérios de inclusão e exclusão considerados para a elaboração do mapa sistemático.

Código	Descrição
CI1	Relacionado ao tema da pesquisa (Técnico)
CI2	Relacionado ao tema da pesquisa (Psicológico)
CE1	Não está escrito em inglês ou português
CE2	Versão anterior de um trabalho já considerado
CE3	Não é um artigo completo, dissertação ou tese
CE4	Trabalho inacessível
CE5	Trabalho fracamente relacionado com a pesquisa
CE6	Trabalho não relacionado com a pesquisa

Ao término do ES1, foram selecionados um total de 27 trabalhos. Uma razão para esse número final ter sido baixo foi o número elevado de ocorrências de resultados nos canais ScienceDirect[1] e SpringerLink[2] que não apresentaram relação com o tema da pesquisa (CE6). Por fim, o ES2 consistiu na leitura completa dos 27 trabalhos. Após análise, foram selecionados 17 trabalhos fortemente relacionados ao tema da pesquisa. Um ponto importante é que, ao término do segundo estágio, não foram encontrados trabalhos na língua portuguesa ou desenvolvidos em centros de pesquisa brasileiros, assim como artigos que estejam relacionados diretamente à competição PAN-2012.

A fim de preencher essas duas lacunas, foram realizadas duas novas buscas no indexador Google Scholar. A primeira busca – (“Sexual” AND (“Predator” OR “Offenders”) AND (“Identification” OR “Detection”) AND (“internet” OR “Online”) AND (“Conversation” OR “Chat”)) **AND “PAN-2012”** – teve como propósito de encontrar trabalhos relacionados à competição PAN-2012. A segunda busca – (“aliciamento” AND “sexual”) AND “pedofilia” AND (“Identificação” OR “Detecção”) AND (“online” OR “internet”) AND “conversas” AND (“Crianças” OR “adolescentes”) – buscou preencher a lacuna de trabalhos relacionados ao domínio da pesquisa e na língua portuguesa do Brasil. As duas novas buscas retornaram um total de 161 trabalhos. Após a avaliação do ES1 foram selecionados 38 trabalhos. Por fim, após a execução do ES2, foram selecionados 27 trabalhos relacionados.

Dentre os 44 trabalhos que apresentaram relevância no domínio da pesquisa (17 destes oriundos dos principais canais de publicação e outros 27 trabalhos selecionados a partir das duas novas buscas realizados no Google Scholar), foram selecionados aqueles

[1] `https://www.sciencedirect.com/`
[2] `https://link.springer.com/`

com os resultados mais significativos. Nesse contexto, foram usados dois critérios para analisar a significância dos trabalhos relevantes: (i) os resultados obtidos de acordo com a medida F_1; (ii) O emprego de características textuais e comportamentais. Dessa forma, foram considerados 11 trabalhos relacionados como base para a presente pesquisa.

2.2- Identificação de atividade predatória sexual em conversas virtuais

Em 2007, foi realizado um estudo piloto considerando as conversas registradas no site PJ para a criação de um conjunto de dados [Pendar, 2007]. A tarefa de identificação dos predadores sexuais foi executada considerando o uso de unigramas, bigramas e trigramas para a geração dos vetores de características. São considerados os algoritmos K vizinhos mais próximos (k-NN) e SVM nos experimentos. A aplicação do algoritmo k-NN, considerando *k* = 30, obteve melhores resultados (F_1 = 94%), superando os resultados obtidos com o classificador SVM (F_1 = 90%).

No primeiro trabalho a fazer uso de características comportamentais que se tem conhecimento, Kontostathis [2009] usam a *Theory of luring communication* (LCT) [Olson et al., 2007]. A LCT apresenta um modelo constituído por quatro estágios que representam a evolução de um aliciamento na internet: (i) ganhar acesso à vítima; (ii) desenvolver falsa relação de confiança; (iii); manter comunicação de cunho sexual com a vítima; (iv) realizar o abuso sexual físico. Kontostathis [2009] considerou 288 conversas extraídas do site PJ. Em seguida, para cada estágio previsto na LCT, foi criado um léxico com os termos que melhor caracterizam a ação predatória. O primeiro experimento considerou 16 conversas. Em seguida, as conversas foram separadas de acordo com as mensagens enviadas por predadores e vítimas para então categorizá-las de acordo com o estágio mais provável da LCT. Para a tarefa de classificação, foi utilizado o algoritmo de árvore de decisão *J48*. Os resultados obtidos são destacados como promissores (F_1 = 60%). No segundo experimento, foram consideradas 16 conversas extraídas do site PJ e 16 conversas extraídas do ChatTracker [Bengel et al., 2004]. Após a fase de pré-processamento, as conversas foram submetidas a uma árvore de decisão, sendo obtidos resultados expressivos (F_1 = 93%). Posteriormente, o trabalho foi expandido [McGhee et al., 2011; Kontostathis et al., 2012] com a adição de características psicolinguísticas

ao modelo de regras. Embora bons resultados tenham sido encontrados no conjunto de testes da competição PAN-2012 (*abrangência* = 87%), a precisão foi afetada com a ocorrência de muitos falso positivos.

Em Villatoro-Tello et al. [2012], o problema de classificação de conversas predatórias é decomposto em duas partes. Na primeira parte o autor busca 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 sequencias longas de caracteres não reconhecíveis (i.e. imagens codificadas em texto). Esse filtro permitiu a otimização do uso de recursos computacionais e focar nos cenários mais propícios para que ocorra um aliciamento, o que resultou em 10% do conjunto de dados da competição PAN-2012. A segunda parte, responsável pela identificação do predador sexual dentre os participantes de uma conversa, fez uso de uma Rede Neural Perceptron e representação binária. O resultado atingido (F_1 = 87,3%, $F_{0,5}$ = 93,4%) conferiu ao autor a primeira colocação na competição PAN-2012.

No trabalho de Bogdanova et al. [2014], buscou-se explorar as características psicolinguísticas e comportamentais presente em conversas virtuais. Dessa forma, foram consideradas três fontes distintas de dados para a comparação do método proposto: o primeiro conjunto de dados foi criado a partir de conversas disponibilizadas no site PJ, o segundo conjunto de dados apresenta apenas conversas variadas, denominado NPSChat [Forsythand and Martell, 2007] e, por último, o terceiro conjunto de dados foi criado a partir de conversas com teor sexual porém não predatórias extraídas do site Cybersex[3]. Com isso, o principal objetivo do trabalho tem por distinguir corretamente as conversas predatórias das conversas não predatórias que apresentam teor sexual. Nesse cenário, ao explorar as características psicolinguísticas e comportamentais, o classificador SVM atingiu 97% de acurácia. A relevância do uso de características psicolinguísticas e comportamentais é comprovada quando apenas as características textuais presentes nas conversas são consideradas. Nesse cenário, o melhor resultado atingido foi 64% de acurácia com a aplicação do classificador SVM.

No trabalho realizado por Cano et al. [2014] a estratégia adotada para identificação da atividade predatória sexual considerou o modelo psicológico composto por O'Connell

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

[2003] para explicar como o aliciamento é realizado na internet. A criação do conjunto de dados considerou 50 conversas predatórias disponibilizadas no site PJ e cada mensagem contida nas conversas escolhidas passou por uma classificação perante os estágios de aliciamento previstos no modelo psicológico de O'Connell [2003]. Para cada estágio do aliciamento previsto no modelo, foram identificados diferentes padrões após a análise das características textuais, sintáticas[4], psicolinguísticas, a polaridade dos sentimentos[5] presente nas conversas estudadas. Para a extração das características textuais o modelo de linguagem BoW foi aplicado. A ferramenta LIWC [Pennebaker et al., 2001] permitiu a extração de características psicolinguísticas. Para uma melhor seleção das características psicolinguísticas das sentenças, foi usada a medida Ganho de Informação para a seleção das cinco características mais relevantes em cada estágio do aliciamento. Os resultados obtidos com o classificador SVM se mostraram satisfatórios (F_1 = 85%) e comprovam o uso de características psicolinguísticas como uma alternativa para a identificação de diferentes estágios de aliciamento.

O trabalho apresentado por Ebrahimi [2016] buscou validar duas hipóteses: 1) É possível identificar conversas predatórias usando abordagens empregadas em detecção de anomalias; 2) Uma arquitetura de aprendizagem profunda é capaz de superar o desempenho do atual estado da arte dentro do domínio da pesquisa. Os experimentos foram realizados em dois conjuntos de dados: o conjunto de dados disponibilizado na competição PAN-2012 e um conjunto de dados privado disponibilizado pelo *Sûreté du Québec* (SQ). Para a validação da primeira hipótese, foi considerado o algoritmo Naïve Bayes como *baseline* e o algoritmo SVM para classificação binária. No contexto de detecção de anomalias no domínio da pesquisa, as conversas predatórias são consideradas as anomalias devido a baixa ocorrência de conversas dessa natureza dentre os mais diversos assuntos e tópicos discutidos na internet. Nesse cenário, foi considerado o algoritmo OC-SVM para a identificação das conversas predatórias. Os resultados mais significativos foram encontrados no conjunto de dados SQ (F_1 = 97,4%), no entanto o autor considera o experimento realizado como uma prova de conceito, sendo apenas válido para a comprovação da hipótese. No segundo experimento, o uso de CNN se mostrou promissor (F_1 = 81,64%) quando comparado ao resultado obtido pelos algoritmos de classificação SVM (F_1 = 61,02%) e MLP (F_1 = 79,96%). Um ponto que vale ressaltar neste trabalho é que ele busca avaliar a aplicação de diferentes algoritmos de

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

classificação no problema de identificação de atividade predatória sexual em conversas na internet e não definir um novo estado da arte para o problema de identificação de conversas predatórias.

O trabalho de Cheong and Jensen [2015] representou a primeira iniciativa documentada para estudo do comportamento predatório na comunicação realizada em jogos *online*. Em parceria com o jogo *MovieStartPlanet*, cujo público-alvo são crianças e adolescentes entre 8 e 15 anos, foram disponibilizados três conjuntos de dados contendo a comunicação escrita de diferentes jogadores: (i) mensagens de *status*; (ii) comentários em vídeos e postagens em fóruns; (iii) conversas públicas e privadas realizadas dentro da plataforma. Dois dos conjuntos consideraram apenas 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 frases sem sentido. Em conjunto a isso, em alguns casos, as conversas realizadas no jogo apresentam similaridades com assuntos introduzidos pelo predador sexual: 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 textuais, sentimentais e comportamentais. Após a definição das características, estas foram submetidos a diferentes algoritmos de aprendizado de máquina: SVM, MLP, DT, NBM, Regressão Logística e K vizinhos mais próximos. Os melhores resultados (F_1 = 78%, $F_{0,5}$ = 86%) foram obtidos com o algoritmo MLP. Ao fim, os resultados foram considerados promissores, tendo em vista a natureza ruidosa dos dados.

O uso de quantificadores da teoria da informação também foi considerado em tarefas de classificação de conversas predatórias. Em trabalho recente, Postal et al. [2017] fez uso de dois quantificadores: a entropia de Shannon e a divergência de Jensen-Shannon. A principal motivação para o uso de quantificadores é a disposição de recursos computacionais reduzidos em dispositivos móveis, o que torna proibitivo o uso de modelos de linguagem para a geração de características a partir de conversas virtuais como, por exemplo, o BoW. Uma amostra do conjunto de dados da competição PAN-2012 é considerado para a aplicação do método proposto. Os melhores resultados foram obtidos (F_1 = 90%) ao combinar o uso de quantificadores com o algoritmo *Naïve Bayes*. Esse resultado representa uma perda de aproximadamente 4% quando comparado ao uso do

modelo de linguagem BoW . No entanto, devido o ganho de 72,9% no tempo de execução quando comparado ao modelo de linguagem BoW se apresentou como uma alternativa válida para a uso embutido em dispositivos móveis.

O trabalho realizado por Cardei and Rebedea [2017] fez uso da mesma estratégia descrita em Villatoro-Tello et al. [2012] ou seja, considerando dois estágios para identificação do predador sexual. O primeiro estágio, responsável pela identificação de conversas predatórias, considerou o uso de características textuais, por meio do uso do modelo de linguagem BoW, e características comportamentais que reflitam como o predador atua em uma conversa na internet. Para a tarefa de classificação foi aplicado o classificador SVM obtendo resultados superiores ao encontrado por Villatoro-Tello et al. [2012] na competição PAN-2012 ($F_{0,5}$ = 93,8%). No segundo estágio, em que se busca identificar o predador sexual em uma conversa, é feito o uso de Florestas Aleatórias. Os resultados encontrados (*acurácia* = 100%,*abrangência* = 81,8%, $F_{0,5}$ = 95,7%, F_1 = 89,90%) superam os encontrados na competição PAN-2012.

Recentemente, Monroy et al. [2018] define um modelo de múltiplas perspectivas para representação de documentos. A intuição por trás da proposta se baseia em que, a medida que novas palavras são adicionadas a um texto novas perspectivas são obtidas. A cada nova perspectiva, de acordo com a semântica presente no texto, compreende-se que as palavras presentes no texto podem apresentar variações na sua relevância. Para a implementação da proposta é adotado *Bag of Centroids* (BoC) e representação vetorial de palavras. Para efeitos de validação do modelo, são considerados como *baseline* foi feito o uso do modelo de linguagem BoW e do esquema de ponderação de termos TF-IDF por serem tipicamente empregados em tarefas de classificação de textos, em conjunto com os algoritmos LSA e LDA, usados em tarefas de modelagem de tópicos. Os modelos foram validados em dois conjuntos de dados: (i) o conjunto de dados da competição PAN-2012 (ii) o conjunto de dados da competição *eRisk 2017*[6] com textos que apresentam teor depressivo. As conversas presentes no conjunto de dados da competição PAN-2012 foram segmentadas em 10 partes e então analisadas de forma iterativa e incremental. Nesse cenário, a cada iteração, uma parte da conversa era adicionada para análise ao método e então avaliado o desempenho do experimento. Ao final dos experimentos, o resultado mais expressivo ($F_{0,5}$ = 97,4%) superou o estado da arte quando proposta a identificação antecipada de atividade predatória sexual em conversas e os resultados da

[6] `http://early.irlab.org/2017/index.html`

competição PAN-2012.

Liu et al. [2017] faz o uso das redes neurais LSTM e RNN em conjunto com modelos de representação vetorial de palavras *Fasttext*[7] de forma a atingir dois objetivos: (i) identificar a atividade predatória em conversas virtuais; (ii) identificar as mensagens enviadas por predadores sexuais dentre todas as mensagens presentes em uma conversa predatória. Em um primeiro momento, foram usadas duas combinações das redes neurais LSTM e RNN. A primeira, explorou a relação entre os termos de uma mensagem presente em uma conversa. Uma vez identificada essa habilidade, a segunda combinação de redes neurais estendeu as capacidades da primeira combinação treinada, aprendendo a relação entre as mensagens de uma dada conversa. O entendimento da relação das mensagens em uma conversa é considerado a principal estratégia para atingir o objetivo (i). Uma vez validada essa capacidade, um modelo de representação vetorial de palavras é treinado a partir de conversas presentes no conjunto de dados PAN-2012 de forma a atuar no objetivo (ii). Os resultados apresentados para a tarefa de identificação de conversas predatórias (*acurácia* = 98,3%, *precisão* = 98,4%, *abrangência* = 98,2%, F_1 = 98,3%, $F_{0,5}$ = 98,4%) são considerados o estado da arte. A última etapa do classificador, responsável por identificar o predador em conversas obteve resultados superiores ao da competição PAN-2012 e muito próximos ao obtido por Cardei and Rebedea [2017] (*acurácia* = 100%, *abrangência* = 81,1%, F_1 = 89,5%, $F_{0,5}$ = 95,5%).

[7] `https://fasttext.cc/`

Tabela 2 – Resumo dos trabalhos relacionados.

Autores	Idioma da pesquisa	Conjunto de dados	Características	Classificador	Resultados (F_1)
Pendar [2007]	Inglês	Perverted-Justice	BoW	k-NN SVM	90%
Kontostathis [2009]	Inglês	Perverted-Justice	Comportamentais	DT	93%
Villatoro-Tello et al. [2012]	Inglês	PAN-2012	Vetores de características	MLP	$87,3\%$
Bogdanova et al. [2014]	Inglês	*Perverted-Justice* NPSChat Cybersex	Psicolinguísticas Comportamentais	SVM	$-x-$
Cano et al. [2014]	Inglês	*Perverted-Justice*	BoW + LIWC Comportamentais	SVM	85%
Ebrahimi [2016]	Inglês	PAN-2012 SQ	Termos mais frequentes	CNN	80%
Cheong and Jensen [2015]	Inglês	*MovieStarPlanet*	BoW + Sentimentais Comportamentais	MLP	78%
Postal et al. [2017]	Português	PAN-2012	H + JSD	Naïve Bayes	90%
Cardei and Rebedea [2017]	Inglês	PAN-2012	BoW Comportamentais	SVM + RF	$89,9\%$
Monroy et al. [2018]	Inglês	PAN-2012 eRisk 2017	Unigramas	MulR + TVT	$97,4\%$
Liu et al. [2017]	Inglês	PAN-2012	Vetores de palavras	LSTM+RNN	$98,3\%$

2.3- Considerações finais

A Tabela 2 apresenta os trabalhos com resultados mais significativos e considerados candidatos ao estado da arte. Ao longo da apresentação dos trabalhos, observar-se o uso do site PJ como a referência mais comum para a extração de conversas predatórias no domínio da pesquisa. Dessa forma, observa-se que a maioria dos trabalhos fizeram o uso de dados na língua inglesa. O uso de dicionários também merece atenção, visto que poucos trabalhos fizeram o uso [Kontostathis, 2009; Cano et al., 2014]. Quando consideradas conversas com vítimas reais, poucos trabalhos foram encontrados (i.e, Cano et al. [2014] e Ebrahimi [2016]), porém os dados não foram disponibilizados publicamente.

A presente pesquisa se difere das demais apresentadas nesse capítulo por construir um método com foco na exploração de características textuais e comportamentais normalmente presentes em conversas predatórias para a língua portuguesa do Brasil, de forma a identificar a ocorrência de atividade predatória entre criminosos e vítimas reais. O uso de léxicos oriundos de variadas fontes de dados é estendido (quando comparado aos trabalhos relacionados) de forma a melhor identificar as características investigadas em conversas predatórias anonimizadas e não predatórias. Após a identificação, cada característica investigada (anonimizada ou não) é representada como conceitos de alto nível com um maior poder discriminativo.

3- Conjunto de dados

O presente capítulo tem por objetivo apresentar o conjunto de dados PRED-2050-ALL. Dessa forma, as seções a seguir tratam o tema da seguinte forma: a Seção 3.1 apresenta a motivação e detalha os passos necessários para a criação dos conjuntos de dados (conversas predatórias e não predatórias) PRED-2050-ALL. Na Seção 3.2 é detalhada a obtenção e o trabalho de adequação realizado nas conversas predatórias. A Seção 3.3 descreve o processo de extração, transformação e seleção das conversas não predatórias oriundas de comunidades virtuais e apresenta as principais informações a respeito do conjunto de dados PRED-2050-ALL. A Seção 3.4 discute a análise estatística realizada no conjunto de dados criado. Por fim, a Seção 3.5 traz as considerações finais.

3.1- Motivação

O levantamento de fontes de dados para o desenvolvimento da presente pesquisa considerou o trabalho de Andrijauskas et al. [2017] como o ponto de partida. O conjunto de dados resultante da pesquisa[1] apresenta um total de 39 conversas predatórias e 137 conversas não predatórias. Após análise das conversas não predatórias disponibilizadas, foram observados dois pontos relacionados à representatividade dos dados:

- Ausência de conversas da categoria adulta (i.e., que podem conter termos sexuais): as conversas predatórias normalmente apresentam a ocorrência de termos sexuais. As conversas não predatórias presentes no conjunto de dados não apresentam essa característica.
- Transcrição de mensagens de áudio: uma proporção não informada de mensagens que compõem as conversas não predatórias são transcritas de áudios. Essa característica impacta diretamente a ocorrência da forma grafo-linguística difundida em textos de conversas virtuais [Komesu and Tenani, 2009].

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

Dessa maneira, compreendeu-se que apenas as conversas predatórias contribuiriam para a atual pesquisa. Com isso, surgiu a necessidade de criação de um conjunto de dados que representasse mais precisamente as conversas que ocorrem em meios virtuais. Nesse cenário, para a construção do conjunto de dados a ser usado na pesquisa, o método proposto por Inches and Crestani [2012] (aplicado na PAN-2012) é considerado como base. A Figura 8 apresenta todos as etapas contempladas na construção do conjunto de dados PRED-2050-ALL.

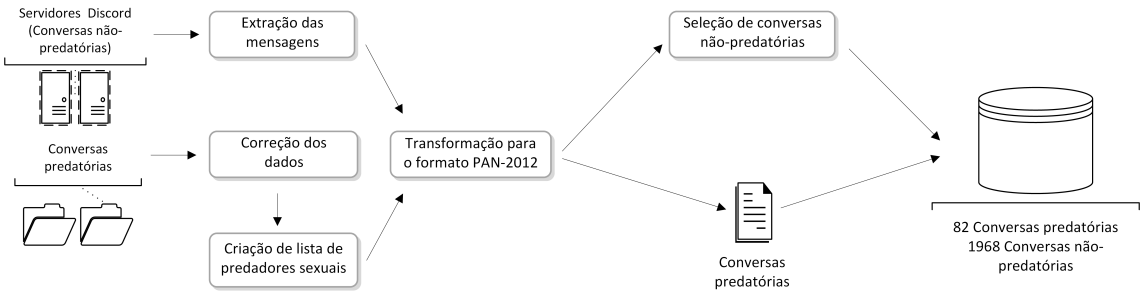


Figura 8 – Etapas para criação do conjunto de dados PRED-2050-ALL: 1) Extração de conversas não predatórias de diferentes categorias, oriundas de comunidades virtuais no Discord; 2) Acerto de dados em conversas predatórias e geração de lista de predadores sexuais; 3) Transformação das conversas extraídas para formato usado na competição "PAN-2012"; 4) Seleção de conversas não predatórias considerando a representatividade de cada categoria; 5) Construção do conjunto de dados PRED-2050-ALL.

3.2- Conversas predatórias

As conversas predatórias foram obtidas por meio de uma parceria acadêmica entre o Ministério Público Federal de São Paulo (MPF-SP) e o Centro Universitário da Fundação Educacional Inaciana (FEI). Inicialmente, o trabalho realizado por Andrijauskas et al. [2017] disponibilizou 39 conversas predatórias para a comunidade científica. Em um momento posterior, foram disponibilizadas 43 conversas predatórias adicionais para a presente pesquisa. No total, 82 conversas predatórias são consideradas para a construção do conjunto de dados PAN-2050-ALL. Essas conversas, antes sob sigilo de justiça, foram anonimizadas para que a identidade dos participantes fosse preservada. As informações sensíveis foram substituídas por marcações que representam um conceito de alto nível.

A Tabela 3 apresenta as marcações consideradas.

Tabela 3 – Marcações inseridas inicialmente na conversas disponibilizadas pelo MPF-SP e FEI 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 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 dos 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[2]. 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. 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 "*predators.txt*" apresentou um total de 82 predadores sexuais.

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

3.3- Conversas não predatórias

A criação de conversas não predatórias foi realizada por meio de um trabalho de extração de mensagens enviadas em comunidades virtuais hospedadas na plataforma Discord[3]. O Discord é uma plataforma constituída por diferentes comunidades que permitem que os participantes se comuniquem de diferentes maneiras: 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 usada para uma grande variedade de propósitos [Webb, 2018].

Uma ferramenta especializada na indexação de comunidades Discord foi usada[4] para a busca de comunidades pertencentes à diferentes categorias e na língua portuguesa do Brasil. As mensagens foram adquiridas por meio de uma ferramenta de código aberto[5]. Uma vez que a ferramenta é inicializada e autenticada em uma comunidade na plataforma Discord, é possível selecionar os canais de bate-papo desejados e então realizar a extração de todo o histórico de mensagens enviadas em um determinado canal. Para o presente trabalho, a categoria das conversas de uma comunidade foi definida com base na descrição dos servidores e o conjunto de *tags* que classificam a comunidade Discord na ferramenta de apoio. Ao todo, foram consideradas 5 categorias (Jogos, Política, Tecnologia, Estudos e Adulto). A Figura 9 ilustra uma das comunidades encontradas por meio da ferramenta de apoio.

Figura 9 – Exemplo de uma comunidade Discord sobre jogos em ambiente virtual.

[3] https://www.discord.com/
[4] https://disboard.org/
[5] https://github.com/Tyrrrz/DiscordChatExporter

Os responsáveis pela PAN-2012 [Inches and Crestani, 2012] sugeriram que a presença de conversas predatórias represente aproximadamente 4% do total de conversas presentes no conjunto de dados da competição [Inches and Crestani, 2012]. A decisão foi justificada pelo desbalanceamento natural das conversas predatórias. 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. No total, foi possível obter um total de 163.632 conversas com até 150 mensagens e pertencentes a 5 categorias consideradas. No entanto, para o presente cenário, considerar todas as conversas não predatórias aumentaria o desbalanceamento inicialmente proposto. Sendo assim, foi realizada uma etapa adicional na criação do conjunto de dados para a seleção das conversas não predatórias.

Nessa etapa, a representatividade de cada categoria foi considerada. Para isso, foi realizada uma amostragem estratificada de todas as conversas não predatórias com até 150 mensagens (total de 163.632 conversas) por meio da biblioteca Scikit-learn [Pedregosa et al., 2011]. O conjunto de dados após a etapa de seleção de conversas não predatórias pode ser observado na Tabela 4. Esse conjunto de dados recebeu o nome de PRED-2050-ALL.

Tabela 4 – Conjunto de dados PRED-2050-ALL.

Classe	Conversas
Predatória	82
Não predatória	1968

3.4- Análise estatística

Para uma melhor compreensão das conversas que compõem o conjunto de dados PRED-2050-ALL, 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 5. 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, em ambas as classes, a quantidade de termos em conversas apresentam 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 seguirem uma distribuição positivamente enviesada (não normal).

Tabela 5 – Análise estatística descritiva do conjunto de dados PRED-2050-ALL.

Característica	Classe predatória	Classe não predatória	Total
Termos	16.829	84.043	100.872
Vocabulário	2.921	17.274	18.742
Número de mensagens	4.157	16.411	20.568
Termos por conversa (μ)	205,23	42,70	49,20
Termos por conversa (σ)	394,36	135,72	157,85
Termos por mensagem (μ)	4,04	5,12	4,90
Termos por mensagem (σ)	4,21	12,40	11,24
Hapax Legomena	1.822	11.348	12.229
Dis Legomena	361	2.384	2.529

As conversas predatórias tendem a ser mais extensas e apresentar um volume maior de mensagens. Ao analisar as 4.157 mensagens enviadas em conversas com predadores sexuais, pode ser observado que 2.299 delas (55,30%) tiveram o predador sexual como o remetente. As demais mensagens (1.858 - 44,70%) foram enviadas majoritariamente por crianças e adolescentes. Em um total de 4 conversas 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. As Figuras 10 e 11 ilustram a distribuição das conversas predatórias e não predatórias no conjunto de dados PRED-2050-ALL, de acordo com a quantidade de mensagens. Em conjunto a isso, gráficos de *rug*[6] foram adicionados para permitir uma melhor interpretação da quantidade de mensagens.

[6]https://seaborn.pydata.org/generated/seaborn.rugplot.html

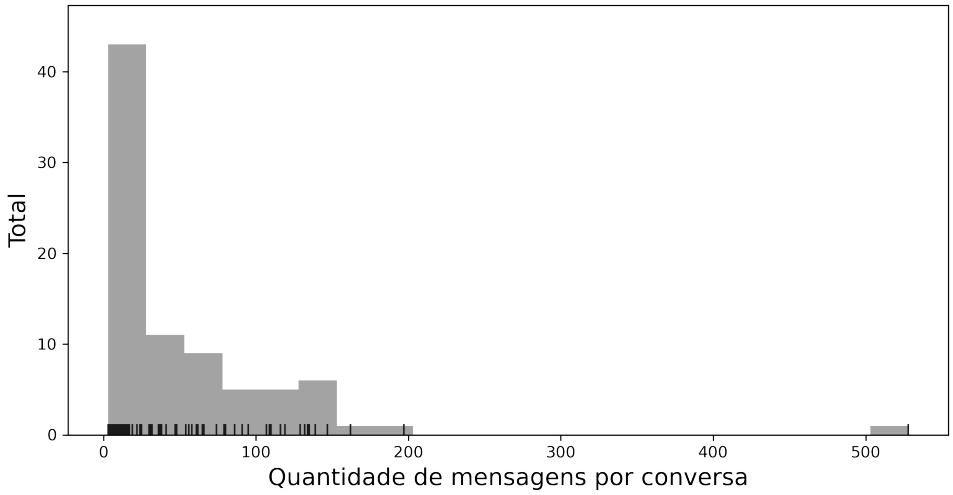


Figura 10 – Distribuição das conversas predatórias no conjunto de dados PRED-2050-ALL de acordo com a quantidade de mensagens. Cada conversa predatória apresentou uma média pouco superior a 50 mensagens trocadas e um desvio padrão superior à 70. A maioria das conversas apresenta apenas 4 mensagens trocadas, porém é possível encontrar conversas com até 528 mensagens.

Figura 11 – Distribuição das conversas não predatórias no conjunto de dados PRED-2050-ALL de acordo com a quantidade de mensagens. Cada conversa não predatória apresentou uma média superior a 8 mensagens trocadas e um desvio padrão superior à 14. A maioria das conversas não predatórias apresenta apenas 4 mensagens trocadas, porém é possível encontrar conversas contendo até 150 mensagens.

3.4.1 Percentual de termos raros

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, os termos que

apresentam a frequência de uso abaixo de um limiar θ 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 Equação 7 em 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 cenário do presente trabalho, são considerados termos raros os *Hapax Legomena* e *Dis Legomena* [Sokolova and Bobicev, 2018]. Sendo assim, o hiperparâmetro θ apresenta o valor igual a 3. Após verificação, é possível observar que 78,74% dos termos sem repetição presentes no conjunto de dados PRED-2050-ALL são considerados raros.

$$PTR = \frac{|\{t \in V | c(t) < \theta\}|}{|V|} \tag{7}$$

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]. No entanto, em determinados domínios de pesquisa, os termos raros proveem informações significativas em tarefas de classificação [Price and Thelwall, 2005].

Ao considerar apenas o vocabulário das conversas predatórias (i.e., $|V| = 2.921$), 74,73% é considerado raro (*Hapax Legomena + Dis Legomena*). Por outro lado, as conversas e mensagens não predatórias apresentam uma riqueza de vocabulário maior (79,49%). 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 6. É possível observar que, embora os percentuais de termos sobrepostos sejam baixos, a presença de termos raros em conversas predatórias é menos impactada com os efeitos de sobreposição entre as classes (2,56%). Considerando os termos com maior frequência no conjunto de dados, ou seja, os termos não raros, é possível observar que dos 738 termos presentes em conversas predatórias, 504

estão presentes no vocabulário de ambas as classes (73,17% dos termos predatórios). Esse fator contribui para o aumento da complexidade 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 6 – Sobreposição de termos presentes no conjunto de dados PRED-2050-ALL.

Característica	Conversas Predatórias	Conversas não predatórias	Sobreposição (% total)
Termos raros	2.183	13.732	409 (2,56%)
Demais termos (c(t)>2)	738	3.542	504 (11,77%)

3.5- Considerações finais

Nesse capítulo apresentamos uma das contribuições previstas para a presente pesquisa: o conjunto de dados PRED-2050-ALL. O conjunto é composto por 82 conversas predatórias e 1968 conversas não predatórias. Em um primeiro momento, foi apresentada a motivação para a criação do conjunto de dados, embora já existisse uma iniciativa publicada e disponível na internet [Andrijauskas et al., 2017]. A partir dessa iniciativa foram obtidas 39 conversas predatórias e, posteriormente, 43 conversas predatórias em acordo com o MPF-SP. A partir disso, todas as conversas escolhidas foram formatadas no padrão PAN-2012. O uso de um único padrão para criação de conjuntos de dados, já conhecido no domínio da pesquisa, fomentaria a adoção e a experimentação. Além das premissas estabelecidas por Inches and Crestani [2012] para a criação do conjunto de dados, foi considerado o conceito de representatividade para a escolha das conversas não predatórias para compor o PRED-2050-ALL. Em seguida, é realizada uma análise estatística nos dados. Nela, observa-se um PTR alto (78,74%), o que tornou oportuno uma melhor compreensão do fenômeno e os impactos na identificação de atividade predatória. Por fim, foi analisada a sobreposição de termos entre as classes visto ser um dos fatores que contribuem para o aumento da complexidade de identificação no domínio da pesquisa. Os resultados encontrados sugerem possíveis ganhos de desempenho uma vez desconsiderados os termos raros presentes no conjunto de dados PRED-2050-ALL.

4- Metodologia

O presente capítulo tem o objetivo de apresentar um método que permita a identificação de uma conversa predatória, no formato de texto, que tenha ocorrido na internet, isto é, uma conversa em que ocorra a presença de um predador sexual e uma vítima, potencialmente, uma criança ou adolescente. Conforme apresentado no Capítulo 2, existe uma lacuna em trabalhos que estudem a identificação de atividade predatória sexual por meio de características textuais e comportamentais para a língua portuguesa do Brasil. Dessa forma, são explorados diferentes conhecimentos presentes no domínio da pesquisa de forma a contribuir para uma representação mais precisa de características normalmente presentes em conversas predatórias. Cada característica identificada pelo método proposto é representada por um conceito de alto nível (CAN), isto é, um termo único que seja capaz de representar um subconjunto de termos com menor capacidade discriminativa e que contribui para melhorar a capacidade de identificação da atividade predatória.

As seções a seguir detalham o método proposto: a Seção 4.1 introduz o método proposto para a presente pesquisa e apresenta as características textuais e comportamentais que serão tratadas e representadas por meio de CANs. A partir desta, as Subseções 4.1.1, 4.1.2 e 4.1.3 descrevem os módulos que compõem o método proposto. A Seção 4.2 apresenta uma proposta de implementação do MDAP. Por fim, na Seção 4.3 são feitas as considerações finais.

4.1- MDAP: Método de Detecção de Atividade Predatória

Conforme apresentado nos Capítulos 1 e 2, a exploração das características textuais e comportamentais no domínio da pesquisa tem se apresentado como uma alternativa válida para a identificação da atividade predatória em conversas textuais na internet. No Capítulo 3, é observado que as conversas disponibilizadas pelo FEI e MPF-SP passaram por um processo de anonimização de informações sensíveis. Toda

informação sensível presente em uma conversa predatória encontra-se representada por meio de CANs e, desta forma, anonimizadas. As informações anonimizadas também são consideradas características importantes dentro do domínio da pesquisa. Diante do cenário descrito, compreende-se que para evitar uma avaliação tendenciosa das conversas e realizar a identificação das características que permitam encontrar uma conversa predatória é necessário que sejam cumpridas algumas premissas:

- Oferecer um método efetivo de normalização das conversas textuais, isto é, aplicar um conjunto de medidas de transformação textual de forma a maximizar o reconhecimento de termos candidatos a indicarem a presença de características que permitam identificar conversas predatórias. Por exemplo, um cenário presente no conjunto do dados PRED-2050-ALL é a identificação do cumprimento “olá” e suas variantes: “ola”, “Olá”, “Ola”. Uma outra situação que motiva a normalização das conversas é a identificação de elogios comuns em uma conversa predatória (e.g., “Linda”, “linda”). Os cenários descritos e que motivam a normalização das conversas textuais são ilustrados por meio de conversas predatórias na Figura 12. Em particular, as conversas 12a e 12b retratam três formas distintas de escrita do cumprimento “olá” pode ser encontrado em conversas, enquanto as conversas 12c e 12d retratam duas formas distintas de se encontrar o elogio “linda”.

- Permitir que as informações anonimizadas e representadas por meio de CANs nas conversas predatórias sejam identificadas de forma automática. Conforme apresentado na Subseção 1.1.1, as informações anonimizadas em conversas predatórias compõem um subconjunto de características a serem exploradas na identificação da atividade predatória.

- Possuir uma base de conhecimento para auxiliar a identificação de características a serem exploradas no processo de identificação de atividade predatória. Em alguns casos, a presença da característica está atrelada ao uso de um conjunto restrito de termos. A identificação de uma menção a parentes, o cômodo da casa em que a vítima se encontra no momento da conversa ou o emprego de elogios para fins predatórios são alguns dos exemplos de características que podem ser reconhecidas por meio dessa estratégia. Desta forma, entende-se que é necessário a criação de diferentes conjuntos de termos, no formato de um léxico, oriundo de fontes externas (e.g., sites, pesquisas, orgãos do governo) e de conhecimento público, assim

como de fontes internas, isto é, léxicos criados a partir das conversas predatórias presentes no conjunto de dados PRED-2050-ALL, para auxiliar o processo de reconhecimento de determinadas características que permitam identificar uma conversa predatória.

- Minimizar os possíveis efeitos de uma generalização indesejada ao realizar o mapeamento de uma característica predatória em conversas textuais. Por exemplo, tão importante quanto verificar se uma foto foi compartilhada em uma conversa, é a possibilidade de identificar o interesse do predador sexual em obter essa informação da vítima. Ao analisar as conversas predatórias, é possível observar que o uso do verbo "tirar" está frequentemente associado a um pedido de foto. No entanto, também é possível encontrar a associação em outras conversas com o termo "férias". A fim de contornar tais efeitos, devem ser introduzidas regras específicas, representadas por meio de algoritmos que, ao considerar o conhecimento do domínio da pesquisa, definam os critérios para a realização do mapeamento de uma determinada característica comportamental predatória.

De forma a atender as premissas expostas acima, é proposto o Método de Detecção de Atividade Predatória (MDAP). Ao todo, o método é composto por três módulos. O primeiro, o Módulo de Padronização de Conteúdo Textual Inicial (MPCTI), é responsável pela normalização e remoção de ruídos de uma determinada conversa textual. O segundo, o Módulo de Identificação de Comportamento Predatório (MICP) atua de forma a mapear diferentes categorias de características presentes em conversas predatórias e então representar as ocorrências mapeadas no formato de CANs. Por fim, o Módulo de Padronização de Conteúdo Textual Final (MPCTF) tem como principal propósito remover eventuais ruídos textuais mantidos após a execução do módulo MICP. A Figura 13 apresenta uma visão geral do MDAP.

Conforme é possível observar no módulo MICP, é explorada a identificação de duas categorias de características: (i) as comuns, isto é, são características consideradas relevantes para a identificação da atividade predatória sexual porém não são dependentes da iniciativa de um predador sexual para que a característica ocorra; (ii) predatórias, ou seja, são comportamentos apresentados pelo predador sexual em uma conversa de texto ao conversar com uma vítima. Para atingir esse objetivo, o MICP se propõe a identificar as duas categorias de características presentes em conversas predatórias

```
<conversation id="66">
  <message line="1">
    <author>951268944efa206879d8e3898b98c793</author>
    <text>Olá [nome]</text>
  </message>
  <message line="2">
    <author>ce9c5d8a12e9f0024a868c75bf80ef77</author>
    <text>ola [nome]</text>
  </message>
  <message line="3">
    <author>951268944efa206879d8e3898b98c793</author>
    <text>Tudo bem, [nome]?</text>
  </message>
  <message line="4">
    <author>ce9c5d8a12e9f0024a868c75bf80ef77</author>
    <text>tudo e com voce?</text>
  </message>
  <message line="5">
    <author>951268944efa206879d8e3898b98c793</author>
    <text>Eu estou bem</text>
  </message>
  ...
</conversation>
```

(a) Conversa Predatória 66

```
<conversation id="56">
  <message line="1">
    <author>f4fd5edf3069539ee4afc927ff22589b</author>
    <text>Ola gato</text>
  </message>
  <message line="2">
    <author>579257d82c944869b699000b4555a3b1</author>
    <text>Oii</text>
  </message>
  <message line="3">
    <author>579257d82c944869b699000b4555a3b1</author>
    <text>Rs</text>
  </message>
  <message line="4">
    <author>579257d82c944869b699000b4555a3b1</author>
    <text>[emoticon]</text>
  </message>
  <message line="5">
    <author>f4fd5edf3069539ee4afc927ff22589b</author>
    <text>Gosto da foto</text>
  </message>
  ...
</conversation>
```

(b) Conversa Predatória 56

```
<conversation id="42">
  <message line="1">
    <author>1fff885f525a16df641fb0ee198c0b25</author>
    <text>Oi linda boa noite</text>
  </message>
  <message line="2">
    <author>b762aedcce29e95c7c438c48ce7bdc2b</author>
    <text>Boa noite;)</text>
  </message>
  <message line="3">
    <author>1fff885f525a16df641fb0ee198c0b25</author>
    <text>Vc mora onde bb</text>
  </message>
  <message line="4">
    <author>b762aedcce29e95c7c438c48ce7bdc2b</author>
    <text>No [local]</text>
  </message>
  <message line="5">
    <author>1fff885f525a16df641fb0ee198c0b25</author>
    <text>Vc mora onde bb</text>
  </message>
  ...
</conversation>
```

(c) Conversa Predatória 42

```
<conversation id="57">
  ...
  <message line="94">
    <author>fb5efb70ad62de2a36a1b6b5e3e2e496</author>
    <text>Linda</text>
  </message>
  <message line="95">
    <author>e37d3f7cee16907edd4630582dd292ff</author>
    <text>Lindi</text>
  </message>
  <message line="96">
    <author>e37d3f7cee16907edd4630582dd292ff</author>
    <text>Lindo*</text>
  </message>
  <message line="97">
    <author>fb5efb70ad62de2a36a1b6b5e3e2e496</author>
    <text>Linda descupa se disse algo qui nao li agradou</text
      >
  </message>
  <message line="98">
    <author>e37d3f7cee16907edd4630582dd292ff</author>
    <text>Foi nada . Nao falo nada de mais</text>
  </message>
  ...
</conversation>
```

(d) Conversa Predatória 57

Figura 12 – Trechos de conversas predatórias em que pode ser observada a variação na escrita de termos. As conversas 12a e 12b apresentam como o cumprimento “olá” pode ser encontrado enquanto as conversas 12c e 12d retratam duas formas distintas de se encontrar o elogio “linda”.

por meio de três estratégias distintas. A primeira estratégia (E1) explora a identificação de diferentes características por meio de padrões textuais predefinidos (e.g., Telefone, Emojis). Na segunda estratégia (E2) é adotado apenas o uso de léxicos criados a partir de fontes internas e externas. Um cenário para o emprego dessa estratégia seria a identificação de nomes próprios em uma conversa. Por fim, a terceira estratégia (E3) explora unicamente as características presentes na literatura que dependem da iniciativa do predador sexual. Nesse contexto, a exploração de cada característica requer o mapeamento do comportamento predatório. Para cada característica mapeada, é definido

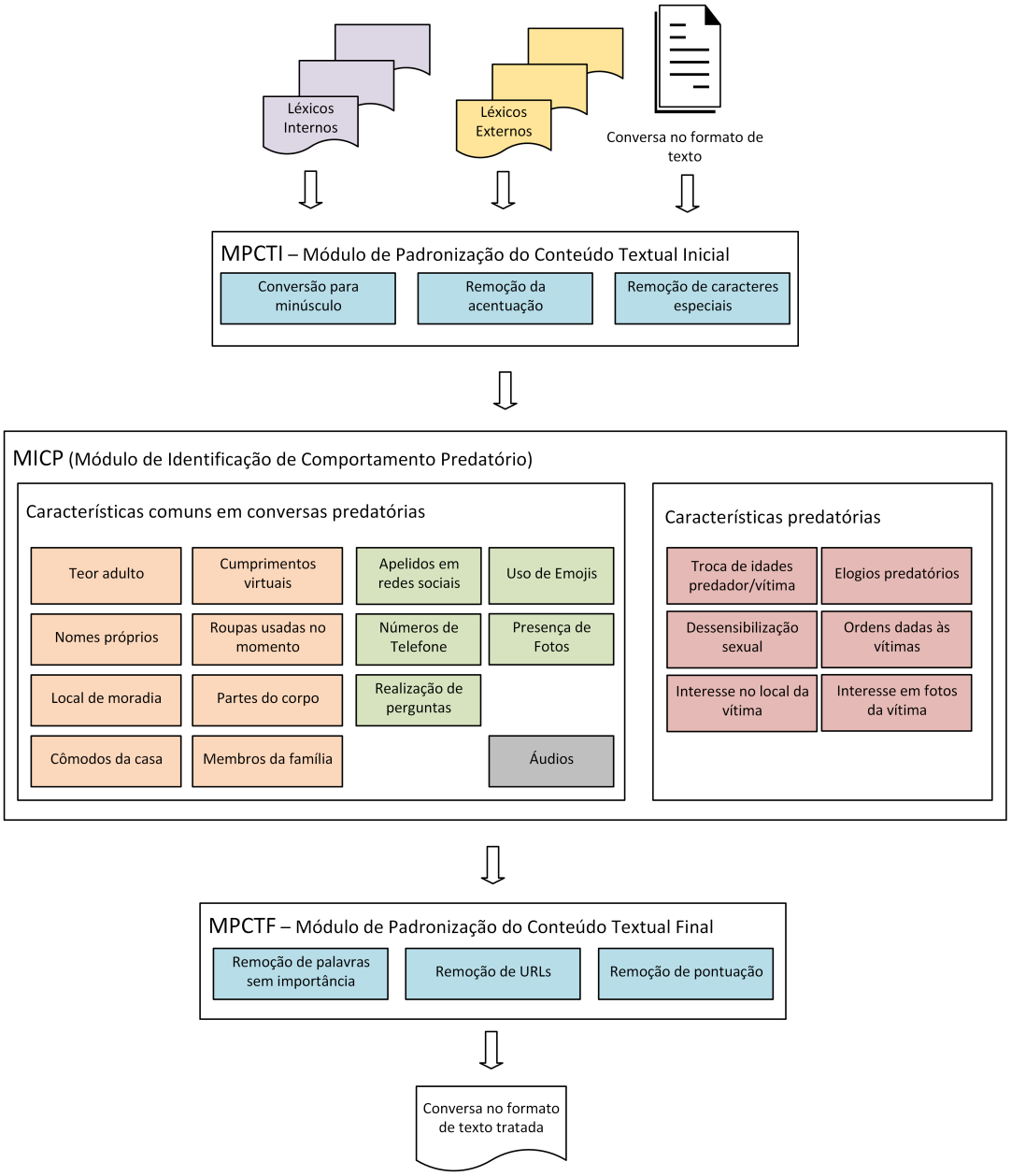


Figura 13 – MDAP: Método de Detecção de Atividade Predatória.

um conjunto de regras para a verificação do MDAP. De forma a permitir uma identificação mais precisa do comportamento predatório, o conjunto de regras faz uso de léxicos de

fontes internas, dada a necessidade de uma melhor compreensão dos termos usados usados por predadores sexuais.

Ao todo, o método proposto se propõe a identificar um total de dezenove características textuais e comportamentais. Dentre as características selecionadas para a criação do MDAP, as menções a áudios foram desconsideradas dado o foco da pesquisa, que explora o estudo da comunicação textual em conversas realizadas na internet. Dessa forma, todas as menções a áudios presentes nas conversas predatórias foram desconsideradas. A tabela 7 apresenta as estratégias e características exploradas no MDAP.

Tabela 7 – Características exploradas na identificação da atividade predatória sexual pelo MDAP. Pode-se notar as estratégias possíveis (E1, E2 e E3) para a identificação e os CANs associados. Dessa forma, cada característica explorada faz uso de uma estratégia e é representada por um CAN.

Estratégia	Característica	CAN
E1	Apelidos em redes sociais	Nome
E1	Números de telefone	Telefone
E1	Presença de fotos	Foto
E1	Realização de Perguntas	Pergunta
E1	Uso de Emojis[1]	Emoticon
E2	Cômodos da casa	Cômodo da casa
E2	Cumprimentos virtuais	Cumprimento
E2	Dessensibilização sexual	Teor adulto Predatório
E2	Elogios predatórios	Elogio
E2	Membros da família	Parente
E2	Nomes próprios	Nome
E2	Local de moradia	Local
E2	Ordens dadas às vítimas	Ordem dada
E2	Partes do corpo	Partes do corpo
E2	Roupas usadas no momento	Peças de roupa
E2	Teor adulto	Teor adulto
E3	Troca de idades predador/vítima	Idade maior/menor
E3	Interesse em fotos da vítima	Interesse em foto
E3	Interesse no local da vítima	Interesse em local

4.1.1 MPCTI: Módulo de Padronização de Conteúdo Textual Inicial

Conforme foi apresentado anteriormente no capítulo 3, as conversas não predatórias consideradas para a composição do conjunto de dados PRED-2050-ALL foram extraídas da plataforma Discord. Essas conversas geralmente apresentam um alto grau de ruído. Nas conversas predatórias o fenômeno também ocorre, porém por diferentes motivações, como por exemplo, o emprego de erros ortográficos pelo predador sexual com o propósito de apresentar uma comunicação mais receptiva pela vítima e, com isso, uma maior intimidade.

Ao considerar o contexto descrito, compreende-se que a identificação de termos e comportamentos e sua posterior representação pode ser prejudicada se não ocorrer uma normalização da escrita presente nas conversas textuais. Desta forma, o MPCTI é aplicado em um primeiro momento com o objetivo de normalizar a escrita presente nas conversas textuais e em léxicos que suportarão o método de forma a tornar mais eficiente a identificação das características exploradas. Em adição a isso, a remoção de caracteres especiais também é considerada devido a ausência de emprego em conversas predatórias. Nesse cenário, o MPCTI atua como um primeiro passo para a normalização e limpeza de ruídos. A seguir, o algoritmo 1 descreve as transformações realizadas nas conversas textuais e nos léxicos.

O algoritmo mod_mpcti recebe dois parâmetros: a conversa C_{xml} oriunda do conjunto de dados PRED-2050-ALL e o conjunto de léxicos DL que atua no suporte à identificação de características presentes em conversas predatórias. O primeiro passo é a transformação da conversa em um documento de texto, isto é, em uma cadeia única de caracteres. A conversão para um documento facilita a aplicação das transformações posteriores, assim como permite obter um melhor desempenho ao método ao desconsiderar a necessidade de interpretar documentos XML. Em seguida, nos Passos 2 e 3, toda a conversa é convertida para minúscula e caracteres especiais são removidos. Neste cenário, foram considerados apenas os caracteres subscritos e superescritos visto a forma como são explorados em conversas não predatórias, por exemplo, em apelidos e em salas de batepapo nas comunidades Discord. No Passo 4, é definido o dicionário DL_n no estado inicial, sem entradas. A seguir, o Passo 5 realiza uma iteração para cada associação de CAN c e léxico l, representadas por entradas (c, l), presentes no dicionário

Algoritmo 1 – $mod\_mpcti(C_{xml}, DL)$

Input :

- C_{xml} = Conversa presente no conjunto de dados PRED-2050-ALL no formato PAN-2012 (XML)
- DL = Dicionário de léxicos não normalizados por CAN

Output : Tupla contendo dois valores:

- C_d = Conversa no formato de um documento normalizada
- DL_n = Dicionário de léxicos normalizados por CAN

1 $C_d \leftarrow convert\_pan12\_to\_document(C_{xml})$
  // Ex. de valor para C_d: ‘‘Oi tudo bem? tudo e vc? tudo bem tb. Vc viu o filme? não. puxa!’’
2 $C_d \leftarrow convert\_to\_lowercase(C_d)$
3 $C_d \leftarrow remove\_special\_chars(C_d)$
4 $DL_n \leftarrow \emptyset$
  // c: CAN presente em DL
  // l: léxico associado ao CAN c
5 **foreach** $(c, l) \in DL$ **do**
6 | $l_n \leftarrow convert\_to\_lowercase(l)$
7 | $l_n \leftarrow remove\_special\_chars(l_n)$
8 | $DL[c]_n \leftarrow l_n$
9 **end**
  // Ex. de valor para C_d: ‘‘oi tudo bem tudo e vc tudo bem tb vc viu o filme não puxa’’
10 **return** (C_d, DL_n)

DL_n. O objetivo com essa iteração é aplicar o mesmo tratamento dado à conversa C_{xml} ao léxico l. A cada léxico l normalizado é gerado um léxico l_n. Por fim, no Passo 8, o léxico l_p é armazenado no dicionário DL_n. O algoritmo finaliza no Passo 10, com o retorno da conversa C_d e o dicionário DL_n.

4.1.2 MICP: Módulo de Identificação de Comportamento Predatório

O módulo MICP foi desenvolvido com o propósito de identificar as características normalmente presentes em conversas predatórias. Para atingir esse objetivo, o módulo MICP apresenta a capacidade de executar as três estratégias distintas (i.e. E1, E2 e E3, detalhadas na Seção 4.1). Em conjunto a isso, o MICP permite desconsiderar um CAN,

se aplicável. A seguir, o algoritmo 2 é apresentado.

Algoritmo 2 – $mod\_micp(C_d, concSet_{inc}, concSet_{exc}, DL_n, DP)$

Input :

- C_d = Conversa no formato de um documento normalizada no módulo MPCTI
- $concSet_{inc}$ = Conjunto de CANs a serem identificados
- $concSet_{exc}$ = Conjunto de CANs a serem desconsiderados
- DL_n = Dicionário de léxicos por CAN normalizados no módulo MPCTI
- DP = Dicionário de padrões textuais por CAN

Output :

- C_d = Conversa após o mapeamento dos CANs encontrados

1 **foreach** $conc$ *in* $concSet_{exc}$ **do**
  `// Áudios`
2 $C_d \leftarrow clean\_up\_concept(C_d, conc)$
3 **end**
4 **foreach** $conc$ *in* $concSet_{inc}$ **do**
  `// Telefones (e.g., ‘‘NNNNN-NNNN’’), Fotos (e.g., ‘‘*.jpg’’)`
  `// Estratégia 1 (E1)`
5 **if** $DP[conc] \neq \emptyset$ **then**
    $C_d \leftarrow apply\_pattern\_based\_concept(C_d, conc, DP[conc])$
  `// Elogios (e.g., ‘‘novinh*’’, ‘‘lind*’’)`
  `// Estratégia 2 (E2)`
6 **if** $DL_n[conc] \neq \emptyset$ **then**
    $C_d \leftarrow apply\_lexical\_based\_concept(C_d, conc, DL_n[conc])$
7 **end**
 `// Estratégia 3 (E3)`
8 $C_d \leftarrow apply\_concept\_for\_age\_verification(C_d, DL_n[\text{“}idade\text{”}])$
9 $C_d \leftarrow apply\_concept\_for\_photo\_interest(C_d, DL_n[\text{“}foto\text{”}], DP[\text{“}foto\text{”}])$
10 $C_d \leftarrow apply\_concept\_for\_location\_interest(C_d, DL_n[\text{“}local\text{”}], DP[\text{“}local\text{”}])$
11 **return** C_d

O algoritmo 2 (mod_micp) recebe seis parâmetros de entrada: a conversa C_d, no formato de um documento após processamento no módulo MPCTI. O segundo e o terceiro parâmetro representam dois conjuntos: O conjunto $concSet_{exc}$ remete aos CANs previamente mapeados e que não devem ser considerados no momento da execução do módulo (e.g., Áudio). Por outro lado, o conjunto $concSet_{inc}$ contempla todos os CANs que devem ser representados na conversa C_d. Por fim, são parametrizados três dicionários (DL_n, DP e DF) que são responsáveis por definir as diretrizes de execução do módulo: O dicionário DL_n armazena todos os léxicos que auxiliam na

identificação de uma determinada característica presente em uma conversa predatória; O dicionário DP auxilia na identificação de todos os padrões textuais previstos e, por último, o dicionário DF permite a verificação de características comportamentais em conjunto com o processamento de regras personalizadas.

A primeira tarefa prevista no algoritmo 2 permite que determinados CANs previamente presentes em uma conversa C_d sejam removidos. Este passo foi considerado devido a necessidade de remoção de um CAN presente nas conversas predatórias: "Áudio". Sendo assim, no Passo 1, é possível observar uma iteração para cada CAN $conc$ no conjunto $concSet_{exc}$. Em seguida, o Passo 2, realiza a remoção do CAN $conc$ por meio da função $clean\_up\_concept$.

A seguir, o Passo 4 realiza uma iteração por cada CAN $conc$ presente no conjunto $concSet_{inc}$. Para cada CAN $conc$ é aplicada a estratégia esperada. As estratégias apresentadas na Seção 4.1 são encontradas a seguir. A estratégia 1 (E1), responsável por mapear os CANs identificáveis por meios de padrões, se encontra no Passo 5. A E1 faz uso da função $apply\_pattern\_based\_concept$. Na função, o CAN $conc$ é considerado internamente para a busca do padrão $DP[conc]$ e posterior substituição das ocorrências encontradas na conversa C_d. Em seguida, o Passo 6 apresenta a estratégia 2 (E2), que explora a identificação de CANs por meio de léxicos de origem interna e externa. Na aplicação da E2 é possível observar uma referência à função $apply\_lexical\_based\_concept$. Esta função é responsável por identificar todos os termos presentes no léxico $DL_n[conc]$ com base no CAN $conc$ em uma conversa C_d. Ainda nesse contexto, cada termo identificado é substituído pelo CAN $conc$. Por fim, a execução da estratégia E3 contempla a aplicação de três funções distintas, conforme pode ser observado nos Passos 8, 9 e 10.

O Passo 8 introduz a função $apply\_concept\_for\_age\_verification$, detalhada no Algoritmo 3. Em seguida, no Passo 9 a função $apply\_concept\_for\_photo\_interest$ (Algoritmo 4) é aplicada na conversa C_d. Por fim, no Passo 10, é aplicada a regra para identificar o interesse em descobrir a localidade de alguma pessoa em uma conversa C_d. Ao término da execução da estratégia E3, no Passo 11, a conversa C_d é retornada após o potencial mapeamento dos CANs.

O algoritmo 3 apresenta os passos necessários para a identificação da troca de idades entre um predador sexual (maior de idade) e uma vítima (menor de idade), um comportamento frequente em conversas predatórias. No algoritmo, são esperados quatro parâmetros na entrada: a conversa C_d e o léxico L_n. O algoritmo se inicia com a definição

das faixas etárias a serem consideradas para os participantes da conversa C_d.

Os Passos 1 e 2 definem a faixa etária considerada para um menor de idade e a definição do CAN associado à presença dessa característica. A faixa etária considerada para os menores de idade é baseada em BARBOSA [2018]. O mesmo procedimento é realizado para a faixa etária adulta, representados nos Passos 3 e 4. O limiar superior para a faixa etária adulta considerou os dados mais recentes disponibilizados no Disque 100 [ONDH, 2020].

Algoritmo 3 – $apply\_concept\_for\_age\_verification(C_d, L_n)$

Input :

- C_d = Conversa em processamento no módulo MICP
- L_n = Léxico normalizado com os termos comuns relacionadas à troca de idades entre o predador sexual e um menor de idade

Output :

- C_{mcp} = Conversa de texto tratada com as regras para verificação de idade

```
1  underage_range = 9..17
2  minor_age_concept = "idade_menor"
3  adult_range = 18..100
4  adult_age_concept = "idade_maior"
5  has_ocurrences ← False
6  foreach l in L_n do
7      if l ∈ C_d then
8          has_ocurrences ← True
9          break
10     end
11 end
12 if has_ocurrences then
13     foreach ua in underage_range do
14         C_d.replace(ua, minor_age_concept)
15     end
16     foreach aa in adult_range do
17         C_d.replace(aa, adult_age_concept)
18     end
19 end
20 return C_d
```

O Passo 5 inicia $has\_ocurrences$ com o valor $False$. A partir de então, no Passo 6, para cada termo l presente no léxico L_n é verificada a sua presença em uma conversa C_d. Uma vez que ocorra a presença de algum dos termo l, essa é registrada em $has\_ocurrences$, conforme descrito no Passo 8.

No Passo 12, é avaliado se algum termo relacionado à troca de idades foi encontrado por meio de $has\_ocurrences$. Caso tenha sido registrado alguma ocorrência, é iniciado no Passo 13 uma iteração para cada idade ua presente na faixa etária considerada para menores de idade $underage\_range$. Em seguida, no Passo 14, é realizada a substituição da idade ua pelo conceito $minor\_age\_concept$ por meio da função $replace$. De forma análoga, os Passos 16, 17 e 18 verificam a presença de idades de predadores sexuais. Por fim, o Passo 20 retorna a conversa C_d.

Algoritmo 4 – $apply\_concept\_for\_photo\_interest(C_d, L_n, P_n)$

Input :

- C_d = Conversa em processamento no módulo MICP
- L_n = Léxico normalizado com os termos comuns relacionadas à pedidos de foto da vítima
- P_n = Padrão textual para o CAN alvo `// (e.g. expressões regulares)`

Output :

- C_d = Conversa de texto tratada com as regras para identificação de pedidos de foto da vítima

1 $ocrsSet_{li} \leftarrow \emptyset$
2 $c_{ngrams} = ngrams(C_d, 1)$
3 **foreach** li *in* L_n **do**
4 | **if** $li \in c_{ngrams}$ **then**
5 | | $ocrsSet_{li} \leftarrow ocrsSet_{li} \cup li$
6 | **end**
7 **end**
8 **if** $ocrsSet_{li} \neq \emptyset$ *and* $search(P_n, C_d)$ **then**
9 | **foreach** ocr_{li} *in* $ocrsSet_{li}$ **do**
10 | | $C_d.replace(ocr_{li}, \text{“}interesse\_foto\text{”})$
11 | **end**
12 **end**
13 **return** C_d

O algoritmo 4, responsável por identificar o interesse do predador sexual em obter fotos da vítima, recebe quatro parâmetros como entrada: a conversa C_d, o CAN $conc$, o léxico L_n e o padrão textual P_n. No Passo 1, o conjunto $ocrsSet_{li}$ é iniciado vazio. Em seguida, no Passo 2, são obtidos todos os unigramas presentes em uma conversa C_d por meio da função $ngrams$. A seguir, no Passo 3, é realizada a busca por termos li, presentes no léxico L_n. Todo termo li identificado dentre o conjunto de unigramas C_{ngrams} é armazenado no conjunto de termos $ocrsSet_{li}$ (Passo 5).

Algoritmo 5 – $apply\_concept\_for\_location\_interest(C_d, L_n, P_n)$

Input :

- C_d = Conversa em processamento no módulo MICP
- L_n = Léxico padronizado com termos comuns relacionadas à pedidos de localização da vítima `// e.g., ‘‘onde mora’’, ‘‘mora aonde’’`
- P_n = Padrão textual para o CAN alvo

Output :

- C_d = Conversa de texto tratada com as regras para identificação de pedidos de localização da vítima

1 $ocrsSet_{ll} \leftarrow \emptyset$
2 $c_{ngrams} = ngrams(C_d, 1) + ngrams(C_d, 2)$
3 **foreach** ll *in* L_n **do**
4 **if** $ll \in c_{ngrams}$ **then**
5 $ocrsSet_{ll} \leftarrow ocrsSet_{ll} \cup ll$
6 **end**
7 **end**
8 **if** $ocrsSet_{ll} \neq \emptyset$ **then**
9 **foreach** ocr_{ll} *in* $ocrsSet_{ll}$ **do**
10 $C_d.replace(ocr_{ll}, \text{“}interesse\_local\text{”})$
11 **end**
12 **end**
13 **return** C_d

No Passo 8, são realizadas duas verificações. A primeira está relacionada quanto a presença de algum termo li no conjunto $ocrsSet_{li}$. A segunda verificação está diretamente relacionada a ocorrência de padrão textual. Esta segunda verificação ocorre por meio da chamada à função $search$ em que são informados dois parâmetros: o padrão textual P_n e a conversa C_d.

Uma vez que seja identificado algum termo associado ao pedido de fotos à vítima e a presença do padrão textual P_n na conversa C_d, cada termo ocr_{li} presente no conjunto $ocrsSet_{li}$ é substituído pelo CAN $conc$ no Passo 10. Ao final, a conversa C_d é retornada (Passo 13).

A última regra implementada para o MDAP, o algoritmo 5 apresenta uma proposta similar ao algoritmo 4 para a identificação do local em que a vítima se encontra. O algoritmo 5 se inicia com quatro parâmetros: a conversa C_d, o CAN $conc$, o léxico L_n e o padrão textual P_n. Tal como o algoritmo 4, no primeiro passo, é definido um conjunto $ocrsSet_{ll}$ para que sejam armazenadas as ocorrências de termos relacionados à pedidos

de localização da vítima. Em seguida, no Passo 2, o conjunto C_{ngrams} recebe todos os unigramas e bigramas gerados a partir da conversa C_d. Em seguida, no Passo 3 ocorre uma iteração para cada termo ll presente no léxico L_n. No Passo 4, é verificado a presença do termo ll. Uma vez identificado como um termo presente no conjunto C_{ngrams}, ele é adicionado ao conjunto $ocrsSet_u$ (Passo 5).

Uma vez que o conjunto $ocrsSet_u$ não esteja vazio, é iniciado o processo de substituição dos termos suspeitos em conceitos. O Passo 9 realiza uma iteração para cada termo ocr_u pertencente á $ocrsSet_u$ e então, no Passo 10, é realizada a substituição do termo ocr_u pelo CAN $conc$ na conversa C_d por meio da função $replace$. Por fim, no Passo 13 é retornada a conversa C_d.

4.1.3 MPCTF: Módulo de Padronização de Conteúdo Textual Final

Em um momento posterior, após o processamento do módulo MICP, algumas informações ruidosas ainda podem estar presentes na conversa textual a ser analisada. Parte dos ruídos são oriundos da manutenção das palavras sem importância e sinais de pontuação por conta do uso de diferentes léxicos. Um caso de exemplo que sofreria impacto seria uma eventual busca por locais como o “Rio de Janeiro” ou “Dr. Sá Fortes”. Outra fonte de ruídos é a manutenção inicial das URLs devido ao fato que podem refletir um envio público de foto para os participantes em conversas não predatórias. Sendo assim, uma extensão do módulo MPCT é aplicada de forma a remover os ruídos restantes. O algoritmo 6 detalha o processo de remoção dos ruídos restantes.

O algoritmo mod_mpctf possui apenas uma entrada: a conversa C_{micp}, isto é, o resultado de uma conversa processada pelo módulo MICP (pode conter características predatórias mapeadas). O Passo 1 realiza a remoção de todas as palavras sem importância por meio da função $remove\_stopwords$. Como referência para a remoção, é considerado o conjunto de palavras sem importância presentes na biblioteca NLTK [Loper and Bird, 2002] para a língua portuguesa. A seguir, no Passo 2, todas as URLs em que não foi possível identificar uma imagem compartilhada na conversa são removidas. Por fim, no Passo 3, os sinais de pontuação restantes também são removidos. Por fim, o Passo 4 finaliza o algoritmo ao retornar a conversa C.

Algoritmo 6 – $mod\_mpctf(C_{micp})$

Input :

- C_{micp} = Conversa no formato de documento após processamento do módulo MICP

Output :

- C = Conversa de texto tratada

$C \leftarrow remove\_stopwords(C_{micp})$
$C \leftarrow remove\_urls(C)$
$C \leftarrow remove\_punctuation(C)$
return C

4.2- MDAP: Proposta de implementação

A seguir, o algoritmo 7 apresenta uma proposta de implementação do MDAP, de acordo com os módulos que o compõe e apresentados ao longo do Capítulo 4 (MPCTI, MCIP, MPCIF).

Algoritmo 7 – $met\_mdap(DS_{xml}, concSet_{inc}, concSet_{exc}, DL, DP)$

Input :

- DS_{xml} = Conversas no formato PAN-2012 (XML)
- $concSet_{inc}$ = Conjunto de CANs a serem identificados
- $concSet_{exc}$ = Conjunto de CANs a serem desconsiderados
- DL = Dicionário de léxicos não normalizados por CAN
- DP = Dicionário de padrões textuais por CAN

Output :

- DS_d = Conversas selecionadas após a aplicação do MDAP

$DS_d \leftarrow \emptyset$
foreach C_{xml} *in* DS_{xml} **do**
  $C_d, DL_n \leftarrow mod\_mpcti(C_{xml}, DL)$
  $C_{micp} \leftarrow mod\_micp(C_d, concSet_{inc}, concSet_{exc}, DL_n, DP)$
  $C \leftarrow mod\_mpctf(C_{micp})$
  $DS_d \leftarrow DS_d \cup C$
end
return DS_d

O algoritmo met_mdap possui cinco parâmetros de entrada. O primeiro parâmetro,

o conjunto de conversas DS_{xml}, possui o formato em XML definido pela competição PAN-2012. Desta forma, cada conversa deve seguir o formato apresentado na Subseção 1.1.2. Em seguida, são informados dois parâmetros de entrada que definem os conjuntos de CANs a serem considerados ou desconsiderados: $concSet_{inc}$ e $concSet_{exc}$. A definição desses dois conjuntos de CANs atende a uma necessidade do módulo MICP. Por fim, são parametrizados dois dicionários (DL e DP). O dicionário DL possui todos os léxicos a serem aplicados no MDAP e cada léxico apresenta um CAN associado. Um comportamento análogo ocorre com o dicionário DP. Nesse dicionário, são registrados todos os padrões textuais que devem ser mapeados para CAN.

A primeira tarefa, retratada no Passo 1 é a criação do conjunto de conversas resultante DS_d. A seguir, no Passo 2, é possível observar uma iteração para cada conversa C_{xml} no conjunto DS_{xml}. Em seguida, o Passo 3, inicia o processo de normalização de todo o conteúdo textual informado (conversas e léxicos) com o auxílio do módulo MPCTI (mod_mpcti). Ao término do processamento do módulo MPCTI, a conversa (C_d) e léxicos (DL_n) normalizados são passados como parâmetros de entrada para o módulo MICP (Passo 4). Além desses dois parâmetros também são informados os demais parâmetros exigidos ($concSet_{inc}$, $concSet_{exc}$ e DP). No Passo 5, é executado o último passo do MDAP: o módulo MPCTF é parametrizado com a conversa C_{micp}.Por fim, no Passo 6, uma vez finalizado o processamento do módulo MPCTF (mod_mpctf), a conversa sem os eventuais ruídos restantes (C) é adicionada ao conjunto DS_d. Uma vez que todas as conversas DS_{xml} são processadas, o algoritmo met_mdap retorna o conjunto de conversas processadas DS_d (Passo 8).

4.3- Considerações finais

No capítulo 2 são apresentados diferentes trabalhos com resultados significativos no domínio da pesquisa. Também foi identificado a ausência de trabalhos ao considerar as conversas predatórias com vítimas reais e a exploração de características presentes nessas conversas na língua portuguesa do Brasil para a elaboração de método que permita a identificação automática de atividade predatória sexual.

De forma a preencher essa lacuna, ao longo do presente capítulo foi apresentado

o MDAP, um método elaborado para a identificação de conversas predatórias realizadas na internet e na língua portuguesa do Brasil. Para tal, diferentes características presentes em conversas predatórias apresentadas no capítulo 1 foram selecionadas para o processo de identificação. Todas as características selecionadas para o estudo apresentaram uma natureza textual ou comportamental.

Após a apresentação, os três módulos integrantes do MDAP foram introduzidos. Primeiramente o MPCTI, responsável por tornar mais eficiente o processo de identificação das características selecionadas para a presente pesquisa. Em seguida, o módulo MICP, detalhou as dezenove características a serem contempladas e as três estratégias exploradas para a identificação das características em conversas textuais submetidas ao MDAP. Por fim, o módulo MPCTF detalha as ações necessárias para a remoção de eventuais ruídos remanescentes dos módulos aplicados anteriormente. Por fim, é apresenta uma proposta de implementação do MDAP em que se faz uso de todos os módulos previamente detalhados.

A seguir, no capítulo 5 é explorada a aplicação do MDAP em conjunto com diferentes algoritmos de aprendizado de máquina. De forma a validar a eficiência do método, ele é comparado com um *baseline*. Ao fim, é analisada a importância dos conceitos aplicados em diferentes limiares de frequência de termos nas conversas textuais.

5- Avaliação Experimental

O presente capítulo descreve os experimentos executados com o conjunto de dados PRED-2050-ALL, apresentado no capítulo 3 e o método proposto denominado MDAP, descrito no capítulo 4. A Seção 5.1, apresenta os léxicos que fornecem suporte para a aplicação do MDAP e a execução dos experimentos. Em seguida, na Seção 5.2 são detalhadas as configurações aplicadas nos experimentos realizados. Por fim, a Seção 5.3 apresenta e discute os resultados obtidos.

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 considerada na criação do ambiente. Os experimentos com os algoritmos SVM, NBM, MLP, DT e RF foram codificados com o auxílio da aplicação Web Jupyter Notebook 6.0.1 [Kluyver et al., 2016] e a biblioteca Scikit-learn na versão 0.21.3. A linguagem Python (versão 3.7.4) foi considerada como a padrão para toda a codificação realizada. Devido a possibilidade de um léxico conter uma grande quantidade de termos, o emprego de expressões regulares foi desconsiderado para a identificação do CAN correspondente. No seu lugar foi aplicado o método FlashText[1] [Singh, 2017].

5.1- Léxicos

Conforme apresentado no capítulo anterior, a identificação de determinadas características presentes em conversas predatórias é restrita à ocorrência de um conjunto específico de termos. Foram consideradas duas origens de termos para a criação dos léxicos a serem aplicados no MDAP:

- Externa: a presença de casos em que a característica presente em conversas predatórias é um conceito bem definido, isto é, de conhecimento público e sem a

[1] https://github.com/vi3k6i5/flashtext

interferência das particularidades do domínio da pesquisa. Uma motivação adicional para o uso de léxicos de origem externa é permitir que alguns dos conceitos de alto nível preenchidos manualmente em conversas predatórias (presentes nos dados do MPF) sejam identificados nas demais conversas presentes no conjunto de dados. Algumas das características predatórias, como a menção a informações pessoais (e.g., os nomes dos participantes e o locais mencionados), se encontram previamente anonimizadas. Outras características, como a menção a cômodos da casa também motivam a escolha dos termos em fontes externas.

- Interna: os termos selecionados para a criação dos léxicos foram extraídos de conversas predatórias. Um exemplo motivador para essa decisão é a forma como os elogios são explorados. Durante a análise dos termos empregados nas conversas predatórias foi possível observar um cenário em que poucos elogios eram explorados com grande frequência, como por exemplo o elogio “linda” - usado 284 vezes e os elogios “gostoso” e “gostosa”. No total, os dois elogios foram empregados em um total de 48 oportunidades. Ao considerar o emprego de flexões diversas no elogio “gostoso” (e.g., “gostosinho”, “gostosos”, “gostosas”) atinge-se um total de 58 ocorrências. Esse comportamento também se estende para outras características predatórias como o envio de ordens às vítimas.

5.1.1 Origem externa

A criação dos léxicos provenientes de origem externa considerou três fontes distintas para a busca dos termos: (i) Dados abertos do governo federal, por meio dos canais oficiais para divulgação de documentos e informações estatísticas. (ii) Diferentes dicionários, como o LIWC 2015 na sua versão para a língua portuguesa do Brasil [Carvalho et al., 2019] e um estudo recente sobre as partes do corpo humano mais citadas na literatura portuguesa [Gil, 2014]; (iii) Sites na internet que se apresentaram como uma fonte confiável para a extração de termos a respeito de uma determinada característica presente em uma conversa predatória.

A identificação de nomes próprios presentes em conversas considerou os dados

oriundos do Instituto Brasileiro de Geografia e Estatística [2] (IBGE). Para a obtenção da lista de nomes próprios de pessoas brasileiras, foi realizada uma solicitação ao sistema e-SIC, parte integrante do Portal de Acesso à Informação [3]. Como resultado, foram disponibilizados todos os nomes próprios brasileiros registrados que apresentaram ao menos 20 registros em cartórios civis. Ao todo, foram considerados 64.459 nomes próprios para a criação do léxico.

Para a representação de nomes de locais mencionados em conversas, o portal Sidra[4], também pertencente ao IBGE, foi usado como referência. Por meio de buscas no portal, foi possível recuperar todos os municípios, distritos, subdistritos e estados do Brasil. Ao todo, além do Distrito Federal, foram recuperados 26 estados, 682 subdistritos, 10.495 distritos e 5.569 municípios para a construção do léxico.

O levantamento dos possíveis graus de parentesco se deu por meio de um documento público presente na internet[5]. Ao todo, foram levantados 31 graus de parentesco possíveis. De forma similar, para a composição do léxico relativo aos cômodos de uma casa, foi considerada uma página na internet destinada ao ensino de vocabulário relacionado a cômodos da casa para crianças[6]. No total foram considerados 20 termos para a composição do léxico.

O léxico para auxílio na identificação de peças de roupa considerou apenas o nicho de roupas para uso íntimo (e.g., cueca, calcinha) e para um maior conforto dentro de casa como por exemplo,os pijamas e suas variações. Esta decisão foi tomada com base no resultado da análise individual das conversas em que se observa a presença de participantes das conversas, principalmente as vítimas, em casa. Logo, como estratégia para a construção do léxico, foram consideradas diversos sites de vendas *online*[7–12] com especialidade na venda de vestuário íntimo e próprio para uso dentro de casa. Após levantamento manual dos termos, foi considerado um total de 42 termos.

Para a criação do léxico relacionado às partes do corpo, foram consideradas duas

[2] https://www.ibge.gov.br/
[3] https://www.gov.br/acessoainformacao/pt-br
[4] https://sidra.ibge.gov.br/territorio
[5] http://www.pmf.sc.gov.br/arquivos/arquivos/pdf/22_05_2014_16.26.46.4d2554168e739ff213de782f7f262238.pdf
[6] https://www.infoescola.com/ingles/vocabulario-comodos-da-casa-mobilia-e-aderecos-rooms-furniture-appliances/
[7] https://www.monthal.com.br/blog/homewear/
[8] https://www.marcyn.com.br/pijamas
[9] https://www.joge.com.br/
[10] https://www.inspirate.com.br/pijamas-homewear
[11] https://www.trussardi.com.br/homewear/xale
[12] https://www.pijamania.net/

Tabela 8 – Léxicos criados a partir de fontes externas.

Nome	Conceito	Entradas
LEX_NOMES_BR	Nomes	64.459
LEX_LOCAIS_BR	Locais	16.772
LEX_PARTES_CORPO_BR	Partes do corpo	265
LEX_ADULTO_BR	Teor adulto	131
LEX_CASA_BR	Cômodos da casa	20
LEX_ROUPA_BR	Peças de roupa	42
LEX_PARENTES_BR	Parentes	31
LEX_SAUDACOES_BR	Saudações	14

contribuições existentes na literatura. A primeira contribuição, composta por 231 termos relacionados, foi a categoria 71 (*Biological Processes - Body*) do dicionário LIWC 2015 na sua versão para o Português do Brasil. Em adição a isso, a segunda contribuição considerou um levantamento dos termos relacionados ao corpo humano com maior frequência na língua portuguesa [Gil, 2014]. Com as duas colaborações, foi possível identificar um total de 265 termos.

Tabela 9 – Exemplo de entradas presentes nos léxicos criados a partir de fontes externas. Para cada um dos léxicos, foram retiradas dez entradas de forma a ter-se uma melhor compreensão dos dados.

Nome	Entradas
LEX_NOMES_BR	“MARIA”, “ANA”, “JOAO”, “GABRIEL”, “LUCAS”, “PEDRO”, “MATEUS”, “JOSE”, “GUSTAVO”, “VITORIA”
LEX_LOCAIS_BR	“Rio de Janeiro”, “Taguatinga”, “Vila Isabel”, “Ceará”, “Rio Branco”, “Palestina de Goiás”, “Panamá”, “Fortaleza”, “Plano Piloto”, “Centro”
LEX_PARTES_CORPO_BR	“cabelo”, “cabeça”, “olho”, “costas”, “pé”, “braço”, “mão”, “orelha”, “cotovelo”, “nariz”
LEX_ADULTO_BR	“amante*”, “consolo”, “orgia”, “perverter”, “pornô”, “promíscu*”, “fetiche*”, “ereto*”, “erótico*”, “gigolô*”
LEX_CASA_BR	“sala”, “sacada”, “cozinha”, “quarto”, “quintal”, “closet”, “banheiro”, “quintal”, “laje”, “telhado”
LEX_ROUPA_BR	“pijama”, “moletom”, “calça”, “vestido”, “sutiã”, “lenço”, “tomara que caia”, “biquini”, “camisola”, “camiseta”
LEX_PARENTES_BR	“pai”, “mãe”, “avô”, “avó”, “irmão”, “irmã”, “primo”, “prima”, “filho”, “filha”
LEX_SAUDACOES_BR	“oi”, “olá”, “opa”, “bom dia”, “boa tarde”, “boa noite”, “eaí”, “e aí”, “fala aí”, “qual é”

De forma diferente ao levantamento de termos relacionados a partes do corpo,

a presença de teor adulto em conversas predatórias remete ao uso de termos com o propósito de estimular sexualmente a vítima (e.g., menções à órgãos sexuais e zonas erógenas) Esses termos estão presentes em conversas informais e normalmente são considerados vulgares. Nesse cenário, a categoria 73 (*Biological Processes - Sexual*) do dicionário LIWC 2015 na sua versão para o Português do Brasil foi considerada como base. Ao todo, foram extraídos 131 termos. Dessa forma, na Tabela 8, é apresentado um resumo dos léxicos criados a partir de fontes externas. Em conjunto a isso, na Tabela 9, é apresentada uma amostra de dez exemplares presentes em cada um dos léxicos de origem externa criados. Os oito léxicos de origem externa encontram-se disponibilizados publicamente[13].

5.1.2 Origem interna

Para a criação dos léxicos de origem interna foram considerados alguns passos: A primeira etapa foi a aplicação do módulo MPCTI. Após a aplicação, foram extraídos os 1.000 unigramas e bigramas mais frequentes em conversas predatórias. Ao analisar os resultados encontrados, foram identificados tanto os termos com variadas flexões que podem refletir a ocorrência do comportamento predatório. Vale destacar que foram encontrados diversos termos em internetês, que são abreviaturas utilizadas em contextos específicos na internet. Nesse cenário, foi considerado o mesmo método empregado no LIWC 2015 [Pennebaker et al., 2015], em que um termo é flexionado até certa parte, em alguns casos até o radical, e posteriormente é acrescido um caractere *wildcard* (*). A Tabela 10 apresenta os léxicos internos e a Tabela 11 detalha as entradas resultantes da análise.

[13]https://github.com/LaCAfe/MDAP

Tabela 10 – Léxicos criados a partir de conversas predatórias (Origem interna).

Nome	Conceito	Entradas
LEX_PRED_ELOGIOS_BR	Elogio	8
LEX_PRED_INT_LOCAL_BR	Interesse em local	4
LEX_PRED_INT_FOTO_BR	Interesse em foto	1
LEX_PRED_ADULTO_BR	Teor adulto Predatório	23
LEX_PRED_ORDENS_BR	Ordem dada	5

Tabela 11 – Entradas presentes nos léxicos criados a partir de conversas predatórias (Origem interna).

Nome	Entradas
LEX_PRED_ELOGIOS_BR	“novinh*”, “princ*”, “gat*”, “bb*”, “menin*”, “bebe*”, “lind*”, “gostos*”,“am*”
LEX_PRED_INT_LOCAL_BR	“mor*”, “mora onde”, “mora aonde”, “vc mor*”
LEX_PRED_INT_FOTO_BR	“tir*”
LEX_PRED_ADULTO_BR	“trans*”, “masturb*”, “namor*”, “beij*”, “virg*”, “sex*”, “pelad*”, “fod*”, “ativo*”, “atv”, “pass”, “passivo*”, “goz*”, “nud*”, “pau”, “bumb*”, “bund*”, “bucet*”, “cuz*”, “chup*”, “safad*”, “xan*”, “faz*”
LEX_PRED_ORDENS_BR	“mand*”, “tir*”, “mostr*”, “respond*”, “env*”

5.2- Configuração dos experimentos

A Subseção 4.2 definiu uma proposta de implementação do MDAP. Dessa forma, para a realização dos experimentos, o algoritmo 7 é usado como base. Assim sendo, foram considerados os seguintes parâmetros e valores como entrada para o algoritmo:

- DS_{xml}: Conversas presentes no conjunto de dados PRED-2050-ALL (Seguem o formato PAN-2012)
- $concSet_{inc}$: Todos os CANs apresentados na Tabela 7 (“Nome”, “Telefone”, “Foto”, “Pergunta”, “Emoticon”, “Cômodo da casa”, “Cumprimento”, “Teor adulto Predatório”, “Elogio”, “Parente”, “Local”, “Ordem dada”, “Partes do corpo”, “Peças de roupa”, “Teor adulto”, “Idade”, “Interesse em foto”, “Interesse em local”)
- $concSet_{exc}$: O CAN “Áudio”.
- DL: Todos os léxicos associados aos conceitos disponibilizados para a presente

pesquisa. Podem ser encontrados nas Tabelas 8 e 10.

- DP: A Tabela 12 detalha todos os padrões textuais considerados para cada um dos CANs associados.

Tabela 12 – Padrões textuais considerados para o mapeamento de CANs.

CAN	Padrão textual
Realização de perguntas	`\?`

Números de telefone	`(\(\d{2}\)`
	`\d{5}-\d{4}|\d{2}-\d{5}-\d{4}|\(\d{2}\)`
	`\d{4}-\d{4}|(\d{2}-\d{4}-\d{4})`
	`|(\d{4}-\d{4})|(\d{9})|(\d{8})|\d{5}-\d{4})`

Presença de Fotos	`.*\.(jpg|png|gif)$`
Apelidos em redes sociais	`[@][^\s]+`
Uso de Emojis	`:\S+:`

5.2.1 Algoritmos de aprendizado de máquina

A escolha dos algoritmos de aprendizado de máquina para a realização dos experimentos considerou a aplicação no domínio da pesquisa e os resultados significativos obtidos na literatura. Nesse cenário, para a realização dos experimentos foram selecionados cinco algoritmos clássicos de aprendizado de máquina: Máquina de vetores de suporte (SVM), Naïve Bayes Multinomial (NBM); Árvores de decisão (DT), Florestas aleatórias (RF) e Redes Neurais Perceptron (MLP).

Para a realização dos experimentos, foram aplicados os hiperparâmetros definidos como padrão pela biblioteca Scikit-Learn 0.21.3. A única alteração considerada foi a escolha da função de kernel linear. A alteração é motivada pelo propósito do experimento realizado - identificar a ocorrência de atividade predatória sexual em uma conversa textual - que normalmente apresenta melhores resultados quando tratado de forma linear [Joachims, 1998].

5.2.2 Critérios de avaliação

De forma a avaliar o impacto da representação de características textuais e comportamentais na identificação de conversas predatórias, o MDAP foi comparado a um método denominado *baseline*. O *baseline* contempla a remoção dos conceitos de alto nível presentes em conversas predatórias. A motivação para a remoção dos conceitos se sustenta ao fato de que a representação dos conceitos tal como foi realizada durante o processo de anonimização não ocorreria em uma conversa extraída diretamente de uma aplicação de conversas da internet. Portanto, a manutenção desses conceitos implicaria na inserção de um viés. Após a remoção dos conceitos, os módulos MPCTI e MPCTF são aplicados em todas as conversas selecionadas para o experimento.

A partir dessa premissa, foram criados nove conjuntos de dados balanceados a partir do conjunto de dados PRED-2050-ALL. Cada um dos conjunto de dados foi criado com um total de 164 conversas. Primeiramente, foram consideradas todas as 82 conversas predatórias disponibilizadas. Em seguida, para a seleção das 82 conversas não predatórias foram considerados nove diferentes sementes (do inglês: seeds). Para cada semente foi realizada uma subamostragem aleatória sem reposição das conversas não predatórias.

5.2.3 Método de validação

A aplicação do processo de validação cruzada e estratificada em grupos foi considerada em todos os experimentos. Após uma análise que considerou os dados estatísticos obtidos no capítulo 3 e diferentes quantidades de grupos (i.e. três, cinco, dez), concluiu-se que um total de cinco grupos seria uma quantidade ideal, tendo em vista a quantidade de conversas predatórias e não predatórias presentes nos conjuntos de dados balanceados. Dessa maneira, o processo de validação cruzada e estratificada apresentou um total de cinco grupos. Uma vez definido o processo de validação cruzada, ele foi repetido por trinta vezes. A cada repetição, as conversas eram organizadas de forma aleatória no conjunto de dados. Esse comportamento tem como propósito explorar

a variabilidade dos possíveis resultados obtidos com o processo de validação cruzada.

Para a análise dos resultados obtidos com a validação cruzada e estratificada em grupos é considerada a medida de avaliação $F_{0,5}$. Ao considerar o valor $\beta = 0,5$, a ocorrência de FP proporciona uma maior penalização na análise do desempenho do experimento. O valor para o parâmetro β segue o proposto na competição PAN-2012.

Após a execução dos experimentos e a coleta dos resultados para cada validação em grupo é aplicado o teste não-paramétrico de Wilcoxon [Wilcoxon, 1992], a fim de verificar a significância estatística dos resultados entre o método proposto (MDAP) e o *baseline*. Em seguida, de forma a avaliar o ranking dos algoritmos de aprendizado de máquina em conjunto com o MDAP, é aplicado o teste não-paramétrico de Friedman [Friedman, 1937]. Em seguida, para verificar se a diferença identificada entre os algoritmos de aprendizado de máquina após o teste não-paramétrico de Friedman é estatisticamente significante, o método de Holm [Holm, 1979] foi utilizado como teste *post-hoc*. A plataforma STAC [Rodríguez-Fdez et al., 2015] foi considerada tanto para a aplicação do teste de Friedman quanto do método de Holm.

5.2.4 Definição do vocabulário e seleção de características

São exploradas duas estratégias distintas para a definição do vocabulário a ser considerado durante os experimentos. A primeira considerou a remoção de termos raros devido ao resultado da análise estatística realizada no conjunto de dados PRED-2050-ALL, descrito na Seção 3.4.1. A segunda estratégia explorou todos os termos presentes nas conversas.

Uma vez definido o vocabulário, o esquema de ponderação de termos TF-IDF é empregado. Após isso, são selecionados os top-k termos relevantes por meio da medida *Information Gain* (IG). Ao todo, foram considerados os seguintes valores para k: 25, 50, 100 e 200. O objetivo com o conjunto de valores apresentados para k é obter uma melhor compreensão da relevância das características textuais e comportamentais para a identificação de atividade predatória sexual. De forma a evitar a interferência entre os dados entre os conjuntos de treinamento e testes, a seleção das top-k considerou o

uso de Pipelines[14], disponibilizada pela biblioteca Scikit-learn. Sendo assim, as top-k características são selecionadas apenas do conjunto de treinamento.

5.3- Resultados

Esta Seção tem por objetivo avaliar o impacto da presença das características frequentemente presentes em conversas textuais e predatórias, identificadas por meio do MDAP, no reconhecimento de conversas predatórias. Desta forma, foram realizados diferentes experimentos. Cada experimento se encontra configurado de acordo com quatro parâmetros: o algoritmo de aprendizado de máquina, o conjunto de dados utilizado, o limiar para a definição do vocabulário empregado e o número de características k consideradas. Ao todo, foram considerados nove conjuntos de dados criados a partir do PRED-2050-ALL, cinco algoritmos de aprendizado de máquina, dois limiares distintos para a definição de vocabulário e quatro valores possíveis para a seleção das k características mais importantes.

A Seção se encontra organizada da seguinte forma: A Subseção 5.3.1 apresenta uma visão dos resultados obtidos com os experimentos de acordo o desempenho dos algoritmos de aprendizado de máquina, as top-*k* características mais relevantes e a significância estatística dos resultados. Ao término da apresentação, são selecionados os conjuntos de dados criados (dentre os nove conjuntos criados a partir do PRED-2050-ALL) em que os métodos MDAP e *baseline* apresentaram os resultados mais significativos. A Subseção 5.3.2 realiza uma análise estatística dos conjuntos de dados selecionados. A Subseção 5.3.3 apresenta os resultados obtidos com os conjuntos de dados selecionados e investiga as possíveis razões que originaram as diferenças de desempenho entre os conjuntos de dados. Por fim, na Subseção 5.3.4 é discutida a diferença de desempenho identificada após análise das medidas de desempenho. Os demais resultados não discutidos se encontram no Apêndice A.

[14] `https://scikit-learn.org/0.21/modules/generated/sklearn.pipeline.Pipeline.html`

5.3.1 Visão geral dos resultados

Esta Subseção tem por objetivo apresentar uma visão geral de todos os experimentos realizados com o MDAP e *baseline*. As análises realizadas consideraram como base os resultados dos experimentos apresentados na Tabela 23. Assim sendo, foram considerados os resultados oriundo dos nove conjuntos de dados criados a partir do PRED-2050-ALL; os dois limiares para definição de vocabulário (e.g., remoção de termos, uso de todos os termos); os cinco algoritmos de aprendizado de máquina (SVM, DT, RF, NBM e MLP); as top-*k* características mais importantes (25, 50, 100 e 200).

A seguir, os resultados dos experimentos realizados são analisados de acordo com alguns critérios. Em um primeiro momento, é analisado o desempenho dos algoritmos de aprendizado de máquina em relação às top-*k* características mais importantes. Em seguida, a significância estatística dos resultados do MDAP é analisada ao ser comparada com os resultados obtidos no *baseline*. Por fim, as características mais importantes são investigadas.

Quanto aos algoritmos de aprendizado de máquina

Ao analisarmos a aplicação do MDAP combinado com os diferentes algoritmos de aprendizado de máquina considerados para a pesquisa, é possível observar alguns fenômenos na distribuição dos resultados. Em todos os cinco algoritmos de aprendizado de máquina foram atingidos resultados superiores ao *baseline* em todos os valores considerados para a seleção de características. As Figuras 14a até a 14e ilustram os resultados alcançados. O algoritmo DT apresentou a ocorrência de uma anomalia dentre os resultados obtidos com o MDAP quando considerado o valor mínimo para a seleção das características (25). Dessa forma, ao ignorá-la observa-se que o *baseline* obteve resultados discretamente superiores apenas para esse cenário de avaliação.

O algoritmo RF apresentou a menor variabilidade dos resultados tanto para a aplicação do MDAP quanto para o *baseline*. Essa baixa variabilidade pode ser observada por meio do intervalo interquartil (IQR). A aplicação do MDAP permitiu atingir resultados

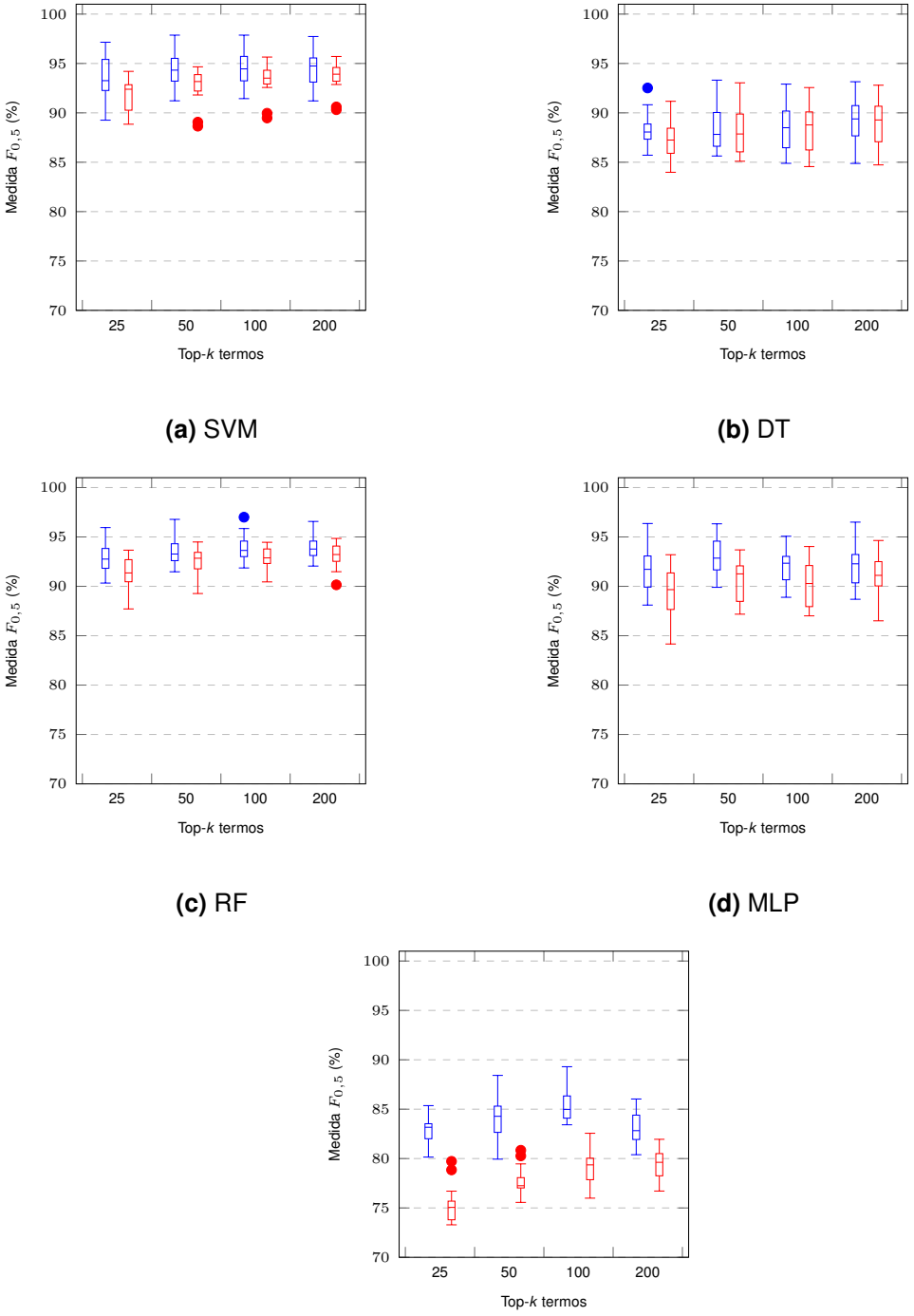
superiores a 97%. Ao comparar os resultados, é possível observar que aproximadamente 75% dos resultados obtidos com o *baseline* são iguais ou inferiores a 50% dos resultados obtidos com o MDAP. Também é possível observar a presença de duas anomalias. Ao considerar os cem termos mais relevantes, o MDAP apresentou 97,20% na medida $F_{0,5}$.

Por outro lado, o algoritmo NBM apresentou os resultados menos expressivos dentre todos os algoritmos de aprendizado de máquina considerados. No entanto, também é possível observar que o algoritmo NBM apresentou um maior ganho de desempenho quanto a aplicação do MDAP. Esse ganho de desempenho é constatada ao compararmos o IQR entre o MDAP e o *baseline*. Em todos os valores *k* considerados para a seleção das características é observada a ausência de sobreposição entre os dois IQR. A aplicação do MDAP em conjunto com o algoritmo NBM atingiu o melhor resultado, aproximadamente 90%, ao considerar os cem termos mais relevantes.

Conforme ilustrado na Figura 14a, o algoritmo SVM apresentou os resultados mais significativos para todas as quantidades de termos consideradas no processo de seleção de características. O uso de 50 termos permitiu atingir o maior resultado dentre todos os experimentos relacionados (97,87%). Os demais experimentos que consideraram 100 ou 200 termos, apresentaram um desempenho próximo. Um ponto importante, o algoritmo MLP apresentou comportamento similar ao algoritmo SVM, com capacidade de atingir resultados superiores a 95%. No entanto, o algoritmo MLP também apresentou uma variabilidade maior nos resultados para todas as diferentes quantidades de termos consideradas.

Quanto a significância estatística dos resultados

A análise da significância estatística dos resultados considera dois diferentes aspectos. Em um primeiro momento, analisa-se o grau de significância estatística entre os resultados dos experimentos com o MDAP e o *baseline*. Dentro desse contexto, foram considerados todos os experimentos sem distinção de parâmetros (e.g. top-k características, algoritmo de aprendizado de máquina e critério para a definição de vocabulário). Em um segundo momento, apenas para os resultados obtidos com o MDAP, é analisado o desempenho geral dos algoritmos de aprendizado de máquina. Como



(a) SVM

(b) DT

(c) RF

(d) MLP

(e) NBM

Figura 14 – Comparação do desempenho dos cinco algoritmos de aprendizado de máquina após aplicação dos métodos MDAP e *baseline* para cada um dos valores *k* considerados.

último ponto da análise, a significância estatística entre a diferença de desempenho para cada um dos algoritmos de aprendizado de máquina também é investigada.

A Figura 15 ilustra os resultados obtidos com o teste não-paramétrico de Wilcoxon. Conforme é possível observar, os resultados obtidos com a aplicação do MDAP para a identificação de atividade predatória sexual apresentaram significância estatística quando comparado aos resultados obtidos com o *baseline* na maioria dos conjuntos de dados criados para a análise. Dentre os nove conjuntos de dados gerados a partir do PRED-2050-ALL, apenas o conjunto de dados denominado "0" não apresentou significância estatística ($p > 0,05$) nos resultados obtidos ($p = 0,2$). Na avaliação dos demais conjuntos de dados, os resultados oriundos da aplicação do MDAP apresentaram significância estatística ($p < 0,001$).

A aplicação do teste não paramétrico de Friedman considerou a seguinte hipótese nula (H0): "A média dos resultados entre dois ou mais algoritmos de aprendizado de máquina é a mesma" e o intervalo de confiança $p = 0,05$. A tabela 13 apresenta a posição média (ranking) dos algoritmos de aprendizado de máquina após a aplicação do teste não paramétrico de Friedman. O resultado obtido, após a análise de todos os resultados dos experimentos com o MDAP, atingiu significância estatística ($p < 0,05$). Dessa forma, a hipótese nula foi rejeitada. Esse resultado permite interpretar que cada um dos algoritmos de aprendizado de máquina apresentou uma média significativamente estatística diferente. Essa conclusão é importante, visto que permite concluir que os algoritmos SVM e RF apresentaram os resultados mais expressivos dentre os experimentos realizados. De forma a melhor compreender se a diferença de desempenho entre os dois algoritmos é significante, foi considerado o método de Holm.

Tabela 13 – Resultados da aplicação do teste de Friedman em todos os experimentos realizados com a aplicação do método MDAP.

Algoritmo	Ranking
SVM	$1,44360$
RF	$1,97183$
MLP	$2,64085$
DT	$4,02817$
NBM	$4,91549$

A aplicação do método de Holm considera a seguinte Hipótese Nula (H0): "A média dos resultados de cada par de algoritmos de aprendizado de máquina é igual."

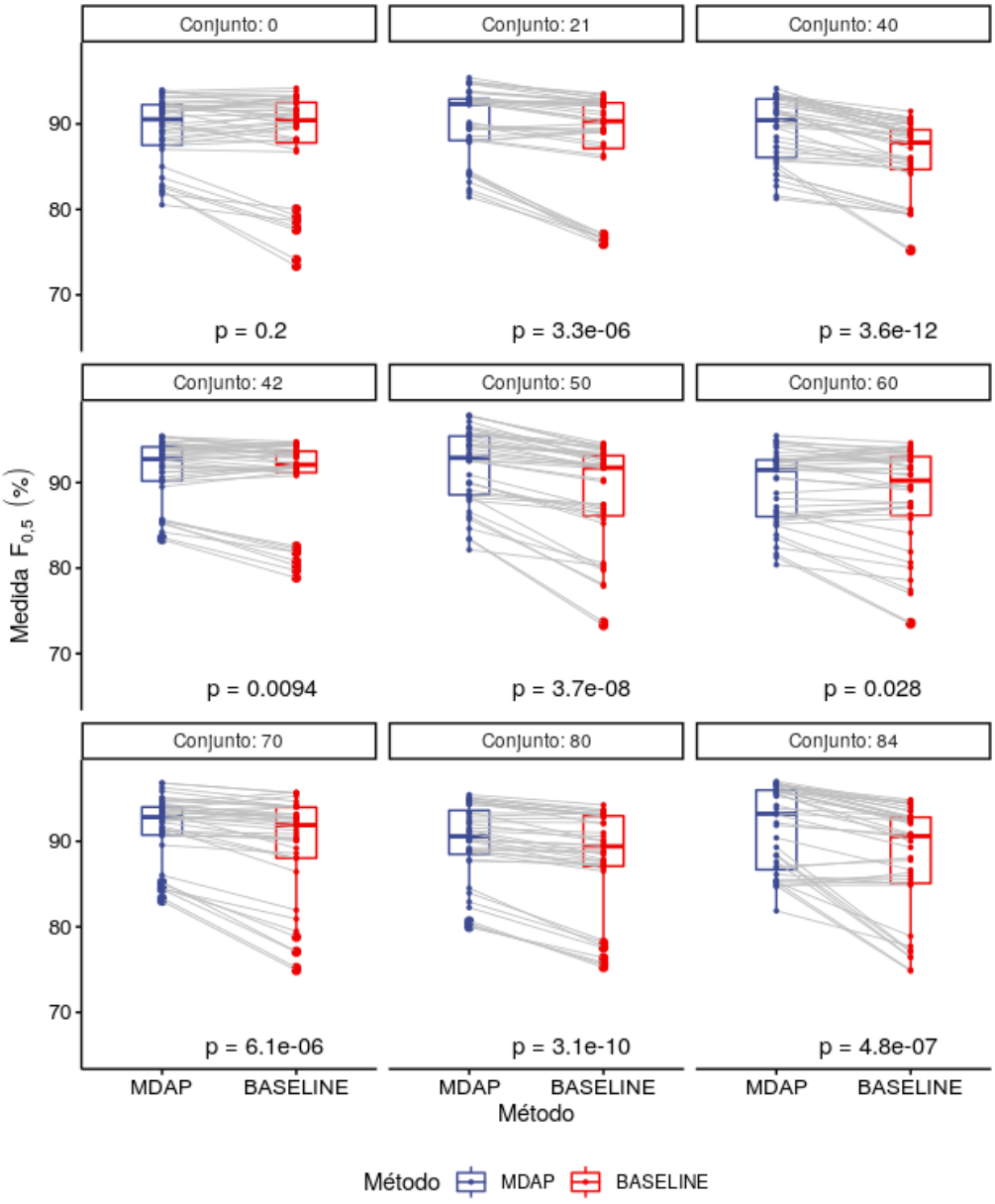


Figura 15 – Resultados obtidos após a aplicação do teste não-paramétrico de Wilcoxon com os diferentes resultados obtidos por meio da aplicação do MDAP e o *baseline*.

Os resultados encontrados (valores críticos e níveis de significância estatística) são apresentados na tabela 14. Ao analisar os valores críticos obtidos da comparação dentre o desempenho dos diferentes algoritmos de aprendizado de máquina em conjunto com a aplicação do MDAP e considerando o intervalo de confiança $p = 0,05$, é possível observar que a H0 foi rejeitada em todos os cenários de comparação. Esse resultado é importante, visto que permite afirmar que o algoritmo SVM, além de apresentar a melhor posição média dentre todos os experimentos realizados, também apresenta a média de

Tabela 14 – Resultados obtidos após a aplicação do teste *post-hoc* de Holm e $p = 0,05$.

Algoritmo	SVM	RF	MLP	DT	NBM
SVM	$0,00000$				
RF	$1,99029^{*}$	$0,00000$			
MLP	$4,51133^{***}$	$2,52104^{*}$	$0,00000$		
DT	$9,73917^{***}$	$7,74888^{***}$	$5,22784^{***}$	$0,00000$	
NBM	$13,08286^{***}$	$11,09257^{***}$	$8,57153^{***}$	$3,34369^{**}$	$0,00000$

NÍVEL DE SIGNIFICÂNCIA: $^{*}p < 0,05$ - $^{**}p < 0,01$ - $^{***}p < 0,001$

resultados estatisticamente significante quando comparado aos demais algoritmos de aprendizado de máquina. Ao analisar os algoritmos SVM e RF, foi possível identificar uma baixa significância estatística. Ao analisar a diferença entre os algoritmos RF e DT é possível observar uma alta significância estatística. Este resultado é importante, visto que consolida o algoritmo RF como uma alternativa promissora a ser explorada em experimentos futuros e uma eventual revisão do MDAP.

De forma a obter uma melhor compreensão do desempenho do MDAP em comparação ao *baseline* foram selecionados dois conjuntos de dados para análise. O primeiro, o conjunto de dados, criado com o auxílio da semente “0”, foi escolhido devido a não significância estatística encontrada após a aplicação do teste não-paramétrico de Wilcoxon. Dentre os demais conjuntos de dados que apresentaram significância estatística. O segundo conjunto de dados, que fez uso da semente “50”, foi selecionado devido aos resultados mais expressivos após a aplicação do MDAP. Desta forma, compreende-se que devem ser melhor investigados os resultados obtidos com os conjuntos de dados “0” e “50”.

Quanto a importância das características

A Figura 16 apresenta uma visão geral das 50 características mais importantes mapeadas durante a aplicação dos métodos *baseline* e MDAP. Para a análise, a importância de cada característica foi calculada com base na mediana da importância obtida em todos os experimentos realizados. Em um primeiro momento, é possível observar a presença de diferentes características mapeadas pelo MDAP (Local, Perguntas, Elogios, Teor adulto predatório, Saudações, Ordens dadas e Emoticons) dentre as dez principais

características, isto é, com maior importância, de acordo com a medida IG.

Ao analisar as dez características mais importantes com a aplicação de cada método, pode-se notar uma diferença evidente. Em geral, os valores de IG alcançados com o MDAP tendem a apresentar valores superiores aos valores alcançados com o *baseline* quando comparados de maneira sequencial. A presença deste comportamento permite algumas interpretações. Em primeiro lugar, o uso de conceitos de alto nível para representar os diferentes termos presentes em conversas virtuais (e normalmente explorados em conversas predatórias) possibilitou a criação de características com uma maior importância e poder discriminativo. Esse cenário fica evidente ao analisarmos as seguintes características: “micp_envio_saudacao”, “micp_local”, “micp_teoradulto_pred”, “micp_elogio” e “micp_ordem_dada”. Uma segunda interpretação remete à identificação de características do atividade predatória no módulo MPCTI. Devido a separação do processo de padronização textual para análise, foram identificadas características relevantes, como a ocorrência de perguntas, representada pela características “micp_pergunta”, assim como o uso de Emoticons (“micp_emoticon”).

Além das sete características previamente apresentadas, também foi possível identificar a presença de outras seis características mapeadas pelo MDAP dentre as cinquenta mais importantes: “micp_idade_menor”,“micp_parte_do_corpo”, “micp_nome”, “micp_micp_foto”, “micp_idade_maior” e “micp_interesse_local”. Vale ressaltar que a característica com a maior importância não foi mapeada pelo MDAP. A característica “vc” já se encontrava presente no conjunto de dados sem que houvesse a necessidade da aplicação dos passos do módulo MPCTF. Além da característica “vc”, outras características também apresentaram alta importância. Nesse contexto, se destacam as características “sim”, “nao”, frequentemente usadas pelas vítimas nas respostas enviadas. A relevância dessas características indica uma das razões para a obtenção de resultados significativos apenas com a aplicação do *baseline*.

Ainda na Figura 16, ao analisar o restante das características em comum, observa-se a presença de diferentes verbos usados ao longo da ocorrência da atividade predatória. Alguns verbos, por exemplo: “gostar”, “querer”, “poder”, “ver”, “estar”, “morar” e “ir” representam uma parte relevante das 50 características mais relevantes. Os verbos se encontram representados principalmente por meio de diferentes formas de flexão, no entanto, também é possível observar o uso de linguagem informal com grande importância(“ta” e “to”). A presença de treze características dentre as cinquenta mais importantes levantam

questionamentos acerca da importância das demais mapeadas pelo MDAP. De forma a buscar uma melhor compreensão, a Figura 17 apresenta as duzentas características mais importantes ao considerar todos os experimentos realizados. Ao estender a análise às 200 características mais relevantes, a característica "micp_teoradulto", prevista no MDAP com o propósito de identificar conversas com teor adulto porém não predatório, apresentou uma importância maior quando comparada a outras características como "micp_interesse_foto", "micp_parentes", "micp_peca_de_roupa" e "micp_comodo_da_casa". Uma das características importantes para se medir o quão avançado se encontra um abuso sexual na internet, a troca de números de telefone (representada pela característica "micp_telefone") apresentou a menor importância dentre as características identificadas pelo MDAP.

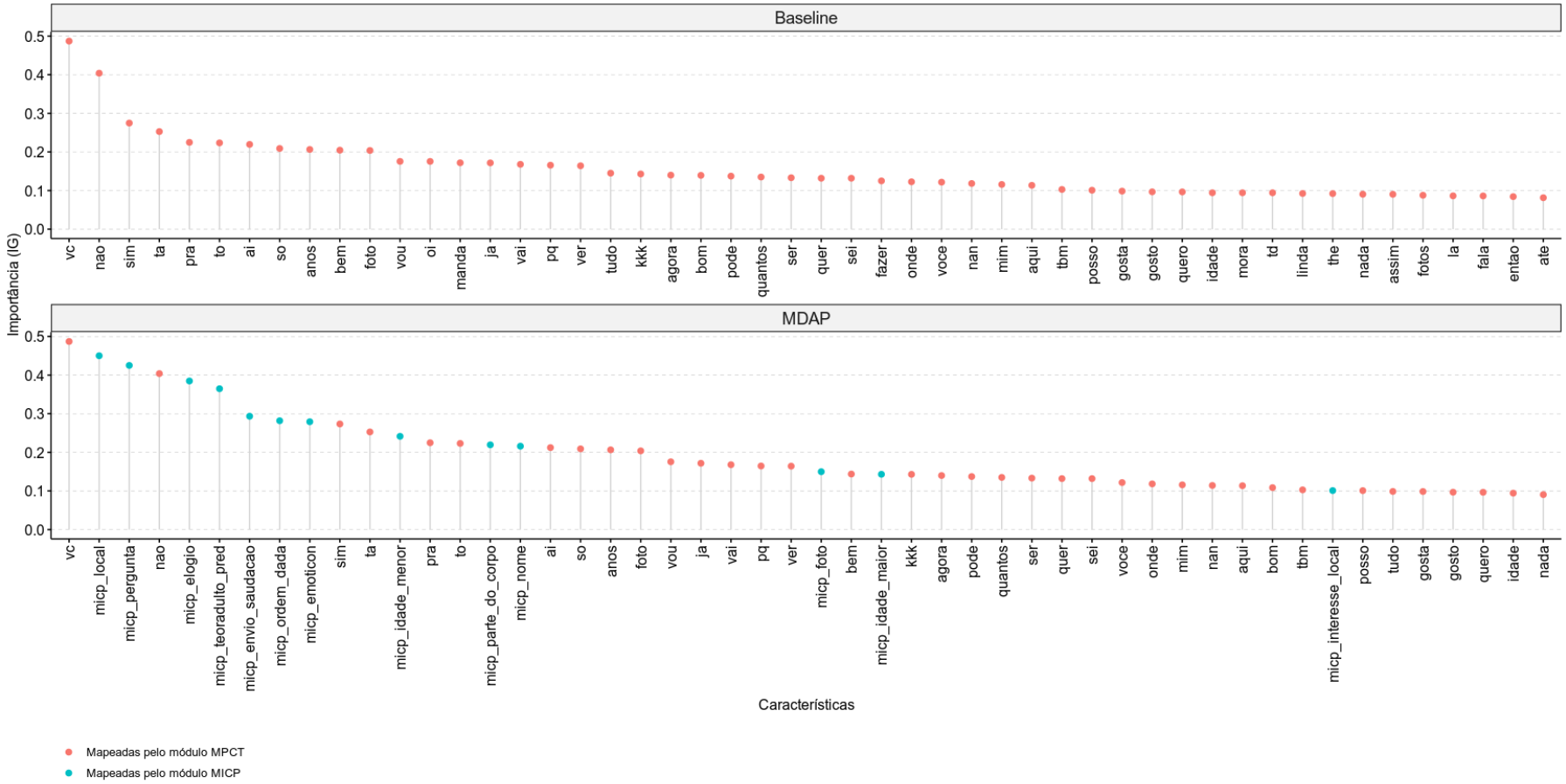


Figura 16 – Média da importância das 50 principais características quanto à aplicação dos métodos MDAP e *baseline*.

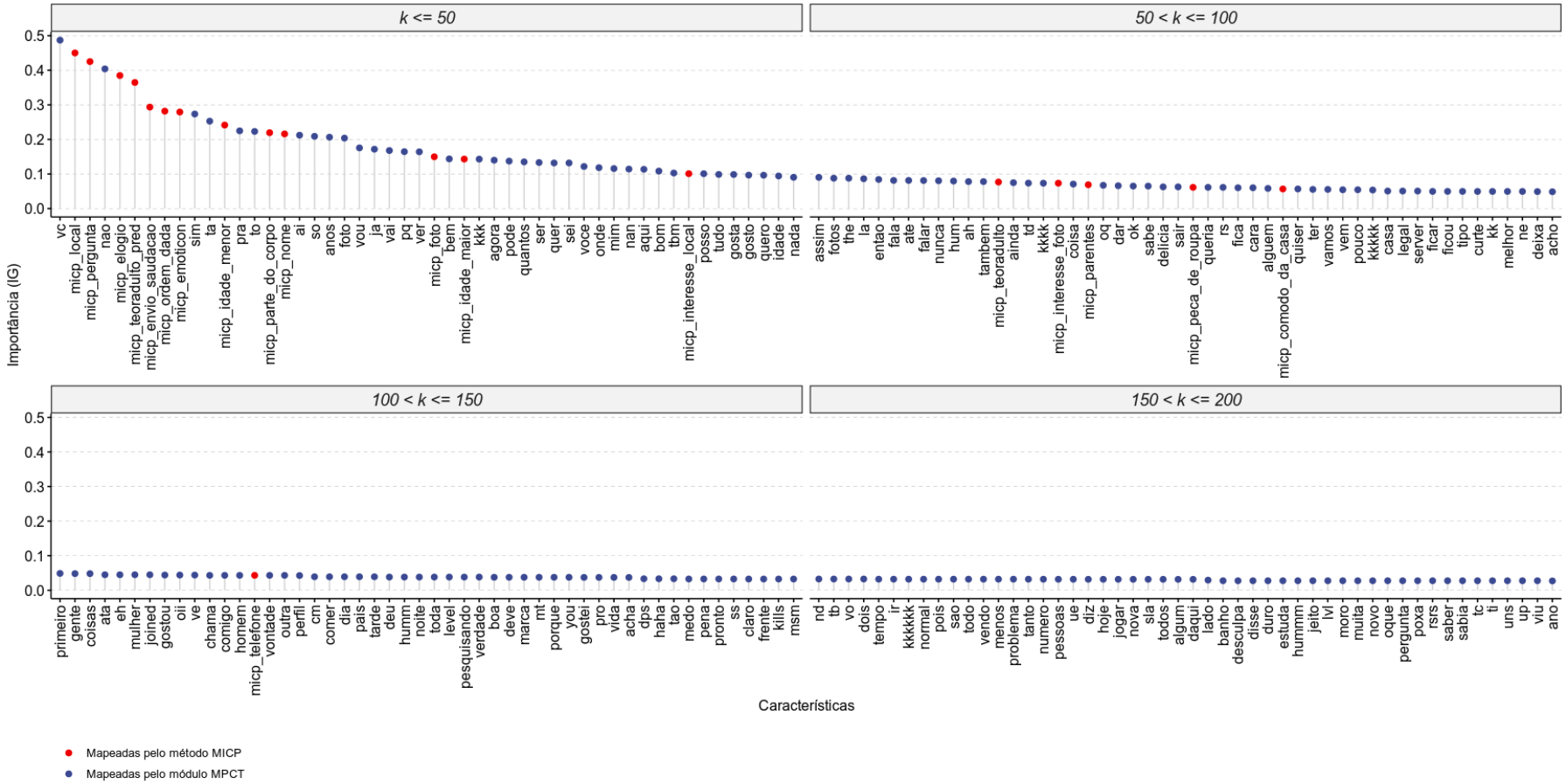


Figura 17 – Duzentas características mais importantes após à aplicação do MDAP e agrupadas em linha com a seleção das Top-*k* características importantes.

Além das características restantes identificadas pelo MDAP, é possível identificar o uso de diferentes características oriundas do internetês, como por exemplo, as risadas (“kkk”,“kk”,“haha”,“rs” etc). Embora não existam registros na literatura que tenham associado a presença de risadas, comportamento manifestado comumente em conversas virtuais, a conversas predatórias, enxerga-se uma possibilidade de explorar a característica em futuros experimentos. Dentre os demais termos oriundos do internetês, é possível observar uma pequena quantidade de características (“mt”, “msm”, “nd”, “tb”, “sla”, “dps”, “ss” e “pq”) com um grau discreta. Diante do cenário, entende-se que o internetês pode contribuir para a identificação da atividade predatória sexual.

5.3.2 Análise dos conjuntos selecionados

Diante da visão geral dos resultados apresentada na Seção anterior, compreende-se que a investigação acerca do comportamento do MDAP nos conjuntos de dados “0” e “50” pode permitir uma melhor compreensão do desempenho do método proposto na presente pesquisa. Dessa forma, como um primeiro passo, foi reproduzida a mesma análise estatística realizada no conjunto PRED-2050-ALL, após a aplicação do métodos MDAP e *baseline*, em ambos os conjuntos, dado que o conjunto de dados “0” não apresentou diferenças significantes entre os resultados obtidos entre os métodos aplicados. Os resultados são apresentados na Tabela 15. Também foi verificada a sobreposição de termos raros e não-raros entre as classes, detalhados na Tabela 16.

Ao analisar a aplicação do MDAP nos conjuntos de dados “0” e “50” é possível observar dois comportamentos: (i) o aumento do número de termos; (ii) a redução do vocabulário; o aumento de termos após a aplicação do MDAP se originou da identificação de características previstas no módulo MICP e a remoção das características inseridas manualmente nas conversas predatórias ao aplicar o *Baseline*. Dentre as características previstas pelo módulo MICP, ocorreu uma maior identificação de menções à locais devido a incorreta identificação de algumas palavras sem importância como por exemplo, a preposição “para” (versão normalizada de “Pará”). Com relação à redução de vocabulário, foram identificadas duas razões. Em primeiro lugar, o MDAP promove a representação de diferentes termos em conceitos de alto nível. Um segundo ponto, não menos importante,

Tabela 15 – Análise estatística dos conjunto de dados “0” e “50” após a aplicação do métodos MDAP e *baseline*.

	Conjunto “0”		Conjunto “50”	
Característica	*Baseline*	MDAP	*Baseline*	MDAP
Termos	18.121	19.373	13.985	15.185
Vocabulário	3.133	2.750	3.483	3.028
Número de mensagens	24.318	24.318	20.523	20.523
Termos por conversa (μ)	110,49	118,34	85,27	92,59
Termos por conversa (σ)	425,45	430,42	205,22	215,41
Termos por mensagem (μ)	5	5	4	4
Termos por mensagem (σ)	11	11	8,06	8,06
Hapax Legomena	1.927	1.739	2.155	1.923
Dis Legomena	465	399	514	437

encontra-se na definição do *baseline* que promove a remoção das características inseridas manualmente.

Ao comparar os dois conjuntos de dados, é possível observar que o conjunto “0” apresenta um volume aproximadamente 20% maior de mensagens quando comparado ao conjunto “50”. Em linha com a diferença no volume de mensagens, também são observadas conversas e mensagens mais extensas. Um ponto importante, é a manutenção do comportamento de não normalidade na distribuição dos termos, de forma que o desvio padrão apresentou um resultado superior ao encontrado na análise do conjunto de dados PRED-2050-ALL. Esse comportamento indica a presença de conversas *outliers* ao analisar a quantidade de termos totais em uma conversa. Por fim, o conjunto “0” apresentou uma quantidade menor de termos raros, em linha com a diferença de volume de vocabulário encontrada (-9%).

A Tabela 16 apresenta uma visão da sobreposição dos termos únicos, isto é, termos comuns em ambas as classes de conversas, após a aplicação do MDAP e o *baseline*. Em ambos os conjuntos de dados, observa-se uma redução aproximada em 9% da sobreposição de termos raros. Por outro lado, a sobreposição dos demais termos isto é, os termos não raros, apresentaram um aumento superior à 10%. Ao observar o conjunto “0”, é registrado um aumento de 16%. A razão para tal comportamento observado se justifica na aplicação do módulo MICP. Nesse cenário, alguns dos termos raros foram identificados como candidatos à representação de conceitos de alto nível presentes em conversas predatórias. Nesse cenário, as menções à idades e perguntas sobre localidade são as principais responsáveis pelo resultado.

Tabela 16 – Sobreposição de termos entre as classes de conversas presentes nos conjunto de dados “0” e “50”. Em cada um dos conjunto de dados é apresentado o número de termos sobrepostos e, ao lado, o percentual correspondente ao total de termos únicos presentes.

	Conjunto “0”		Conjunto “50”	
Sobreposição	*baseline*	MDAP	*baseline*	MDAP
Termos raros	89 (3,33%)	81 (3,41%)	134 (4,41%)	126 (4,69%)
Demais termos (c(t)>2)	44 (5,86%)	51 (8,05%)	65 (7,88%)	72 (10,40%)

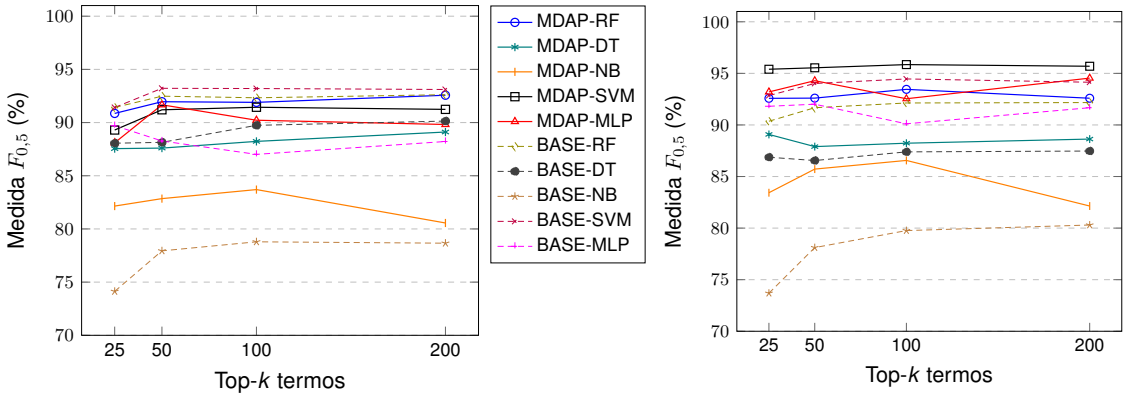
5.3.3 Experimentos realizados

A seguir são apresentados os experimentos realizados com a aplicação dos métodos MDAP e *baseline* nos conjuntos de dados “0” e “50”. A Figura 18 ilustra o desempenho dos experimentos ao considerar todos os termos presentes no vocabulário como o limiar escolhido. A Figura 19 ilustra a diferença entre os resultados obtidos ao considerar como limiar para o vocabulário a remoção de todos os termos raros. As tabelas 17 e 18 apresentam os resultados obtidos com os experimentos. A seguir, são analisados os resultados obtidos em cada um dos limiares.

Análise considerando todos os termos presentes em conversas

Na Figura 18a, um comportamento notável após a aplicação dos métodos MDAP e *baseline* é a melhora do desempenho ao considerar as 50 características mais importantes para a maioria dos experimentos. A única exceção encontra-se com a aplicação do *baseline* em conjunto com o algoritmo MLP. Ao analisar a seleção de um número maior que 50 características mais importantes, não é possível identificar ganhos significativos. A seleção de mais do que 50 características contribuiu de forma expressiva apenas na aplicação do MDAP para os algoritmos de aprendizado de máquina DT e RF.

Ainda com relação ao conjunto “0”, observa-se que os algoritmos de aprendizado máquina SVM e RF apresentaram os resultados mais significativos na maioria dos experimentos. Os melhores resultados foram obtidos com a aplicação do *baseline* e a



(a) Conjunto “0”

(b) Conjunto “50”

Figura 18 – Resultados obtidos com a aplicação do MDAP e o *baseline* ao considerar todos os termos presentes nos conjuntos de dados “0” e “50”.

seleção das 50 características mais importantes. Com esses parâmetros, o algoritmo SVM atingiu 93,22% na medida $F_{0,5}$. O algoritmo RF apresentou resultado inferior (92,48%), porém ao considerar as 200 características (92,62%) foi possível obter um ganho de 0,14%. Ao analisar a aplicação do MDAP, o algoritmo RF apresentou resultado superior ao SVM em todos os experimentos realizados. Por meio do uso das 200 características mais relevantes atingiu-se 92,57% na medida $F_{0,5}$, o que representou um ganho de 0,61% quanto ao uso de 50 características.

A Figura 18b ilustra um cenário em que a aplicação do MDAP obteve resultados superiores ao *baseline* em todos os experimentos realizados. Os ganhos de desempenho ao considerar o uso de 50 características são mais discretos quando comparados ao conjunto “0”, no entanto, identifica-se a presença de resultados satisfatórios ao considerar apenas 25 características para diferentes algoritmos de aprendizado de máquina. Ao analisarmos os resultados com os algoritmo SVM, observa-se um ganho inferior à 0,5% quando comparado os resultados obtidos com as 25 características mais importantes (95,40%) ao demais valores considerados para *k* (50, 100, 200). Esse comportamento que se repete com o algoritmo RF. Esse fenômeno é relevante, visto que indica a importância das características identificadas pelo MDAP e o critério de seleção por meio do uso da medida IG. O algoritmo NBM apresentou um comportamento promissor após a aplicação dos métodos MDAP e *baseline* nos conjuntos de dados “0” e “50”. Em ambos os conjuntos, identificam-se ganhos significativos de desempenho com a aplicação do MDAP para até as 50 características mais importantes.

Conj	*k*	*Baseline* (%)					MDAP (%)				
		SVM	DT	RF	MLP	NBM	SVM	DT	RF	MLP	NBM
0	25	**91,42**	88,06	91,15	89,71	74,13	89,30	87,55	**90,86**	88,10	82,15
	50	**93,22**	88,14	92,48	88,27	77,94	91,20	87,60	**91,96**	91,67	82,85
	100	**93,20**	89,74	92,33	87,01	78,79	91,44	88,23	**91,90**	90,22	83,70
	200	**93,11**	90,14	92,62	88,22	78,66	91,24	89,11	**92,57**	89,82	80,56
50	25	**92,85**	86,87	90,36	91,81	73,69	**95,40**	88,47	92,29	93,21	83,43
	50	**94,02**	86,55	91,67	92,01	78,11	**95,54**	87,88	92,88	94,29	85,73
	100	**94,46**	87,39	92,13	90,11	79,76	**95,85**	87,93	93,13	92,54	86,56
	200	**94,14**	87,47	92,17	91,67	80,30	**95,69**	88,36	92,86	94,94	82,19

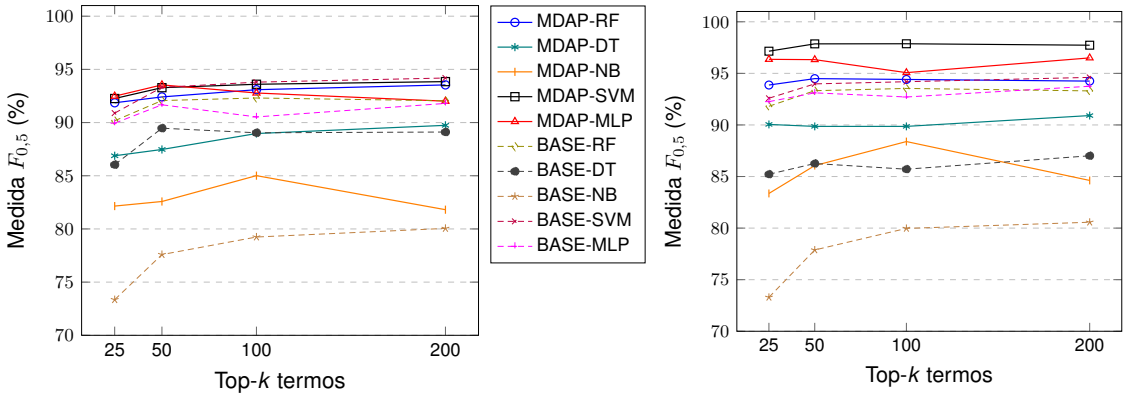
Tabela 17 – Desempenho dos algoritmos de aprendizado de máquina após a aplicação dos métodos MDAP e *baseline* ao considerar como limiar para a seleção de características todos os termos presentes nos conjuntos de dados.

Análise da remoção de termos raros

A Figura 19a ilustra dois comportamentos quanto a aplicação do métodos MDAP e *baseline* no conjunto “0”. O primeiro comportamento, ao considerarmos o uso de até 50 características mais importantes, nota-se que os algoritmos de aprendizado de máquina SVM, RF e MLP apresentaram desempenho superior à 90%. Dentre os três algoritmos de aprendizado de máquina, o algoritmo MLP apresentou o melhor desempenho nos experimentos, atingindo 92,48% na medida $F_{0,5}$ ao considerar as 25 características mais importantes. Quando analisado o desempenho com o uso das 50 características mais importantes, o uso do MLP atingiu 93,54%, seguido do algoritmo SVM (93,28%) e, por fim, o algoritmo RF (92,41%).

Ao analisar os resultados obtidos com a aplicação do *baseline*, o algoritmo SVM apresentou o resultado mais expressivo (93,31%), seguido do algoritmo RF (92,07%). Um segundo comportamento, observado a partir do uso de 100 características mais importantes, é a melhora do desempenho do algoritmo SVM em conjunto com a aplicação do *baseline*. Nesse cenário, o algoritmo SVM atingiu 94,19% quando foram consideradas as 200 características mais importantes. Em contrapartida, a aplicação do MDAP atingiu resultados próximos, porém inferiores. O resultado mais expressivo com o MDAP também foi alcançado com o algoritmo SVM (93,75%), seguido do algoritmo RF (93,54%).

Ao analisar a Figura 19b observa-se que o emprego dos algoritmos de aprendizado de máquina SVM, MLP e RF em conjunto com o MDAP apresentou desempenho superior ao compará-los a mesma configuração, porém aplicada ao *baseline*. Ao considerar apenas o uso das 25 características mais importantes, ambos os algoritmos apresentaram



(a) Conjunto “0” **(b)** Conjunto “50”

Figura 19 – Resultados obtidos com a aplicação do MDAP e o *baseline* quando considerado o vocabulário sem a presença de termos raros oriundos dos conjuntos de dados “0” e “50”.

superior à 95% (97,15% e 96,36%, respectivamente). Ao considerar as 50 características mais importantes, o algoritmo SVM atingiu 97,87% na medida $F_{0,5}$. Esse resultado representa um aumento de desempenho aproximado em 4% perante ao *baseline*. O algoritmo RF atingiu resultados superiores à 95% ao considerar as 100 características mais importantes.

Conj	k	*Baseline* (%)					MDAP (%)				
		SVM	DT	RF	MLP	NBM	SVM	DT	RF	MLP	NBM
0	25	**90,89**	86,04	90,17	89,94	73,35	92,26	86,88	91,86	**92,48**	82,15
	50	**93,31**	89,47	92,07	91,67	77,59	93,28	87,47	92,41	**93,54**	82,58
	100	**93,80**	89,04	92,31	90,53	79,24	**93,60**	88,97	93,09	92,78	85,01
	200	**94,19**	89,11	92,05	91,80	80,05	**93,85**	89,74	93,54	92,01	81,81
50	25	**92,59**	85,22	91,81	92,26	73,29	**97,15**	90,82	93,92	96,36	83,33
	50	**93,98**	86,26	93,32	93,11	77,88	**97,87**	90,16	94,40	96,28	86,04
	100	**94,19**	85,71	93,54	92,71	79,97	**97,87**	89,83	95,04	95,81	88,39
	200	**94,61**	87,01	93,30	93,75	80,57	**97,76**	90,57	94,18	96,13	84,62

Tabela 18 – Desempenho dos algoritmos de aprendizado de máquina após a aplicação dos métodos MDAP e *baseline* ao considerar como limiar para a seleção de características a remoção dos termos raros presentes nos conjuntos de dados.

5.3.4 Discussão dos experimentos

A compreensão do comportamento dos métodos MDAP e *baseline* nos conjuntos de dados “0” e “50” teve como ponto de partida a análise estatística dos conjuntos de

dados. Por meio da análise realizada, observou-se uma sobreposição maior entre os termos presentes nas conversas predatórias e não predatórias no conjunto de dados “50”. A análise realizada no conjunto de dados “0” identificou uma quantidade maior de termos em detrimento de um menor vocabulário. Nesse cenário, também observou-se uma sobreposição menor dos termos. Uma sobreposição menor de termos comuns, isto é, não raros, implica em uma quantidade maior de termos em caso de sobreposição assim como uma concentração de termos repetidos em apenas uma classe de conversa.

Em um segundo momento, o MDAP e o *baseline* foram aplicados em conjunto com diferentes algoritmos de aprendizado de máquina. A avaliação de desempenho considerou o uso de dois limiares distintos para a seleção de vocabulário: (i) o uso de todos os termos presentes nos conjuntos de dados; (ii) a remoção dos termos raros. Ao analisar os resultados encontrados para cada limiar em cada um dos conjuntos de dados selecionados (“0” e “50”), a estratégia para a remoção de termos raros apresentou os melhores resultados. Esse comportamento permite concluir que a presença de termos raros em conversas predatórias atua como um ruído, reduzindo o desempenho dos métodos MDAP e *baseline*.

Ao analisar os resultados obtidos com a remoção de termos raros, foi possível observar que a aplicação do MDAP obteve resultados superiores em comparação ao *baseline* ao considerarmos até as 50 características mais importantes nos conjuntos de dados “0” e “50”. Ao considerar apenas o conjunto de dados “50”, os resultados mais expressivos foi atingido com o algoritmo SVM no MDAP e *baseline*. Dado que o algoritmo SVM apresentou os melhores resultados com ambos os métodos, cabe uma análise com relação ao desempenho desse algoritmo com relação às medidas: acurácia, precisão e abrangência. A tabela 19 apresenta os resultados obtidos. Ao analisar os resultados, é possível observar que o principal fator para a diferença dos resultados entre os conjuntos de dados foi a capacidade de identificar as conversas predatórias corretamente. Os experimentos realizados com o conjunto de dados “50” apresentaram uma precisão superior à 5% quando comparado ao conjunto de dados “0”. Ao analisar a abrangência dos experimentos realizados, ambos os conjuntos de dados apresentaram resultados similares. Esse comportamento é melhor observado nos experimentos em que foram considerados 25 ou 50 características mais relevantes. Desta forma, compreende-se que a diferença na incidência de FP foi o principal fator para a diferença de resultados entre os conjuntos de dados “0” e “50”.

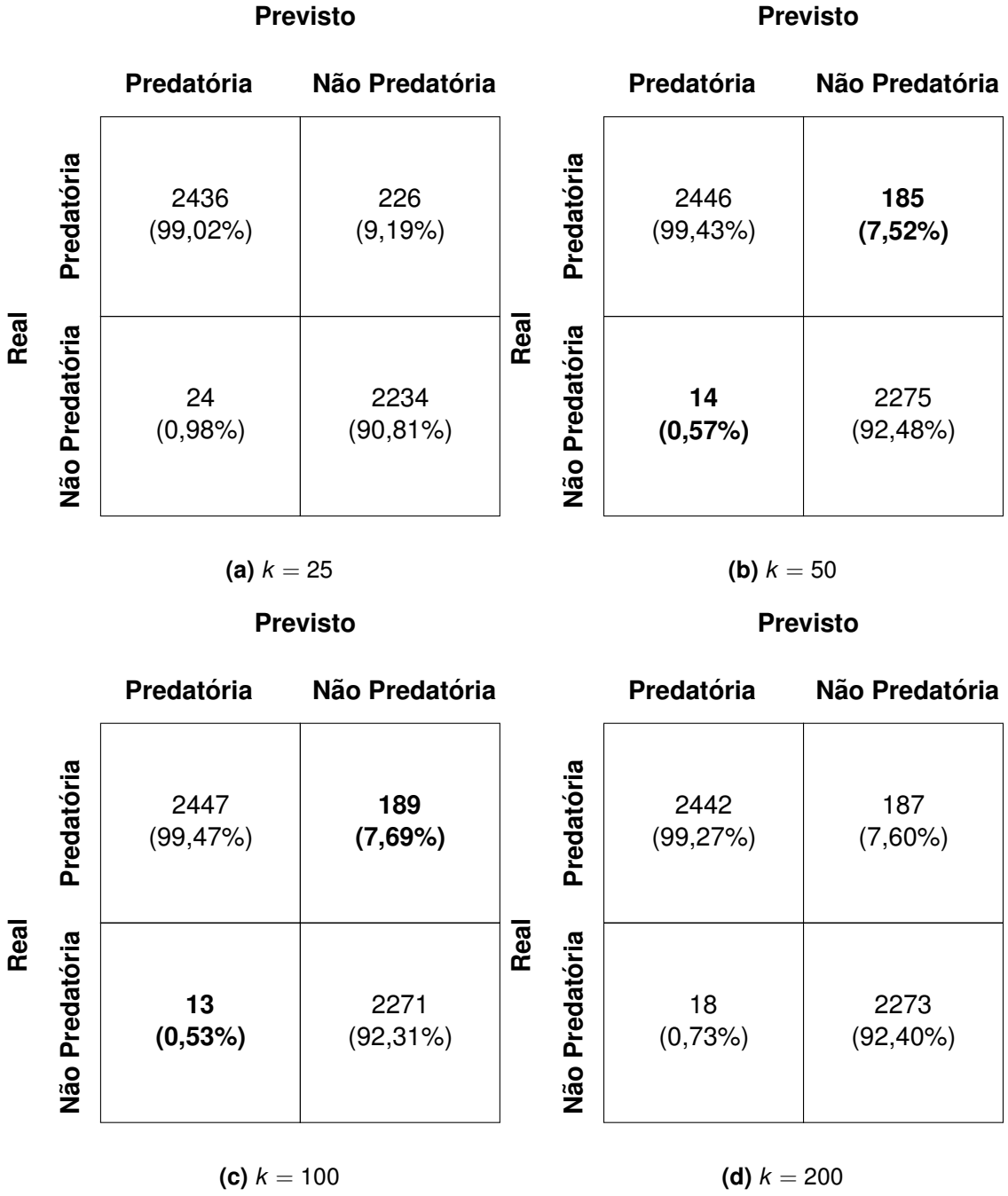
Conj	*k*	Termos raros removidos (%)				Todos os termos (%)			
		Acurácia	Precisão	Abrangência	$F_{0,5}$	Acurácia	Precisão	Abrangência	$F_{0,5}$
0	25	91,74	**92,78**	90,81	**92,26**	89,91	**88,79**	91,87	**89,26**
	50	92,94	**93,63**	92,48	**93,28**	91,82	**90,81**	93,54	**91,22**
	100	93,47	**93,77**	93,42	**93,60**	92,19	**90,92**	94,23	**91,44**
	200	93,77	**93,94**	93,87	**93,85**	92,12	**90,54**	94,59	**91,21**
50	25	94,92	**99,00**	90,82	**97,15**	94,17	**96,44**	91,92	**95,40**
	50	95,96	**99,42**	92,48	**97,87**	94,92	**96,07**	93,90	**95,54**
	100	95,89	**99,46**	92,31	**97,87**	94,96	**96,59**	93,41	**95,85**
	200	95,83	**99,27**	92,39	**97,73**	94,90	**96,34**	93,58	**95,69**

Tabela 19 – Desempenho do algoritmo SVM após a aplicação dos métodos MDAP nos conjuntos de dados “0” e “50”.

A relação entre a quantidade de características *k* e a medida de precisão destaca os resultados obtidos no conjunto de dados “50”. O uso de apenas 25 características, em que 10 são mapeadas pelo MDAP, permitiu atingir um resultado médio de 99% na medida precisão. Ao considerar uma quantidade maior de características, é possível identificar um ganho aproximado de 0,5% na medida precisão. Com base na precisão obtida nos experimentos, o MDAP apresenta os melhores resultados quando *k* = 100.

Em contrapartida, o conjunto de dados “0” apresentou um comportamento distinto do conjunto “50” quando analisada a evolução da medida precisão em relação ao número de características. Ao considerar a remoção dos termos raros como o limiar para a definição de vocabulário, as 50 características mais importantes proporcionaram um ganho aproximado de 0,9% na medida precisão. O ganho é superior à 2% quando todos os termos são considerados como o limiar escolhido. Embora as características mapeadas pelo MDAP se apresentem entre as mais importantes, a importância presente nessas características não foi capaz gerou ganhos discretos quando é analisado o uso das características mapeadas pelo MDAP em conjunto com as demais características presentes nas conversas.

Diante da diferença de precisão obtida pela aplicação do MDAP nos conjuntos de dados “0” e “50”, foi realizada uma investigação de forma a melhor compreender quais as conversas contribuíram para atingir os resultados encontrados. Desta forma, em um primeiro momento foram coletadas as matrizes de confusão dos experimentos considerados com o algoritmo SVM após a realização do processo de validação cruzada com cinco grupos repetida trinta vezes. O cálculo das matrizes de confusão considerou o método proposto em Bestagini et al. [2017]. A seguir, a Figura 20 exibe a matriz de confusão para o conjunto de dados “50”e a Figura 21 ilustra a matriz de confusão para o conjunto de dados “0”.

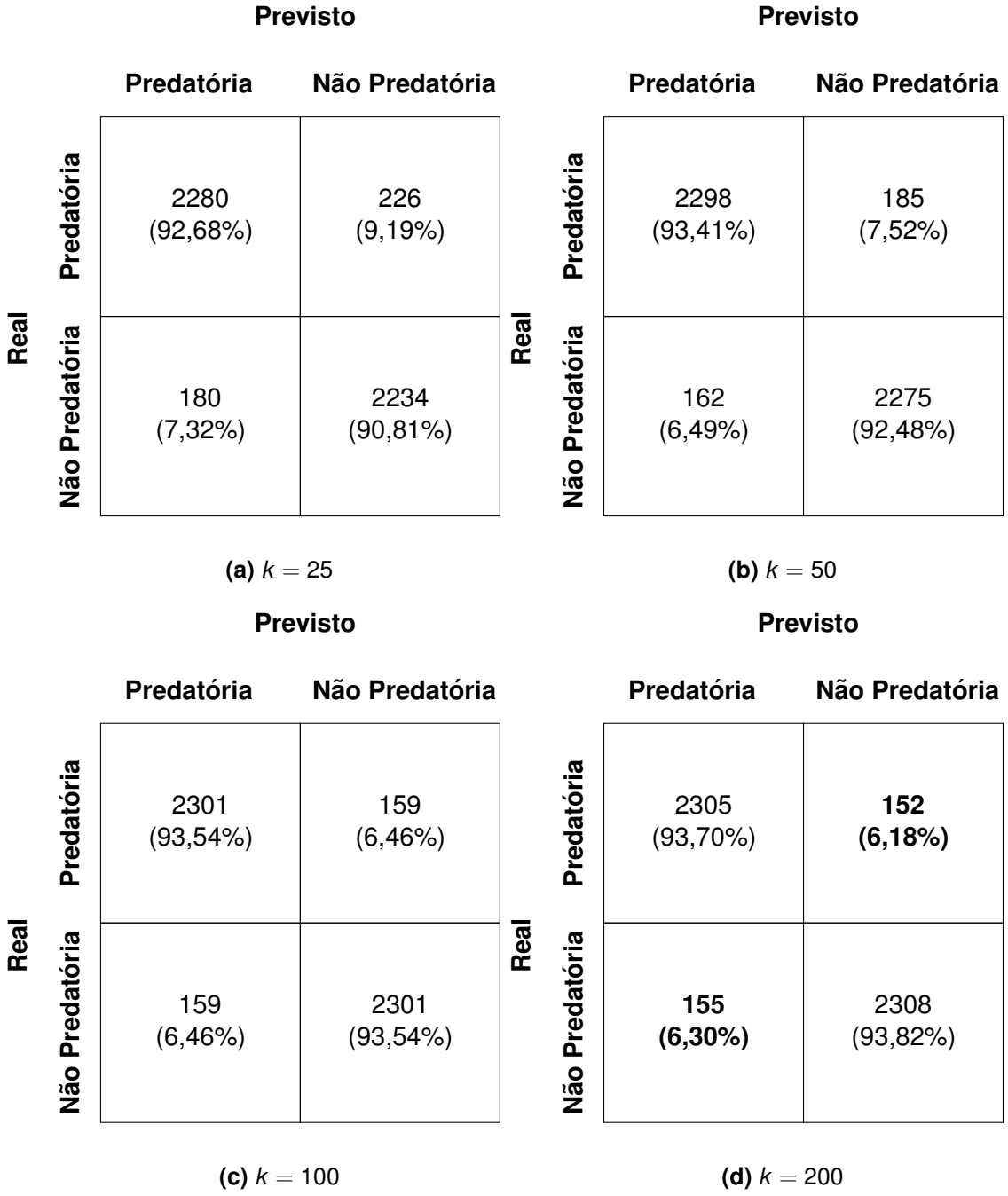


(a) $k = 25$ **(b)** $k = 50$

(c) $k = 100$ **(d)** $k = 200$

Figura 20 – Matriz de confusão dos experimentos realizados com a aplicação do MDAP em conjunto com o algoritmo SVM no conjunto de dados “50”. Os resultados encontram-se agrupados pela quantidade de características *k* mais relevantes consideradas.

A Figura 20 apresenta uma pequena diferença na quantidade de FPs quando é realizada uma comparação entre as diferentes top-*k* características consideradas para o conjunto de dados “50”. Observa-se que menos de 1% das conversas foram classificadas incorretamente como conversas predatórias em todos os cenários de avaliação. Dentre



(a) $k = 25$ **(b)** $k = 50$

(c) $k = 100$ **(d)** $k = 200$

Figura 21 – Matriz de confusão dos experimentos realizados com a aplicação do MDAP em conjunto com o algoritmo SVM no conjunto de dados "0". Os resultados encontram-se agrupados pela quantidade de características *k* mais relevantes consideradas.

as conversas predatórias classificadas incorretamente (FN) nota-se um resultado inferior à 10%. Ao analisar as Figuras 20b e 20c observa-se uma quantidade similar de FPs e FNs encontrados.

Na Figura 21 ilustra um comportamento similar ao exibido na Figura 20. Nota-se

que em ambos os conjuntos de dados atingiram um mesma quantidade de FNs para *k* = 25 (Figura 21a) e *k* = 50 (Figura 21b). Essa descoberta é importante, visto que pode demonstrar uma dificuldade comum do MDAP na identificação de determinadas conversas predatórias em ambos os conjuntos de dados. Em linha com o apresentado na Tabela 19, é identificada a ocorrência aproximadamente dez vezes maior de FPs quando comparado ao conjunto “50”. Um outro comportamento observado é a redução de FNs a medida em que são consideradas mais características. Uma interpretação para esse fenômeno seria a dificuldade de distinguir determinadas conversas predatórias apenas com o uso das características oriundas do MDAP, sem o auxílio das demais características com menor importância.

Diante dos resultados constatados nas Figuras 20 e 21, iniciou-se um levantamento das conversas predatórias e não predatórias classificadas incorretamente. Desta forma, busca-se uma melhor compreensão das conversas que ofereceram maior dificuldade durante os experimentos. Em conjunto com o levantamento das conversas, foi calculada a taxa de erro média (TEM). O cálculo da medida TEM pode ser compreendido como a razão entre a quantidade de vezes em que uma conversa foi classificada incorretamente e o total de vezes em que a conversa foi submetida para classificação (quando presente no conjunto de testes) ao longo do processo de validação cruzada. Sendo assim, as Tabelas 20 e 21 apresentam as conversas classificadas como FP enquanto as as Tabelas 24 e 25 exibem as conversas classificadas como FN durante a realização dos experimentos com o MDAP com o algoritmo SVM. Para cada uma das conversas classificadas incorretamente são destacadas as características mapeadas pelo módulos MPCTF (vermelho) e MICP (azul).

A Tabela 20 apresenta as três conversas que foram responsáveis pelo total de ocorrências de FP ao longo da realização dos experimentos com o conjunto de dados “50”. Ao analisarmos o primeiro exemplar de conversa, “desisto lol cd vc tioooo”, observa-se que apenas o termo “vc”, mesmo ao considerar as 200 características mais importantes, foi identificado como uma característica importante a ser considerada para a identificação de atividade predatória. Dessa forma, a presença do termo “vc” e a associação à ocorrência de atividade predatória se torna perceptível, o que justifica a classificação incorreta. Em conjunto a isso, também entende-se que a pouca quantidade de termos presentes em uma conversa a ser analisada pode implicar em informações insuficientes para a análise quanto a ocorrência de atividade predatória, o que torna a conversa em questão um

TEM (%)	Conversa
36,67	“desisto lol cd vc tioooo”
16,67	“mamae micp_interesse_local la ae talvez consiga queria micp_interesse_local r micp_local n sei bahiano acho q micp_interesse_local micp_local ah br nah ate hoje nao sei micp_local jambao fake gringo nao lol pq nao pode micp_teoradulto_pred contato q fica birra pq acho q micp_local acha presidente micp_local micp_local tambem nao achasse ultima bolacha pacote micp_local nao macaco convidaria pra ca pena liberal sojado fudido jambao micp_elogio jaja pego fofo micp_ordem_dada pra mim desenha bem negocios ai uniformes exercito ve pagina jambao cancro micp_nome nan”
4,17	“micp_nome voce nao presta decada 90 bem ruim agora nao seipor disse regime ha catedral ativa pyongyang micp_local nao engano tipo apoiar algum regime afeganistao q considera cristaos micp_elogio coreia micp_local proibido ter religiao nao catolico lanao q muitos catolicos micp_local micp_nome micp_local micp_pergunta desrespeito religiao micp_pergunta sinceramente deixa meio micp_teoradulto micp_nome ”

Tabela 20 – Conversas classificadas incorretamente como pertencentes à classe “Predatória” (FP) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “50”.

exemplar ruidoso.

Os dois exemplares seguintes apresentaram uma característica diferente da primeira conversa analisada. Cada um apresentou uma quantidade significativa de características identificadas pelo módulo MICP dentre todos os termos que compõem a conversa. Nesse cenário, observa-se uma maior ocorrência da característica relacionada à menção de locais (“micp_local”). Outras características fortemente associadas à ocorrência de atividade predatória (“micp_teoradulto_pred” e “micp_ordem_dada”) também são encontradas. Em ambos os cenários, é compreendido que a ocorrência de conversas não predatórias em conjunto com o alto de teor de características identificadas pelo módulo MICP contribuíram de forma significativa para a classificação incorreta dos exemplares.

A Tabela 21 indica um total de treze conversas não predatórias distintas foram classificadas incorretamente durante a realização dos experimentos com o conjunto de dados “0”. Dentre essas conversas, observa-se que os nove exemplares iniciais (responsáveis por ao menos 10% na medida TEM) apresentaram duas características interessantes: (i) As conversas classificadas incorretamente como predatórias pelo MDAP, em sua maioria, possuem um número reduzido de termos para a análise; (ii) A alta ocorrência de características identificadas pelo módulo MICP. Dentre elas, uma maior ocorrência de menções à locais, a realização de perguntas, e a identificação de termos associados à teor adulto usado por predadores sexuais.

Esse fenômeno apresenta um cenário consistente com a interpretação realizada a partir das conversas não predatórias incorretamente classificadas no conjunto de dados

TEM (%)	Conversa
100,00	“micp_elogio sei questao vcs queren micp_local micp_pergunta micp_local sim conheco micp_local tanto micp_teoradulto_pred recomendo bom”
95,84	“sim nao desbanir pato”
92,50	“ ta micp_parte_do_corpo vou micp_teoradulto_pred micp_local texto micp_local vcs micp_ordem_dada jogos crimonosos”
73,34	“ micp_emoticon micp_teoradulto_pred ja postei micp_pergunta vjesc3yvzsws”
67,50	“firma micp_teoradulto_pred la vai micp_teoradulto_pred micp_envio_saudacao senhores”
40,00	“micp_comodo_da_casa aparentemente bem organizada micp_elogio nan”
28,33	“micp_ordem_dada convite vc n ta banido micp_nome micp_emoticon desban micp_ordem_dada convite ainda to micp_nome”
11,66	“alguem ps4 micp_pergunta micp_local ja lancou crossplay pro apexmicp_pergunta alguem ai sabe”
10,00	“micp_local nao vc sistema puro assim dizer micp_pergunta wm gt so interface grafica gt praticamente suite pronta micp_local usuario sistema praticamente inteiro”
6,67	“libertario nao gosto palpitar escolhas alheias tendo achar okay voce nao precisa definir isso relativo muitas coisas pessoas diferentes pra relevar micp_nome nao sei to justamente procurando saber sobre micp_interesse_local disso micp_nome micp_teoradulto micp_teoradulto_pred visao mundo segundo afirmacao mulher nao pessoa objetivos individualidade micp_local sim meramente quotbrinquedosquot servico homem”
5,00	“tenta navegador pra ver s aplicativ micp_pergunta abri pc pra ver ta mesma coisa so servee windowns ai mudo pro aplicativo funfa normal kkkkkkkk costuma dar problema usando navegador problema rota aplicativo celular micp_nome vc aplicativo navegadorusando discord micp_pergunta outro servers participo conectando normalmente ta acontecendo alguem nao to conseguindo conectar salas servidor”
0,83	“vem vem pra caixa voce tambem micp_emoticon ”
0,83	“kkkkkkkkkkkkkkkkkkkk so entra micp_local virus”

Tabela 21 – Conversas classificadas incorretamente como pertencentes à classe “Predatória” (FP) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “0”.

“50”. Dessa forma, conclui-se que a principal justificativa para a diferença de desempenho, especialmente no que se refere à ocorrência de FP, entre os conjuntos de dados “0” e “50” é a maior ocorrência de conversas não predatórias com poucos termos e que apresentem características importantes e relacionadas a ocorrência de atividade atividade predatória, previstas no módulo MICP.

Em linha com a análise realizada sobre a maior ocorrência de FPs no conjunto de dados “0” e “50”, também foram investigadas as principais conversas predatórias classificadas incorretamente (FN). Dessa forma, pretende-se obter um melhor entendimento sobre quais as conversas predatórias o MDAP apresentou uma maior dificuldade de identificação da atividade predatória. No primeiro momento, ao comparar o levantamento das FNs realizado em ambos os conjuntos de dados, observa-se um conjunto restrito de conversas predatórias as quais o MDAP apresentou alta dificuldade na identificação de atividade predatória. Ainda nesse conjunto restrito de conversas, é possível encontrar

conversas predatórias em que o MDAP não apresentou a capacidade de classificar corretamente em todos os experimentos executados (100% na medida TEM), por exemplo: “aqui quero ve micp_local vale pena hum sabia micp_parte_do_corpo deliciosa”. Dentre essas conversas, um total de sete exemplares distintos de conversas predatórias apresentaram uma elevada taxa de erro médio superior à 50%. Sendo assim, compreende-se que esses exemplares contribuíram para o comportamento observado nas Figuras 20b e 20c, limitando o desempenho do MDAP de forma significativa. A Tabela 22 apresenta as conversas predatórias selecionadas para a análise. As demais conversas predatórias são exibidas nas Tabelas 24 e 25 e estão disponibilizadas no Apêndice A.

#	Conversa
1	“ai bom coisa fofa ainda micp_teoradulto_pred agora vim entender porque voce timido pois vou quebrar tabu teu”
2	“aqui quero ve micp_local vale pena hum sabia micp_parte_do_corpo deliciosa”
3	“fico cheio vontades aqui vontade d nao sei micp_local falo ummm vontade muita fala ta vontade tocar todo micp_parte_do_corpo micp_teoradulto_pred micp_parte_do_corpo la naquele lugar”
4	“vai ser primeira vez pouco medo pois nunca fiz tbm primeira vez micp_pergunta micp_pergunta sim fica boa sera micp_local resto vida ta bem”
5	“voce deita migo cama micp_teoradulto_pred nao posso durmi ai nao pra dumir so fica migo ate 3 horas tardi ai levo voce micp_local tah”
6	“micp_envio_saudacao micp_elogio voce micp_parentes bem micp_elogio pra brincar homem micp_idade_maior anos micp_pergunta pedofilo bem bonzinho pra brincar ela voce sabe ideia micp_parte_do_corpo ne cara micp_pergunta tipo exatamente tipo crime pf permissao micp_local quebrar protocolos privacidade localizar usuarios sei apenas zuando afinal pedofilo verdade nao tao burro”
7	“vc micp_interesse_foto foto micp_teoradulto_pred tambem nao sei q horas vergonha entao micp_peca_de_roupa nao sei nunca fiz”

Tabela 22 – Conversas predatórias classificadas incorretamente que apresentaram a medida TEM superior à 50% após a aplicação MDAP nos conjuntos de dados “0”ou “50”.

O primeiro exemplar da Tabela 22 apresenta características comuns em uma conversa predatória, como o uso de termos adultos predatórios. Nesse caso em particular, o termo adulto predatório está associado ao questionamento sobre a virgindade da vítima. Sendo a única característica identificada pelo módulo MICP, observa-se uma prevalência de características mapeada pelo módulo MPCTF. Outro ponto importante, o vocabulário empregado pelo predador sexual (i.e. timido, quebrar tabu) e a forma como fez o uso de elogios, evitando o uso de termos mais comuns em atividades predatórias em conjunto com o emprego de elogios mais simples para uma maior compreensão de menores de idade (e.g., fofa). Essas características contribuíram para o não reconhecimento de características tanto no módulo MICP e o posterior descarte dos termos em virtude da adoção da remoção de termos raros como o limiar para a definição de vocabulário.

O exemplar 2 remete a um comportamento observado durante a análise das conversas não predatórias classificadas incorretamente. A conversa possui poucos termos, embora tenha duas características mapeadas pelo módulo MICP, apresentou um potencial discriminativo insuficiente para que fosse classificada corretamente. Em conjunto a isso, observa-se a não identificação do elogio com intenções de aliciamento "deliciosa". A correta identificação do elogio poderia contribuir para a correta classificação da conversa dada a importância identificada pela característica "micp_elogio", presente entre as dez características mais importantes.

No terceiro exemplar, o uso do termo "vontade" é explorado em três momentos ao longo da conversa predatória, assim como também é observada a identificação de diferentes características mapeada pelo módulo MICP. Também observa-se que o predador sexual evita se referir diretamente às partes íntimas da vítima, as referindo de forma subjetiva ("naquele lugar"). Esse comportamento, que remete à exploração de termos que atenuam a agressividade presente na comunicação predatória, também pode ser observado no primeiro exemplar (i.e. exemplo). As informações presentes nessa comunicação tendem a ser perdidas, dada a ausência dos dois termos dentre as 200 características mais importantes. Nesse cenário, uma possível interpretação para a dificuldade de classificação seria a baixa importância do termo "vontade" em conjunto com as demais características mapeadas pelo módulo MPCT, o que reduziu a capacidade discriminativa das demais características mapeadas pelo módulo MICP.

Em seguida, no exemplar 4, observa-se a exploração de uma comunicação mais subjetiva em uma conversa entre o predador sexual e a vítima, em um estágio avançado de aliciamento. Ao longo da conversa é possível identificar que o tema da conversa se refere a experiência sexual de ambos os participantes. Durante a conversa, o tema "virgem" ou "virgindade" não é mencionado diretamente. Em seu lugar, são usadas outras formas de expressar, por exemplo: "primeira vez", "nunca fiz" e "tbm primeira vez". Nesse cenário, a identificação de características mapeadas pelo módulo MICP não mostrou capacidade de discriminar corretamente a conversa predatória.

No exemplar 5, a conversa reflete um estágio avançado de aliciamento em que o predador sexual busca o contato físico com a vítima. Nessa conversa, o predador sexual faz uso de um vocabulário com variados erros ortográficos ("migo", "dumir", "tardi"), de forma a assemelhar a sua comunicação com a vítima e com isso adquirir uma maior intimidade [Olson et al., 2007]. Em virtude desse comportamento, nota-se a perda de

informações que poderiam contribuir para a identificação ao considerarmos a remoção de termos raros como o limiar para a remoção de vocabulário. A conversa predatória em questão retrata comportamento observado em exemplares anteriores, em que as características mapeadas pelo módulo MICP não apresentaram capacidade de discriminação suficiente.

No sexto exemplar, observa-se uma conversa entre um adulto, possivelmente um responsável legal da vítima ou um agente da lei e um criminoso. Nessa conversa, embora existam diferentes características mapeadas pelo módulo MICP, não foi possível classificá-la corretamente na maioria dos testes. Uma característica similar a observada em outras conversas analisadas, é a presença de vocábulos incomuns quando comparada às demais conversas predatórias. Esses vocábulos não são oriundos da atividade de aliciamento e sim relacionados à atividade de tráfico de menores de idade para fins de abuso sexual e, uma vez exposto a intenção criminosa, a posterior advertência do adulto sobre a prática. Compreende-se que embora seja possível extrair características importantes sobre o comportamento do predador sexual, o uso da conversa em questão promove um ruído durante a busca de atividade predatória visto que o assunto principal da discussão não ocorreria diretamente entre um predador sexual e um menor de idade.

Por fim, o exemplar 7 apresenta uma conversa em que ocorre um pedido de foto para um menor de idade. Nessa conversa, que apresenta um comportamento similar aos exemplares 3 e 4, retrata uma característica comportamental normalmente presente durante a ocorrência da atividade predatória e não explorada na presente pesquisa. Além das características identificadas pelo módulo MICP, é possível identificar uma relutância da vítima em compartilhar fotos com o predador sexual. Esse comportamento é previsto em modelos de aliciamento [Olson et al., 2007]. Compreende-se que a identificação desse comportamento, em conjunto com as demais características normalmente presentes em conversas predatórias contribuirá para a identificação correta.

Após a análise das principais conversas predatórias selecionadas, conclui-se que três principais fatores foram responsáveis para a classificação incorreta: (i) O predador sexual tende a adaptar a sua comunicação de forma a melhor dissuadir a vítima e, com isso, atingir os objetivos inicialmente planejados. A adaptação da comunicação pode resultar na presença de termos mais incomuns e com isso a perda de informação relevante para contribuir na identificação correta das conversas; (ii) A presença de conversas predatórias com poucos termos cuja a identificação das características mapeadas pelo

módulo MICP não apresentou a capacidade discriminativa suficiente em conjunto com outras características mapeadas pelo MPCTF; (iii) A ocorrência de conversas predatórias sem a presença de menores de idade dentre os participantes.

6- Conclusão

Os predadores sexuais são uma ameaça grave à crianças e adolescentes na internet brasileira. Nesse cenário, o presente trabalho buscou responder a seguinte pergunta de pesquisa: é possível melhorar o desempenho de algoritmos de aprendizado de máquina utilizando características textuais e comportamentais, representadas por meio de conceitos de alto nível, para a identificação de atividade predatória em conversas virtuais na língua portuguesa do Brasil? A partir dessa pergunta de pesquisa foi realizado um estudo acerca da ocorrência de diferentes características textuais e comportamentais presentes na comunicação de predadores sexuais, identificadas na literatura internacional e no âmbito do domínio da pesquisa.

A atividade de coleta de dados para a pesquisa teve início com a disponibilização de 39 conversas predatórias disponibilizadas no trabalho de Andrijauskas et al. [2017]. Em seguida, após um acordo com o MPF-SP e o FEI, foram inseridas mais 43 conversas predatórias. Dessa maneira, o total de conversas predatórias alcançado foi de 82 conversas. Em seguida, iniciou-se a criação de uma base de conversas não predatórias com foco na comunicação de adolescentes e jovens adultos. Nesse cenário, a plataforma Discord se apresentou como uma relevante fonte de conversas não predatórias para o estudo da comunicação de adolescentes e jovens adultos na internet brasileira. Por meio das conversas da plataforma, foi possível encontrar uma grande variedade de assuntos conversados na internet, o que promove um enriquecimento do conjunto de dados, dado que esse trabalho considera o conceito de representatividade.

Ao final da análise das conversas predatórias disponibilizadas, foram consideradas 19 características (comuns em conversas predatórias) para o desenvolvimento do MDAP. O método é apresentado no Capítulo 4. Para o mapeamento das características, foram consideradas três estratégias distintas: (i) o reconhecimento de padrões textuais predefinidos; (ii) o uso de léxicos; (iii) o uso de regras definidas para a identificação de comportamentos específicos.

Em seguida, o Capítulo 5 apresentou os resultados da aplicação do MDAP em comparação a um *baseline* predefinido. Ao todo, a realização de um experimento considerou nove subconjuntos de dados criados a partir do PRED-2050-ALL, cinco algoritmos

de aprendizado de máquina, dois limiares distintos para a definição de vocabulário a ser explorada nas conversas dos subconjuntos e quatro valores possíveis para a seleção das Top-*k* características importantes.

Após o término dos experimentos, os resultados foram analisados sob diferentes perspectivas. A aplicação do MDAP apresentou resultados estatisticamente significativos em oito dos subconjuntos de dados aplicados. É importante notar que, a remoção de termos raros das conversas possibilitou atingir os resultados mais expressivos. Dentre os cinco algoritmos de aprendizado de máquina, o uso do MDAP em conjunto com o algoritmo SVM apresentou os melhores resultados. Observando os resultados mais expressivos, o MDAP foi capaz de atingir 97,87% na medida $F_{0,5}$ ao considerar as 100 características mais importantes. Os algoritmos RF e MLP também apresentaram resultados significativos. Por fim, os algoritmos DT e NBM apresentaram resultados promissores, respectivamente 93,31% e 89,30% na medida $F_{0,5}$.

Dessa forma, conclui-se que, com base nos resultados obtidos ao longo da avaliação experimental, o MDAP melhorou o desempenho do algoritmos de aprendizado de máquina em comparação ao *baseline* apresentado para a identificação de atividade predatória em conversas virtuais na língua portuguesa do Brasil. Desta forma, o MDAP se apresenta como uma alternativa válida e eficiente para a aplicação no domínio da pesquisa.

6.1- Resumo das contribuições

Ao longo da presente pesquisa, foram apresentadas as diferentes contribuições anunciadas na Introdução. A seguir, elas são apresentadas:

- Criação de um método para a identificação de atividade predatória sexual em conversas virtuais na língua portuguesa do Brasil: o Capítulo 4 apresenta o MDAP. O método explora a identificação de dezenove características textuais e comportamentais normalmente presentes em conversas predatórias. Os resultados observados no Capítulo 5 apresentam o MDAP como um método para a identificação do comportamento predatório em conversas na língua portuguesa do Brasil.

- Criação de um conjunto de dados de conversas entre predadores sexuais brasileiros e vítimas: no Capítulo 3 é apresentado o conjunto de dados PRED-2050-ALL. O conjunto de dados apresenta um total de 82 conversas predatórias (disponibilizadas pelo FEI e MPF-SP) e 1.968 conversas não predatórias (oriundas de salas de bate-papo presentes comunidades criadas no aplicativo Discord). É importante ressaltar que as conversas extraídas da plataforma são oriundas apenas de conversas textuais. Com relação às conversas, foram consideradas diferentes categorias, de forma a buscar uma maior representatividade dos assuntos conversados em comunidades virtuais por menores de idade.
- Criação de léxicos para apoiar a aplicação do MDAP: para atingir tal objetivo, duas fontes de dados foram consideradas. Primeiramente, as fontes externas, isto é, sem considerar o vocabulário empregado em conversas predatórias. A segunda fonte, de origem interna, levou em consideração o conhecimento do domínio da pesquisa, presente no vocabulário empregado pelo predador sexual como a fonte primária para a sua composição. Ao todo, foram criados oito léxicos a partir de fontes externas e cinco léxicos de origem interna. Os léxicos se encontram na Seção 5.1.

6.2- Limitações

A seguir, são apresentadas diferentes limitações identificadas ao longo do desenvolvimento da presente pesquisa:

- Ausência de conversas privadas não predatórias: ao longo da presente pesquisa não foi possível obter conversas não predatórias ocorridas em fórum privado. A decisão por não coletar e considerar as conversas privadas para a criação do PRED-2050-ALL, levou em consideração a necessidade de disponibilizar meios válidos (de acordo com a Lei Geral de Proteção de Dados – LGPD), para o registro do consentimento dos participantes. O desenvolvimento desse processo não se mostrou viável durante a execução da pesquisa.
- Conversas não predatórias com poucos termos e com características que indicam

atividade predatória: o MDAP encontrou dificuldades no reconhecimento de conversas não predatórias que apresentaram as seguintes características: (i) curtas (i.e., com poucos termos); (ii) com características mapeadas pelo módulo MICP.

- Conversas predatórias com linguagem subjetiva e vocabulário incomum: foi observado que, em alguns casos, o predador sexual faz uso de uma linguagem menos agressiva e direta para atingir o objetivo desejado. O MDAP apresentou dificuldade na identificação de conversas com essas características.

6.3- Trabalhos futuros

Ao longo da presente pesquisa, diferentes técnicas foram identificadas como possibilidades para evolução do MDAP:

- Aprimorar a identificação de locais em conversas: nas conversas disponibilizadas pelo MPF-SP foram encontradas marcações para anonimizar as menções à locais. Na presente pesquisa, de forma a não ignorar a presença dessa característica, foram consideradas as informações disponibilizadas pelo IBGE sobre os diferentes tipos de regiões administrativas. Embora sejam referências válidas que devem ser consideradas na identificação dessa característica (i.e., locais), as menções à locais em textos na internet apresentam diferentes níveis de granularidade. Uma possibilidade a ser explorada é o uso dos dados presentes nos arquivos de base setorial do IBGE. A principal motivação para o uso é a existência de dados em diferentes granularidades para a referência a um local ou ponto de interesse (e.g., regiões administrativas, tipos de logradouro, tipos de complemento de logradouro, referências em texto livre para a localização de logradouros)

- Extensão da presente pesquisa para contemplar a identificação precoce de potencial atividade predatória: o MDAP, assim como outros métodos presentes na literatura, considera a conversa inteira entre os participantes como a amostra a ser analisada. No entanto, essa abordagem, embora apresente resultados significativos na identificação da atividade predatória sexual, apresentaria uma grande desvantagem, pois o tempo de exposição do menor de idade ao predador sexual seria maior,

o que poderia aumentar os riscos. Reduzindo esse tempo, diminui-se as chances de sextorsão, do abuso sexual e do desenvolvimento de distúrbios mentais. Desta forma, considera-se explorar a aplicação do MDAP em conjunto com modelos de aliciamento presentes na literatura, como por exemplo, o LCT [Olson et al., 2007].

- Exploração de características psicolinguísticas: a presente pesquisa explorou a presença de características textuais e comportamentais na identificação da atividade predatória sexual. A exploração das características psicolínguisticas na identificação da atividade predatória sexual contribuiu para a criação de métodos candidatos ao estado da arte na língua inglesa, por meio de ferramentas como o LIWC [Pennebaker et al., 2015]. Dentro desse contexto, entende-se como oportunidade para a evolução do MDAP, o uso da ferramenta LIWC 2015 para a língua portuguesa [Carvalho et al., 2019].

- Revisão do método de seleção de conversas não predatórias: na criação do conjunto de dados PRED-2050-ALL, buscou-se uma maior representatividade dos dados, de tal forma que as conversas predatórias obtidas pudessem ser analisadas em conjunto com as conversas oriundas de categorias normalmente presentes na internet brasileira. No entanto, uma consequência dessa representatividade foi a presença de uma quantidade de conversas não predatórias cujas as características não contribuíram para uma melhor identificação da atividade predatória sexual. Um exemplo de conversas com essa característica são aquelas representadas por apenas uma ou poucas mensagens trocadas. Também foram identificadas algumas conversas totalmente em inglês. Assim, compreende-se uma lacuna para a melhoria do processo de seleção de conversas não predatórias. Da mesma forma que a baixa sobreposição de termos entre as classes permitiu atingir resultados significativos em alguns cenários, a pouca ocorrência de características oriundas do MDAP em conversas não predatórias apresentaram um comportamento ruidoso.

- Exploração de novas características textuais e comportamentais: ao analisar as conversas predatórias classificadas incorretamente, foram encontrados dois exemplares em que o assunto discutido remetia à maturidade sexual da vítima, como por exemplo, se a vítima era virgem ou já havia tido relações sexuais. Em ambas as conversas, o MDAP não reconheceu o contexto discutido devido as diferentes abordagens aplicadas pelo predador sexual para a condução da conversa. Um outro

comportamento identificado durante a análise foi a resistência da vítima durante a ocorrência da atividade predatória. Dentre as características textuais, é possível identificar a ocorrência de diferentes vocábulos, normalmente usados na internet, denominado *internetês*.

- Revisão da padronização aplicada pelos módulos MPCTI e MPCTF: conforme foi possível observar na avaliação experimental, o algoritmo NBM apresentou os resultados mais modestos dentre os cinco algoritmos de aprendizado de máquina considerados. No entanto, o algoritmo NBM foi o único algoritmo que apresentou resultados superiores em todos os experimentos realizados com a aplicação do MDAP (quando comparado ao *baseline*). Desta forma, compreende-se que devem ser realizados novos experimentos com o algoritmo, uma vez que sejam identificadas novas estratégias para a redução de dimensionalidade das conversas e a exploração de mais características comportamentais e textuais. Dentre as possibilidades, ao observar as 200 características mais importantes, compreende-se que a lematização e o *stemming* são boas alternativas a serem exploradas em testes futuros.

- Aumento de amostras de conversas predatórias: a base de conversas predatórias apresenta diferentes características documentadas na literatura sobre o domínio da pesquisa, no entanto, algumas delas ainda carecem de mais exemplos para que possam apresentar uma maior importância e assim contribuir para resultados mais significativos. Desta forma, entende-se que para a evolução do MDAP, é necessário a disponibilização de mais conversas predatórias.

- Disponibilização de API para a identificação de conversas predatórias: a principal motivação para essa iniciativa seria permitir que as plataformas que ofereçam o serviço de conversas entre os usuários, por exemplo, as redes sociais e as salas de bate-papo, tenham a possibilidade de incluir uma camada adicional de segurança na comunicação virtual ocorrida.

A- Apêndice

Tabela 23 – Resultados obtidos com a avaliação experimental.

Configuração	SVM (%)	DT (%)	RF (%)	MLP (%)	NBM (%)
0-MIN1-25	89,26	87,55	90,86	88,10	82,15
0-MIN1-50	95,32	87,60	91,96	91,67	82,85
0-MIN1-100	95,40	88,23	91,90	90,22	83,70
0-MIN1-200	95,85	89,11	92,57	89,82	80,56
0-TR-25	92,26	86,88	91,86	92,48	82,15
0-TR-50	93,28	87,47	92,41	93,54	82,58
0-TR-100	93,60	88,97	93,09	92,78	85,01
0-TR-200	93,85	89,74	93,54	92,01	81,81
21-MIN1-25	92,24	88,27	92,48	89,85	82,32
21-MIN1-50	95,03	87,76	92,84	95,11	84,16
21-MIN1-100	95,14	87,94	92,63	93,76	83,95
21-MIN1-200	95,09	88,00	92,86	94,99	81,45
21-TR-25	95,43	87,85	92,91	92,3	81,96
21-TR-50	94,69	89,13	93,68	92,31	84,48
21-TR-100	94,68	88,98	94,04	93,03	84,24
21-TR-200	94,85	89,66	94,03	94,87	83,20
42-MIN1-25	94,56	88,80	92,79	94,60	83,56
42-MIN1-50	95,45	90,22	93,88	95,35	85,18
42-MIN1-100	95,32	90,31	93,39	93,29	85,43
42-MIN1-200	95,34	90,79	94,05	92,44	84,23
42-TR-25	94,56	90,43	93,74	94,61	83,34
42-TR-50	95,45	91,74	94,23	95,35	85,33
42-TR-100	95,28	90,88	93,80	93,29	85,67
42-TR-200	95,29	91,20	94,03	92,32	85,63
84-MIN1-25	96,33	87,35	94,14	90,57	85,17
84-MIN1-50	96,75	85,89	95,41	96,69	88,29
84-MIN1-100	96,71	84,89	96,10	94,86	87,58
84-MIN1-200	96,63	84,99	95,68	95,08	81,86
84-TR-25	96,82	87,33	95,48	95,68	84,79
84-TR-50	96,74	85,63	96,81	95,39	88,42
84-TR-100	96,67	85,48	97,2	93,89	89,30
84-TR-200	96,50	84,88	96,23	93,60	85,21
40-MIN1-25	92,85	86,01	90,98	89,72	84,83
40-MIN1-50	93,49	86,24	92,04	92,14	81,25
40-MIN1-100	93,74	85,73	92,13	91,56	83,42
40-MIN1-200	93,46	87,11	92,26	90,97	82,74
40-TR-25	94,15	87,82	92,66	91,67	85,36
40-TR-50	93,19	87,24	92,82	91,17	81,58
40-TR-100	93,55	86,82	93,25	91,65	84,05
40-TR-200	93,42	88,20	93,29	89,92	84,10
50-MIN1-25	95,40	88,47	92,29	93,21	83,43
50-MIN1-50	95,54	87,88	92,88	94,29	85,73
50-MIN1-100	95,85	87,93	93,13	92,54	86,56
50-MIN1-200	95,69	88,36	92,86	94,94	82,19
50-TR-25	97,15	90,82	93,92	96,36	83,33
50-TR-50	97,87	90,16	94,40	96,28	86,04
50-TR-100	97,87	89,83	95,04	95,81	88,39
50-TR-200	97,76	90,57	94,18	96,13	84,62
60-MIN1-25	90,43	86,38	91,04	88,45	81,27
60-MIN1-50	95,53	86,41	90,98	96,51	83,40
60-MIN1-100	95,72	86,22	91,78	95,59	84,95
60-MIN1-200	95,83	87,56	92,66	96,35	80,41
60-TR-25	92,63	85,71	91,77	91,93	81,63
60-TR-50	94,12	85,80	92,69	95,06	83,89
60-TR-100	94,40	86,35	92,81	94,42	86,71
60-TR-200	94,88	86,53	93,20	95,28	82,32
70-MIN1-25	92,29	89,63	93,20	91,43	83,00
70-MIN1-50	96,04	90,83	93,70	95,00	84,37
70-MIN1-100	95,65	91,16	93,98	94,82	84,50
70-MIN1-200	95,70	91,03	94,02	95,54	84,46
70-TR-25	95,85	92,52	94,14	93,15	83,34
70-TR-50	96,27	93,31	95,11	93,81	85,24
70-TR-100	96,80	92,91	94,64	92,29	85,00
70-TR-200	96,83	93,15	95,05	92,83	86,11
80-MIN1-25	91,63	88,37	93,25	88,85	80,14
80-MIN1-50	95,63	87,94	94,47	94,93	79,96
80-MIN1-100	95,42	88,76	94,15	92,75	83,93
80-MIN1-200	95,71	90,34	94,72	92,03	82,24
80-TR-25	93,42	88,89	94,24	90,91	80,67
80-TR-50	94,69	89,66	94,94	91,82	80,36
80-TR-100	94,78	90,3	95,27	90,14	84,48
80-TR-200	94,54	90,96	94,99	90,04	82,90

(a) MDAP

Configuração	SVM (%)	DT (%)	RF (%)	MLP (%)	NBM (%)
0-MIN1-25	91,42	88,06	91,15	89,71	74,13
0-MIN1-50	93,22	88,14	92,48	88,27	77,94
0-MIN1-100	93,20	89,74	92,33	87,01	78,79
0-MIN1-200	93,11	90,14	92,62	88,22	78,66
0-TR-25	90,89	86,04	90,17	89,94	73,35
0-TR-50	93,31	89,47	92,07	91,67	77,59
0-TR-100	93,80	89,04	92,31	90,53	79,24
0-TR-200	94,19	89,11	92,05	91,80	80,05
21-MIN1-25	92,19	86,04	90,17	87,39	76,70
21-MIN1-50	92,24	89,47	92,07	89,01	77,07
21-MIN1-100	92,57	89,04	92,31	87,73	76,41
21-MIN1-200	92,87	89,11	92,05	90,47	76,71
21-TR-25	92,85	86,35	91,05	90,47	75,92
21-TR-50	92,50	90,06	92,57	92,45	76,60
21-TR-100	93,07	89,45	92,73	91,42	75,99
21-TR-200	93,54	89,20	92,50	93,35	77,10
42-MIN1-25	92,89	91,27	93,31	91,72	79,72
42-MIN1-50	93,12	91,75	93,58	92,13	80,84
42-MIN1-100	94,23	91,99	93,91	91,61	82,56
42-MIN1-200	94,55	92,06	94,26	91,00	81,88
42-TR-25	93,02	90,81	93,66	93,20	78,85
42-TR-50	93,40	91,25	94,49	93,69	80,27
42-TR-100	94,34	91,72	94,46	92,27	82,24
42-TR-200	94,78	91,62	94,67	91,22	81,86
84-MIN1-25	92,78	87,78	90,77	86,65	74,95
84-MIN1-50	94,35	84,96	90,88	89,21	76,44
84-MIN1-100	94,41	85,12	92,26	90,44	76,44
84-MIN1-200	94,70	84,89	92,73	90,76	77,71
84-TR-25	92,69	88,08	92,46	89,99	74,83
84-TR-50	93,54	85,52	93,95	92,05	77,08
84-TR-100	92,91	85,92	94,41	92,93	77,18
84-TR-200	93,55	86,32	94,85	92,38	78,90
40-MIN1-25	88,85	84,84	87,71	84,16	75,16
40-MIN1-50	89,06	85,49	89,27	87,18	79,46
40-MIN1-100	89,48	84,78	90,45	87,92	79,49
40-MIN1-200	90,32	85,83	90,15	88,68	79,39
40-TR-25	88,93	85,13	89,02	85,55	75,32
40-TR-50	88,65	85,97	90,74	88,30	79,44
40-TR-100	89,96	84,36	90,75	88,60	80,06
40-TR-200	90,60	86,07	91,48	90,07	79,85
50-MIN1-25	92,85	86,87	90,36	91,81	73,69
50-MIN1-50	94,02	86,55	91,67	92,01	78,11
50-MIN1-100	94,46	87,39	92,13	90,11	79,76
50-MIN1-200	94,14	87,47	92,17	91,67	80,30
50-TR-25	92,59	85,22	91,81	92,26	73,29
50-TR-50	93,98	86,26	93,32	93,11	77,88
50-TR-100	94,19	85,71	93,54	92,71	79,97
50-TR-200	94,61	87,01	93,30	93,75	80,57
60-MIN1-25	89,56	84,13	89,13	89,62	73,47
60-MIN1-50	92,76	87,25	91,57	90,85	77,01
60-MIN1-100	93,74	86,30	92,53	92,56	80,63
60-MIN1-200	94,35	87,67	92,99	93,61	78,57
60-TR-25	89,35	85,80	91,03	91,65	73,63
60-TR-50	91,81	87,11	93,19	91,93	77,40
60-TR-100	93,29	87,14	93,80	94,03	81,90
60-TR-200	93,69	87,64	94,07	94,64	80,06
70-MIN1-25	94,20	88,59	91,71	86,43	74,89
70-MIN1-50	94,07	90,47	92,76	88,03	77,07
70-MIN1-100	95,30	90,32	92,63	88,03	79,48
70-MIN1-200	95,71	90,87	93,17	90,03	80,92
70-TR-25	94,01	91,28	92,78	89,19	75,22
70-TR-50	94,66	92,80	94,10	92,08	77,10
70-TR-100	95,65	92,24	93,97	90,13	78,80
70-TR-200	95,65	93,13	94,08	92,54	81,95
80-MIN1-25	91,00	87,17	93,23	88,71	75,81
80-MIN1-50	92,20	87,41	93,17	87,90	76,39
80-MIN1-100	92,78	87,65	93,03	87,22	78,40
80-MIN1-200	93,31	89,55	93,52	86,51	77,53
80-TR-25	90,06	88,83	93,05	88,45	75,25
80-TR-50	92,03	89,26	93,48	90,03	75,56
80-TR-100	92,95	89,56	93,73	87,66	77,68
80-TR-200	93,16	90,58	94,25	86,83	78,15

(b) *baseline*

TEM (%)	Conversa
100,00	“ai bom coisa fofa ainda micp_teoradulto_pred agora vim entender porque voce timido pois vou quebrar tabu teu”
100,00	“aqui quero ve micp_local vale pena hum sabia micp_parte_do_corpo deliciosa”
100,00	“fico cheio vontades aqui vontade d nao sei micp_local falo ummm vontade muita fala ta vontade tocar todo micp_parte_do_corpo micp_teoradulto_pred micp_parte_do_corpo la naquele lugar”
100,00	“vai ser primeira vez pouco medo pois nunca fiz tbm primeira vez micp_pergunta micp_pergunta sim fica boa sera micp_local resto vida ta bem”
98,33	“voce deita migo cama micp_teoradulto_pred nao posso durmi ai nao pra dumir so fica migo ate 3 horas tardi ai levo voce micp_local tah”
80,83	“micp_envio_saudacao micp_elogio voce micp_parentes bem micp_elogio pra brincar homem micp_idade_maior anos micp_pergunta pedofilo bem bonzinho pra brincar ela voce sabe ideia micp_parte_do_corpo ne cara micp_pergunta tipo exatamente tipo crime pf permissao micp_local quebrar protocolos privacidade localizar usuarios sei apenas zuando afinal pedofilo verdade nao tao burro”
41,67	“aqui perto casa pq senao nao vai querer deixar sair kkk pode ser marca lugar encontro vc ai vai pr casa ai micp_teoradulto_pred dia depos vc vai pr casa deixa micp_local depos carnaval ai fas micp_elogio aqui zap”
20,01	“hi micp_envio_saudacao micp_nome micp_elogio he hello dont know how to speak spanish micp_elogio voce quer fotos micp_teoradulto_pred duro quero fotos micp_foto micp_envio_saudacao delicioso realmente quero lamber ha fotos micp_pergunta quer ver micp_parte_do_corpo micp_idade_menor anos garanto micp_teoradulto_pred delicioso bem voce quer jogar quero micp_elogio tambem quero micp_foto gostou micp_teoradulto_pred”
12,25	“vc micp_interesse_foto foto micp_teoradulto_pred tambem nao sei q horas vergonha entao micp_peca_de_roupa nao sei nunca fiz”
1,67	“entao cara procuro outro micp_teoradulto_pred sigiloso pra fuder mlk junto comigo micp_parentes 7 quantos anos micp_pergunta 7 cara nao certo vc forca micp_teoradulto_pred micp_teoradulto_pred pq quer micp_pergunta pq quer”

Tabela 24 – Conversas classificadas incorretamente como pertencentes à classe “Não Predatória” (FN) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “50”.

TEM (%)	Conversa
100,00	“aqui quero ve micp_local vale pena hum sabia micp_parte_do_corpo deliciosa”
100,00	“fico cheio vontades aqui vontade d nao sei micp_local falo ummm vontade muita fala ta vontade tocar todo micp_parte_do_corpo micp_teoradulto_pred micp_parte_do_corpo la naquele lugar”
100,00	“vai ser primeira vez pouco medo pois nunca fiz tbm primeira vez micp_pergunta micp_pergunta sim fica boa sera micp_local resto vida ta bem”
95,00	“ai bom coisa fofa ainda micp_teoradulto_pred agora vim entender porque voce timido pois vou quebrar tabu teu”
86,67	“voce deita migo cama micp_teoradulto_pred nao posso durmi ai nao pra dumir so fica migo ate 3 horas tardi ai levo voce micp_local tah”
55,83	“vc micp_interesse_foto foto micp_teoradulto_pred tambem nao sei q horas vergonha entao micp_peca_de_roupa nao sei nunca fiz”
28,34	“micp_envio_saudacao micp_elogio voce micp_parentes bem micp_elogio pra brincar homem micp_idade_maior anos micp_pergunta pedofilo bem bonzinho pra brincar ela voce sabe ideia micp_parte_do_corpo ne cara micp_pergunta tipo exatamente tipo crime pf permissao micp_local quebrar protocolos privacidade localizar usuarios sei apenas zuando afinal pedofilo verdade nao tao burro”
16,66	“hi micp_envio_saudacao micp_nome micp_elogio he hello dont know how to speak spanish micp_elogio voce quer fotos micp_teoradulto_pred duro quero fotos micp_foto micp_envio_saudacao delicioso realmente quero lamber ha fotos micp_pergunta quer ver micp_parte_do_corpo micp_idade_menor anos garanto micp_teoradulto_pred delicioso bem voce quer jogar quero micp_elogio tambem quero micp_foto gostou micp_teoradulto_pred”
10,00	“aqui perto casa pq senao nao vai querer deixar sair kkk pode ser marca lugar encontro vc ai vai pr casa ai micp_teoradulto_pred dia depos vc vai pr casa deixa micp_local depos carnaval ai fas micp_elogio aqui zap”
10,00	“oiee micp_emoticon micp_local micp_local valeu kkkk micp_emoticon curte fotos curtir micp_pergunta micp_local vc curte ver fotos ah ta depende micp_local fotos normais sim normal sim nao to dizendo micp_teoradulto_pred ah ta sim rsrsr micp_emoticon micp_teoradulto_pred pra ganhar pontos micp_pergunta sei la puts hehe he he quantos anos micp_idade_menor humvc gosta ver micp_teoradulto_pred quer micp_ordem_dada sim micp_ordem_dada micp_teoradulto_pred vai”
7,50	“ate hj espero foto sinto vc nao ira micp_ordem_dada r sinta q vou micp_ordem_dada r micp_ordem_dada agr ja to micp_parte_do_corpo bom vc so deve ta micp_peca_de_roupa nada micp_local micp_comodo_da_casa ja apagada micp_emoticon ok”
3,33	“entao cara procuro outro micp_teoradulto_pred sigiloso pra fuder mlk junto comigo micp_parentes 7 quantos anos micp_pergunta 7 cara nao certo vc forca micp_teoradulto_pred micp_teoradulto_pred pq quer micp_pergunta pq quer”
2,50	“micp_ordem_dada vc frente micp_ordem_dada piupiu micp_parte_do_corpo micp_ordem_dada nao so voce micp_ordem_dada outra micp_elogio micp_ordem_dada entenda micp_ordem_dada”
1,67	“ta ond vamos micp_pergunta onde quer ir ond vc quiser nao sei vamos micp_teoradulto_pred vamos micp_teoradulto_pred q vc vontade onde vai encontrar pq ninguem pode ver ne q ver”

Tabela 25 – Conversas classificadas incorretamente como pertencentes à classe “Não Predatória” (FN) após a aplicação do MDAP e o algoritmo SVM no conjunto de dados “0”.

Referências Bibliográficas

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Akter, S. and Aziz, M. T. (2016). Sentiment analysis on facebook group using lexicon based approach. *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1:1–4.

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*.

Anguita, D., Ghio, A., Ridella, S., and Sterpi, D. (2009). K-fold cross validation for error rate estimate in support vector machines. *DMIN*, 1:291–297.

Apté, C. and Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3):197–210.

Babchishin, K. M., Karl Hanson, R., and Hermann, C. A. (2011). The characteristics of online sex offenders: A meta-analysis. *Sexual Abuse*, 23(1):92–123.

Bagley, C. and King, K. (2003). *Child sexual abuse: The search for healing*. Routledge.

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. *ISI*.

Bestagini, P., Lipari, V., and Tubaro, S. (2017). A machine learning approach to facies classification using well logs. *Seg technical program expanded abstracts 2017*, 1:2137–2142.

Black, P. J., Wollis, M., Woodworth, M., and Hancock, J. T. (2015). A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect*, 44:140–149.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. *Proceedings of the 2006 conference on empirical methods in natural language processing*, 1:120–128.

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

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 1:144–152.

Boyer, M. and Lapalme, G. (1985). Generating paraphrases from meaning-text semantic networks. *Computational Intelligence*, 1(1):103–117.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cano, A. E., Fernandez, M., and Alani, H. (2014). Detecting child grooming behaviour patterns on social media. *International conference on social informatics*, 1:412–427.

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

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., and Guedes, G. P. (2019). Evaluating the brazilian portuguese version of the 2015 liwc lexicon with sentiment analysis in social networks. *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, 1:24–34.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., and Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

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.

Clark, J. L., Algoe, S. B., and Green, M. C. (2018). Social network sites and well-being: the role of social connection. *Current Directions in Psychological Science*, 27(1):32–37.

Crowell, C. R., Narvaez, D., and Gomberg, A. (2008). Moral psychology and information ethics: Psychological distance and the components of moral behavior in a digital world. *Information Security and Ethics: Concepts, Methodologies, Tools, and Applications*, 1:3269–3281.

Dhouioui, Z., Alqahtani, A. A., and Akaichi, J. (2016). Social networks security policies. *Intelligent Interactive Multimedia Systems and Services 2016*, 1:395–403.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

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

Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York.

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

Ellison, N. B. and Boyd, D. M. (2013). Sociality through social network sites. *The Oxford handbook of internet studies*.

Eyheramendy, S., Lewis, D. D., and Madigan, D. (2003). On the naive bayes model for text categorization.

Fathi, E. and Shoja, B. M. (2018). Deep neural networks for natural language processing. *Handbook of statistics*, 38:229–316.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Feldman, R., Sanger, J., et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Forsythand, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. *International Conference on Semantic Computing (ICSC 2007)*, 1:19–26.

Fortin, F., Paquette, S., and Dupont, B. (2018). From online to offline sexual offending: Episodes and obstacles. *Aggression and Violent Behavior*.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Gabrilovich, E. and Markovitch, S. (2005). Feature generation for text categorization using world knowledge. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1:1048–1053.

García, S., Luengo, J., and Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

García, V., Mollineda, R. A., and Sánchez, J. S. (2008). On the k-nn performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, 11(3-4):269–280.

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. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 124.

Gil, B. D. (2014). O léxico do corpo humano em livros didáticos de português para falantes de outras línguas. *Linha D'Água*, 27(2):9–23.

Gillam, L. and Vartapetiance, A. (2012). Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. *LNCS*.

Gudivada, V. N., Rao, D. L., and Gudivada, A. R. (2018). Information retrieval: Concepts, models, and systems. *Handbook of statistics*, 38:331–401.

Gunawan, F. E., Ashianti, L., Candra, S., and Soewito, B. (2016). Detecting online child grooming conversation. *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, 1:1–6.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hasebrink, U., Livingstone, S., Haddon, L., and Olafsson, K. (2009). Comparing children's online opportunities and risks across europe: Cross-national comparisons for eu kids online.

Haykin, S. S., Haykin, S. S., Haykin, S. S., Elektroingenieur, K., and Haykin, S. S. (2009). *Neural networks and learning machines*, volume 3. Pearson education Upper Saddle River.

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.

Hidalgo, J. M. G. and Díaz, A. A. C. (2012). Combining predation heuristics and chat-like features in sexual predator identification. *CLEF (Online Working Notes/Labs/Workshop)*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 1:65–70.

Hu, Y.-J. and Kibler, D. (1996). Generation of attributes for learning algorithms. *Conference on Innovative Applications of Artificial Intelligence - AAAI/IAAI*, 1:806–811.

Huber, P. and Ronchetti, E. (1981). Robust statistics, ser. *Wiley Series in Probability and Mathematical Statistics. New York, NY, USA, Wiley-IEEE*, 52:54.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63.

Hunt, E. B., Marin, J., and Stone, P. J. (1966). Experiments in induction.

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

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning*, 1:137–142.

Joachims, T. (2002). *Learning to classify text using support vector machines*, volume 668. Springer Science & Business Media.

Kao, A. and Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Karan, M. and Šnajder, J. (2018). Cross-domain detection of abusive language online. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137.

Keipi, T. (2018). Relatedness online: an analysis of youth narratives concerning the effects of internet anonymity. *Young*, 26(2):91–107.

Khurana, U., Nargesian, F., Samulowitz, H., Khalil, E., and Turaga, D. (2016). Automating feature engineering. *Transformation*, 10(10):10.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 26(9):1011–1013.

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

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

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

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

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283.

Kotu, V. and Deshpande, B. (2019). Chapter 4 - classification. *Data Science (Second Edition)*, 1:65–163.

Larose, D. T. and Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*, volume 4. John Wiley & Sons.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45.

Leary, M. R. and Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. *Advances in experimental social psychology*, 32:1–62.

Liu, D., Suen, C. Y., and Ormandjieva, O. (2017). A novel way of identifying cyber predators. *arXiv preprint arXiv:1712.03903*.

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.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Lorenzo-Dus, N. and Izura, C. (2017). "cause ur special": Understanding trust and complimenting behaviour in online grooming discourse. *Journal of Pragmatics*, 112:68–82.

Lorenzo-Dus, N., Izura, C., and Pérez-Tattam, R. (2016). Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media*, 12:40–50.

Manning, C. D., Schütze, H., and Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press.

Mayfield, E. and Penstein-Rosé, C. (2010). Using feature construction to avoid large feature spaces in text classification. *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 1299–1306.

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., and Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122.

McLaughlin, J. H. (2010). Crime and punishment: Teen sexting in context. *Penn St. L. Rev.*, 115:135.

Misra, S. and Li, H. (2019). Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine Learning for Subsurface Characterization*, page 243.

Monroy, A. P. L., González, F. A., Montes, M., Escalante, H. J., and Solorio, T. (2018). Early text classification using multi-resolution concept representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, 1:1216–1225.

Morris, C. (2013). *Identifying online sexual predators by svm classification with lexical and behavioral features*. PhD thesis.

MSD (2018). Transtorno de conversão - distúrbios de saúde mental - manual msd versão saúde para a família. `https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/transtornos-de-sintomas-som%C3%A1ticos-e-transtornos-relacionados/transtorno-de-convers%C3%A3o` (Acessado em 21 de novembro de 2020).

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)*.

Olowu, D. (2014). Cyber-based obscenity and the sexual exploitation of children via the internet: Implications for africa. *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.

ONDH (2020). Disque direitos humanos - relatório 2019. *National Center for Missing & Exploited Children Web site., https://www.gov.br/mdh/pt-br/acesso-a-informacao/ouvidoria/Relatorio_Disque_100_2019_.pdf . (Acessado em 22 de setembro de 2020)*.

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

Patel, K., Fogarty, J., Landay, J. A., and Harrison, B. (2008). Investigating statistical machine learning as a tool for software development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1:667–676.

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, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. *International Conference on Semantic Computing (ICSC 2007)*, 1:235–241.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

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, N. and Thelwall, M. (2012). Biographies or blenders: Which resource is best for cross-domain sentiment analysis? *International Conference on Intelligent Text Processing and Computational Linguistics*, 1:488–499.

Postal, J. G. et al. (2017). Avaliação do uso de quantificadores de teoria da informação para identificação de conversas online de pedofilia.

Price, L. and Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology*, 56(8):883–888.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 1:532–538.

Rennie, J. D. (2001). *Improving Multi-class Text Classification with Naive Bayes*. PhD thesis, Massachusetts Institute of Technology.

Rodríguez-Fdez, I., Canosa, A., Mucientes, M., and Bugarín, A. (2015). STAC: a web platform for the comparison of algorithms using statistical tests. *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.

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

Say, G. N., Babadağı, Z., Karabekiroğlu, K., Yüce, M., and Akbaş, S. (2015). Abuse characteristics and psychiatric consequences associated with online sexual abuse. *Cyberpsychology, Behavior, and Social Networking*, 18(6):333–336.

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

Seto, M. C., Karl Hanson, R., and Babchishin, K. M. (2011). Contact sexual offending by men with online sexual offenses. *Sexual Abuse*, 23(1):124–145.

Singh, V. (2017). Replace or Retrieve Keywords In Documents at Scale. *ArXiv e-prints*.

Sinnamon, G. (2017). The psychology of adult sexual grooming: Sinnamon's seven-stage model of adult sexual grooming. *The Psychology of Criminal and Antisocial Behavior*, 1:459–487.

Slonje, R. and Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154.

Soh, P. C.-H., Chew, K. W., Koay, K. Y., and Ang, P. H. (2018). Parents vs peers' influence on teenagers' internet addiction and risky online activities. *Telematics and Informatics*, 35(1):225–236.

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

Specia, L., Srinivasan, A., Joshi, S., Ramakrishnan, G., and Nunes, M. d. G. V. (2009). An investigation into feature construction to assist word sense disambiguation. *Machine Learning*, 76(1):109–136.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 1:3104–3112.

Tener, D., Wolak, J., and Finkelhor, D. (2015). A typology of offenders who use online communications to commit sex crimes against minors. *Journal of Aggression, Maltreatment & Trauma*, 24(3):319–337.

van der Hof, S. and Koops, B.-J. (2011). Adolescents and cybercrime: Navigating between freedom and control. *Policy & Internet*, 3(2):1–28.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. and Vapnik, V. (1998). Statistical learning theory. *New York*, 1(624):2.

Verbaeten, S. and Van Assche, A. (2003). Ensemble methods for noise elimination in classification problems. *International Workshop on Multiple classifier systems*, 1:317–325.

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. *CLEF (Online Working Notes/Labs/Workshop)*, 1178.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th annual meeting of the association for computational linguistics (Short papers)*, 2:90–94.

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

Wilcoxon, F. (1992). Individual comparisons by ranking methods. *Breakthroughs in statistics*, 1:196–202.

Winters, G. M., Kaylor, L. E., and Jeglic, E. L. (2017). Sexual offenders contacting children online: an examination of transcripts of sexual grooming. *Journal of sexual aggression*, 23(1):62–76.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolak, J., Finkelhor, D., and Mitchell, K. (2004). Internet-initiated sex crimes against minors: Implications for prevention based on findings from a national study. *Journal of adolescent health*, 35(5):424–e11.

Wolak, J., Finkelhor, D., and Mitchell, K. J. (2009). Trends in arrests of ”online predators”.

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, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *ICML*, 1:412–420.

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis lectures on human language technologies*, 1(1):1–141.

Zhang, H. (2004). The optimality of Naïve Bayes. *Florida Artificial Intelligence Research Society Conference*.

Zhao, X., Wu, Y., Lee, D. L., and Cui, W. (2018). iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics*, 25(1):407–416.