ESTRATÉGIAS PARA DETECÇÃO PRECOCE DE PREDADORES SEXUAIS EM CONVERSAS REALIZADAS NA INTERNET

Marcelle Ramos Panzariello

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2022

ESTRATÉGIAS PARA DETECÇÃO PRECOCE DE PREDADORES SEXUAIS EM CONVERSAS REALIZADAS NA INTERNET

Marcelle Ramos Panzariello

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Geraldo Bonorino Xexéo

Aprovada por: Prof. Geraldo Bonorino Xexéo
Prof. Geraldo Zimbrão da Silva
Prof. Gustavo Paiva Guedes e Silva

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2022

Panzariello, Marcelle Ramos

Estratégias para Detecção Precoce de Predadores Sexuais em Conversas realizadas na Internet/Marcelle Ramos Panzariello. – Rio de Janeiro: UFRJ/COPPE, 2022.

XVII, 101 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2022.

Referências Bibliográficas: p. 97 – 101.

1. early detection of sexual predators. 2. sexual predator identification. 3. text classification. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico este trabalho a todas as crianças. Quisera o mundo ser um lugar seguro.

Agradecimentos

Primeiramente, agradeço a Deus por me permitir chegar até aqui, me dando forças e sabedoria.

Agradeço muito aos meus pais, Elizabeth e Marcelo, que foram muito compreensivos durante toda a minha jornada, apoiando e respeitando minhas escolhas.

Agradeço imensamente ao meu orientador, Geraldo Xexéo, por todos os conhecimentos passados, pelas reuniões, pelas oportunidades concedidas, pela paciência e pelas palavras gentis em momentos de desespero. Agradeço profundamente por não ter desistido de mim, e ter me permitido e incentivado a concluir meu mestrado.

Um agradecimento especial a alguns amigos que me ajudaram significativamente na minha caminhada: Arthur Ferreira, Débora Andrade, Airine Carmo, Eduardo Mangeli, Luan Barbosa, Gabriel Almeida e em especial, Izandro Monteiro, que esteve comigo na reta final do mestrado dividindo conhecimentos, medos, erros e acertos.

Agradeço também ao Fellipe Duarte por ter me indicado o desafio do PAN 2012, que me abriu diversas possibilidades que culminaram nesta dissertação.

Agradeço aos professores Geraldo Zimbrão e Gustavo Guedes por aceitarem fazer parte da minha banca e por suas valiosas considerações.

Agradeço ao PESC pela oportunidade e a CAPES pelo apoio financeiro.

A todos que contribuíram de alguma forma, muito obrigada.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ESTRATÉGIAS PARA DETECÇÃO PRECOCE DE PREDADORES SEXUAIS EM CONVERSAS REALIZADAS NA INTERNET

Marcelle Ramos Panzariello

Setembro/2022

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Crianças e adolescentes estão expostas a riscos na internet. Predadores sexuais podem entrar em contato com suas vítimas através de *chats* em redes sociais e jogos *online*. Partindo da premissa que predadores sexuais precisam manter um vínculo com suas vítimas antes do encontro presencial, esta dissertação objetiva detectar precocemente predadores sexuais em conversas virtuais entre duas pessoas através do desenvolvimento de três estratégias distintas utilizando algoritmos de classificação de textos para auxiliar a reduzir o número de casos de abuso sexual infantil. Como conjunto de dados é utilizada a base do PAN 2012 e são utilizados algoritmos de classificação de textos como *Naive Bayes*, KNN, Floresta Aleatória, SVM, Rede Neural MLP e BERT. Para cada estratégia foram realizados experimentos sem balanceamento dos dados e utilizando técnicas de *undersampling*, que obtiveram resultados superiores ao estado da arte. A primeira estratégia obteve melhores resultados que as demais, atingindo $F_{0.5} = 85,96\%$ já para as primeiras 10 mensagens para o experimento sem balanceamento dos dados e $F_{0.5} = 99,89\%$ para as primeiras 10 mensagens com o experimento com técnicas de *undersampling*.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

STRATEGIES FOR EARLY DETECTION OF SEXUAL PREDATORS IN INTERNET CONVERSATIONS

Marcelle Ramos Panzariello

September/2022

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Children and teenagers are exposed to risks on the internet. Sexual predators can contact their victims through social media chats and online games. Starting from the premise that sexual predators need to maintain a bond with their victims before the face-to-face meeting, this dissertation aims to early detect sexual predators in online conversations between two people through the development of three distinct strategies using text classification algorithms to help reduce the number of cases of child sexual abuse. The 2012 PAN base is used as a dataset and text classification algorithms such as Naive Bayes, KNN, Random Forest, SVM, MLP Neural Network and BERT are used. For each strategy, experiments were performed without data balancing and using undersampling techniques, which obtained better results than the state of the art. The first strategy obtained better results than the others, reaching $F_{0.5} = 85.96\%$ for the first 10 messages for the experiment without data balancing and $F_{0.5} = 99.89\%$ for the first 10 messages with the experiment with undersampling techniques.

Sumário

Lista de Figuras **x**

Lista de Tabelas **xv**

Lista de Abreviaturas **xvii**

1 Introdução **1**

1.1 Motivação 1

1.2 Objetivos 4

1.3 Contribuições 4

1.4 Organização 5

2 Revisão Bibliográfica **6**

2.1 Pesquisa Bibliográfica 6

2.2 Trabalhos Correlatos 8

2.2.1 Detecção de *Grooming* 9

2.2.2 PAN 2012 11

2.2.3 Trabalhos mais Recentes 13

2.2.4 Detecção Precoce de Predadores Sexuais 14

2.2.5 Trabalhos com Dados Reais Brasileiros 15

2.2.6 Problemas da Área 17

3 Proposta **20**

3.1 Definição da Proposta 20

3.2 Definição dos Experimentos 21

3.3 Estratégia 1 - Distinguir Conversas Predatórias e Gerais 22

3.4 Estratégia 2 - Distinguir Predador e Vítima 23

3.5 Estratégia 3 - Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima 23

4 Experimentos **24**

4.1 Ambiente 24

4.2 Base do PAN 2012 . . . 25
4.3 Pré-Processamento . . . 26
4.4 Pré-Filtro . . . 31
4.5 Extração de Dados . . . 32
4.6 Configuração dos Experimentos . . . 32
4.6.1 Configurações dos Algoritmos . . . 33
4.7 Estratégia 1 - Distinguir Conversas Predatórias e Gerais . . . 34
4.7.1 Sem Balanceamento dos Dados . . . 35
4.7.2 Com *Undersampling* . . . 46
4.8 Estratégia 2 - Distinguir Predador e Vítima . . . 55
4.8.1 Sem Balanceamento dos Dados . . . 55
4.8.2 Com *Undersampling* . . . 65
4.9 Estratégia 3 - Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima . . . 73
4.9.1 Sem Balanceamento dos Dados . . . 73
4.9.2 Com *Undersampling* . . . 83
4.10 Considerações Finais . . . 92

5 Conclusão **94**
5.1 Contribuições . . . 95
5.2 Limitações e Trabalhos Futuros . . . 95

Referências Bibliográficas **97**

Lista de Figuras

2.1 Quantidade de publicações por ano das bases *Scopus*, IEEE e *Web of Science*. 8

4.1 Exemplo de conversa da base de treinamento do PAN 2012. 25
4.2 Trecho de conversa que continha entidades de caracteres HTML. As mensagens originais estão à direita, e as mensagens alteradas por (3) estão à esquerda. 27
4.3 Exemplo de mensagem que possuía muitas pontuações. As mensagens originais estão à direita, e as mensagens alteradas por (5) estão à esquerda. 28
4.4 Trecho de conversa que possuía símbolos e letras desconhecidas. As mensagens originais estão à direita, e as mensagens alteradas por (5) estão à esquerda. 29
4.5 Trecho de conversa que continha mensagens com letras repetidas. As mensagens originais estão à direita, e as mensagens alteradas por (8) estão à esquerda. 30
4.6 Trecho de conversa que continha mensagens com palavras repetidas. As mensagens originais estão à direita, e as mensagens alteradas por (9) estão à esquerda. 30
4.7 *Wordcloud* das mensagens predatórias da base de treinamento. 31
4.8 Quantidade de valores positivos e negativos para a classe alvo utilizando as primeiras 50 mensagens. 35
4.9 Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1. . . . 36
4.10 Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1. . . . 37
4.11 Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1. 37
4.12 Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1. 38

4.13 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1. 39
4.14 Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 1. 40
4.15 Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 1. 40
4.16 Resultado final do *recall* para o experimento sem balanceamento dos dados para a estratégia 1. 41
4.17 Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 1. 41
4.18 Resultado final do $F_{0.5}$ para o experimento sem balanceamento dos dados para a estratégia 1. 42
4.19 Quantidade de conversas predatórias corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 1. . . . 42
4.20 Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 1. 43
4.21 Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 1. 45
4.22 Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1. 46
4.23 Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1. 47
4.24 Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1. 47
4.25 Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1. 48
4.26 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1. 48
4.27 Resultado final da acurácia para o experimento com *undersampling* para a estratégia 1. 49
4.28 Resultado final da precisão para o experimento com *undersampling* para a estratégia 1. 49
4.29 Resultado final do *recall* para o experimento com *undersampling* para a estratégia 1. 50
4.30 Resultado final do F_1 para o experimento com *undersampling* para a estratégia 1. 50
4.31 Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 1. 51

4.32 Quantidade de conversas predatórias corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 1. . . . 52
4.33 Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 1. 52
4.34 Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 1. 54
4.35 Quantidade de valores positivos e negativos para a classe alvo utilizando as primeiras 50 mensagens. 56
4.36 Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2. . . . 56
4.37 Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2. . . . 57
4.38 Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2. 57
4.39 Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2. 58
4.40 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2. 58
4.41 Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 2. 59
4.42 Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 2. 60
4.43 Resultado final *recall* para o experimento sem balanceamento dos dados para a estratégia 2. 60
4.44 Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 2. 61
4.45 Resultado final do $F_{0.5}$ para o experimento sem balanceamento dos dados para a estratégia 2. 61
4.46 Quantidade de predadores (tuplas) corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 2. 62
4.47 Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 2. 63
4.48 Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 2. 64
4.49 Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2. 65
4.50 Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2. 65

4.51 Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2. 66
4.52 Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2. 67
4.53 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2. 67
4.54 Resultado final da acurácia para o experimento com *undersampling* para a estratégia 2. 68
4.55 Resultado final da precisão para o experimento com *undersampling* para a estratégia 2. 68
4.56 Resultado final do *recall* para o experimento com *undersampling* para a estratégia 2. 69
4.57 Resultado final do F_1 para o experimento com *undersampling* para a estratégia 2. 69
4.58 Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 2. 70
4.59 Quantidade de predadores (tuplas) corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 2. 70
4.60 Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 2. 71
4.61 Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 2. 72
4.62 Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3. . . . 73
4.63 Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3. . . . 74
4.64 Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3. 75
4.65 Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3. 75
4.66 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3. 76
4.67 Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 3. 76
4.68 Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 3. 77
4.69 Resultado final do *recall* para o experimento sem balanceamento dos dados para a estratégia 3. 77

4.70 Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 3. 78
4.71 Resultado final do $F_{0.5}$ para para o experimento sem balanceamento dos dados para a estratégia 3. 78
4.72 Quantidade de predadores (tuplas) corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 3. 79
4.73 Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 3. 80
4.74 Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 3. 82
4.75 Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3. 83
4.76 Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3. 83
4.77 Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3. 84
4.78 Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3. 85
4.79 Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3. 85
4.80 Resultado final da acurácia para o experimento com *undersampling* para a estratégia 3. 86
4.81 Resultado final da precisão para o experimento com *undersampling* para a estratégia 3. 86
4.82 Resultado final do *recall* para o experimento com *undersampling* para a estratégia 3. 87
4.83 Resultado final do F_1 para o experimento com *undersampling* para a estratégia 3. 87
4.84 Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 3. 88
4.85 Quantidade de predadores (tuplas) corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 3. 88
4.86 Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 3. 89
4.87 Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 3. 91

Lista de Tabelas

1.1 Como crianças e adolescentes brasileiros utilizavam a internet segundo a *TIC Kids Online Brasil* 2019. 2

2.1 Palavras-chave selecionadas e seus sinônimos. 7
2.2 *Strings* de busca intermediárias e a *string* final utilizada para buscas nas bases de dados. 7
2.3 Quantidade de documentos capturados por base de dados. Entre parênteses estão as quantidades de documentos manualmente selecionados. Busca realizada em 08 de abril de 2021. 8
2.4 Exemplos de frases categorizadas por KONTOSTATHIS *et al.* (2009). Traduzido de KONTOSTATHIS *et al.* (2009). 11
2.5 Quantidade de conversas do conjunto de dados PRED-2050-ALL. Retirado de DOS SANTOS (2021). 16
2.6 Tabela comparativa dos principais problemas da área e os trabalhos já realizados para cada problema. 18

4.1 Quantidade de dados das bases de treinamento e teste do PAN 2012. 25
4.2 Dados das bases de treinamento e teste após execução de (1). 27
4.3 Dados da base de treinamento após execução de (7). 29
4.4 Melhores parâmetros encontrados para os algoritmos. 34
4.5 Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 1. 44
4.6 Resultados detalhados das métricas para o experimento com *undersampling* dos dados para a estratégia 1. 53
4.7 Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 2. 63
4.8 Resultados detalhados das métricas para o experimento com *undersampling* dos dados para a estratégia 2. 71
4.9 Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 3. 81

4.10 Resultados detalhados das métricas para o experimento com *under-sampling* dos dados para a estratégia 3. 90

Lista de Abreviaturas

LCT	Luring Communication Theory, p. 10
LIWC	Linguistic Inquiry and Word Count, p. 13
MDAP	Método de Detecção de Atividade Predatória, p. 15
NIC.br	Núcleo de Informação e Coordenação do Ponto BR, p. 2
PJ	Perverted Justice, p. 9

Capítulo 1

Introdução

Esta dissertação busca fornecer estratégias eficientes que apoiem a redução do número de casos de abuso sexual infantil. Casos de pedofilia são um problema global (BEECH *et al.*, 2008; HILLMAN *et al.*, 2014) e crianças e adolescentes estão expostas diariamente a diferentes tipos de riscos na internet (NIC.BR, 2020), podendo ser abordadas por predadores sexuais em *chats* de redes sociais e jogos *online*.

Segundo OLSON *et al.* (2007), predadores sexuais executam certo comportamento padrão para abordar suas vítimas antes de conseguir contato sexual. Partindo desta premissa, é possível inferir que, se identificado precocemente este comportamento padrão, é possível evitar o encontro físico e, assim, o abuso sexual.

Dessa forma, este trabalho propõe o desenvolvimento de três estratégias distintas para detectar o mais cedo possível se existe um predador sexual em uma conversa realizada na internet entre duas pessoas. Para o desenvolvimento das estratégias são utilizados algoritmos de classificação de textos. Como conjunto de dados, são utilizados os dados provenientes da base disponibilizada pelo desafio *Sexual Predator Identification* da competição do PAN 2012, em virtude de sua diversidade de conversas, grande quantidade de dados e boa aceitação na literatura.

Nas seções seguintes são apresentadas as motivações para a realização deste trabalho, os objetivos e questões de pesquisa elencados, as contribuições e a organização geral deste documento.

1.1 Motivação

Com o avanço da tecnologia e popularização dos celulares e computadores domésticos, as gerações têm começado a utilizar a internet cada vez mais cedo. Segundo dados do *International Telecommunication Union*, agência especializada em tecnologia de informação e comunicação das Nações Unidas, 4.9 bilhões de pessoas no mundo usaram a internet em 2021, um aumento de aproximadamente 123% em comparação com 10 anos antes (ITU, 2022).

Em 2019, no Brasil, 89% da população de crianças e adolescentes entre 9 e 17 anos (aproximadamente 24 milhões de indivíduos) utilizava a internet (NIC.BR, 2020).

De acordo com a pesquisa *TIC Kids Online Brasil* 2019, realizada pelo Núcleo de Informação e Coordenação do Ponto BR (NIC.br) com crianças e adolescentes brasileiros de 9 a 17 anos, entre outubro de 2019 e março de 2020, as principais formas de utilização da internet neste período incluíam enviar mensagens instantâneas e acessar redes sociais. O estudo revelou também que a faixa etária de 15 a 17 anos enviou mais mensagens instantâneas e realizou mais chamadas de vídeo do que as demais faixas etárias (NIC.BR, 2020). As principais formas de utilização da internet obtidas na pesquisa e a porcentagem de crianças e adolescentes correspondentes podem ser observados na Tabela 1.1.

Tabela 1.1: Como crianças e adolescentes brasileiros utilizavam a internet segundo a *TIC Kids Online Brasil* 2019.

Principais Usos	**Porcentagem**
Assistir vídeos, programas, filmes ou séries	83%
Enviar mensagens instantâneas	79%
Usar redes sociais	68%
Ler ou assistir notícias	55%
Conversar com pessoas de outras cidades, países ou culturas	39%

Além de compreender como crianças e adolescentes utilizavam a internet, a pesquisa *TIC Kids Online Brasil* 2019 teve por objetivo entender como elas lidavam com os riscos e oportunidades da internet. Segundo a pesquisa, 15% já viram imagem ou vídeo de conteúdo sexual na internet e 6% se sentiram incomodados por conta disso (NIC.BR, 2020).

Na faixa etária de 11 a 17 anos, meninos relataram ter recebido mais mensagens de conteúdo sexual do que meninas (18% e 12%, respectivamente). Por outro lado, a porcentagem de meninas que receberam pedidos para enviar fotos e vídeos em que apareciam sem roupa foi maior do que a de meninos (13% e 8%, respectivamente) (NIC.BR, 2020).

Apesar dos riscos relatados pela *TIC Kids Online Brasil* 2019, como exposição a conteúdos sensíveis e contato com adultos de má conduta, é importante mencionar que existem outros riscos, como o *cyberbulling* (UNICEF, 2019) e o contato com predadores sexuais (pedófilos) (MOURA e CANGUÇU, 2022; UN, 2020).

Por conta destes riscos, é essencial que os pais analisem as atividades que seus filhos realizam *online*. Muitas crianças não tem consciência do risco que estão correndo e, na maioria das vezes, acreditam que não estão fazendo nada de errado. Segundo a *TIC Kids Online Brasil* 2019, existe uma incoerência entre o que as

crianças acham que podem fazer sozinhas na internet e o que os pais realmente permitem que elas façam. Por exemplo, 18% das crianças e adolescentes disseram ter permissões para dar informações pessoais para outras pessoas enquanto que os pais disseram que apenas 7% delas poderiam fazer isso. O mesmo ocorreu para postar fotos ou vídeos em que apareciam (60% para crianças *versus* 39% para os pais) e enviar mensagens instantâneas (77% para crianças *versus* 61% para os pais). Em todas as categorias da pesquisa as crianças informaram ter mais permissões do que, de fato, tinham (NIC.BR, 2020).

Outro estudo, com mais de 25 mil crianças de 25 diferentes países, revelou que 34% das crianças adicionam pessoas que não conhecem nas redes sociais e 15% enviam informações pessoais à estranhos (LIVINGSTONE *et al.*, 2011).

Além disso, o número de casos de exploração sexual de crianças e adolescentes tem crescido pelo mundo. Em 2017, em São Paulo, por exemplo, houve o maior número de denúncias de casos de exploração sexual de crianças e adolescentes desde 2013 (RBA, 2017). A pandemia do novo coronavírus, SARS-CoV-2, também aumentou o número de casos de pedofilia virtual no Brasil (BILCHES, 2020). De acordo com a reportagem, o isolamento social pode ter deixado as crianças mais vulneráveis com tanta exposição a internet, onde abusadores podem induzir crianças e adolescentes a fazerem algum ato sexual ou pedir fotos e/ou vídeos através de perfis falsos em redes sociais. Em casos mais graves, eles podem tentar marcar encontros pessoalmente, se passando por outros adolescentes.

Chats de redes sociais e jogos *online* podem ser ambientes propícios para pedófilos (CUNHA, 2017), já que crianças muitas vezes não têm maturidade suficiente para compreender riscos. Segundo OLSON *et al.* (2007), predadores sexuais executam certo comportamento padrão para abordar suas vítimas antes de conseguir contato sexual. Em algumas situações, os predadores podem criar um vínculo de amizade com suas vítimas, prometer presentes em troca de vídeos íntimos e até usar vídeos já enviados como ameaça para conseguir novos vídeos e fotos (MOURA e CANGUÇU, 2022).

Quase sempre, o encontro presencial com a vítima é o objetivo final do predador, pois é onde ocorre o abuso sexual. Conseguir detectar este comportamento pedofílico no início pode impedir que crianças e adolescentes tenham suas vidas destruídas ou interrompidas.

Assim, este trabalho parte desta premissa em busca de detectar, o mais cedo possível, se existe um predador sexual em uma conversa realizada na internet entre duas pessoas.

1.2 Objetivos

Partindo da premissa que a internet é um ambiente propício para os predadores sexuais e que eles precisam entrar em contato e manter um vínculo com suas vítimas antes do encontro presencial, foram levantadas as seguintes questões de pesquisa:

1. Detectar precocemente predadores sexuais é relevante para a sociedade?
2. Existem dados disponíveis para trabalhar com detecção precoce de predadores sexuais?
3. É possível identificar precocemente se existe um predador sexual em uma conversa realizada na internet?
4. Qual o estado da arte do problema de detecção precoce de predadores sexuais?
5. Quais as dificuldades, oportunidades e lacunas ainda em aberto para se identificar predadores sexuais em conversas realizadas na internet?
6. Utilizar algoritmos de classificação de texto são suficientes para detectar precocemente predadores sexuais em conversas realizadas na internet?
7. É possível que estratégias mais simples, como classificar conversas em predatórias e não predatórias, consigam melhores resultados do que estratégias mais complexas, como classificadores em dois estágios, para detectar precocemente predadores sexuais?

1.3 Contribuições

As principais contribuições deste trabalho são:

1. Criar estratégias que possam auxiliar a reduzir o número de casos de abuso sexual infantil;
2. Desenvolver três estratégias distintas que sejam capazes de detectar precocemente predadores sexuais em conversas realizadas na internet entre duas pessoas;
3. Validar se estratégias mais simples, que utilizam apenas um classificador, podem obter melhores resultados que estratégias mais complexas, que utilizam dois classificadores em sequência.

1.4 Organização

O capítulo seguinte apresenta o estudo bibliográfico realizado para esta dissertação, assim como os problemas relacionados à detecção precoce de predadores sexuais e os trabalhos relacionados mais relevantes. O capítulo 3 apresenta a proposta deste trabalho, descrevendo as premissas consideradas e quais as três estratégias desenvolvidas. O capítulo 4 detalha as etapas realizadas de pré-processamento, pré-filtro e extração de dados, assim como as configurações dos experimentos e algoritmos. Para cada estratégia, são apresentados os experimentos realizados com seus resultados. Por fim, o capítulo 5 apresenta as conclusões finais, com contribuições, limitações e trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Este capítulo está dividido em duas seções. A primeira seção revela como foi realizado o estudo bibliográfico para esta dissertação. Utilizando como metodologia a revisão narrativa da literatura, é apresentado como foram capturadas as publicações utilizadas como referencial teórico, além de serem mostradas as bases escolhidas para captura dos documentos e as *strings* de busca utilizadas para pesquisa.

A segunda seção apresenta os trabalhos correlatos mais relevantes, resultado da pesquisa bibliográfica realizada. O objetivo desta seção é apresentar o estado da arte do problema de identificação de predadores sexuais, assim como as dificuldades, oportunidades e lacunas da área.

2.1 Pesquisa Bibliográfica

Para construção do referencial teórico, optou-se por uma revisão narrativa da literatura. Essa metodologia foi escolhida por ter uma temática mais ampla, não necessitando de protocolos rígidos para sua confecção (CORDEIRO *et al.*, 2007). O principal objetivo desta pesquisa foi reunir dados suficientes para entender quais são as lacunas existentes relacionadas ao problema de identificação de predadores sexuais.

Com a finalidade de criar uma *string* de busca para procurar artigos em bases de dados, foram definidas as palavras-chave listadas na Tabela 2.1. Inicialmente, foram utilizadas as palavras-chave dos artigos encontrados isoladamente, e posteriormente, a tabela foi sendo atualizada de acordo com a relevância dos resultados trazidos pelas pesquisas nas bases de dados.

Com as palavras-chave definidas, foram criadas três *strings* de busca, uma para cada palavra-chave. Para montar a *string* final, as palavras "*child pornography*" e "*sexual grooming*" e seus sinônimos foram unidos na mesma sentença, pois têm a mesma finalidade. A palavra "*text classification*" e seus sinônimos foram incluídos na *string* final, pois ajudaram a refinar a pesquisa, evitando trabalhos que não

Tabela 2.1: Palavras-chave selecionadas e seus sinônimos.

Palavra-chave	**Sinônimos**
child pornography	*child sexual abuse* *child sexual exploitation* *child grooming*
sexual grooming	*sexual exploitation* *sexual predation* *sexual predatory* *sexual predator* *sexual predator detection/online predator detection* *sexual predator identification/online predator identification*
text classification	*text categorization* *cluster* *data mining* *deep learning* *machine learning* *natural language processing*

utilizavam nenhum algoritmo. As *strings* intermediárias e a *string* final podem ser visualizadas na Tabela 2.2.

Tabela 2.2: *Strings* de busca intermediárias e a *string* final utilizada para buscas nas bases de dados.

	Strings
String 1	*"child pornography" OR "child sexual abuse" OR "child sexual exploitation" OR "child grooming"*
String 2	*"sexual grooming" OR "sexual exploitation" OR "sexual predat*" OR "sexual predat* detection" OR "online predat* detection" OR "sexual predat* identification" OR "online predat* identification"*
String 3	*"text categorization" OR "classif*" OR "cluster*" OR "data mining" OR "deep learning" OR "machine learning" OR "natural language processing"*
String final	*("text categorization" OR "classif*" OR "cluster*" OR "data mining" OR "deep learning" OR "machine learning" OR "natural language processing") AND ("child pornography" OR "child sexual abuse" OR "child sexual exploitation" OR "child grooming" OR "sexual grooming" OR "sexual exploitation" OR "sexual predat*" OR "sexual predat* detection" OR "online predat* detection" OR "sexual predat* identification" OR "online predat* identification")*

As bases de dados escolhidas para captura dos artigos foram *Scopus*, IEEE e *Web of Science*. Foram considerados artigos publicados entre janeiro de 2012 e abril de 2021 e que estivessem escritos em inglês ou português. O tipo de busca utilizado foi busca em título, resumo e palavras-chave dos documentos.

Na Tabela 2.3 pode ser observada a quantidade de documentos resultantes das buscas nas bases de dados utilizando a *string* final, e, entre parênteses, a quantidade de documentos manualmente selecionados.

Tabela 2.3: Quantidade de documentos capturados por base de dados. Entre parênteses estão as quantidades de documentos manualmente selecionados. Busca realizada em 08 de abril de 2021.

Scopus	**IEEE**	***Web of Science***
88 (40)	37 (17)	53 (26)

Para seleção dos artigos foram utilizados alguns critérios, como manter apenas artigos que utilizavam algum algoritmo de classificação de textos, e descartar artigos que tratavam pornografia infantil em fotos ou vídeos. Este trabalho não lida com imagens, vídeos ou áudios, apenas com o texto das conversas.

Além destes artigos, outros 55 artigos foram capturados isoladamente.

Na Figura 2.1 é possível observar como os documentos resultantes da pesquisa nas bases de dados *Scopus*, IEEE e *Web of Science* estão distribuídos ao longo dos anos.

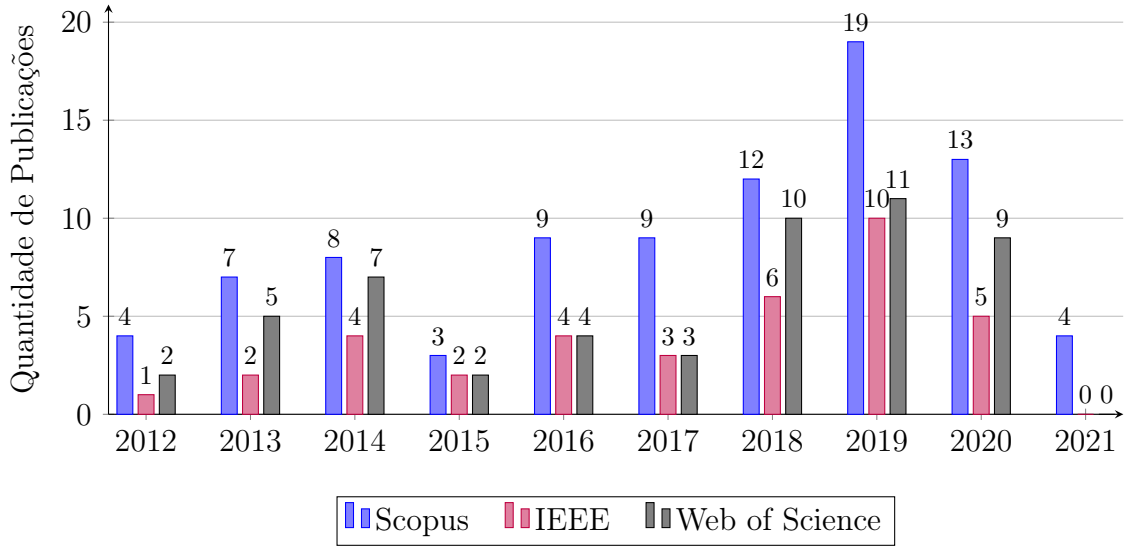


Figura 2.1: Quantidade de publicações por ano das bases *Scopus*, IEEE e *Web of Science*.

2.2 Trabalhos Correlatos

Provavelmente um dos primeiros trabalhos de classificação sobre identificação de predadores sexuais foi realizado por PENDAR (2007). Seu objetivo era conseguir distinguir entre a vítima e o predador sexual em uma conversa. Para isso, ele reuniu

701 conversas do site *Perverted Justice* (PJ)[1] e treinou classificadores SVM e KNN. Utilizando trigramas, foi obtido $F_1 = 94,3\%$ com o algoritmo KNN e $F_1 = 90,8\%$ com o algoritmo SVM.

Segundo PENDAR (2007), existem dois tipos de conversas com conteúdo sexual que são fundamentais para este problema: as conversas que envolvem predadores sexuais e as conversas consensuais entre adultos. As conversas que envolvem predadores sexuais são divididas ainda em três subtipos:

1. Conversas entre predadores e vítimas reais;
2. Conversas entre predadores e voluntários se passando por crianças;
3. Conversas entre predadores e oficiais da lei se passando por crianças.

Conversas do primeiro e terceiro tipo são muito difíceis de serem encontradas, pois envolvem dados sigilosos e questões legais (PENDAR, 2007).

PENDAR (2007) encontrou conversas para o segundo tipo no site PJ, que é uma fundação sem fins lucrativos que recrutava voluntários para se passarem por crianças em salas de bate-papo pela internet em busca de predadores sexuais. Quando um predador era encontrado e condenado, as conversas eram postadas no site. Atualmente, existem 622 conversas reais entre predadores sexuais condenados e voluntários se passando por crianças.

Dada a dificuldade de se obter dados reais entre vítimas e predadores sexuais, o site PJ acabou tornando-se referência na literatura, sendo utilizado posteriormente por diversos autores.

2.2.1 Detecção de *Grooming*

Em 2006, CRAVEN *et al.* (2006) propuseram uma nova definição para o termo "*sexual grooming of children*":

> Um processo pelo qual uma pessoa prepara uma criança, adultos importantes e o ambiente para o abuso dessa criança. Os objetivos específicos incluem obter acesso à criança, obter a conformidade da criança e manter o sigilo da criança para evitar a divulgação. Esse processo serve para fortalecer o padrão abusivo do infrator, pois pode ser usado como meio de justificar ou negar suas ações (CRAVEN *et al.*, 2006, tradução nossa).[2]

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

[2] "*A process by which a person prepares a child, significant adults and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance and maintaining the child's secrecy to avoid disclosure. This process serves to strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions.*"

No ano seguinte, OLSON *et al.* (2007) definiram a "*Luring Communication Theory*" (LCT), que apresenta um modelo do processo de comunicação que predadores sexuais utilizam para enganar suas vítimas. Segundo os autores, predadores sexuais executam certo comportamento padrão para abordar suas vítimas.

O primeiro trabalho de classificação utilizando a LCT foi realizado por KONTOSTATHIS *et al.* (2009), que criaram um dicionário contendo 475 termos e frases para cada uma das seguintes categorias: atividades, abordagem, dessensibilização comunicativa, elogio, isolamento, informação pessoal, reenquadramento e relacionamento. Estas oito categorias fazem parte dos três estágios principais da LCT, que são:

1. Obter acesso à vítima;
2. Prender a vítima em um relacionamento enganoso;
3. Iniciar e manter um relacionamento sexualmente abusivo.

Também foi criado um manual de codificação com regras e instruções específicas para atribuir os termos e frases à cada categoria. Os autores ainda criaram um *software*, denominado *ChatCoder*, para automatizar este processo de codificação. O artigo informa, também, que foram criados guias para traduzir abreviações de internet em inglês padrão e para conversão de *emoticon* em texto. Segundo os autores, estas transformações não distorceram o significado das palavras e frases originais. A Tabela 2.4 apresenta alguns exemplos de conversas classificadas por KONTOSTATHIS *et al.* (2009) e suas respectivas categorias.

Com as frases já classificadas pelo *ChatCoder*, KONTOSTATHIS *et al.* (2009) buscaram resolver três problemas: distinguir entre predador e vítima, distinguir entre conversas predatórias e conversas gerais, e utilizar *cluster* para determinar os diferentes tipos de comunicação de predadores sexuais que existem.

Para o primeiro experimento, foram utilizadas 16 conversas retiradas do site PJ e o algoritmo árvore de decisão J48, que previu a classe corretamente 60% das vezes, segundo o artigo. Para o segundo experimento, foram utilizadas 15 conversas do site PJ e 14 conversas do projeto *ChatTrack*[3] com o algoritmo árvore de decisão C4.5 que acertou 93% das vezes. Por fim, os autores utilizaram apenas as mensagens enviadas por predadores sexuais de 288 conversas retiradas do site PJ para conseguir identificar os diferentes tipos de predadores sexuais que existem. Para isso, foi utilizado o algoritmo *k-means* para clusterizar as conversas e foi concluído que o número ideal de *clusters* era igual a quatro ($k = 4$), sugerindo que existem quatro tipos de

[3]Projeto da Dra. Susan Gauch, da Universidade do Arkansas, que coletava dados de salas de bate-papo. O *software ChatTrack* era o *crawler* utilizado para fazer *download* destas conversas (KONTOSTATHIS *et al.*, 2009).

Tabela 2.4: Exemplos de frases categorizadas por KONTOSTATHIS *et al.* (2009). Traduzido de KONTOSTATHIS *et al.* (2009).

Frase	**Categoria**
are you safe to meet	Abordagem
i just want to meet	Abordagem
i just want to meet and mess around	Abordagem
how cum	Dessensibilização comunicativa
if i don't cum right back	Dessensibilização comunicativa
i want to cum down there	Dessensibilização comunicativa
i just want to gobble you up	Dessensibilização comunicativa
you are a really cute girl	Elogio
you are a sweet girl	Elogio
are you alone	Isolamento
do you have many friends	Isolamento
let's have fun together	Reenquadramento
let's play a make believe game	Reenquadramento
there is nothing wrong with doing that	Reenquadramento

predadores sexuais na internet, e que cada um utiliza uma estratégia diferente para abordar suas vítimas (KONTOSTATHIS *et al.*, 2009).

2.2.2 PAN 2012

Em 2012, a competição do PAN[4] lançou o desafio *Sexual Predador Identification*. O desafio foi baseado em trabalhos anteriores como PENDAR (2007), MCGHEE *et al.* (2011) e KONTOSTATHIS *et al.* (2010) e consistiu na realização de duas tarefas:

1. Identificar quem são os predadores sexuais dentre todos os usuários;
2. Identificar quais são as linhas das conversas que indicam comportamento característico dos predadores sexuais.

Para realizarem as tarefas, os competidores receberam bases de dados volumosas em XML, que continham conversas na língua inglesa. Para a primeira tarefa foram fornecidos dados de treinamento e teste, e para a segunda tarefa foram fornecidos apenas dados de teste (INCHES e CRESTANI, 2012).

A falta de base de treinamento dificultou a segunda tarefa que acabou tendo resultados mais baixos. A maioria dos competidores acabou priorizando a primeira tarefa como se fosse um passo necessário para poder resolver a segunda. Os avaliadores da competição receberam 16 submissões para a primeira tarefa e 14 para a segunda (INCHES e CRESTANI, 2012).

[4]Competição que possui diversos desafios sobre detecção de plágio, *author identification* e questões morais, como os desafios *Sexual Predator Identification* (em 2012) e *Hyperpartisan News Detection* (em 2019). Disponível em: `https://pan.webis.de/`.

O resultado mais famoso da competição do PAN 2012 foi o de VILLATORO-TELLO *et al.* (2012), que conseguiu o melhor resultado para a primeira tarefa ($F_{0.5} = 93,46\%$). Sua solução consistia na criação de um classificador de dois estágios. Primeiramente, os dados passavam por uma etapa de pré-filtro, onde eram eliminadas conversas que tinham apenas 1 participante, conversas que tinham menos de 6 interações para cada usuário e conversas que continham longa sequência de caracteres não reconhecíveis. Depois, os dados eram submetidos ao classificador de Identificação de Conversas Suspeitas, onde as conversas eram classificadas em conversas suspeitas (envolvem um predador sexual) e conversas normais (não envolvem um predador sexual). Por fim, as conversas suspeitas eram submetidas ao último classificador, que identificava quem eram os predadores sexuais presentes nas conversas.

Por conta do excelente resultado, esta estratégia de classificador de dois estágios foi adotada por diversos trabalhos em anos posteriores, como CARDEI e REBEDEA (2017), KULSRUD (2019), BORJ *et al.* (2020) e FAUZI e BOURS (2020).

Outro ponto de impacto na área foi a base criada para o desafio. Segundo INCHES e CRESTANI (2012), a base possui milhares de conversas para que pudesse servir como um ponto de referência comum para pesquisadores de diferentes áreas. Ela foi criada com propriedades realísticas voltadas para o problema de identificação de predadores sexuais, ou seja, o número de verdadeiros positivos é muito baixo e o número de falsos positivos e falsos negativos é muito alto. Segundo os autores, em um cenário realista a porcentagem de conversas com predadores sexuais deve ser bem inferior a porcentagem de conversas normais (INCHES e CRESTANI, 2012).

Para garantir uma grande variedade de conversas, a base do PAN foi formada a partir de algumas fontes. O conjunto de dados verdadeiros positivos foi construído com as conversas provindas do site PJ, correspondendo ao segundo subtipo de conversas envolvendo predadores descrito por PENDAR (2007), as conversas entre predadores e voluntários se passando por crianças (INCHES e CRESTANI, 2012).

O conjunto de falsos positivos foi formado com conversas do site *Omegle*, que permite que estranhos possam conversar anonimamente *online*, e, consequentemente, tem potencial para conter conversas de conteúdo sexual entre adultos. Por fim, as conversas gerais, não relacionadas a conteúdos sexuais, foram formadas a partir de IRC *logs*, disponíveis publicamente na internet (INCHES e CRESTANI, 2012).

De fato, a base foi muito utilizada por diversos autores em anos posteriores ao desafio (BORJ e BOURS, 2019; CARDEI e REBEDEA, 2017; DOS SANTOS e GUEDES, 2018; EBRAHIMI *et al.*, 2016a,b; ESCALANTE *et al.*, 2013; KULSRUD, 2019; LIU *et al.*, 2017; MISRA *et al.*, 2019), tornando-se referência na literatura.

Um ponto fraco da competição foi a criação do *ground truth* para a segunda

tarefa. De acordo com INCHES e CRESTANI (2012), todas as submissões recebidas para a segunda tarefa foram avaliadas manualmente, e o *ground truth* foi gerado a partir destas submissões, conforme o que o especialista da competição considerou como um comportamento característico de predadores sexuais. Um dos motivos para isto ter ocorrido foi que não havia como rotular as conversas de forma distribuída, já que não haviam avaliadores treinados que estivessem aptos a isso. Assim, o trabalho de rotulagem ficou a cargo de apenas um especialista, o que, segundo os autores, pode conter certo grau de subjetividade.

2.2.3 Trabalhos mais Recentes

DOS SANTOS e GUEDES (2018) tiveram por objetivo analisar se existem traços de narcisismo nas mensagens enviadas por predadores sexuais às suas vítimas. Uma de suas premissas era que predadores sexuais tendem a ter traços de narcisismo e que existe uma correlação entre o uso de pronomes e o narcisismo, ou seja, narcisistas costumam utilizar uma mesma classe de pronomes. Assim, eles utilizaram a ferramenta *Linguistic Inquiry and Word Count* (LIWC) e a base do PAN 2012 para fazer esta análise. Por fim, elem concluíram que não existe correção entre narcisistas e o uso de pronomes na primeira pessoa do plural, nem pronomes na terceira pessoa.

Em um trabalho mais recente, FAUZI e BOURS (2020) desenvolveram um classificador de dois estágios que obteve melhores resultados que os melhores competidores do PAN. De forma similar a VILLATORO-TELLO *et al.* (2012), a solução contava com uma etapa de pré-filtro e duas etapas de classificação. A etapa de pré-filtro manteve apenas conversas que possuíam apenas duas pessoas envolvidas e conversas que tinham mais de 6 interações por usuário. O primeiro classificador era responsável por classificar as conversas em conversas normais e conversas suspeitas. As conversas suspeitas eram enviadas ao segundo classificador, que tinha por objetivo distinguir quem era o predador e quem era a vítima. Os autores fizeram experimentos com combinação de oito algoritmos e três *features* e dois *ensemble methods* e o melhor resultado foi de $F_{0.5} = 93,48\%$, obtido usando *soft voting ensemble method* para o primeiro classificador e *Naive Bayes* para o segundo classificador.

FAUZI e BOURS (2020) citaram como trabalhos futuros a utilização de *word embeddings*, como *Word2Vec* e BERT, como *features*, e a utilização de Redes Neurais Recorrentes e Redes de Memória de Longo Prazo. Ele também sugeriu que trabalhos futuros se preocupassem com a detecção precoce de predadores sexuais, ou seja, a partir da menor quantidades de mensagens possíveis, e com a aplicação da tarefa em plataformas de *chat* reais.

Outro trabalho recente que utiliza um classificador em dois estágios é o trabalho realizado por BORJ *et al.* (2020). O trabalho consiste em: uma etapa de pré-filtro,

onde são removidas as conversas que não possuíam apenas 2 autores e todas as conversas com menos de 7 mensagens; uma etapa de classificação, onde são distinguidas conversas predatórias e não predatórias; e uma última etapa de classificação, onde são identificados quem são os predadores e quem são as vítimas das conversas predatórias. Foram feitos experimentos utilizando *Bag of Words* e *GloVe* na etapa de extração de *features* e SVM, Floresta Aleatória e *Naive Bayes* na etapa de classificação. Os resultados foram levemente inferiores aos melhores resultados do PAN 2012.

2.2.4 Detecção Precoce de Predadores Sexuais

Provavelmente, o primeiro trabalho na área de detecção precoce de predadores sexuais foi realizado por ESCALANTE *et al.* (2017). Os autores propuseram o uso de *profile-based representations* para detecção precoce de predadores sexuais e identificação precoce de textos agressivos. Para a detecção de predadores, a base do PAN 2012 foi utilizada como conjunto de dados. Com 50% das mensagens lidas, os autores obtiveram aproximadamente 60% de F_1 com o algoritmo *Naive Bayes Multinomial* enquanto o modelo com Rede Neural e *Profile Specific Representation* obteve aproximadamente 80% de F_1.

O objetivo dos trabalhos relacionados à detecção precoce de predadores sexuais é identificar, o mais cedo possível, que existe um predador sexual em uma conversa. Não foi encontrado na literatura nenhum trabalho que defina quantas mensagens são realmente necessárias ou suficientes para identificar predadores sexuais em conversas. É possível que para diferentes conversas, sejam necessárias diferentes quantidades de mensagens.

Em um trabalho mais recente, KULSRUD (2019) utilizou a técnica de dois classificadores em sequência proposta por VILLATORO-TELLO *et al.* (2012) para detectar precocemente predadores sexuais. Neste trabalho, a base do PAN 2012 também foi utilizada como conjunto de dados.

Na etapa de pré-processamento, KULSRUD (2019) optou por remover *stopwords* e fazer algumas substituições de caracteres, deixando apenas letras, números, espaços em branco, #, + e _. Na etapa de pré-filtro, foram removidas conversas que não tivessem apenas 2 autores, conversas com menos de 6 iterações por usuário e excluídas mensagens vazias ou com mais de 8 caracteres que não fossem letras ou números.

Para os experimentos, KULSRUD (2019) utilizou o algoritmo SVM na primeira etapa do classificador, e os algoritmos Regressão Logística, *Ridge*, SVM, Rede Neural MLP e *Naive Bayes Bernoulli* na segunda etapa. O algoritmo *Naive Bayes* atingiu $F_{0.5} = 89,8\%$ que foi o maior valor de $F_{0.5}$ encontrado durante o experimento. Os

outros algoritmos alcançaram mais de 80% de $F_{0.5}$ após 24 mensagens e foram capazes de detectar 200 predadores com até 30 mensagens. Este trabalho é considerado o estado da arte do problema de detecção precoce de predadores sexuais. KULSRUD (2019) não forneceu matrizes de confusão ou outros dados que pudessem ser utilizados como comparação para esta dissertação.

2.2.5 Trabalhos com Dados Reais Brasileiros

Em 2017, foi criada a primeira base de dados em língua portuguesa com dados verídicos entre crianças e predadores sexuais brasileiros. O trabalho foi uma parceria do Centro Universitário FEI e o Ministério Público Federal de São Paulo (ANDRIJAUSKAS *et al.*, 2017). A base, que inicialmente continha 39 conversas predatórias e 137 conversas não predatórias, foi criada com estrutura semelhante a base do PAN 2012 e está disponível publicamente no GitHub[5].

Dois anos mais tarde, esta base foi utilizada por DOS SANTOS e GUEDES (2019) para classificar conversas em culpadas e não culpadas utilizando Redes Neurais Convolucionais. Na maioria dos experimentos foi obtido $F_{0.5} = 99\%$, com alguns chegando a $F_{0.5} = 100\%$. Possivelmente, este foi o primeiro trabalho de classificação utilizando dados de predadores sexuais brasileiros.

Em 2021, DOS SANTOS (2021) criou o Método de Detecção de Atividade Predatória (MDAP), que teve por objetivo identificar características textuais e comportamentais em conversas de texto para auxiliar a identificar atividade predatória brasileira. O método possui três módulos e se propôs a identificar 19 características textuais e comportamentais:

1. Módulo de Padronização do Conteúdo Textual Inicial: módulo inicial que teve por objetivo a normalização das conversas e remoção de ruídos;

2. Módulo de Identificação de Comportamento Predatório: módulo principal que teve por objetivo mapear as categorias de características das conversas predatórias e representar no formato de conceito de alto nível;

3. Módulo de Padronização do Conteúdo Textual Final: módulo final que teve por objetivo remover ruídos restantes do segundo módulo.

Para o conjunto de dados, DOS SANTOS (2021) utilizou os dados das conversas predatórias da base de dados criada por ANDRIJAUSKAS *et al.* (2017) e gerou dados para as conversas não predatórias. Inicialmente, ANDRIJAUSKAS *et al.* (2017) disponibilizou 39 conversas, e, posteriormente disponibilizou mais 43, totalizando 82

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

conversas predatórias brasileiras. Embora a base contenha 137 conversas brasileiras não predatórias, DOS SANTOS (2021) optou por não utilizá-las por observar que elas não continham assuntos da categoria adulta, e que, uma vez que algumas conversas foram transcritas do áudio, a estrutura das conversas poderia não ser semelhante à estrutura de conversas virtuais reais.

Assim, para criação do conjunto de dados das conversas não predatórias, DOS SANTOS (2021) optou por extrair mensagens das comunidades virtuais da plataforma *Discord*. Para obter uma grande diversidade de assuntos, foram extraídas 1.968 conversas das categorias jogos, política, tecnologia, estudos e adulto. A Tabela 2.5 apresenta a quantidade de conversas resultantes do conjunto de dados PRED-2050-ALL.

Tabela 2.5: Quantidade de conversas do conjunto de dados PRED-2050-ALL. Retirado de DOS SANTOS (2021).

Classe	**Conversas**
Predatória	82
Não predatória	1.968

Para identificação das características comuns e predatórias do módulo principal do MDAP foram utilizadas três estratégias. A primeira estratégia envolveu o uso de padrões textuais para identificar características como telefones, fotos, apelidos em redes sociais, perguntas e *emoticon*. A segunda envolveu a utilização de dicionários léxicos criados a partir de fontes internas e externas, para identificar características como cômodos da casa, cumprimentos virtuais, elogios predatórios, nomes próprios, e outros. E a terceira estratégia fez uso de padrões textuais e léxicos para identificar troca de idades entre predadores sexuais e vítimas e interesse em fotos e no local da vítima (DOS SANTOS, 2021).

Os léxicos de origem externa foram obtidos através de fontes externas, como sites, LIWC e IBGE, e eles são de conhecimento público. Já os léxicos de origem interna foram obtidos através das conversas predatórias, como o dicionário léxico de elogios predatórios. Ao todo foram criados 8 léxicos de origem externa e 5 léxicos de origem interna (DOS SANTOS, 2021).

Para a realização dos experimentos foram utilizados cinco algoritmos de aprendizado de máquina: SVM, *Naive Bayes Multinomial*, Árvore de Decisão, Floresta Aleatória e Rede Neural *Perceptron*. Com os experimentos, DOS SANTOS (2021) concluiu que a remoção dos termos raros possibilitou atingir melhores resultados e o algoritmo SVM apresentou o melhor resultado ao considerar as 100 características mais importantes e remover os termos raros ($F_{0.5} = 97,87\%$). Para finalizar, ele concluiu que o MDAP melhorou o desempenho dos algoritmos de aprendizado de

máquina e é uma alternativa válida e eficiente.

Como trabalhos futuros, DOS SANTOS (2021) sugeriu aumentar a quantidade de conversas predatórias da base de dados, aplicar o MDAP visando a detecção precoce de predadores sexuais e disponibilizar uma API para identificação de conversas predatórias para que possa ser utilizada por redes sociais e salas de bate-papo.

2.2.6 Problemas da Área

ESCALANTE *et al.* (2017) classificaram os problemas existentes na área em três categorias principais:

1. Identificar linhas de chat predatórias: para este problema os autores tiveram que associar cada mensagem a um estágio do *grooming*. Para ESCALANTE *et al.* (2017), existem três estágios principais: ganhar acesso, relacionamento enganoso e caso sexual;

2. Distinguir conversas predatória: para este problema os autores tiveram que distinguir entre conversas predatórias e não predatórias, como foi o caso de VILLATORO-TELLO *et al.* (2012), ou então distinguir conversas em mais categorias, como foi o caso de RAHMANMIAH *et al.* (2011), que classificou as conversas em casos de exploração infantil, fantasias sexuais e conversas gerais;

3. Distinguir entre ofensor e vítima: para este problema os autores tiveram que distinguir quem eram os autores vítimas e quem eram os autores predadores sexuais dentro de uma mesma conversa.

Baseado nas definições de ESCALANTE *et al.* (2017) e na revisão bibliográfica apresentada neste capítulo, foi elaborada a Tabela 2.6, com os principais problemas da área identificados.

Tabela 2.6: Tabela comparativa dos principais problemas da área e os trabalhos já realizados para cada problema.

Problema	**Descrição**	**Autores**
Identificar estágios do *grooming* em conversas predatórias	Identificar o estágio do *grooming* para cada mensagem da base de dados.	KONTOSTATHIS *et al.* (2009)
Distinguir entre conversas predatórias e outras conversas	Distinguir conversas entre conversas predatórias e outras conversas. As outras conversas podem se dividir em conversas gerais, conversas consensuais entre adultos e outros.	KONTOSTATHIS *et al.* (2009); BORJ e BOURS (2019); DOS SANTOS e GUEDES (2019); DOS SANTOS (2021)
Identificar mensagens características do comportamento dos predadores sexuais	Identificar quais mensagens indicam comportamento malicioso característico dos predadores sexuais.	VILLATORO-TELLO *et al.* (2012); GROZEA e POPESCU (2012); PEERSMAN *et al.* (2012)
Distinguir entre predador e vítima	Identificar quais autores são vítimas e quais autores são predadores sexuais.	PENDAR (2007); KONTOSTATHIS *et al.* (2009); VILLATORO-TELLO *et al.* (2012); GROZEA e POPESCU (2012); PEERSMAN *et al.* (2012); CARDEI e REBEDEA (2017); FAUZI e BOURS (2020); BORJ *et al.* (2020);
Detectar precocemente predadores sexuais	Detectar com poucas mensagens se uma conversa contém um predador sexual.	ESCALANTE *et al.* (2017); KULSRUD (2019)

Para o problema de identificar estágios do *grooming*, é importante considerar que a base só contém conversas predatórias, já que conversas gerais não se encaixam nos estágios do *grooming*. Este problema, que foi introduzido por KONTOSTATHIS *et al.* (2009), não é tão explorado na literatura pelos autores, possivelmente pela falta de base de dados que atenda a complexidade do problema, uma vez que os dados precisariam ser classificados por especialistas a fim de evitar subjetividade nos rótulos. Como informado ao longo da seção, existe uma grande falta de dados reais para se trabalhar com predadores sexuais, já que envolvem dados sigilosos e questões legais.

Os problemas de distinguir entre conversas predatórias e outras conversas e distinguir entre predador e vítima têm sido os problemas mais buscados pelos autores, conforme é possível observar na Tabela 2.6. Geralmente, os problemas são utilizados em conjunto, onde primeiro problema é uma etapa para realizar o segundo (BORJ *et al.*, 2020; CARDEI e REBEDEA, 2017; FAUZI e BOURS, 2020; VILLATORO-TELLO *et al.*, 2012).

O problema de identificação de mensagens características do comportamento dos predadores sexuais proposto pela competição do PAN 2012 não tem sido muito explorado na literatura. Um dos possíveis motivos é a falta de uma base de dados adequada com dados reais suficientes ou classificados por especialistas. Afinal, o que é uma mensagem característica do comportamento dos predadores sexuais? É um problema difícil de ser resolvido e sem um especialista treinado para realizar esta tarefa, os rótulos ficam subjetivos, que foi o que INCHES e CRESTANI (2012) relataram que aconteceu com os rótulos da segunda tarefa da competição do PAN 2012.

Por fim, o problema de detectar precocemente predadores sexuais é um problema relativamente novo e ainda não muito explorado na literatura, tendo seu primeiro trabalho realizado por ESCALANTE *et al.* (2017). Apesar de novo, é um caminho natural para os trabalhos já existentes, já que é interessante que se descubra o mais cedo possível que existe um predador sexual em uma conversa. Como dito anteriormente nesta seção, predadores sexuais executam certo comportamento padrão para abordar suas vítimas antes de conseguir contato sexual (OLSON *et al.*, 2007). Desta forma, é possível inferir que identificando precocemente que existe um predador sexual em uma conversa, é possível evitar o encontro físico do predador com a vítima, e assim, evitar o abuso sexual.

Capítulo 3

Proposta

Tendo em vista os pontos apresentados no capítulo anterior, este capítulo apresenta a proposta desta dissertação.

A primeira seção apresenta a definição da proposta, informando as premissas do trabalho e quais as estratégias desenvolvidas. A segunda seção apresenta quais experimentos foram realizados e os algoritmos escolhidos. Por fim, são descritas as três estratégias deste trabalho, cada uma em uma seção.

3.1 Definição da Proposta

Conforme relatado no capítulo 2, o problema de detecção precoce de predadores sexuais é um problema relativamente novo, tendo poucos trabalhos relacionados.

Seu desenvolvimento contribui diretamente para a sociedade, uma vez que pode auxiliar a reduzir o número de casos de abuso sexual infantil. Esta premissa pode ser inferida a partir do trabalho de OLSON *et al.* (2007), que afirma que predadores sexuais executam certo comportamento padrão para abordar suas vítimas antes de conseguir contato sexual. O objetivo é conseguir identificar este comportamento padrão antes que o encontro físico ocorra, e portanto, o contato sexual.

Assim, compreendendo que é relevante identificar o mais cedo possível que existe um predador sexual em uma conversa, este trabalho propõe o desenvolvimento de três estratégias distintas para detectar precocemente predadores sexuais em conversas realizadas na internet entre duas pessoas. São elas:

1. Distinguir Conversas Predatórias e Gerais;
2. Distinguir Predador e Vítima;
3. Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima;

Estas três estratégias foram definidas a partir dos problemas da área identificados na Tabela 2.6 e já foram implementadas de diferentes formas por diversos autores na

literatura, como descrito no capítulo 2. Porém, elas foram utilizadas para resolverem seus problemas correspondentes, voltadas para a identificação de predadores sexuais, que considera a conversa completa, não tendo sido exploradas ainda para o problema de detecção precoce de predadores sexuais, que considera partes das conversas, de forma a avaliar o desempenho dos algoritmos com poucas mensagens.

Até a conclusão da revisão bibliográfica, KULSRUD (2019) foi o único trabalho encontrado que utilizou uma das estratégias para detecção precoce. Ele utilizou a estratégia do classificador de dois estágios, e com a maioria dos algoritmos obteve mais de 80% de $F_{0.5}$ após 24 mensagens.

Optou-se por desenvolver três estratégias já conhecidas na literatura para comparar e avaliar se para o problema de detecção precoce de predadores sexuais estratégias mais simples, que envolvam apenas um classificador, podem obter melhores resultados do que estratégias mais complexas, como a estratégia dos dois classificadores em sequência.

Em particular, presume-se que a estratégia "Distinguir Conversas Predatórias e Gerais" consiga melhores resultados que a estratégia "Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima", por não haver necessidade de identificar quem são os autores predadores sexuais, uma vez que já foi identificado que trata-se de uma conversa predatória.

3.2 Definição dos Experimentos

Este é um problema de classificação de textos com classes binárias. Para os experimentos, propõe-se a utilização dos seguintes algoritmos de classificação de textos: *Naive Bayes Multinomial*, KNN, Floresta Aleatória, SVM, Rede Neural MLP e BERT.

O algoritmo KNN obteve o melhor resultado no trabalho de PENDAR (2007) e depois foi esquecido na literatura. Propõe-se avaliar o desempenho deste algoritmo para o problema de detecção precoce de predadores sexuais. Por outro lado, o BERT ainda não foi explorado na literatura para este problema.

Para o conjunto de dados, optou-se por utilizar a base do desafio *Sexual Predator Identification*[1] da competição do PAN 2012, apresentada no capítulo 2. Esta escolha foi baseada na grande quantidade e diversidade de conversas presente na base e sua grande aceitação na literatura.

Como mencionado também no capítulo 2, a base contém um número de verdadeiros positivos muito baixo e o número de falsos positivos e falsos negativos muito alto. Isso porque, em um cenário realista, a porcentagem de conversas com predadores sexuais deve ser bem inferior a porcentagem de conversas normais (INCHES e

[1] `https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html`

CRESTANI, 2012). Observa-se, portanto, que este é um problema naturalmente de classes desbalanceadas. Dessa forma, para cada estratégia propõe-se experimentos sem balanceamento dos dados e com utilização de técnicas de balanceamento de classes, especificamente, técnicas de *undersampling*.

Apesar de ser um problema originalmente de classes desbalanceadas, alguns trabalhos encontrados na literatura não utilizaram técnicas de balanceamento. Isto deve-se ao fato de removerem conversas com menos de 6 mensagens por usuário na fase de pré-filtro (KULSRUD, 2019; VILLATORO-TELLO *et al.*, 2012). Esta ação garantiu uma redução de mais de 80% no número de conversas (KULSRUD, 2019), o que permitiu bons resultados mesmo sem uso de técnicas de balanceamento.

Como o foco deste trabalho é a detecção precoce, não achou-se interessante implementar esta remoção. Assim, optou-se por avaliar os resultados com e sem técnicas de balanceamento.

O trabalho de KULSRUD (2019), considerado o estado da arte e utilizado como comparação nesta dissertação, implementou a remoção de conversas com menos de 6 mensagens por usuário no pré-filtro. Assim, é importante informar que apesar de utilizar a mesma base de dados utilizada por KULSRUD (2019), a quantidade de dados utilizados nos experimentos é diferente. Para exemplificar, após execução das etapas de pré-filtro desta dissertação, a base de treinamento do PAN 2012 que tinha originalmente 66.927 conversas passou para 66.702 conversas. No trabalho de KULSRUD (2019), restaram apenas 8.692 conversas após execução do pré-filtro.

Para o desenvolvimento das estratégias foi necessário adaptar a base do PAN 2012, agrupando as mensagens de acordo com a abordagem individual de cada estratégia. Por exemplo, no caso da estratégia 1 as mensagens foram agrupados considerando as conversas, e para a estratégia 2 as mensagens foram agrupadas considerando os autores.

É importante frisar que o conjunto de verdadeiros positivos da base do PAN 2012 foi formado a partir de conversas do site PJ, que possui conversas entre predadores sexuais condenados e voluntários se passando por crianças. Portanto, não são dados entre predadores e vítimas reais.

3.3 Estratégia 1 - Distinguir Conversas Predatórias e Gerais

A primeira estratégia, denominada "Distinguir Conversas Predatórias e Gerais", utiliza a abordagem de identificar quais são as conversas predatórias dentre todas as conversas da base de dados.

Para implementação desta estratégia, as mensagens da base do PAN 2012 foram

agrupadas por conversas. Ou seja, cada tupla da base de dados corresponde a todas as mensagens enviadas por todos os autores em uma mesma conversa. É um problema de classificação binário, onde as classes possíveis são conversas predatórias e conversas não predatórias.

A base do PAN 2012 fornece apenas quem são os autores predadores sexuais. Portanto, para o desenvolvimento desta estratégia são consideradas conversas predatórias as conversas que um dos autores é um predador sexual.

3.4 Estratégia 2 - Distinguir Predador e Vítima

A segunda estratégia, denominada "Distinguir Predador e Vítima", utiliza a abordagem de identificar quem são os autores predadores sexuais e quem são os autores vítimas de cada conversa.

Para implementação da estratégia, as mensagens da base do PAN 2012 foram agrupadas por autores em uma mesma conversa. Ou seja, cada tupla da base corresponde a todas as mensagens enviadas por um autor dentro de uma conversa.

Também é um problema de classificação binário, onde as classes possíveis são predadores ou vítimas.

3.5 Estratégia 3 - Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima

A terceira estratégia, denominada "Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima", é o modelo de dois classificadores proposto pela primeira vez por VILLATORO-TELLO *et al.* (2012).

Para implementação desta estratégia, as duas estratégias anteriores são combinadas em sequência. Primeiramente, as conversas são submetidas ao primeiro classificador (Distinguir Conversas Predatórias e Gerais) para que se possa obter as conversas predatórias. Em seguida, apenas as conversas classificadas como predatórias são submetidas ao segundo classificador (Distinguir Predador e Vítima), para que se possa obter os autores predadores sexuais.

Esta estratégia considera que para identificar corretamente qual autor é o predador e qual é a vítima, primeiro é preciso reconhecer quais são as conversas predatórias dentre todas as conversas. Ao mesmo tempo, espera-se que o primeiro classificador tenha um ótimo desempenho, evitando repassar ao segundo classificador conversas que não sejam predatórias.

Capítulo 4

Experimentos

Este capítulo descreve os experimentos realizados utilizando as estratégias descritas no capítulo anterior. O código desenvolvido para esta dissertação está disponível publicamente no GitHub[1].

A primeira seção apresenta o ambiente utilizado para o desenvolvimento do código. A segunda seção apresenta a base do PAN 2012. A terceira e a quarta seções apresentam, respectivamente, as etapas realizadas de pré-processamento e pré-filtro para as bases de treinamento e teste. Os modelos de extração dos dados utilizados são introduzidos na quinta seção. A sexta seção explica como foram realizados os experimentos e quais as configurações utilizadas para cada algoritmo.

Nas seções seguintes, são detalhados os resultados dos experimentos sem balanceamento dos dados e com uso da técnica *undersampling* para cada estratégia. Por fim, a última seção encerra o capítulo com as considerações finais sobre os resultados dos experimentos.

4.1 Ambiente

O código desenvolvido para esta dissertação foi escrito em *Python* 3 e utilizou o ambiente *Google Colaboratory*[2]. Este ambiente, produto da *Google Research*, permite a escrita e execução de códigos em *python* através da criação de *notebooks Jupyter*, sem necessidade de configuração para uso.

Para execução dos experimentos, utilizou-se o ambiente *Collab Pro* que forneceu *notebooks* de 12.68GB ou 25.46GB de memória RAM com 225.89GB de espaço em disco. Para geração dos embeddings do BERT e alguns experimentos, foi utilizado GPU Tesla P100-PCIE-16GB ou Tesla T4.

[1] `https://github.com/marcellepanzariello/early-detection-sexual-predators`
[2] `https://research.google.com/colaboratory/`

4.2 Base do PAN 2012

Como informado na proposta desta dissertação, no capítulo 3, a base do PAN 2012 foi escolhida como conjunto de dados.

O desafio *Sexual Predator Identification* do PAN 2012 forneceu bases de treinamento e teste em XML contendo milhares de conversas entre um ou mais autores. Cada conversa é identificada por um identificador único que possui uma ou mais mensagens, cada uma contendo identificação do autor que enviou a mensagem, horário e texto da mensagem. Na Figura 4.1, é possível visualizar um exemplo de conversa da base de treinamento do PAN 2012 com todos os atributos informados.

```
<conversation id="93667634acded68f53ede6e2dadac950">
  <message line="1">
    <author>fa0151a7b675a6740e09c3e7c6f46d33</author>
    <time>00:04</time>
    <text>hi</text>
  </message>
  <message line="2">
    <author>c8c577c2298b16bcd6c04324f60f06cb</author>
    <time>00:04</time>
    <text>hi</text>
  </message>
  <message line="3">
    <author>c8c577c2298b16bcd6c04324f60f06cb</author>
    <time>00:04</time>
    <text>m or f ?</text>
  </message>
  <message line="4">
    <author>fa0151a7b675a6740e09c3e7c6f46d33</author>
    <time>00:04</time>
    <text>...</text>
  </message>
</conversation>
```

Figura 4.1: Exemplo de conversa da base de treinamento do PAN 2012.

Na tabela 4.1 é possível observar a quantidade de conversas, mensagens e autores das bases de treinamento e teste do PAN 2012.

Tabela 4.1: Quantidade de dados das bases de treinamento e teste do PAN 2012.

	Treinamento	**Teste**
Conversas	66.927	155.128
Mensagens	903.607	2.058.781
Autores	97.689	218.702
Conversas predatórias	2.016	3.737
Mensagens predatórias	40.978	65.239
Predadores sexuais	142	254

4.3 Pré-Processamento

Alguns trabalhos encontrados na literatura não realizaram nenhuma etapa de pré-processamento, como VILLATORO-TELLO *et al.* (2012). Outros, no entanto, como KULSRUD (2019) e DOS SANTOS (2021), optaram por substituir alguns caracteres e remover *stopwords*.

Para os experimentos deste trabalho, foram escolhidas as 9 etapas listadas a seguir:

1. Excluir mensagens nulas;
2. Converter mensagens para minúsculo;
3. Remover entidades de caracteres HTML e `\n`, `\r` e `\t`;
4. Remover URLs;
5. Remover pontuação, letras desconhecidas e com acentos;
6. Remover espaços em branco em sequência;
7. Excluir mensagens vazias;
8. Remover letras repetidas em uma palavra;
9. Remover palavras repetidas em sequência;

Na primeira etapa (1), foram removidas 2.978 mensagens nulas da base de treinamento, gerando uma redução de 13 conversas e 18 autores. Destas, 1 conversa era considerada uma conversa predatória. Das mensagens excluídas, 249 eram mensagens enviadas por predadores sexuais.

Na Tabela 4.2 é possível observar a quantidade de dados antes e após a exclusão de mensagens nulas das bases de treinamento e teste. Como informado anteriormente, são consideradas conversas predatórias as conversas em que um dos autores é um predador sexual. As mensagens predatórias são todas as mensagens enviadas por predadores sexuais.

Em seguida, todas as mensagens foram convertidas para minúsculo para padronização (2).

A terceira etapa do pré-processamento consistiu em remover as entidades de caracteres HTML e `\n`, `\r` e `\t` encontradas na base (3). Nas conversas foram encontradas muitas entidades, como pode ser observado no exemplo de conversa inserido na Figura 4.2.

Tabela 4.2: Dados das bases de treinamento e teste após execução de (1).

Base	Dados	Original	(1)
Treinamento	Conversas	66.927	66.914
	Mensagens	903.607	900.629
	Autores	97.689	97.671
	Conversas predatórias	2.016	2.015
	Mensagens predatórias	40.978	40.729
	Predadores sexuais	142	142
Teste	Conversas	155.128	155.101
	Mensagens	2.058.781	2.052.320
	Autores	218.702	218.681
	Conversas predatórias	3.737	3.723
	Mensagens predatórias	65.239	64.982
	Predadores sexuais	254	254

message	original_message
its nice, isnt it?	It's nice, isn't it?
yeah. always enjoy visiting.	Yeah. Always enjoy visiting.
which island you on oahu?	Which Island you on Oahu?
married?	married?
im assuming since you went on a family trip :p	i'm assuming since you went on a 'family' trip :p
yeah. just found this site a few days ago.	yeah. Just found this site a few days ago.
yeah, oahu.	Yeah, Oahu.
curious to the whole random thing	Curious to the whole "random thing"
pretty crazy the individuals you meet, isnt it?	Pretty crazy the individuals you meet, isn't it?
its been eye opening for sure.	It's been eye opening for sure.

Figura 4.2: Trecho de conversa que continha entidades de caracteres HTML. As mensagens originais estão à direita, e as mensagens alteradas por (3) estão à esquerda.

Depois, optou-se por remover URLs (4). Um possível trabalho futuro é identificar se é possível extrair algum significado das URLs, correlacionando-as com as mensagens predatórias.

Optou-se também por excluir as pontuações. Como a base possui conversas que foram trocadas em *chats online*, muitas mensagens são apenas pontuações, representando *emoticons* ou desenhos, ou mesmo sequências de pontuações, como interroga-

ções. Assim, a quinta etapa do pré-processamento consistiu em excluir pontuações, letras desconhecidas e com acentos para melhor limpeza dos textos (5). Na Figura 4.3 é possível observar um exemplo de mensagem que possuía pontuações em excesso e a mensagem gerada por (5).

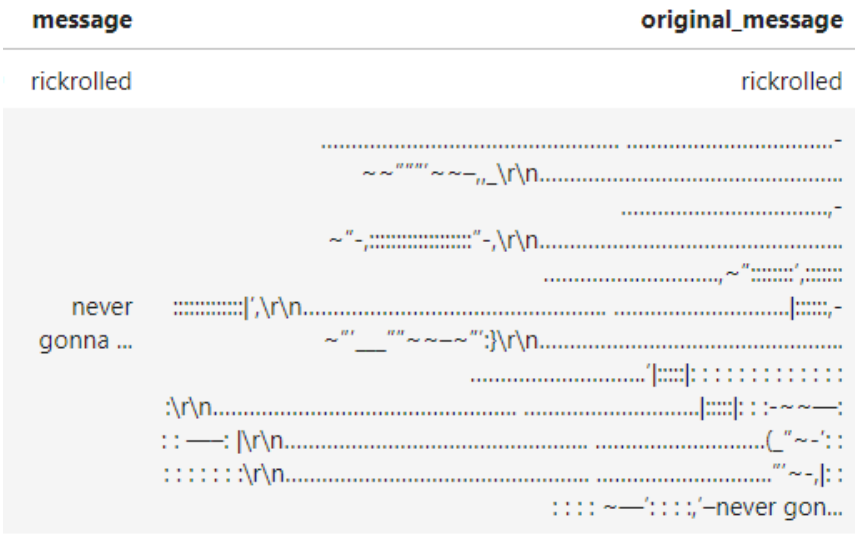


Figura 4.3: Exemplo de mensagem que possuía muitas pontuações. As mensagens originais estão à direita, e as mensagens alteradas por (5) estão à esquerda.

Outro trecho de uma conversa que sofreu alterações de (5) pode ser observado na Figura 4.4. O trecho possuía símbolos e letras desconhecidas. À direita estão as mensagens originais e à esquerda as mensagens alteradas.

Após a conclusão das etapas anteriores, notou-se que muitas mensagens ficaram com espaços em branco em excesso. Assim, foram removidos também (6).

O passo seguinte foi buscar e excluir as mensagens que ficaram vazias por conta da execução das etapas anteriores (7). Ao todo, foram excluídas 27.427 mensagens vazias da base de treinamento, gerando uma redução de 212 conversas e 384 autores, em relação à (1). Também foram perdidas 43 conversas predatórias e 931 mensagens predatórias. Nenhum predador sexual foi perdido. Na Tabela 4.3 pode-se observar em detalhes a quantidade de dados restantes nas bases de treinamento e teste após execução de (7).

A maioria das mensagens enviadas por predadores sexuais que foram excluídas representavam *emojis*.

Ao longo do desenvolvimento do pré-processamento, percebeu-se também que muitas mensagens estavam escritas incorretamente ou com letras repetidas, como, por exemplo "*helloooo*" e "*ffffaaakkkkerrrr*". Para diminuir a variedade de palavras com grafia errada da base, optou-se por remover as letras repetidas, mantendo apenas duas repetições por letra (8), já que na língua inglesa existem palavras com duas repetições, como "*cool*" e "*weekend*".

message	original_message
	ÔÓ ÚÓ˜ÌÓ ÛÚ?Â ?Â ?‡Á·Â?‡
lord sithius blagodarsko	Lord_Sithius blagodarsko
luda glawa edno diablo	Luda_Glawa: edno diablo ?
deathmaster	[Deathmaster], ÒÚ‡,‡
	Ó·‡˜Â ‰‡È ‰‡ ÒÂ Ó·Á‡ÎÓÊËÏ Ô?Â‰Ë ÚÓ,‡
	??
	ÒÔÓ?Â‰ ÚÂ· ÒÎÂ‰ ÍÓÎÍÓ ,?ÂÏÂ ?Â Á‡·ËÂ Ò??,?? ‡?
	:-)))
20	20 ÏËÌÛÚË?
	Ì?Ï‡ ‰‡ Á‡·ËÂ

Figura 4.4: Trecho de conversa que possuía símbolos e letras desconhecidas. As mensagens originais estão à direita, e as mensagens alteradas por (5) estão à esquerda.

Tabela 4.3: Dados da base de treinamento após execução de (7).

Base	Dados	Original	(1)	(7)
Treinamento	Conversas	66.927	66.914	66.702
	Mensagens	903.607	900.629	873.202
	Autores	97.689	97.671	97.287
	Conversas predatórias	2.016	2.015	1.972
	Mensagens predatórias	40.978	40.729	39.798
	Predadores sexuais	142	142	142
Teste	Conversas	155.128	155.101	154.732
	Mensagens	2.058.781	2.052.320	1.990.556
	Autores	218.702	218.681	217.851
	Conversas predatórias	3.737	3.723	3.690
	Mensagens predatórias	65.239	64.982	63.327
	Predadores sexuais	254	254	254

Na Figura 4.5 pode-se observar um trecho de conversa com mensagens que continham letras repetidas.

Notou-se também que algumas mensagens possuíam o mesmo texto repetido inúmeras vezes em sequência, que acabava aumentando drasticamente a quantidade de palavras por mensagens. Optou-se então por remover palavras repetidas em sequência (9) para diminuir a quantidade de texto repetido desnecessário. A Figura 4.6 apresenta um trecho contendo mensagens com palavras repetidas.

message	original_message
wee	weeeeee
im happy	im happy
ru	ru?
hlo	hlo?
ru there	ru there?
i gotta p brb	i gotta p brb
kay bak	kay bak
heelloo	heeellllooooooooooooo
shawn	shawn
wered u go	wered u go?
ur messin with me arnt u	ur messin with me arnt u?

Figura 4.5: Trecho de conversa que continha mensagens com letras repetidas. As mensagens originais estão à direita, e as mensagens alteradas por (8) estão à esquerda.

message	original_message
hey	hey
asl	asl?
gingers do have souls	GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HAVE SOULS! GINGERS DO HA...

Figura 4.6: Trecho de conversa que continha mensagens com palavras repetidas. As mensagens originais estão à direita, e as mensagens alteradas por (9) estão à esquerda.

Com a execução de (9), a quantidade de palavras por mensagem reduziu drasticamente. Para exemplificar, a mensagem que possuía mais palavras antes do pré-processamento na base de treinamento reduziu de 55.855 palavras para apenas 47 palavras após o pré-processamento. Outro exemplo, foi uma mensagem que possuía 8.397 palavras antes do pré-processamento e passou para apenas 4 palavras após

execução de (9).

Após a conclusão de todas as etapas do pré-processamento, a mensagem que contém mais palavras no conjunto de treinamento possui 2.000 palavras, uma diferença significativa.

Por fim, a Figura 4.7 ilustra uma *wordcloud* com as palavras mais frequentes enviadas por predadores sexuais da base de treinamento. Na imagem é possível visualizar palavras como "*want*", "*love*", "*sorry*", "*sex*", "*girl*" e "*kiss*".

Figura 4.7: *Wordcloud* das mensagens predatórias da base de treinamento.

4.4 Pré-Filtro

Para a etapa de pré-filtro, foram apenas excluídas as conversas que não continham somente 2 autores. Conforme apresentado no capítulo anterior, algumas conversas da base do PAN 2012 possuíam apenas 1 participante e outras possuíam muitos participantes, chegando até 30 autores em uma mesma conversa.

Esta estratégia já foi utilizada por KULSRUD (2019) em sua dissertação, e VILLATORO-TELLO *et al.* (2012) implementaram uma estratégia semelhante excluindo as conversas com apenas 1 participante.

É importante informar que para esta etapa a quantidade de conversas excluídas é muito variável. Isto acontece pois a quantidade participantes nas conversas varia em relação a quantidade de mensagens enviadas. Por exemplo, uma conversa que contém 4 participantes no total pode ter as 10 primeiras mensagens enviadas por apenas 1 ou 2 participantes. O quarto participante pode enviar sua primeira mensagem após 50 outras já terem sido enviadas.

Como o objetivo é detectar os predadores sexuais precocemente, os experimentos são realizados com partes das conversas. Ou seja, são feitos experimentos com 5 mensagens, 6 mensagens, 7 mensagens, etc. Portanto, foram excluídas todas as conversas que tivessem menos ou mais de 2 participantes considerando o limite da quantidade de mensagens enviadas. Assim, em um experimento com as primeiras 5 mensagens, uma conversa pode ser sido excluída, e, em um experimento com as primeiras 30 mensagens, a mesma conversa pode não ter sido excluída.

Optou-se por deixar este filtro dinâmico para garantir que só iriam para o classificador conversas que tivessem exatamente 2 autores dentro do limite de mensagens enviadas.

4.5 Extração de Dados

Com exceção dos experimentos realizados com o algoritmo BERT, todos os experimentos utilizaram o modelo *Bag of Words* na etapa de extração de dados.

Para os experimentos com o BERT, foram utilizados os próprios *embeddings* gerados pelo BERT, utilizando o modelo pré-treinado *bert-base-uncased* (DEVLIN *et al.*, 2018).

4.6 Configuração dos Experimentos

Como o desafio do PAN 2012 forneceu bases de treinamento e teste separadas, foi possível realizar o treinamento com a base de treinamento completa e a base de teste foi utilizada apenas para os testes finais.

Para cada uma das três estratégias houve dois experimentos: com a base desbalanceada e utilizando a técnica *undersampling*. Ambos os experimentos foram executados nas bases de treinamento e teste.

A fim de evitar *overfitting* durante o treinamento e garantir que a base de testes ficaria separada para os testes finais, optou-se por utilizar a técnica de validação cruzada *k-fold cross validation* durante o treinamento. Esta técnica consiste em dividir a base em k partes, denominadas *folds*, onde os dados são treinados k vezes com $k-1$ *folds* e a parte restante é usada para validação, e os conjuntos de treino e validação variam a cada iteração. Isto permite que toda a base seja utilizada para

treinamento sem deixar que os dados sejam treinados e validados na mesma porção de dados. Os resultados finais da validação cruzada são a média dos valores obtidos em cada iteração (SCIKIT-LEARN, 2022). Para os experimentos deste trabalho, foram utilizados 10 *folds* ($k = 10$).

Os testes foram feitos utilizando os melhores algoritmos de cada experimento. Os dados foram retreinados na base de treinamento, desta vez sem validação cruzada e utilizando todos os dados, e testados na base de testes do PAN 2012.

4.6.1 Configurações dos Algoritmos

Todas as três estratégias utilizaram os mesmos algoritmos: *Naive Bayes Multinomial*, KNN, Floresta Aleatória, SVM, Rede Neural MLP e BERT.

Os algoritmos *Naive Bayes Multinomial*, KNN, Floresta Aleatória, SVM e Rede Neural MLP utilizaram implementações padrões da biblioteca *scikit-learn*[3]. A implementação utilizada para o BERT foi adaptada do CICIAR, projeto da Defensoria Pública do Estado do Rio de Janeiro, que utiliza o algoritmo BERT para classificar intimações (PARREIRAS *et al.*, 2022).

Para todos os experimentos foram utilizados os parâmetros descritos na Tabela 4.4, que foram obtidos depois de diversos de testes iniciais em busca dos melhores parâmetros realizados com alguns tamanhos de mensagens, como 24, 30, 50 e 100.

Para o KNN, foram realizados testes variando a quantidade de vizinhos, mas em todos os testes os resultados apresentaram queda após 5 vizinhos.

Para a Floresta Aleatória, foram realizados testes variando a quantidade de árvores (100, 150, 200, 250, 300), critério (*gini*, *entropy*) e profundidade da árvore (*None*, 5, 10, 20, 30, 50).

Para o SVM, foram realizados testes com diferentes *kernels* (*linear*, *poly*, *rbf*, *sigmoid*).

Para a Rede Neural MLP, foram realizados testes com 2, 3 e 4 camadas ocultas de 50 e 100 neurônios. Também foram realizados testes com o parâmetro *early_stopping*, *validation_fraction* (0,1 e 0,2) e quantidade de épocas (10, 30, 50, 100, 200).

Por fim, para o BERT foram realizados testes variando a quantidade de épocas (20, 50, 80, 100) e a taxa de *learning rate* ($1e-3$, $1e-4$).

[3] `https://scikit-learn.org/stable/index.html`

Tabela 4.4: Melhores parâmetros encontrados para os algoritmos.

Classificador	**Parâmetros**
Naive Bayes Multinomial	
KNN	$n\_neighbors = 5$
Floresta Aleatória	$n\_estimators = 150$ $max\_depth = None$
SVM	$kernel = linear$
Rede Neural MLP	$hidden\_layer\_sizes = (100, 100, 100,)$ $activation = relu$ $solver = adam$ $max\_iter = 50$ $early\_stopping = True$ $validation\_fraction = 0.2$
BERT	$epochs = 50$ $activation = relu$ $optimizer = adam$ $learning\_rate = 1e - 4$

4.7 Estratégia 1 - Distinguir Conversas Predatórias e Gerais

Para estruturar os dados para a primeira estratégia, foi necessário agrupar as mensagens já pré-processadas por conversas, de tal modo em que cada tupla correspondesse a uma conversa inteira, já que a finalidade da estratégia é classificar conversas em predatórias e não predatórias, conforme citado na proposta, no capítulo 3. A classe alvo foi chamada de "*predatory_conversation*" e possuía valores "*True*" ou "*False*" para cada tupla/conversa.

Como o objetivo da dissertação é a detecção precoce de predadores sexuais, os experimentos não foram realizados com as conversas inteiras, mas com partes das conversas. Por conta disso, este agrupamento de mensagens levou em consideração a quantidade de mensagens por conversas.

Depois de agrupar as mensagens, foram removidas as conversas que não tivessem apenas 2 autores após o agrupamento. Conforme explicado na seção 4.4 deste capítulo, a quantidade de conversas removidas leva em consideração a quantidade de mensagens por conversas utilizada no agrupamento.

Por exemplo, em um experimento com as primeiras 10 mensagens das conversas, foram identificadas 20.176 conversas que não continham apenas 2 autores, e em um experimento com as primeiras 30 mensagens das conversas, foram identificadas 21.218 conversas que não continham apenas 2 autores. Este aumento ocorreu porque alguns autores entraram na conversa após as 10 primeiras mensagens.

Foram realizados experimentos com e sem técnicas de balanceamento, que são detalhados a seguir.

4.7.1 Sem Balanceamento dos Dados

Para esta estratégia, cada tupla correspondia a uma conversa inteira, evidenciando o desbalanceamento entre as classes, conforme pode ser observado na Figura 4.8, que mostra a quantidade de classes positivas e negativas da base de treinamento após agrupamento e remoção de conversas que não continham apenas dois autores com as primeiras 50 mensagens das conversas, por exemplo. Nesta situação, existem 44.316 conversas não predatórias para 1.085 conversas predatórias.

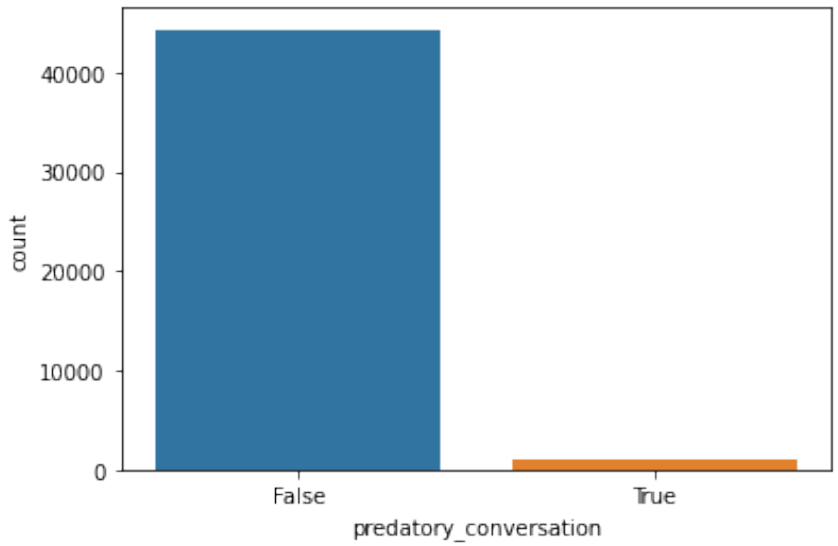


Figura 4.8: Quantidade de valores positivos e negativos para a classe alvo utilizando as primeiras 50 mensagens.

Com a finalidade de mostrar o desempenho dos algoritmos com diferentes tamanhos de conversas, foram gerados gráficos com valores crescentes de quantidades de mensagens por conversas. O intuito foi avaliar como os algoritmos se comportariam lendo partes das conversas, como se estivessem lendo-as em tempo real.

Os gráficos consideraram um intervalo de 2 a 100 mensagens por conversa. Foram necessárias ao menos duas mensagens para que houvesse a possibilidade de existirem dois autores, por conta do pré-filtro que remove conversas que não tenham apenas dois autores. Com apenas uma mensagem só tem como existir necessariamente um autor. O limite de 100 mensagens foi colocado pois percebeu-se que a partir de 100 mensagens a maioria dos algoritmos já estabilizavam os resultados. Além disso, a maioria das conversas das bases de treinamento e teste têm até 100 mensagens e seria mais interessante mostrar o desempenho dos algoritmos logo nas mensagens iniciais.

Como informado anteriormente, os experimentos foram realizados com partes das conversas, e não houveram experimentos com as conversas inteiras. Cada quantidade de mensagem do experimento gerou um modelo diferente. No total, foram gerados 99 modelos (2 a 100 mensagens) para cada algoritmo de cada experimento.

Treinamento

Para este primeiro experimento, a acurácia, apresentada na Figura 4.9, foi a única métrica que basicamente não apresentou alterações entre os algoritmos e manteve-se sempre alta, no geral, acima de 97%. Isto ocorreu por conta da grande quantidade de classes negativas na base de dados que já garantem acurácia elevada, independentemente da quantidade de acertos da classe positiva.

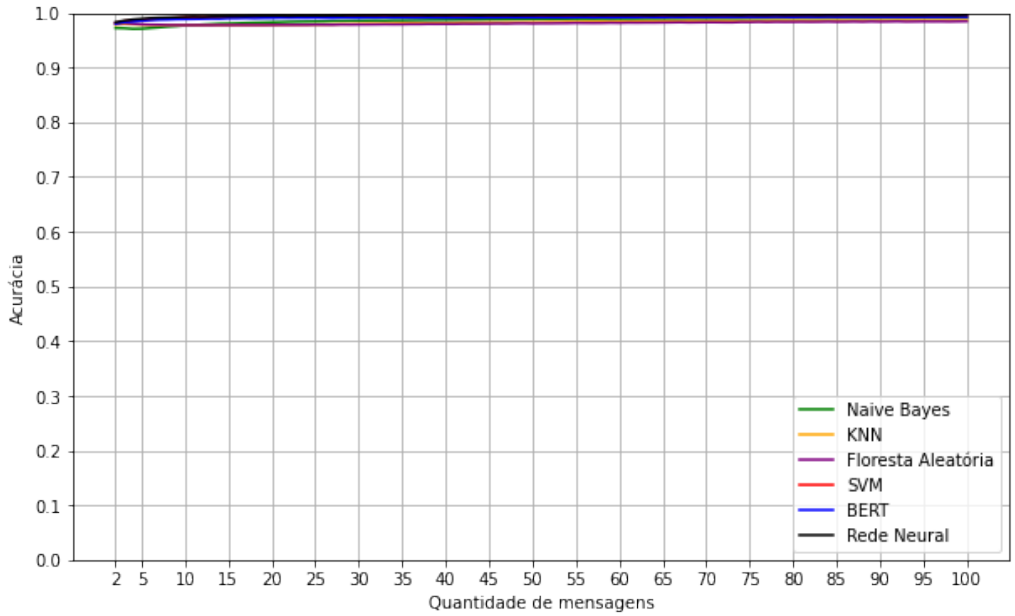


Figura 4.9: Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1.

A precisão apresentou valores mais variáveis, conforme pode ser observado na Figura 4.10. Os algoritmos KNN e Floresta Aleatória apresentaram valores de precisão mais altos para todas as quantidades de mensagens, quase todos acima de 90% de precisão. Isto deve-se ao fato de ambos inicialmente terem classificado quase todas as conversas como não predatórias, aumentando a precisão, porém, diminuindo o *recall*, que é apresentado a seguir.

A Rede Neural MLP e o BERT apresentaram os valores mais altos de precisão inicialmente, com 2 mensagens. A Rede Neural, porém, conseguiu manter seus resultados acima de 90% de precisão a partir de 17 mensagens, enquanto que o BERT precisou de 75 mensagens para chegar a 90% de precisão.

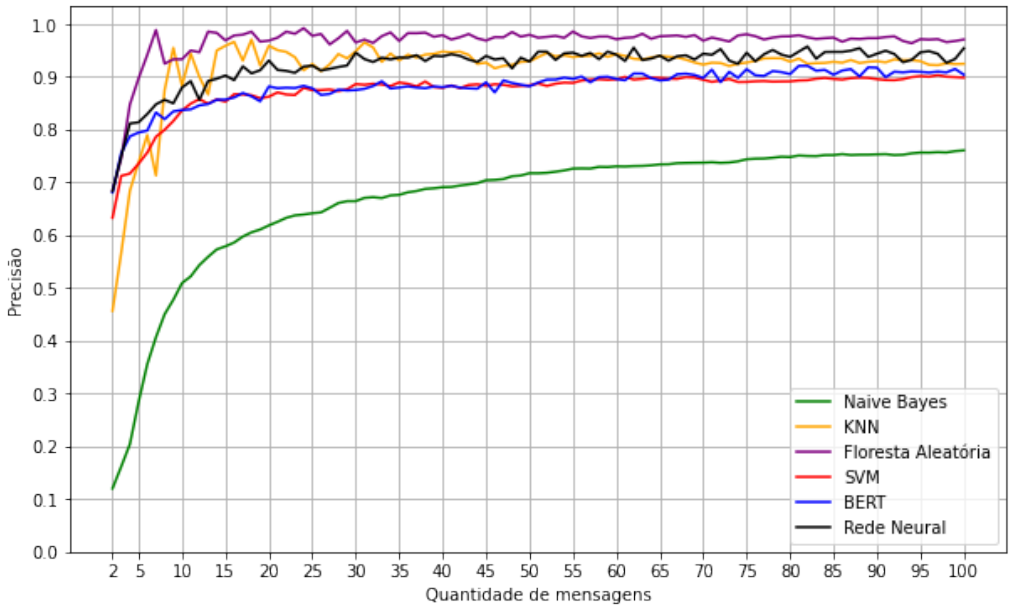


Figura 4.10: Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1.

Para o *recall* (Figura 4.11), o algoritmo *Naive Bayes Multinomial* teve resultados surpreendentes e apresentou uma grande curva de aprendizado, inicialmente com 4,92% de *recall* para as primeiras 2 mensagens, chegando até 91,43% com 100 mensagens. Com aproximadamente 40 mensagens, o *Naive Bayes* conseguiu superar o SVM e com 55 mensagens conseguiu resultados semelhantes à Rede Neural.

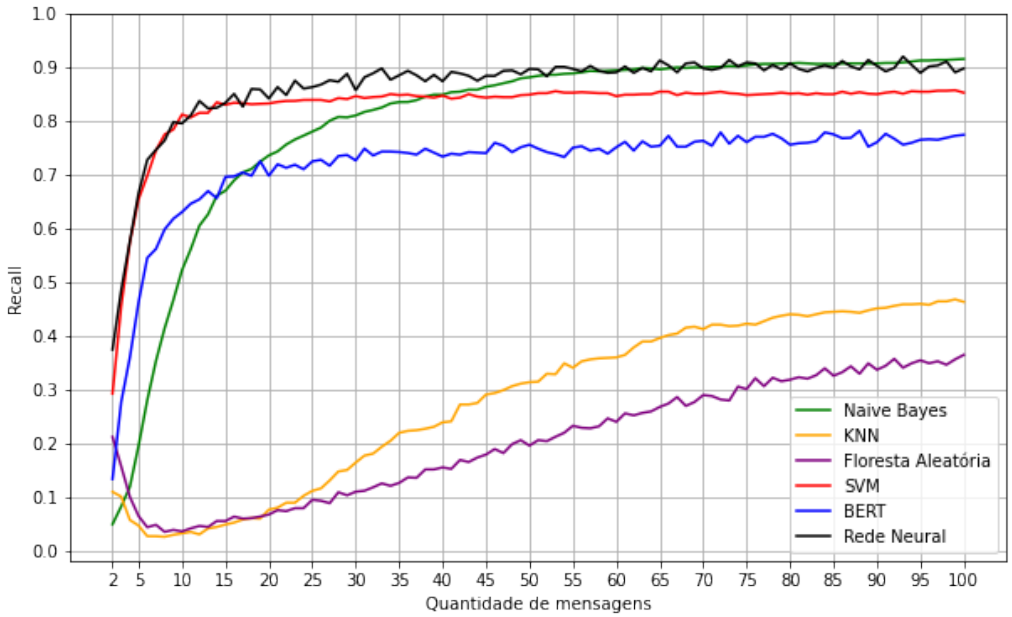


Figura 4.11: Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1.

Mas os melhores resultados ficaram com o SVM e a Rede Neural, que apenas

com 10 mensagens conseguiram *recall* de 81,11% e 79,45%, respectivamente. O Naive Bayes precisou de 28 mensagens para atingir resultados superiores à 80%, obtendo neste ponto 80,66% de *recall*.

O KNN e a Floresta Aleatória tiveram dificuldades para classificar corretamente as amostras positivas. Portanto, seus resultados não chegaram a 50% de *recall*. Foram feitos testes com diversas quantidades de vizinhos e de árvores, mas ambos não conseguiram evoluir bem os resultados, sendo os parâmetros utilizados os que obtiveram melhor equilíbrio entre as métricas.

Observando o gráfico da Figura 4.12, é possível notar que os algoritmos que apresentaram maior F_1 foram a Rede Neural MLP e o SVM, que atingiram, respectivamente, 83,45% e 82,31% já nas primeiras 10 mensagens. Com 30 mensagens a Rede Neural conseguiu $F_1 = 89,78\%$, chegando a $F_1 = 92,39\%$ com 100 mensagens.

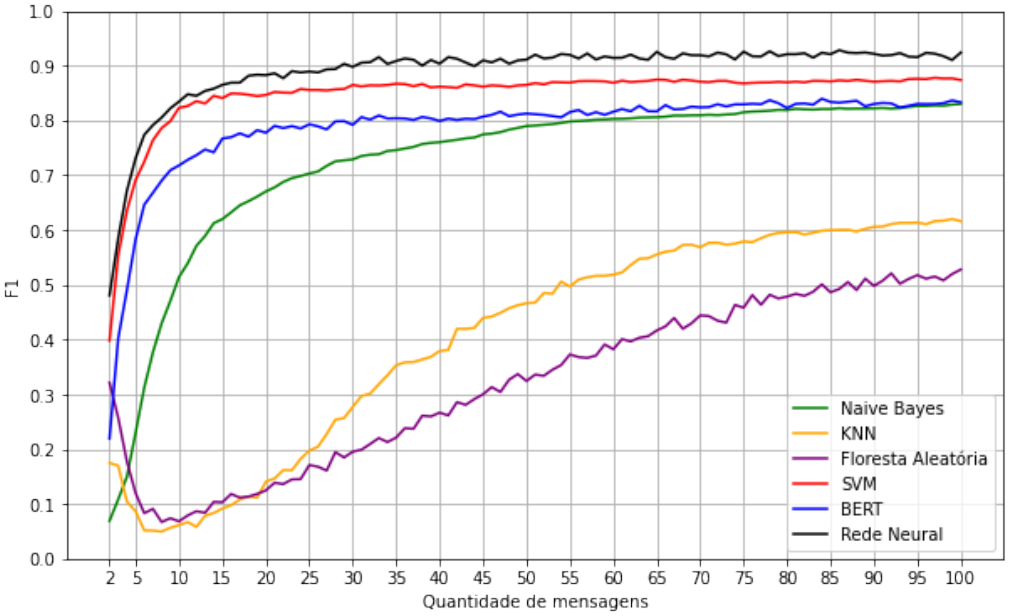


Figura 4.12: Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1.

O BERT apresentou resultados interessantes, com uma curva de aprendizado semelhante a do SVM. Com 10 mensagens, o BERT obteve $F_1 = 71,80\%$.

O *Naive Bayes* precisou de aproximadamente 55 mensagens para obter resultados semelhantes ao BERT e algoritmos KNN e Floresta Aleatória apresentaram os valores mais baixos do experimento, em virtude do *recall*, que não chegou a 50%.

Por fim, os resultados do treinamento com validação cruzada para a métrica $F_{0.5}$ podem ser observados na Figura 4.13. Os melhores resultados foram, respectivamente, da Rede Neural MLP, do SVM e do BERT.

Com apenas 10 mensagens a Rede Neural MLP obteve $F_{0.5} = 86,09\%$, atingindo $F_{0.5} = 91,08\%$ com 20 mensagens. O SVM e o BERT conseguiram $F_{0.5} = 83,09\%$ e

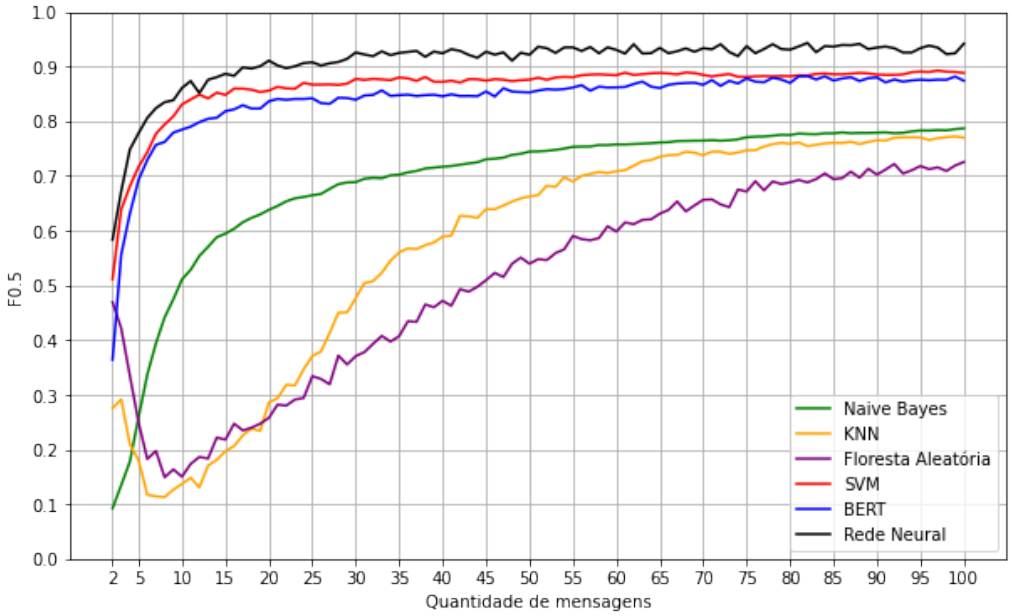


Figura 4.13: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 1.

$F_{0.5} = 78,48\%$, respectivamente, para as primeiras 10 mensagens.

Os algoritmos *Naive Bayes Multinomial*, KNN e Floresta Aleatória necessitaram de mais mensagens para obter bons resultados.

Para comparar os algoritmos e estratégias deste trabalho, assim como os resultados encontrados na literatura, foi escolhida a medida $F_{0.5}$, que é amplamente utilizada nos trabalhos encontrados na área.

Para este experimento, o algoritmo que apresentou melhores resultados de $F_{0.5}$ foi a Rede Neural MLP, tendo superado todos os outros algoritmos em todas as quantidades de mensagens. Sendo assim, ele foi escolhido para ir para a fase de testes.

Teste

O resultado do teste para a métrica acurácia está apresentada da Figura 4.14. Com 10 mensagens foi obtido 99,46% de acurácia.

Os valores de precisão estão apresentados na Figura 4.15. Com 20 mensagens foi obtido 94,29% de precisão, e com 100 mensagens foi obtido 92,46% de precisão.

Com 10 mensagens, o classificador obteve 80,38% de *recall*, chegando a 86,87% com 50 mensagens. Os dados podem ser observados na Figura 4.16

Os resultados do teste para a métrica F_1 podem ser observados na Figura 4.17. Com apenas 10 mensagens, a Rede Neural MLP obteve $F_1 = 83,78\%$, chegando a $F_1 = 88,99\%$ para 20 mensagens.

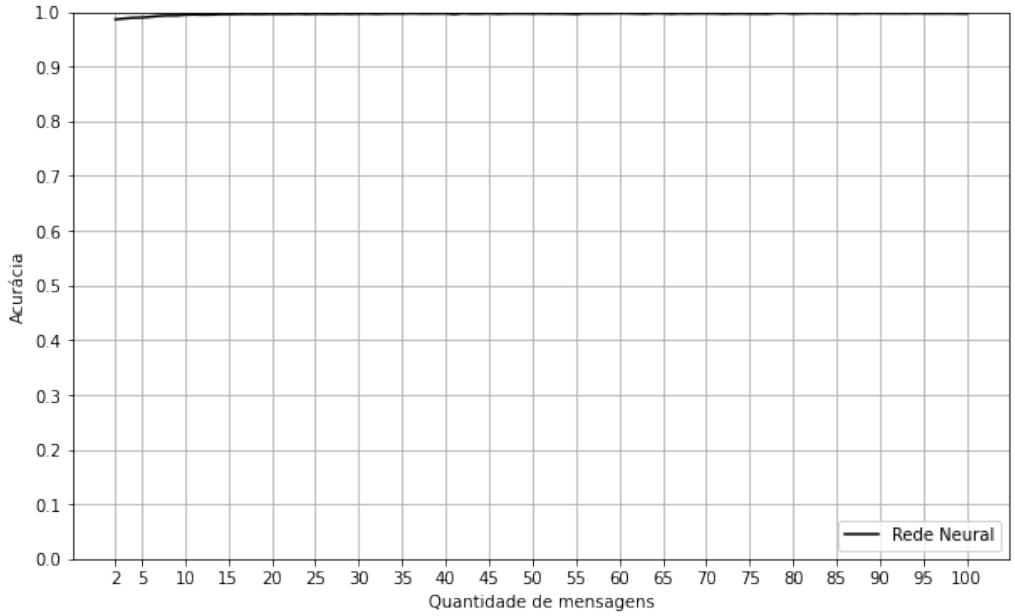


Figura 4.14: Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 1.

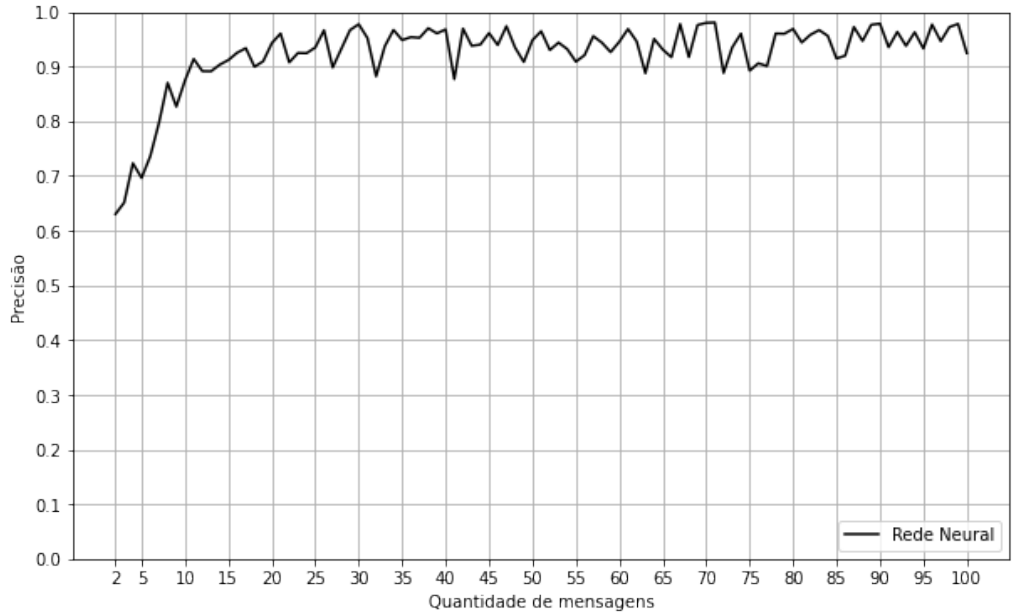


Figura 4.15: Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 1.

Por fim, os resultados obtidos para a métrica $F_{0.5}$ superaram o resultado de KULSRUD (2019), considerado o estado da arte, que obteve mais de 80% de $F_{0.5}$ após 24 mensagens.

Com apenas 10 mensagens, a Rede Neural MLP foi capaz de obter $F_{0.5} = 85,96\%$. Com 24 mensagens, o classificador obteve $F_{0.5} = 90,60\%$, e com 100 mensagens, $F_{0.5} = 91,68\%$. Os resultados podem ser observados na Figura 4.18.

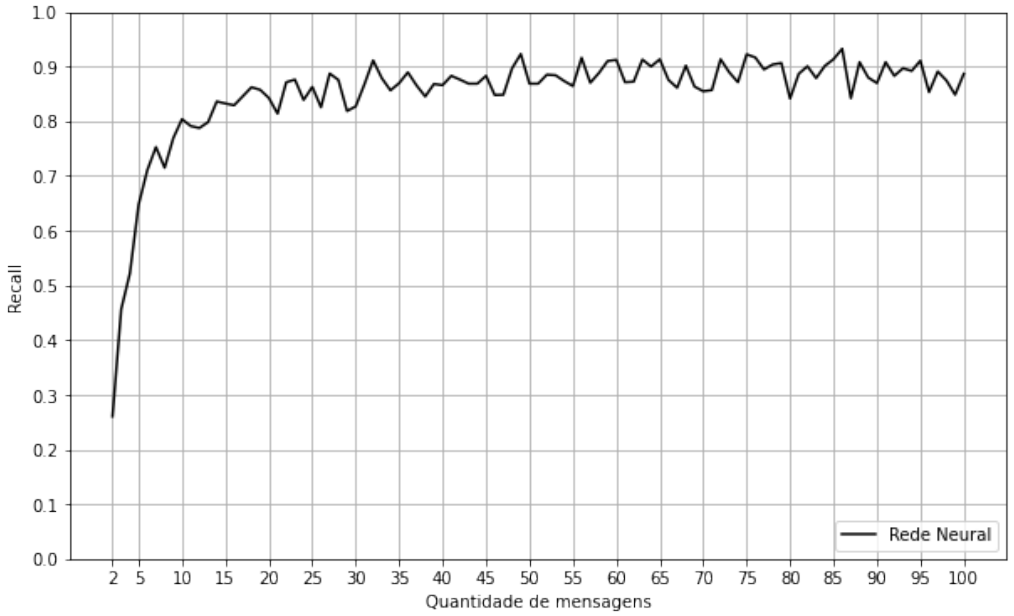


Figura 4.16: Resultado final do *recall* para o experimento sem balanceamento dos dados para a estratégia 1.

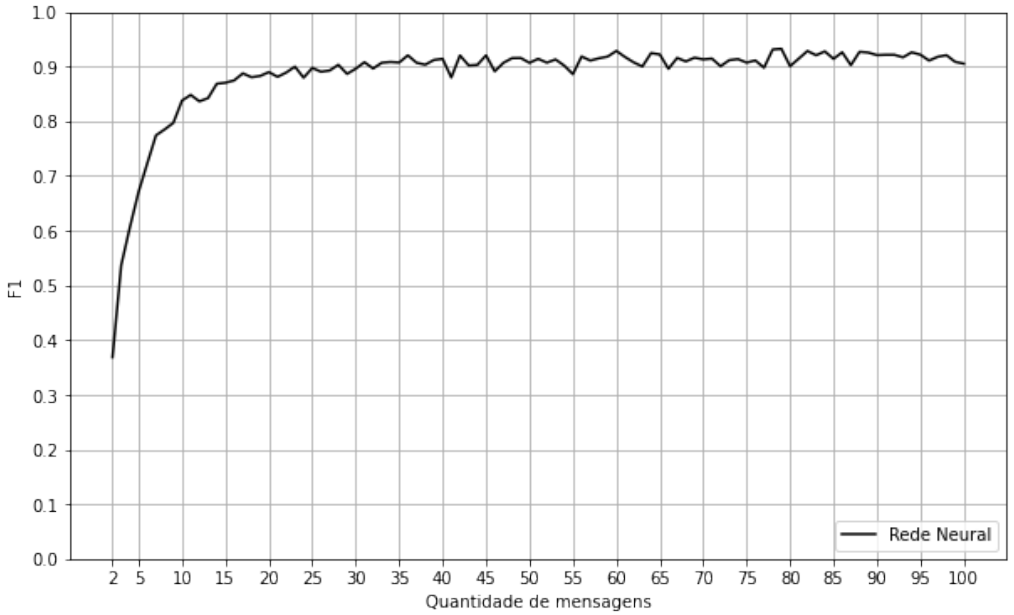


Figura 4.17: Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 1.

A Figura 4.19 exibe a quantidade de conversas predatórias corretamente identificadas para cada quantidade de mensagem. Na linha tracejada estão as quantidades totais de conversas predatórias existentes para cada quantidade de mensagem, representando assim, as quantidades máximas que poderiam ser alcançadas pelo algoritmo para cada mensagem.

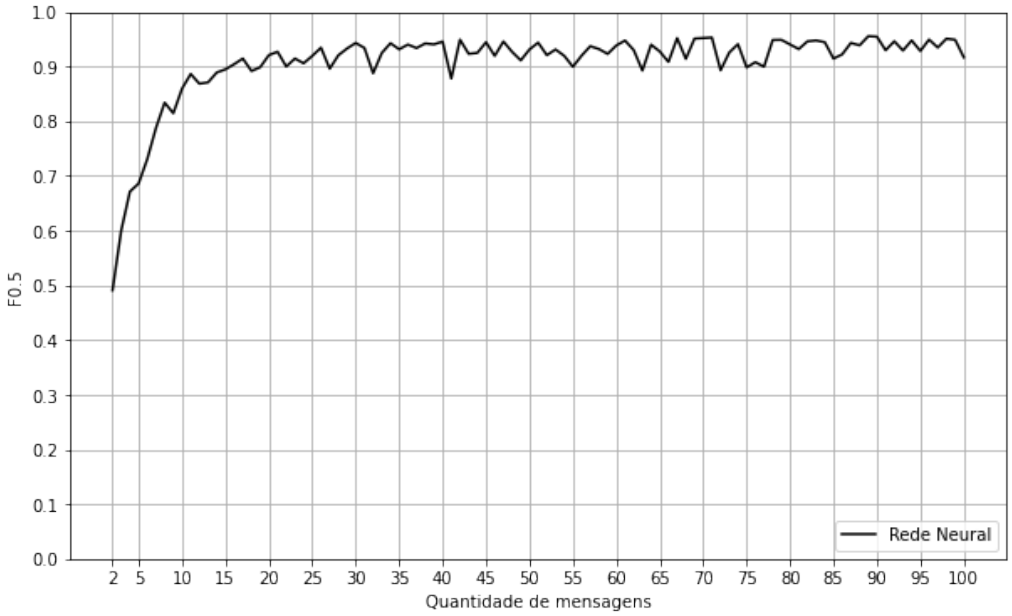


Figura 4.18: Resultado final do $F_{0.5}$ para o experimento sem balanceamento dos dados para a estratégia 1.

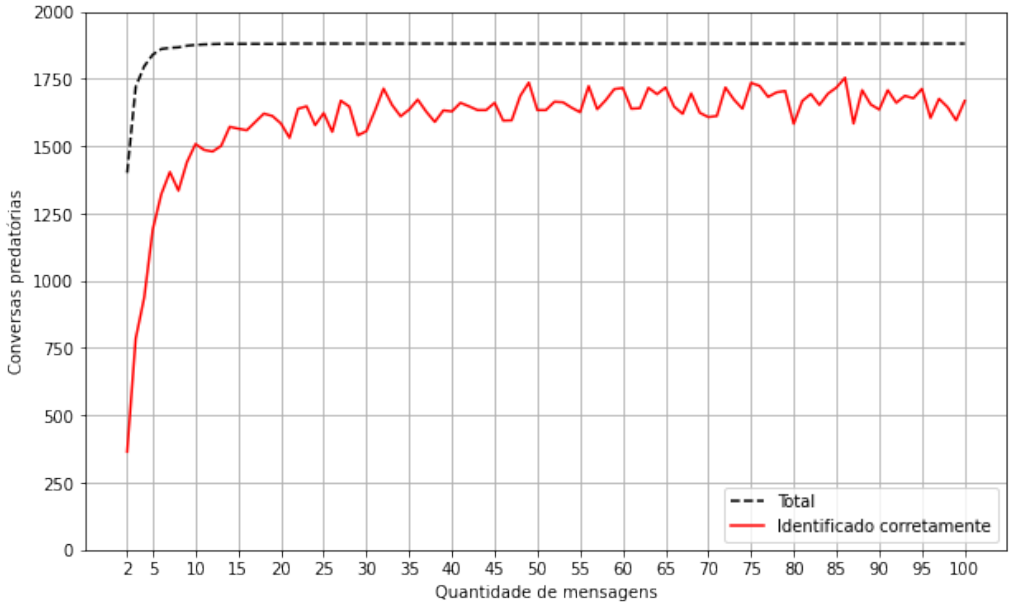


Figura 4.19: Quantidade de conversas predatórias corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 1.

Após execução das etapas de pré-processamento e pré-filtro, restaram 1.881 conversas predatórias na base de testes que continham 2 autores. Porém, para algumas quantidades de mensagens este número pode ser menor, pois, após o filtro pela quantidade de mensagens, ainda podem não haver dois autores conversando. Precisamente, só existem 1.881 conversas predatórias identificáveis a partir de 21 mensagens.

Com 10 mensagens, o algoritmo classificou corretamente 1.508 conversas preda-

tórias, de um total de 1.876 conversas predatórias possíveis de serem identificadas.

Com 20 mensagens, o algoritmo classificou corretamente 1.584 conversas predatórias, de um total de 1.880 conversas predatórias possíveis de serem identificadas. E com 50 mensagens foram classificadas corretamente 1.634 conversas predatórias de um total de 1.881 conversas predatórias possíveis de serem identificadas.

Assim, em relação ao total de 1.881 conversas predatórias da base, é possível afirmar que com 10 mensagens foi possível identificar 80,17% das conversas predatórias totais. Com 20 mensagens, foi possível identificar 84,21% das conversas predatórias, e com 50 mensagens, foi possível identificar 86,87% das conversas predatórias.

O gráfico da Figura 4.20 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem. Na base do PAN, os autores podem estar presentes em mais de uma conversa, então, a quantidade de conversas predatórias não é igual a quantidade de predadores.

Enquanto existem 1.881 conversas predatórias na base de testes após execução das etapas de pré-processamento e pré-filtro, predadores existem apenas 236. Para que hajam 236 predadores únicos identificáveis, são necessárias apenas 6 mensagens.

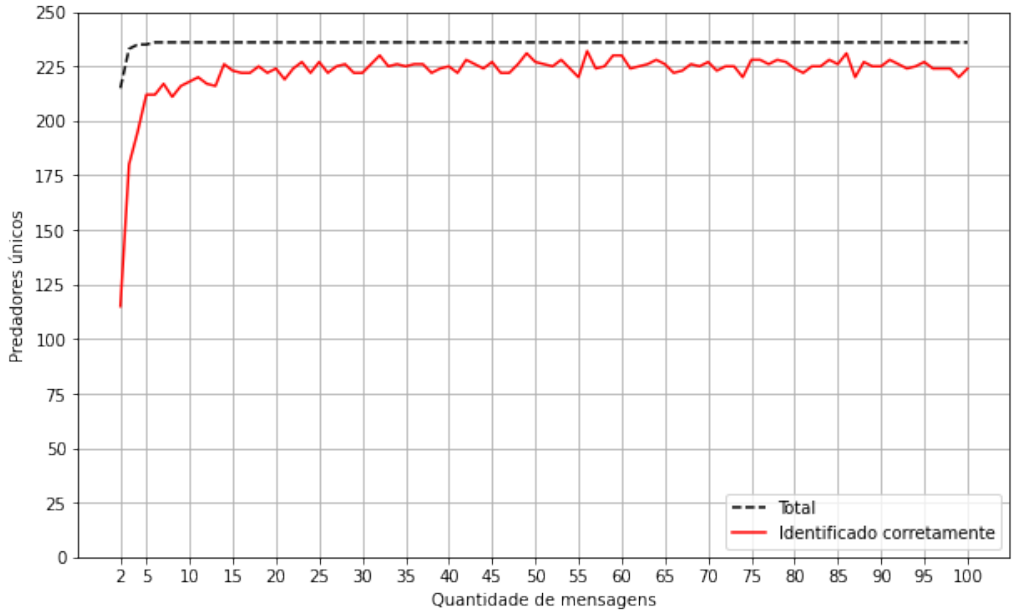


Figura 4.20: Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 1.

Com 10 mensagens, o algoritmo foi capaz de detectar 218 predadores dos 236 possíveis. Com 20 mensagens, foram detectados 224 predadores, e com 50 mensagens, 227 predadores.

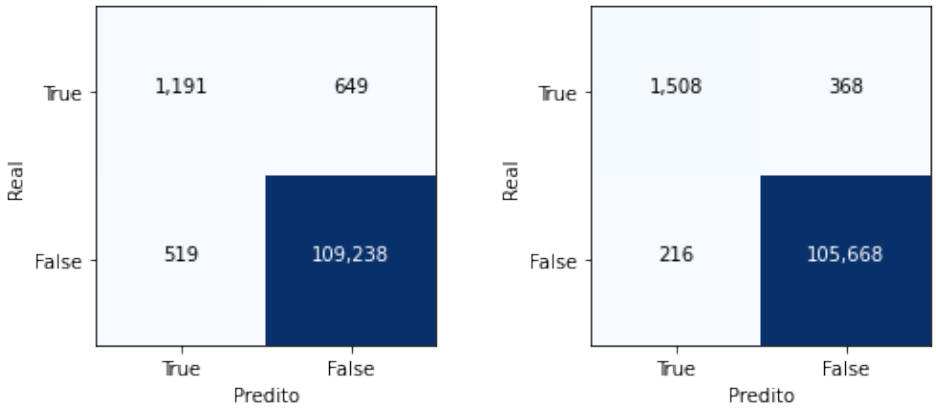
Em relação ao total de 236 predadores da base, 92,37% dos predadores puderam ser identificados com apenas 10 mensagens. Com 20 mensagens, foram detectados 94,92% dos predadores, e com 50 mensagens, foram detectados 96,19% dos predadores sexuais.

A Tabela 4.5 exibe os resultados detalhados para algumas quantidades de mensagens.

Tabela 4.5: Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 1.

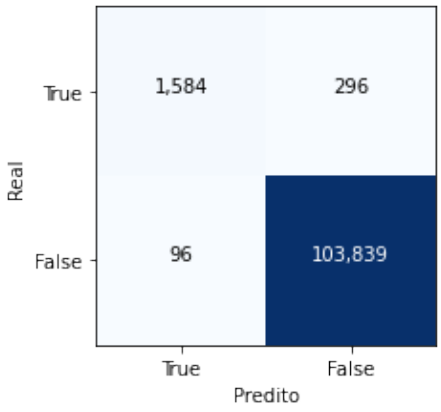
Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,9895	0,6965	0,6473	0,6710	0,6861
10	0,9946	0,8747	0,8038	0,8378	0,8596
20	0,9963	0,9429	0,8426	0,8899	0,9209
24	0,9959	0,9244	0,8389	0,8796	0,9060
50	0,9968	0,9483	0,8687	0,9068	0,9313
100	0,9967	0,9246	0,8868	0,9053	0,9168

Por fim, a Figura 4.21 exibe as matrizes de confusão para algumas quantidades de mensagens.

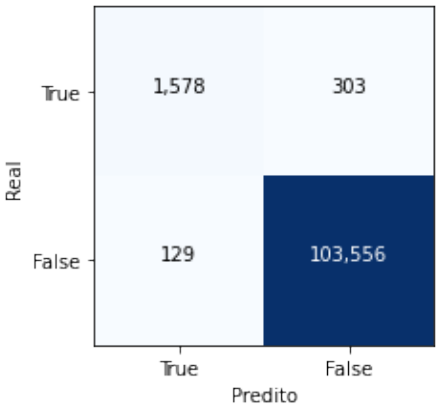


(a) 5 mensagens

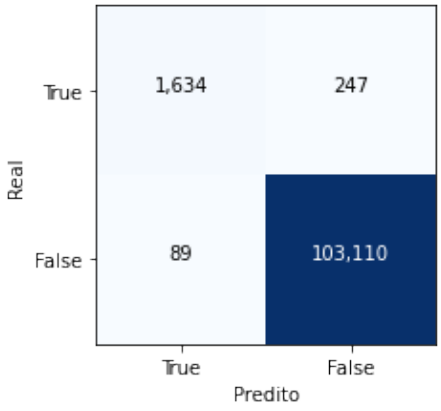
(b) 10 mensagens



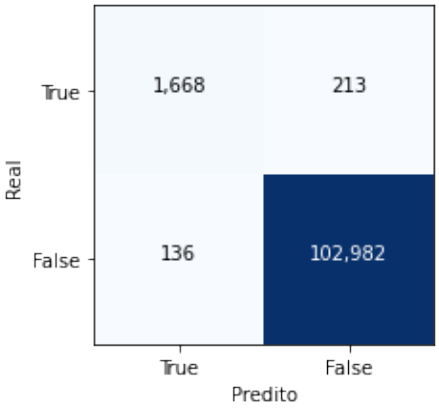
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.21: Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 1.

4.7.2 Com *Undersampling*

Para os experimentos com *undersampling*, utilizou-se o algoritmo NearMiss da biblioteca *imbalanced-learn*[4] com as configurações padrões, que baseia-se nos vizinhos mais próximos para selecionar as amostras. A implementação utilizada seleciona as amostras da classe negativa que tem menor distância média para as amostras da classe positiva.

Treinamento

A acurácia apresentou maior variação e valores mais baixos em relação à validação cruzada do experimento sem balanceamento para os algoritmos *Naive Bayes*, KNN e BERT. A Floresta Aleatória, SVM e Rede Neural MLP apresentaram resultados semelhantes, conforme pode ser visualizado na Tabela 4.22.

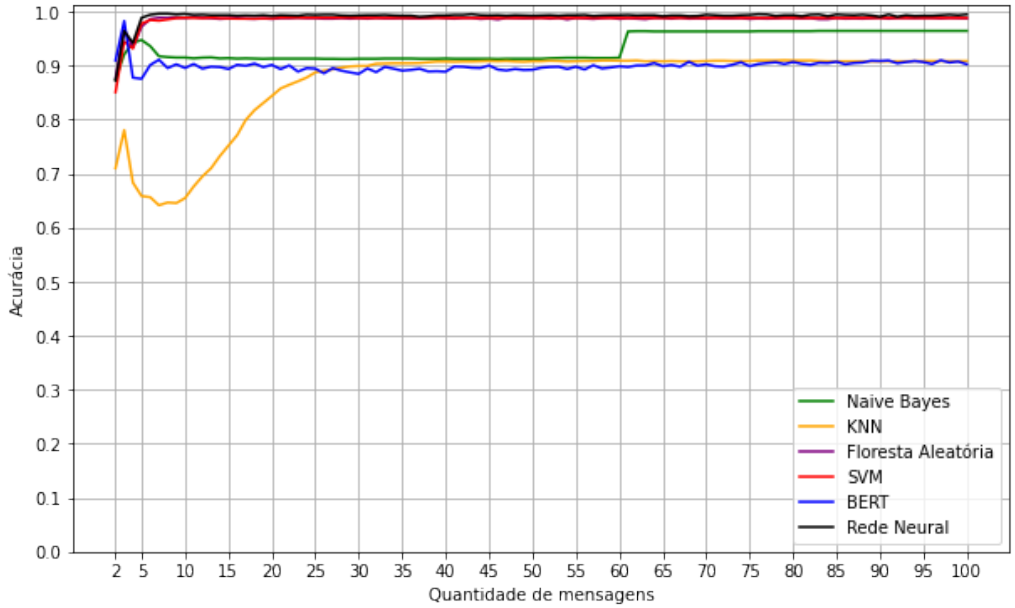


Figura 4.22: Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1.

Com exceção do *Naive Bayes Multinomial* e do BERT, todos os algoritmos apresentaram quase 100% de precisão após 6 mensagens, conforme pode ser visualizado na Figura 4.23.

Para o *recall*, apresentado na Figura 4.24, o algoritmo *Naive Bayes Multinomial* apresentou os melhores resultados, tendo atingido 99,72% de *recall* a partir das primeiras 8 mensagens.

Diferentemente do experimento sem balanceamento dos dados, para este experimento a Floresta Aleatória e o KNN conseguiram aprender a classificar corretamente

[4] `https://imbalanced-learn.org/stable/under_sampling.html`

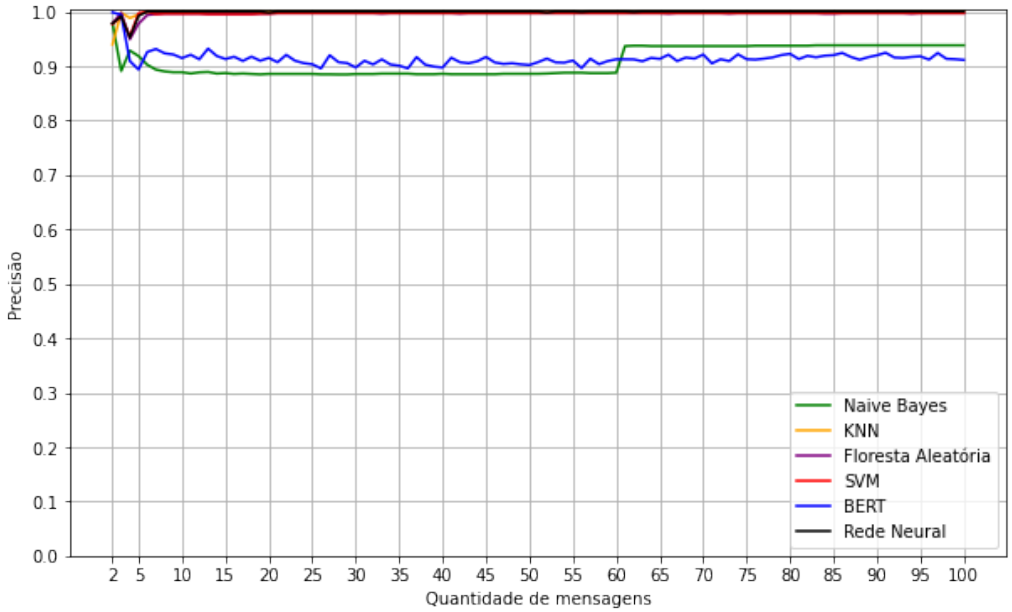


Figura 4.23: Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1.

as dados. O KNN atingiu 80,75% de *recall* com 32 mensagens e a Floresta Aleatória teve resultados compatíveis com o SVM e a Rede Neural MLP.

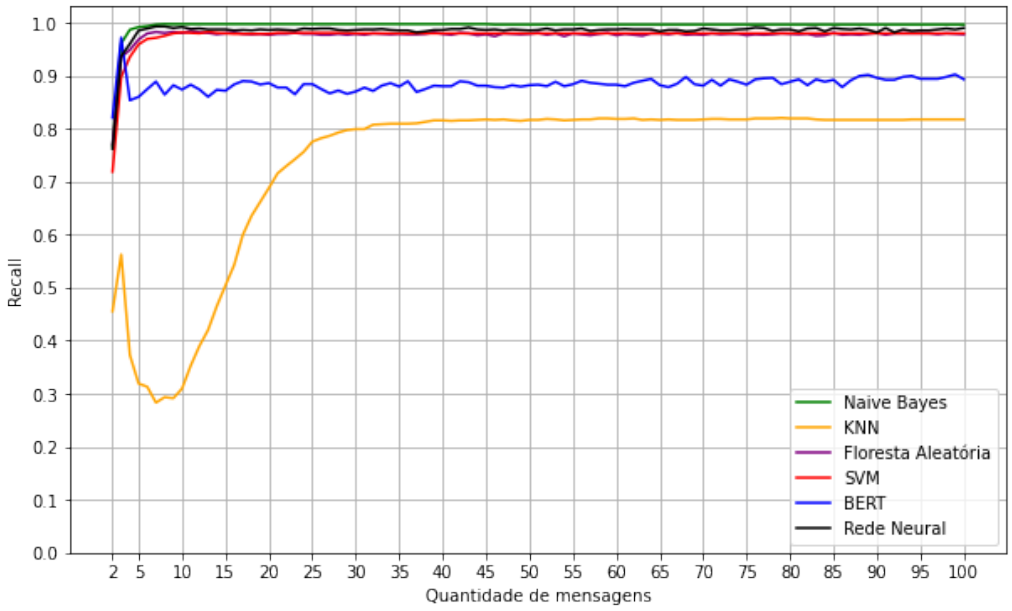


Figura 4.24: Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1.

Os valores de F_1 (Figura 4.25) e $F_{0.5}$ (Figura 4.26) foram muito semelhantes, com os algoritmos SVM, Floresta Aleatória e Rede Neural apresentando os melhores resultados. Com apenas 10 mensagens o SVM obteve $F_1 = 98,89\%$ e $F_{0.5} = 99,34\%$, a Floresta Aleatória obteve $F_1 = 98,84\%$ e $F_{0.5} = 99,32\%$ e a Rede Neural obteve

$F_1 = 99,58\%$ e $F_{0.5} = 99,83\%$.

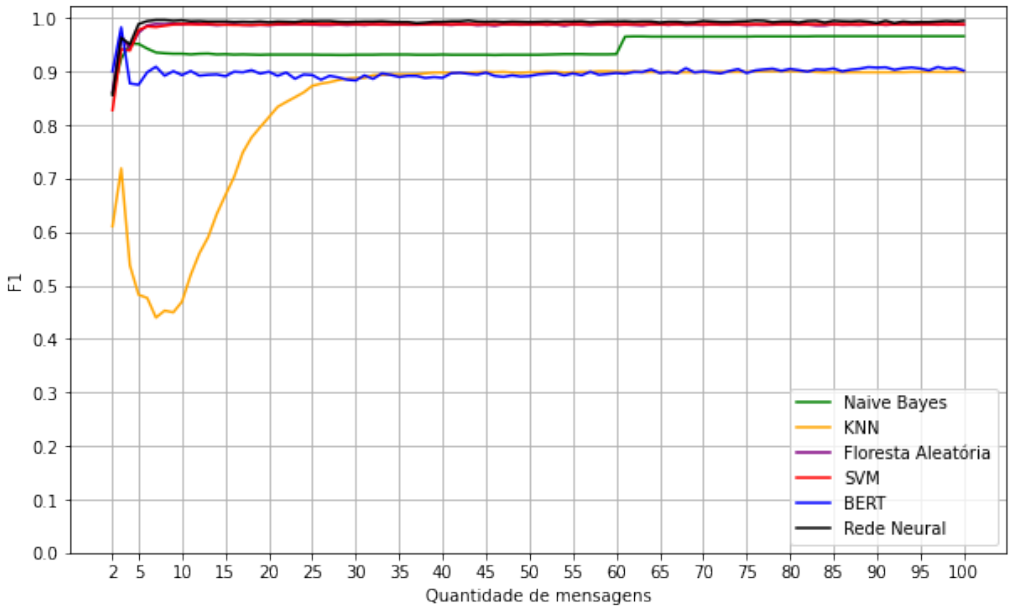


Figura 4.25: Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1.

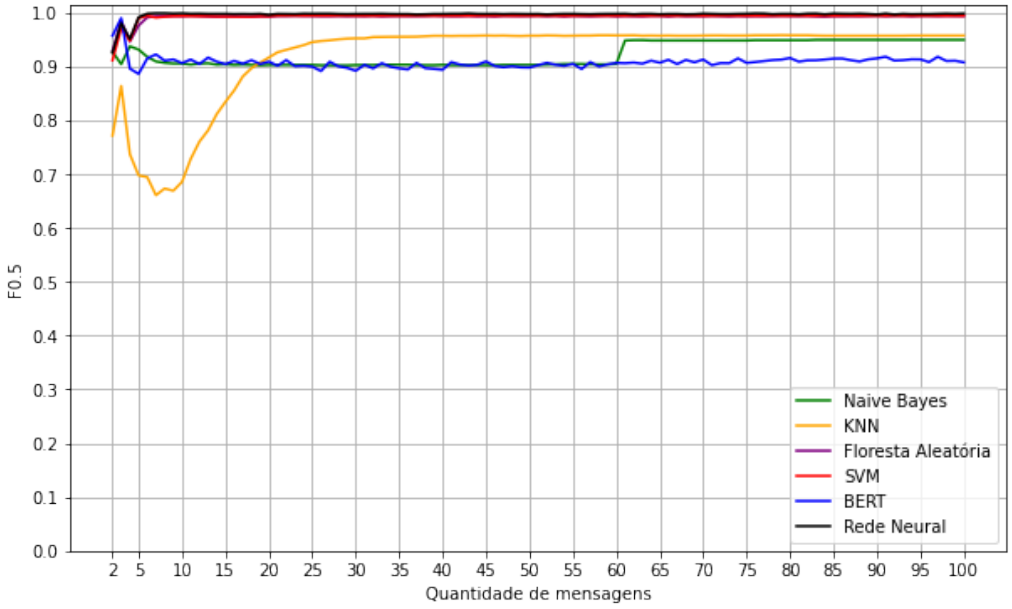


Figura 4.26: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 1.

O KNN apresentou uma queda nos resultados logo nas primeiras mensagens, recuperando-se com aproximadamente 10 mensagens e só apresentando estabilidade por volta de 30 mensagens para as duas métricas.

Por fim, o algoritmo BERT apresentou resultados satisfatórios, obtendo $F_1 = 89,33\%$ e $F_{0.5} = 90,59\%$ com as 10 primeiras mensagens.

Para este experimento, a Rede Neural MLP também apresentou os melhores resultados de $F_{0.5}$, sendo escolhida para ir para a fase de testes.

Teste

Os resultados de acurácia e precisão podem ser visualizados na Figura 4.27 e Figura 4.28, respectivamente.

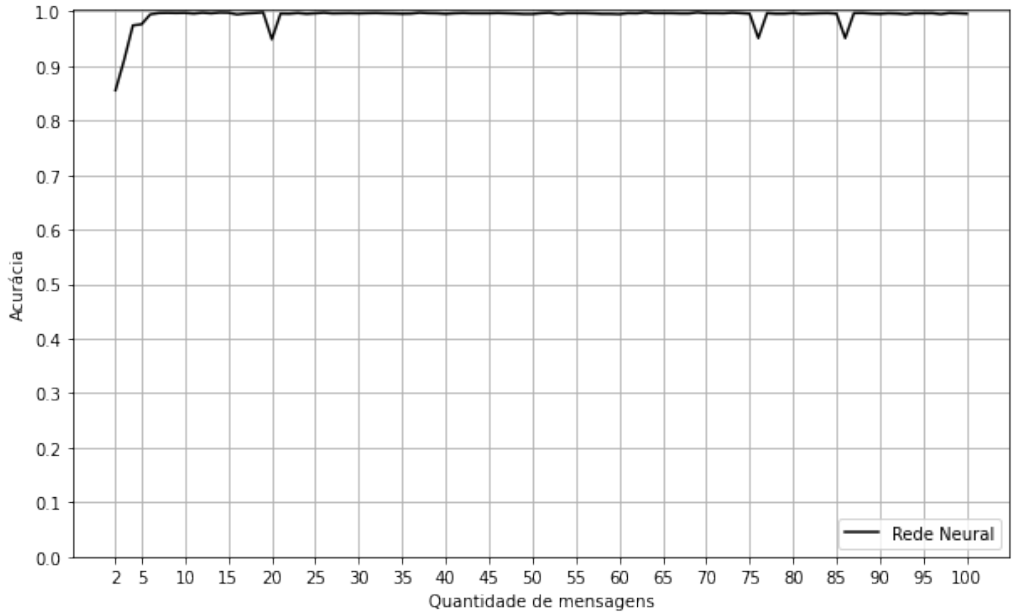


Figura 4.27: Resultado final da acurácia para o experimento com *undersampling* para a estratégia 1.

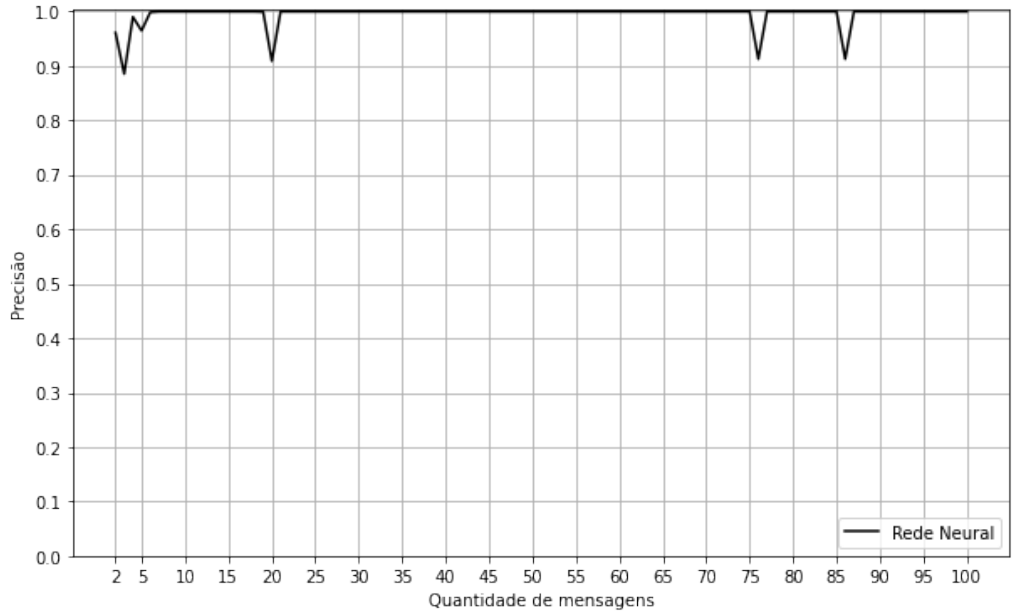


Figura 4.28: Resultado final da precisão para o experimento com *undersampling* para a estratégia 1.

Com 10 mensagens, o algoritmo obteve 99,73% de acurácia e 100% de precisão.

Os resultados para a métrica *recall* mantiveram-se muito estáveis e podem ser visualizados na Figura 4.29. Com 10 mensagens, o algoritmo obteve 99,47% de *recall*, e com 20 mensagens obteve 99,73% de *recall*.

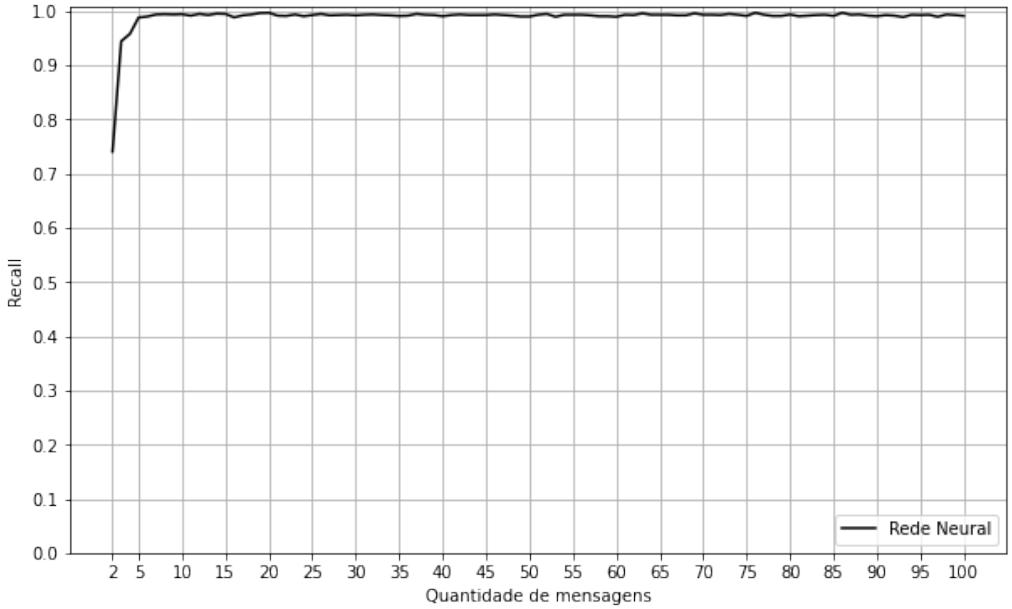


Figura 4.29: Resultado final do *recall* para o experimento com *undersampling* para a estratégia 1.

As métricas F_1 (Figura 4.30) e $F_{0.5}$ (Figura 4.31) também apresentaram o mesmo padrão.

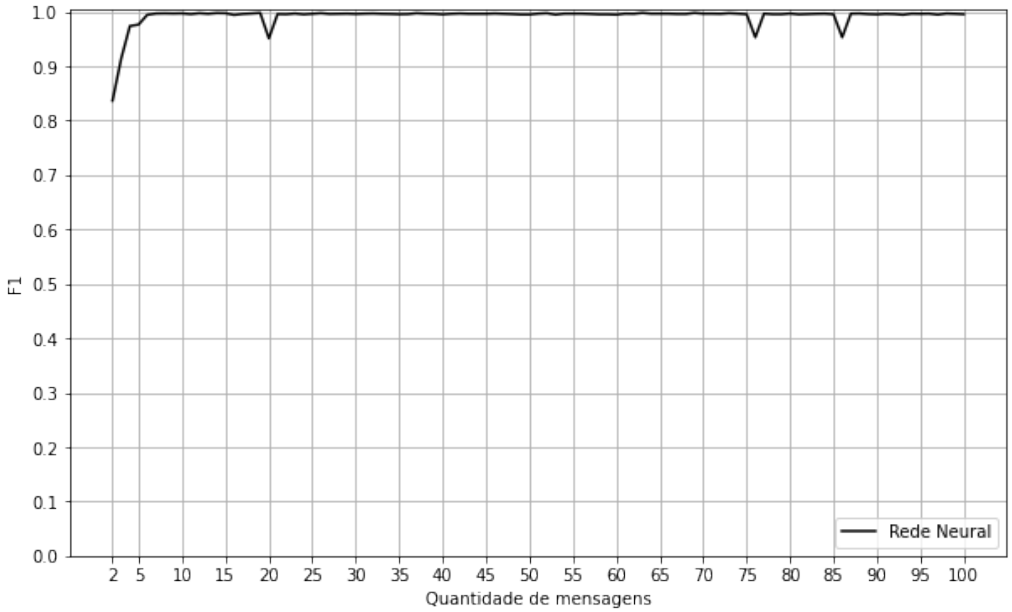


Figura 4.30: Resultado final do F_1 para o experimento com *undersampling* para a estratégia 1.

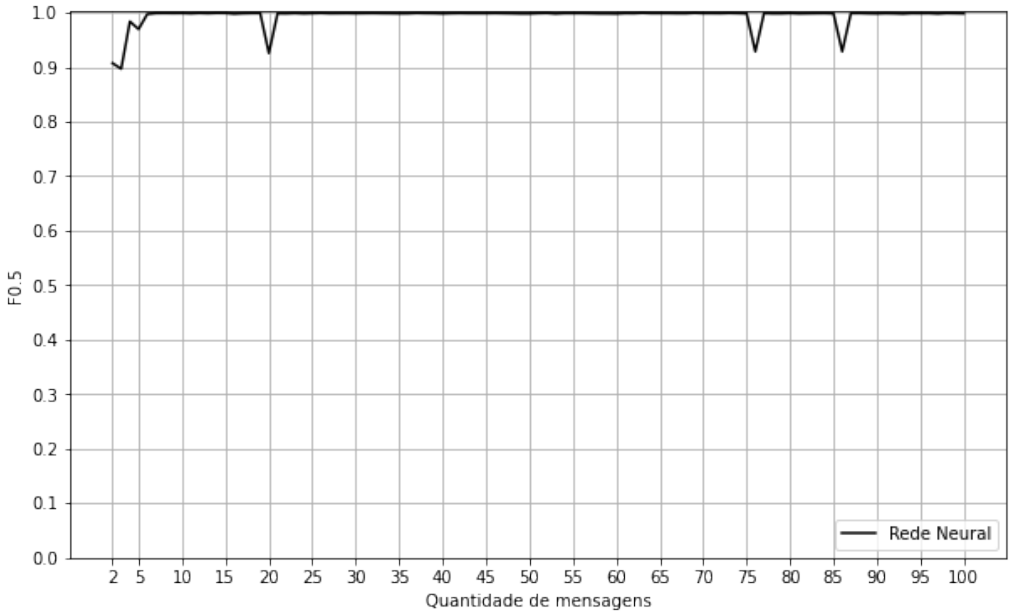


Figura 4.31: Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 1.

Com 10 mensagens, a Rede Neural MLP obteve $F_1 = 99,73\%$ e $F_{0.5} = 99,89\%$. Com 20 mensagens, o algoritmo obteve $F_1 = 95,13\%$ e $F_{0.5} = 92,57\%$, e com 30 mensagens obteve $F_1 = 99,63\%$ e $F_{0.5} = 99,85\%$.

Os resultados deste experimento também superaram os resultados obtidos por KULSRUD (2019).

A quantidade de conversas predatórias corretamente identificadas para cada quantidade de mensagem pode ser visualizada na Figura 4.32.

Como informado anteriormente, após execução das etapas de pré-processamento e pré-filtro, restaram 1.881 conversas predatórias na base de testes que continham 2 autores, e são necessárias 21 mensagens para que existam 1.881 conversas predatórias identificáveis.

Com 10 mensagens, o algoritmo classificou corretamente 1.866 conversas predatórias, de um total de 1.876 conversas predatórias possíveis de serem identificadas. Com 20 mensagens, o algoritmo classificou corretamente 1.875 conversas predatórias, de um total de 1.880 conversas predatórias. E com 50 mensagens foram classificadas corretamente 1.863 conversas predatórias de um total de 1.881 conversas predatórias possíveis de serem identificadas.

Em relação ao total de 1.881 conversas predatórias da base, 99,20% das conversas predatórias foram detectadas com 10 mensagens, 99,68% foram detectadas com 20 mensagens e 99,04% foram identificadas com 50 mensagens.

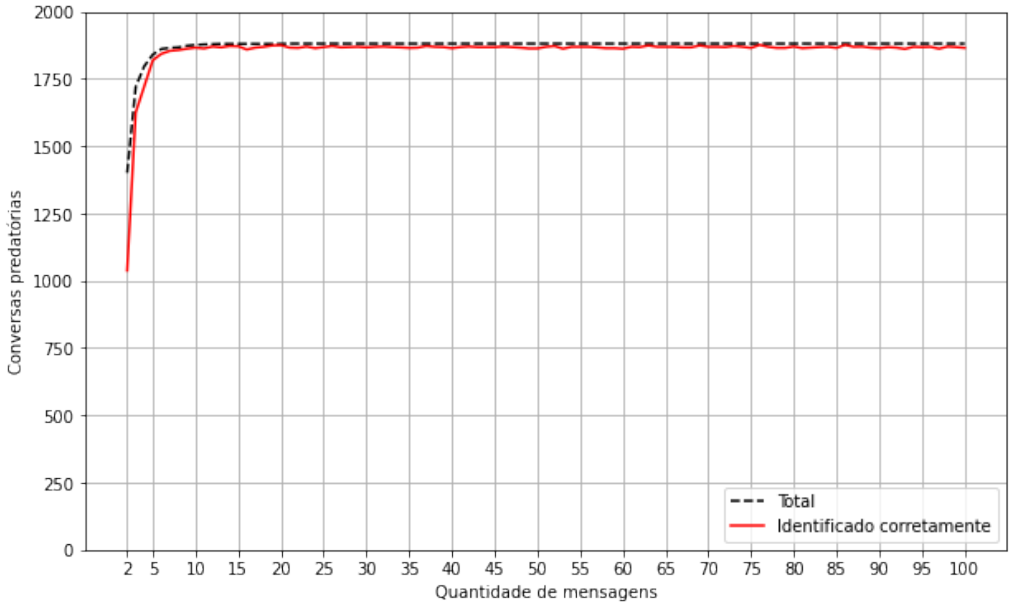


Figura 4.32: Quantidade de conversas predatórias corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 1.

O gráfico da Figura 4.33 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem.

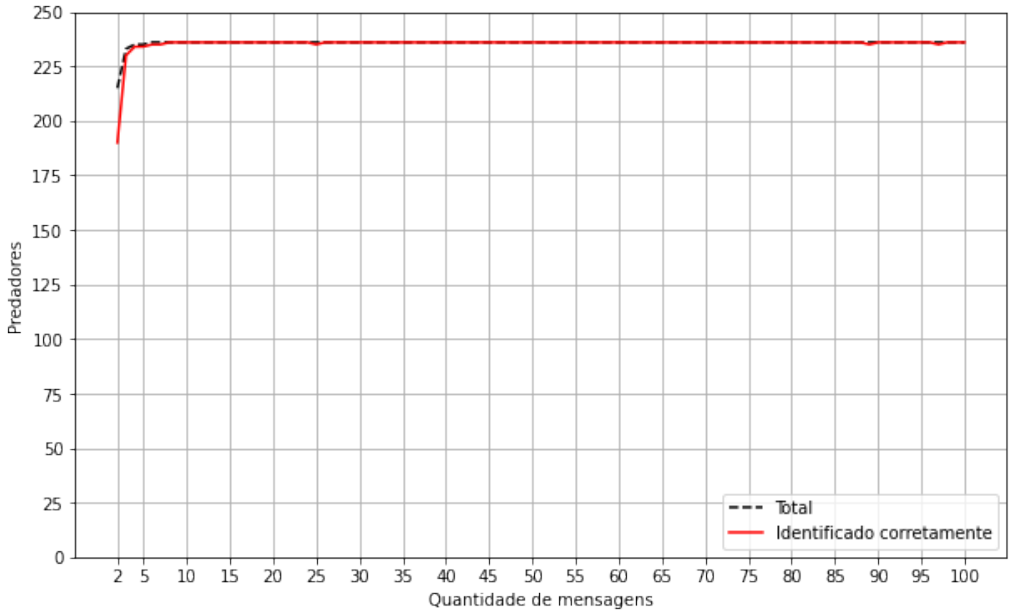


Figura 4.33: Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 1.

Como informado anteriormente, existem 236 predadores na base após execução das etapas de pré-processamento e pré-filtro e são necessárias 6 mensagens para que haja 236 predadores únicos identificáveis.

Com 8 mensagens, o algoritmo já foi capaz de detectar os 236 predadores possí-

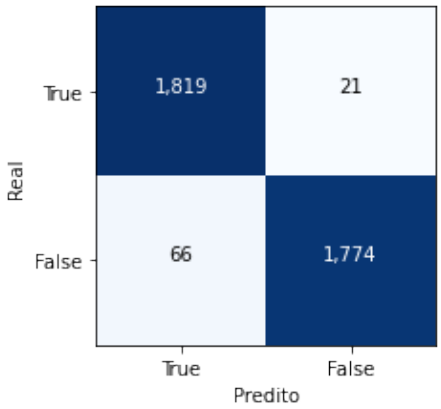
veis. Assim, em relação ao total de 236 predadores da base, 100% dos predadores puderam ser identificados com apenas 8 mensagens.

A Tabela 4.6 exibe os resultados detalhados para algumas quantidades de mensagens.

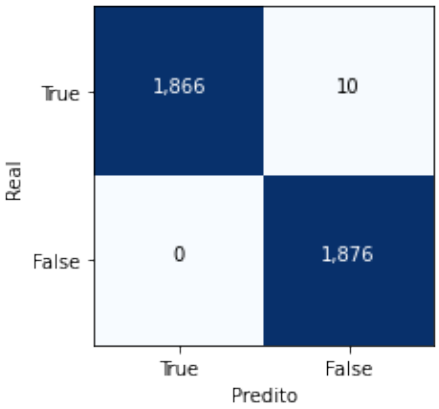
Tabela 4.6: Resultados detalhados das métricas para o experimento com *undersampling* dos dados para a estratégia 1.

Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,9764	0,9650	0,9886	0,9766	0,9696
10	0,9973	1,0000	0,9947	0,9973	0,9989
20	0,9489	0,9093	0,9973	0,9513	0,9257
24	0,9955	1,0000	0,9910	0,9955	0,9982
50	0,9952	1,0000	0,9904	0,9952	0,9981
100	0,9957	1,0000	0,9915	0,9957	0,9983

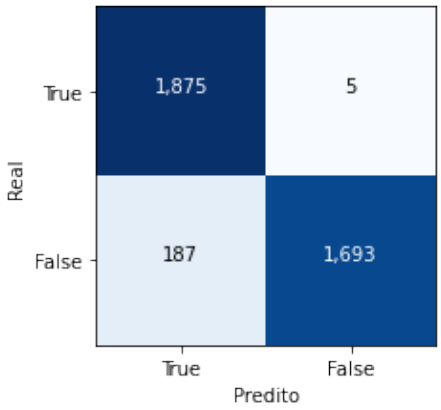
Por fim, a Figura 4.34 exibe as matrizes de confusão para algumas quantidades de mensagens.



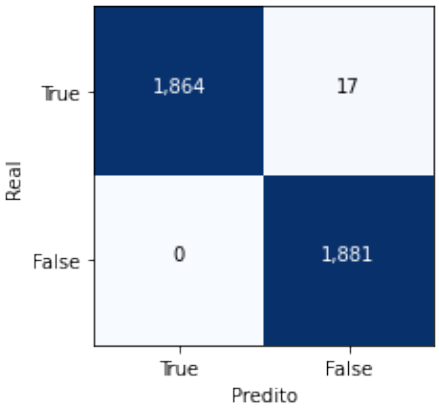
(a) 5 mensagens



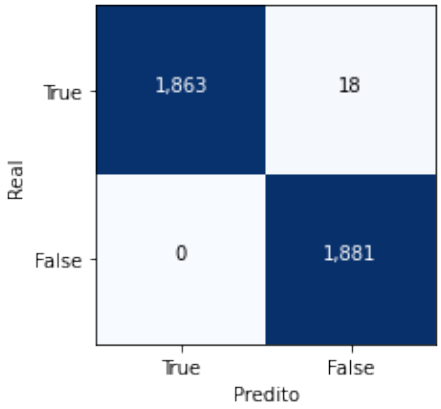
(b) 10 mensagens



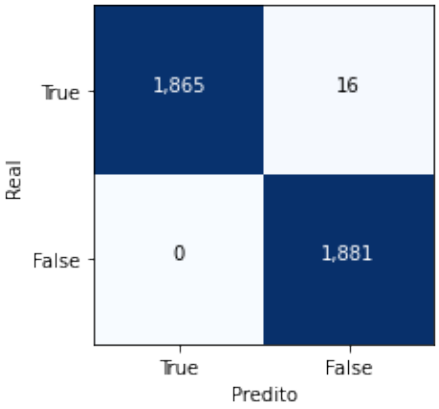
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.34: Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 1.

4.8 Estratégia 2 - Distinguir Predador e Vítima

Diferente da estratégia anterior, para esta estratégia as mensagens foram agrupadas por autores e conversas, de forma que cada tupla correspondesse a todas as mensagens enviadas por um autor em uma conversa. Assim, se uma conversa possuísse 2 autores, ela teria 2 tuplas, e se tivesse 30 autores, teria 30 tuplas. Isto foi necessário já que a classe alvo, denominada "*predator*" possui valores "*True*" ou "*False*" para cada autor em uma conversa, justamente porque esta estratégia baseia-se nos autores, ao contrário da estratégia anterior, que baseia-se nas conversas.

Já nesta etapa do agrupamento alguns autores foram perdidos, a depender da quantidade de mensagens utilizada no experimento. Isto ocorreu pois alguns autores entraram na conversa após o limite definido de quantidade de mensagens, como explicado anteriormente.

O mesmo pré-filtro de remover conversas que não tivessem apenas 2 autores foi utilizado para esta estratégia. Assim, depois do agrupamento, estas conversas foram removidas, fazendo com que a base tivesse sempre duas tuplas para cada conversa, cada uma representando um autor.

Foram realizados experimentos com e sem técnicas de balanceamento, que são detalhados a seguir.

4.8.1 Sem Balanceamento dos Dados

Nesta estratégia, como o objetivo foi classificar os autores em predadores ou vítimas, cada tupla referia-se a todas as mensagens enviadas por um autor em uma conversa. E como cada conversa só pode conter dois autores após a execução do pré-filtro, então cada conversa possuía duas tuplas, o que basicamente duplicou a quantidade de dados da base em relação à estratégia 1 e aumentou a diferença entre as classes.

Na Figura 4.35 é possível visualizar a quantidade de classes positivas e negativas da base de treinamento após agrupamento e remoção de conversas que não continham apenas dois autores, com as primeiras 50 mensagens das conversas, por exemplo. Nesta situação, existem 89.717 rótulos para vítimas e 1.085 rótulos para predadores.

Treinamento

Por conta da quantidade elevada de classes positivas da base, todos os algoritmos apresentaram acurácia acima de 97%, assim como no mesmo experimento da estratégia 1, conforme pode ser observado no gráfico da Figura 4.36.

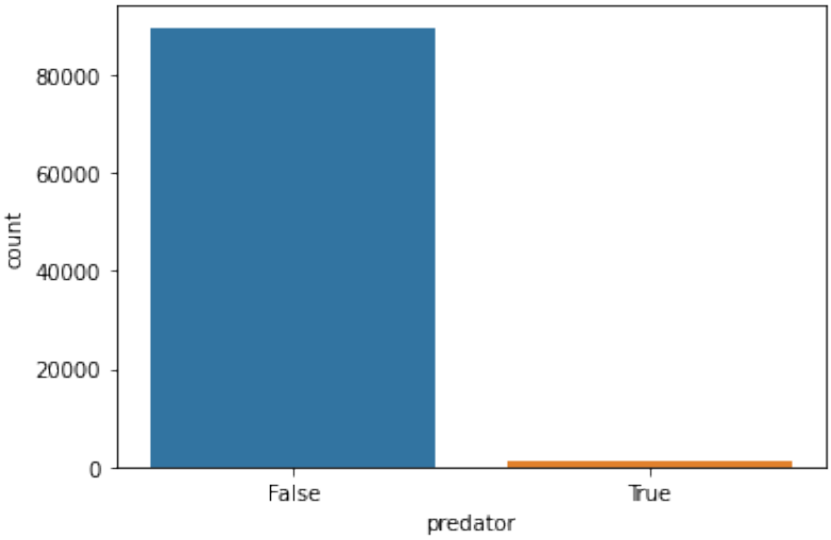


Figura 4.35: Quantidade de valores positivos e negativos para a classe alvo utilizando as primeiras 50 mensagens.

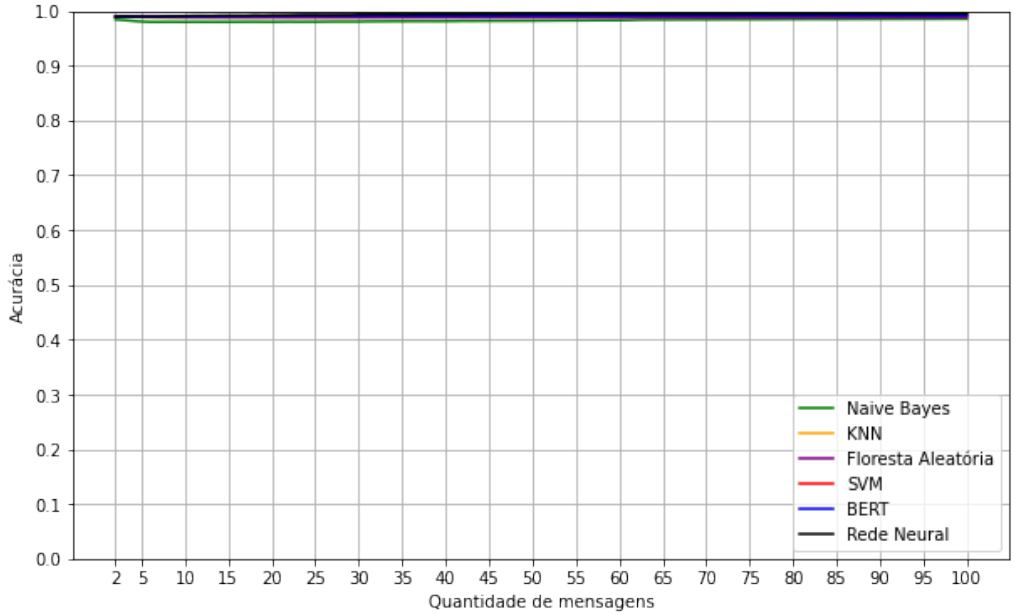


Figura 4.36: Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2.

A precisão apresentou resultados mais baixos em comparação com o experimento sem balanceamento da estratégia 1, com o algoritmo Rede Neural MLP e BERT apresentando os melhores resultados.

Com 30 mensagens, a Rede Neural obteve 78,88% de precisão, e o BERT obteve 74,38% de precisão. Os resultados podem ser visualizados na Figura 4.37.

O *recall* (Figura 4.38) apresentou resultados ainda mais baixos. O resultado mais alto do BERT foi 48,57% com 74 mensagens. A Rede Neural MLP alcançou 68,86% com 100 mensagens.

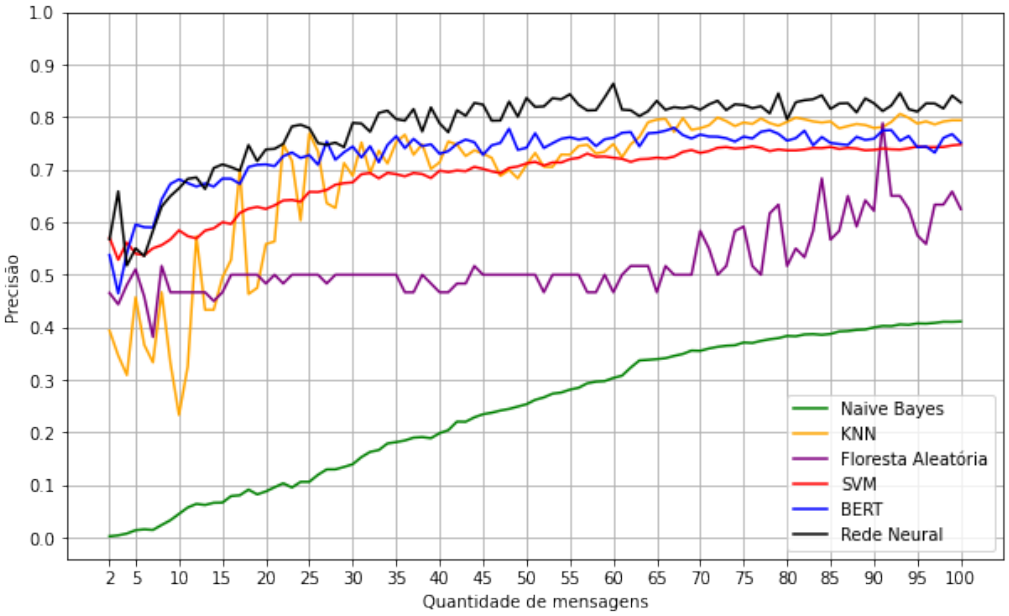


Figura 4.37: Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2.

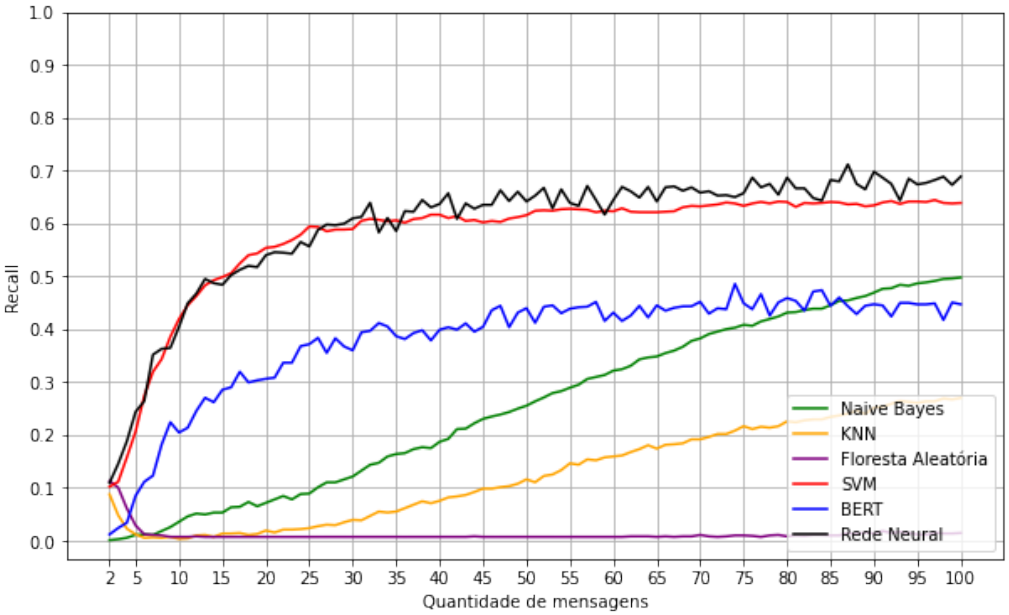


Figura 4.38: Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2.

Os resultados de F_1 e $F_{0.5}$ foram semelhantes, conforme pode ser observado na Figura 4.39 e na Figura 4.40.

A Floresta Aleatória teve dificuldades para lidar com as amostras positivas, classificando quase todos os dados como negativos. Assim, os resultados de F_1 e $F_{0.5}$ para a Floresta Aleatória não chegaram a 0,1%.

O KNN e o *Naive Bayes* tiveram resultados intermediários, conseguindo bom

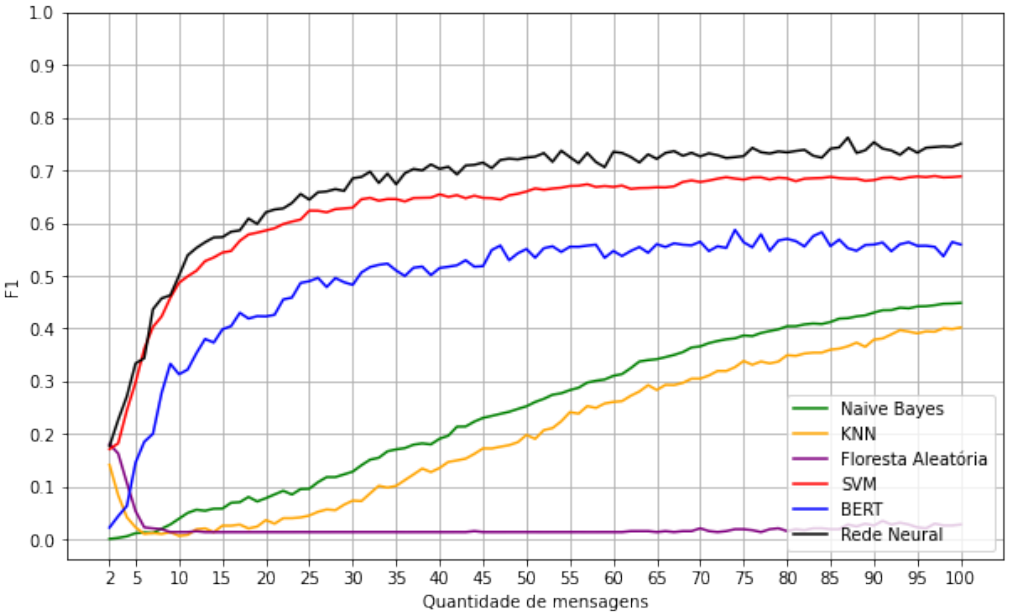


Figura 4.39: Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2.

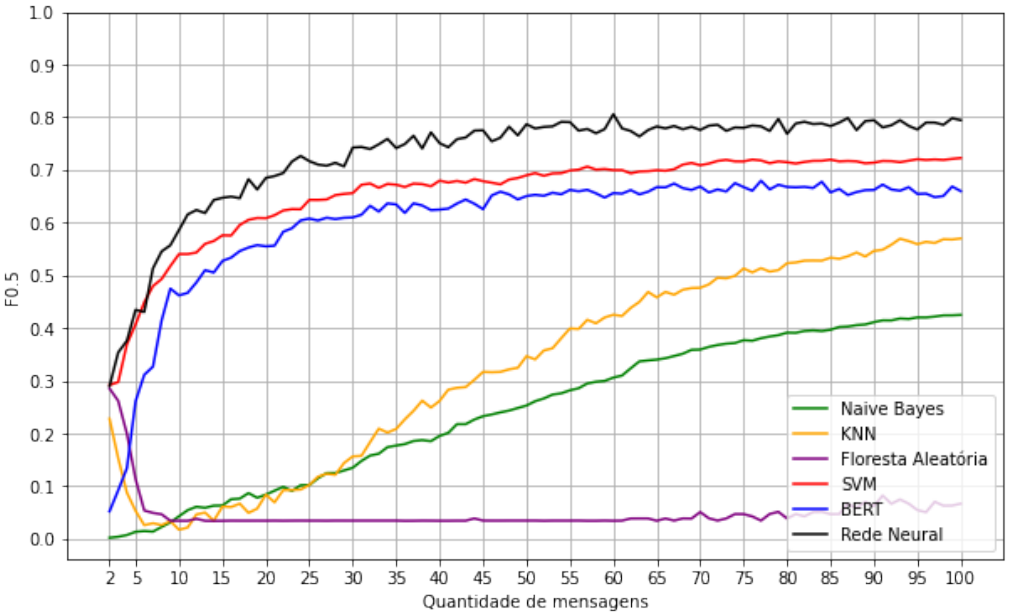


Figura 4.40: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 2.

equilíbrio entre *recall* e precisão.

Os resultados mais altos ficaram, respectivamente, com a Rede Neural MLP, o SVM e o BERT.

Com 50 mensagens, a Rede Neural conseguiu $F_1 = 72,39\%$ e $F_{0.5} = 78,67\%$. O SVM obteve $F_1 = 65,97\%$ e $F_{0.5} = 68,98\%$ para a mesma quantidade de mensagens, e o BERT obteve $F_1 = 55,06\%$ e $F_{0.5} = 65,05\%$.

Em comparação com o mesmo experimento da estratégia 1, a estratégia 2 precisou de mais mensagens para obter bons resultados, e os resultados máximos ainda são inferiores aos da estratégia 1.

Para a fase de testes, o algoritmo Rede Neural MLP foi escolhido por apresentar novamente os melhores resultados de $F_{0.5}$ para todas as quantidades de mensagens.

Teste

Para a fase de testes, a acurácia manteve resultados elevados e estáveis. Com 5 mensagens a Rede Neural MLP já atingiu 99,20%, conforme mostra a Figura 4.41.

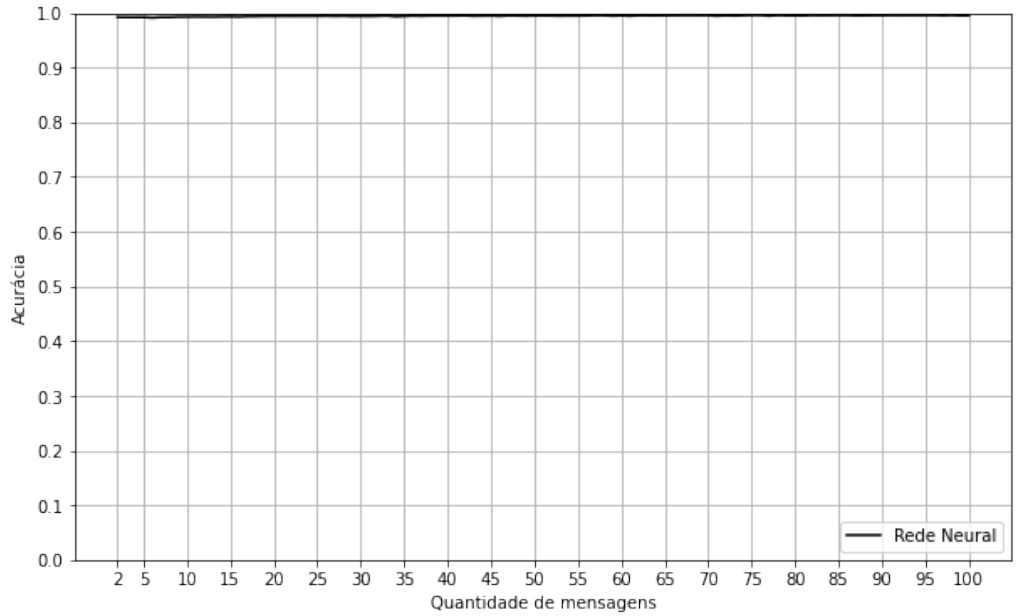


Figura 4.41: Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 2.

A precisão, apresentada na Figura 4.42, obteve resultados muito variáveis ao longo das diferentes quantidades de mensagens. Para as 10 primeiras mensagens, foi obtido 63,95% de precisão, enquanto que com 100 mensagens foi obtido 78,07%. Com 49 mensagens a Rede Neural obteve o resultado mais alto para a precisão, 90,22%.

O *recall*, apresentado na Figura 4.43, manteve valores abaixo de 70%. Com 30 mensagens, o algoritmo obteve 55,40% de *recall*.

Com 10 mensagens, o classificador obteve 48,05% de F_1. Com 20 mensagens foi obtido $F_1 = 54,12\%$ e com 30 mensagens foi obtido $F_1 = 62,43\%$. Os resultados estão apresentados na Figura 4.44.

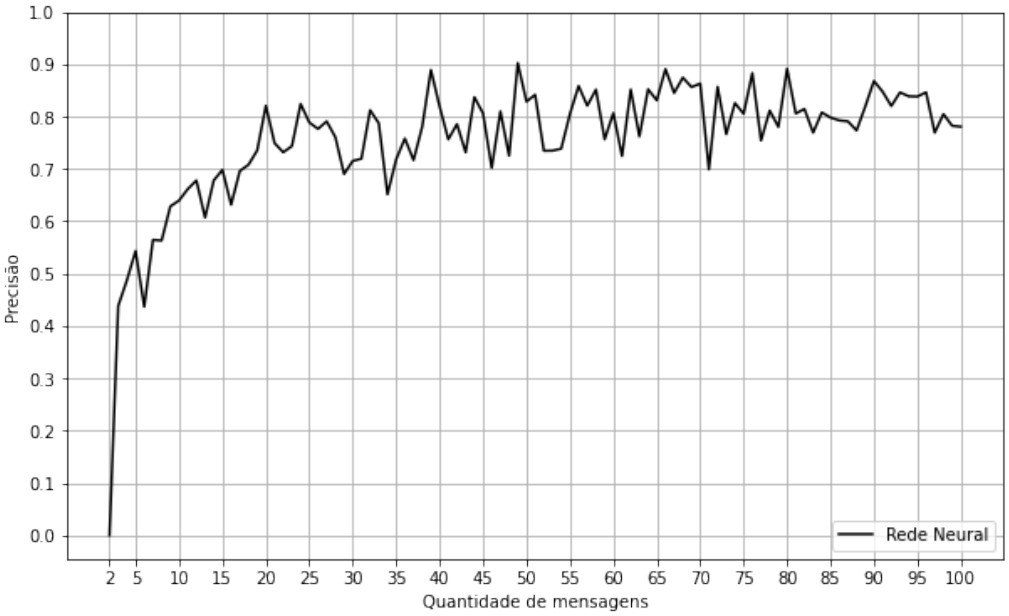


Figura 4.42: Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 2.

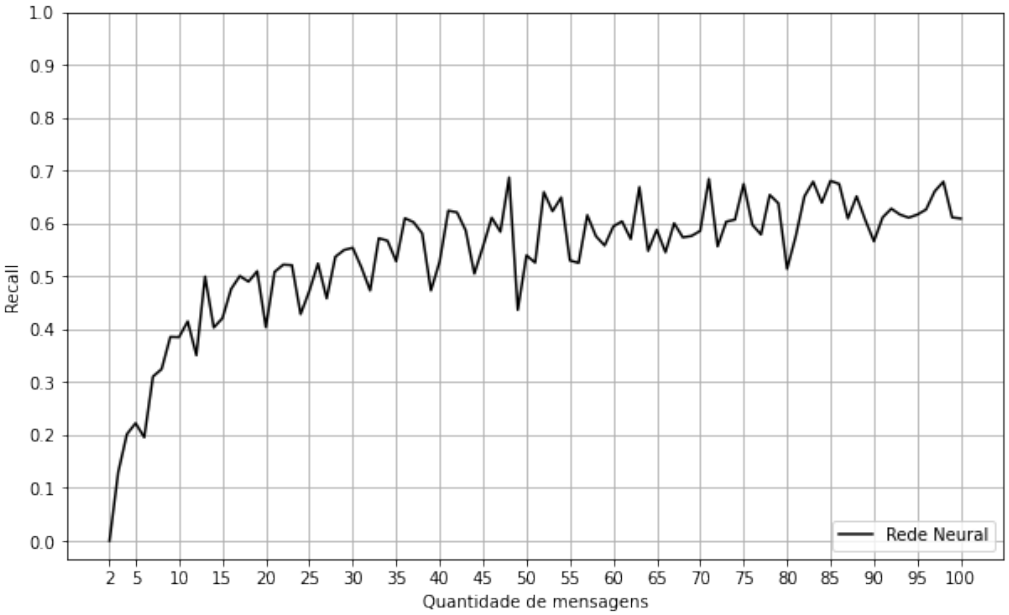


Figura 4.43: Resultado final *recall* para o experimento sem balanceamento dos dados para a estratégia 2.

Finalmente, os resultados para a métrica $F_{0.5}$ podem ser visualizados na Figura 4.45. Com 10 mensagens, a Rede Neural MLP obteve $F_{0.5} = 56,48\%$. Com 20 mensagens foi obtido $F_{0.5} = 68,01\%$ e com 100 mensagens foi obtido $F_{0.5} = 73,91\%$.

Em comparação com o mesmo experimento da estratégia 1, os resultados obtidos por este experimento foram muito inferiores.

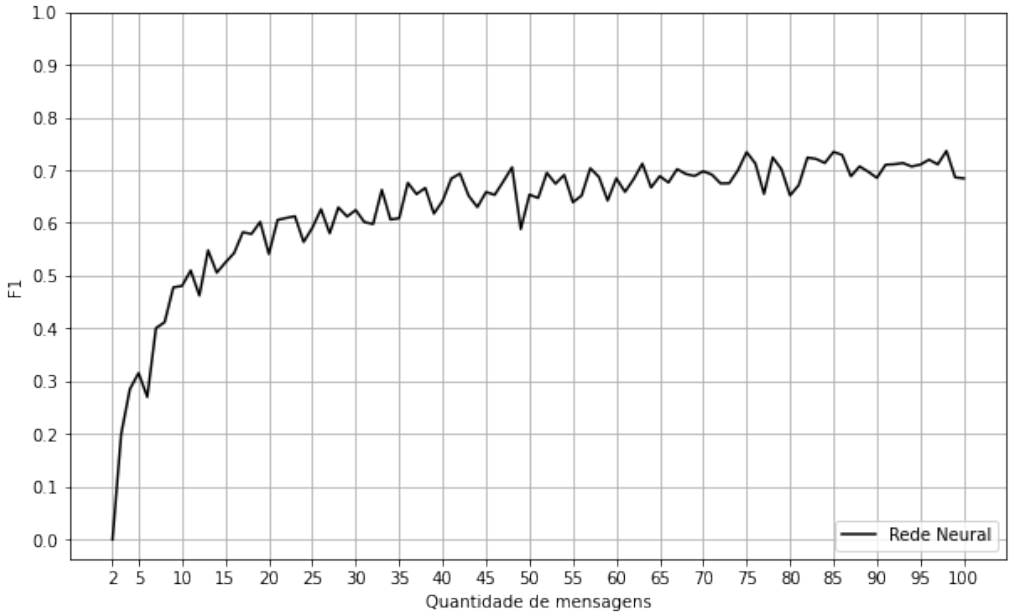


Figura 4.44: Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 2.

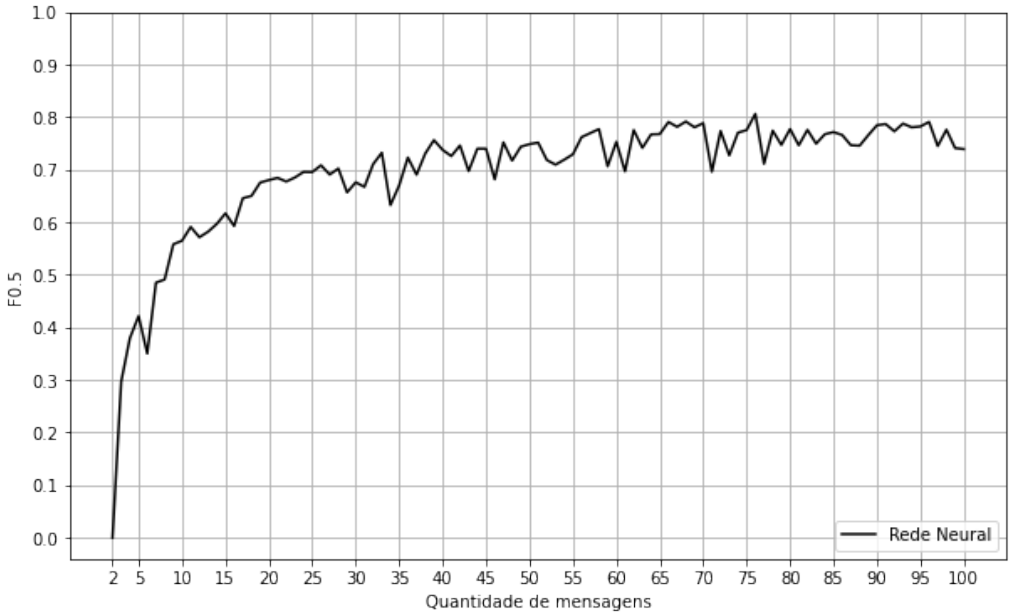


Figura 4.45: Resultado final do $F_{0.5}$ para o experimento sem balanceamento dos dados para a estratégia 2.

Como informado anteriormente, os predadores sexuais participam de mais de uma conversa na base do PAN 2012. Isto significa que a quantidade de tuplas de predadores não é igual a quantidade de predadores únicos. Assim, foram gerados gráficos para as quantidades de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem, e para as quantidades de predadores únicos corretamente identificados para cada quantidade de mensagem.

A Figura 4.46 exibe a quantidade de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem. Na linha tracejada estão as quantidades totais de predadores (tuplas) existentes para cada quantidade de mensagem, representando assim, as quantidades máximas que poderiam ser alcançadas pelo algoritmo para cada mensagem.

Com 10 mensagens, o algoritmo classificou corretamente 722 autores como predadores sexuais, de um total de 1.876 possíveis de serem identificados.

Com 20 mensagens, o algoritmo classificou corretamente 759 autores como predadores sexuais, de um total de 1.880 tuplas possíveis de serem identificadas. Com 50 mensagens foram classificados corretamente 1.015 autores como predadores sexuais, de um total de 1.881 tuplas possíveis de serem identificadas.

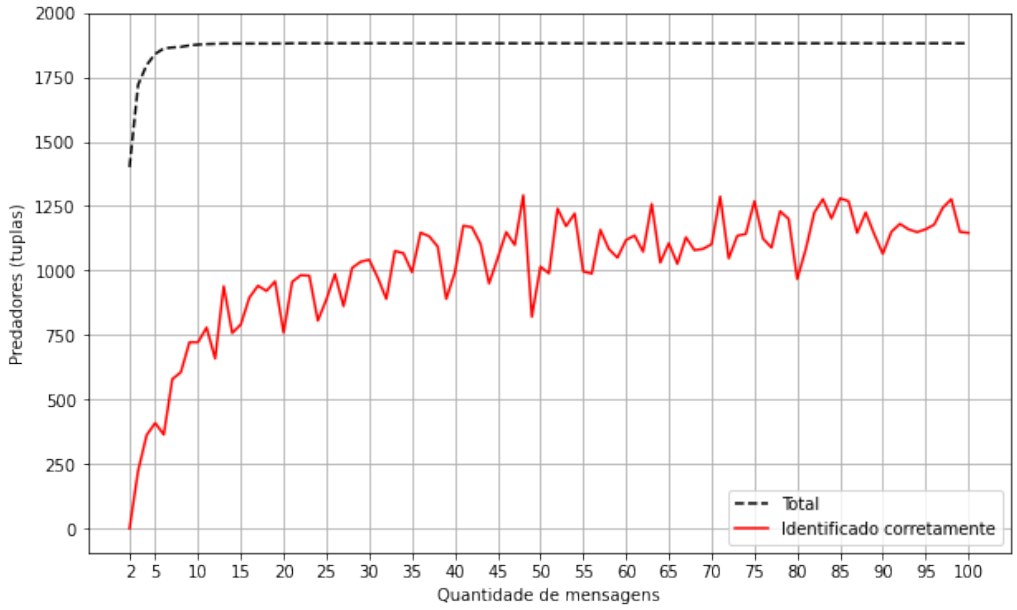


Figura 4.46: Quantidade de predadores (tuplas) corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 2.

A Figura 4.47 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem. Após execução das etapas de pré-processamento e pré-filtro, restaram 236 predadores sexuais e são necessárias 6 mensagens para que hajam 236 predadores únicos identificáveis.

Com 10 mensagens, o algoritmo foi capaz de detectar 171 predadores dos 236 possíveis. Com 20 mensagens, foram detectados 179 predadores, e com 50 mensagens, 206 predadores.

Em relação ao total de 236 predadores da base, 72,46% dos predadores puderam ser identificados com apenas 10 mensagens. Com 20 mensagens, foram detectados 75,85% dos predadores, e com 50 mensagens, foram detectados 87,28% dos predadores sexuais.

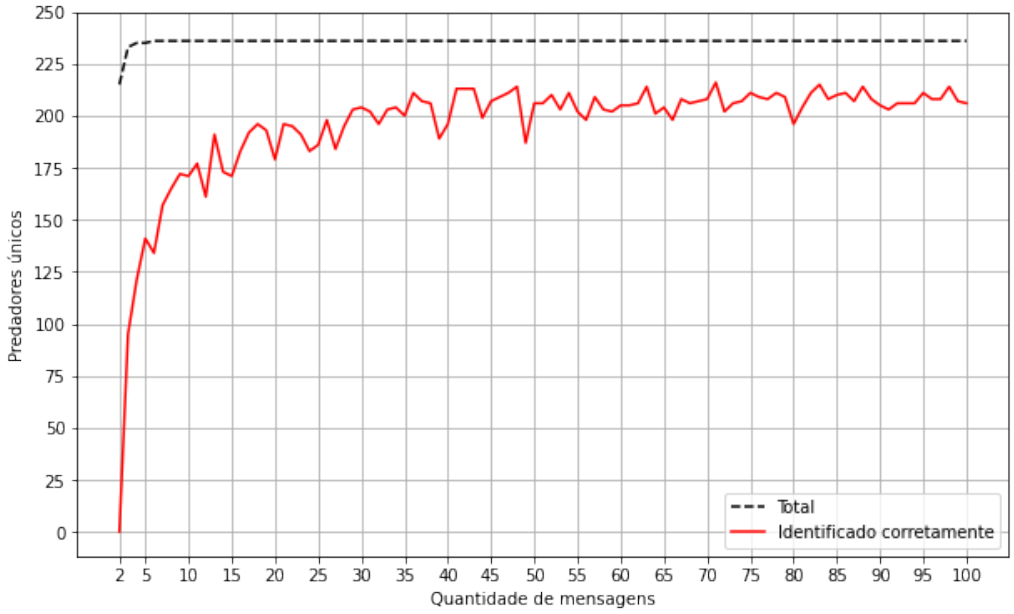


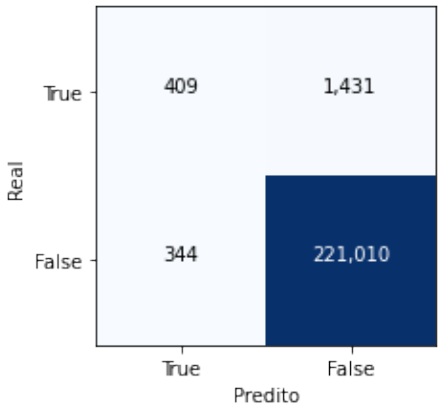
Figura 4.47: Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 2.

A Tabela 4.7 exibe os resultados detalhados para algumas quantidades de mensagens.

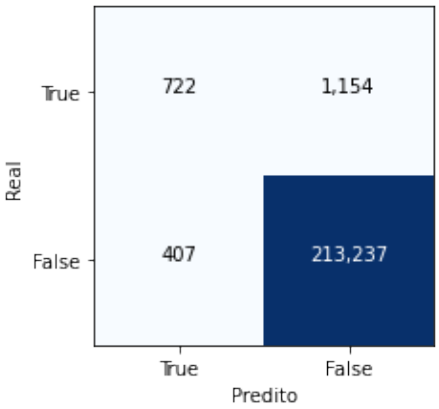
Tabela 4.7: Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 2.

Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,9920	0,5432	0,2223	0,3155	0,4215
10	0,9928	0,6395	0,3849	0,4805	0,5648
20	0,9939	0,8205	0,4037	0,5412	0,6801
24	0,9941	0,8241	0,4285	0,5638	0,6957
50	0,9949	0,8286	0,5396	0,6536	0,7484
100	0,9950	0,7807	0,6093	0,6844	0,7391

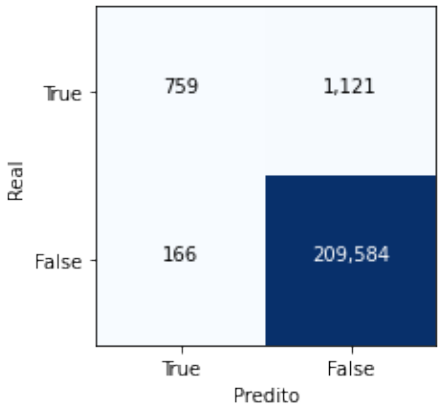
Por fim, a Figura 4.48 exibe as matrizes de confusão para algumas quantidades de mensagens.



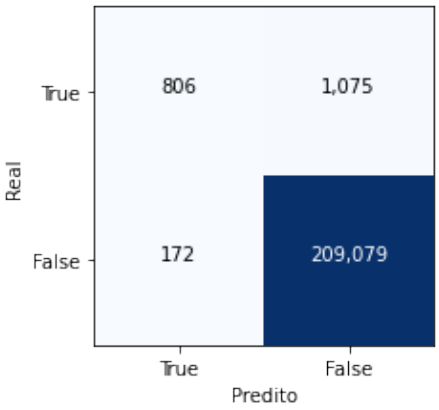
(a) 5 mensagens



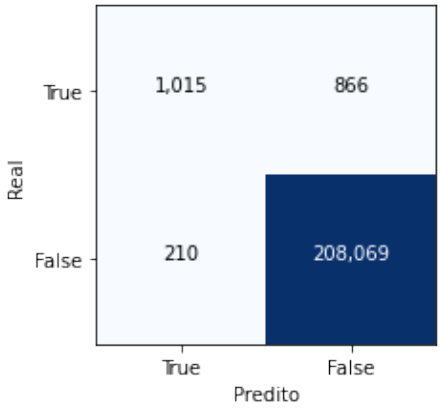
(b) 10 mensagens



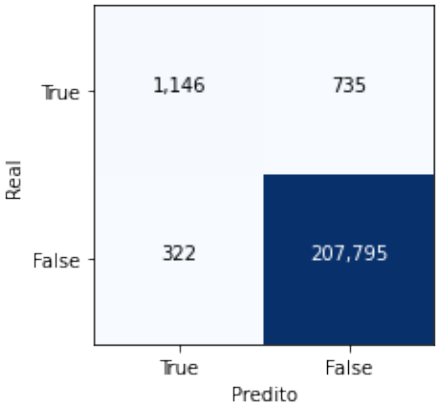
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.48: Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 2.

4.8.2 Com *Undersampling*

Treinamento

Os resultados de acurácia e precisão podem ser visualizados na Figura 4.49 e Figura 4.50.

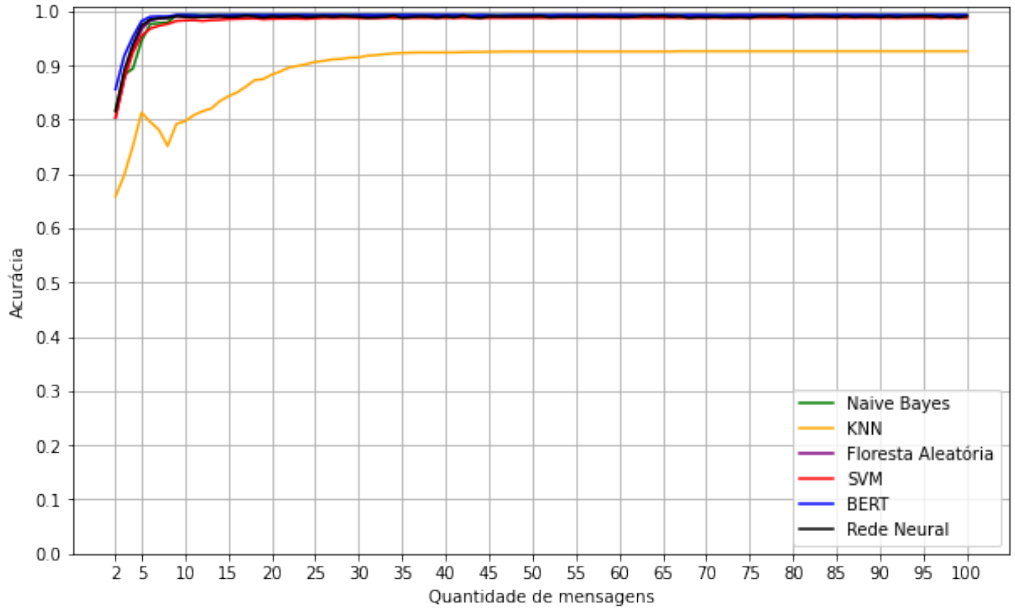


Figura 4.49: Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2.

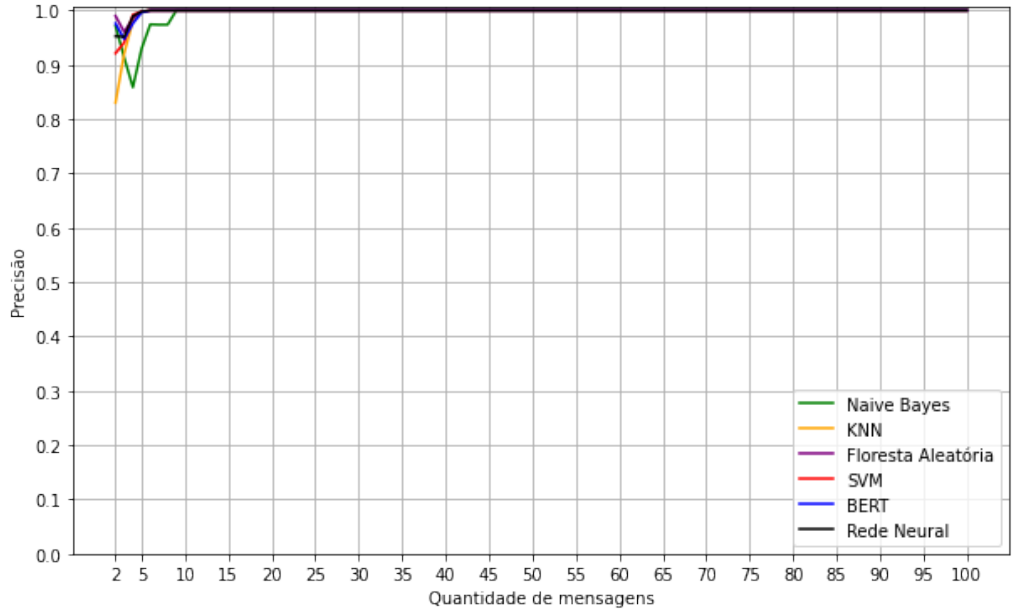


Figura 4.50: Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2.

No geral, a partir de 10 mensagens a acurácia manteve-se acima de 98%. O algoritmo KNN apresentou resultados mais baixos, estabilizando por volta de 92% de acurácia a partir de 33 mensagens.

A precisão manteve-se estável em 100% para todos os algoritmos a partir de 10 mensagens.

Para a métrica *recall* (Figura 4.51), com exceção do KNN, todos os algoritmos mantiveram-se estáveis a partir de 10 mensagens. O BERT e a Rede Neural MLP atingiram 98,52% e 97,79% de *recall* com 10 mensagens, respectivamente, enquanto que o KNN precisou de 37 mensagens para estabilizar os resultados em 84,70%.

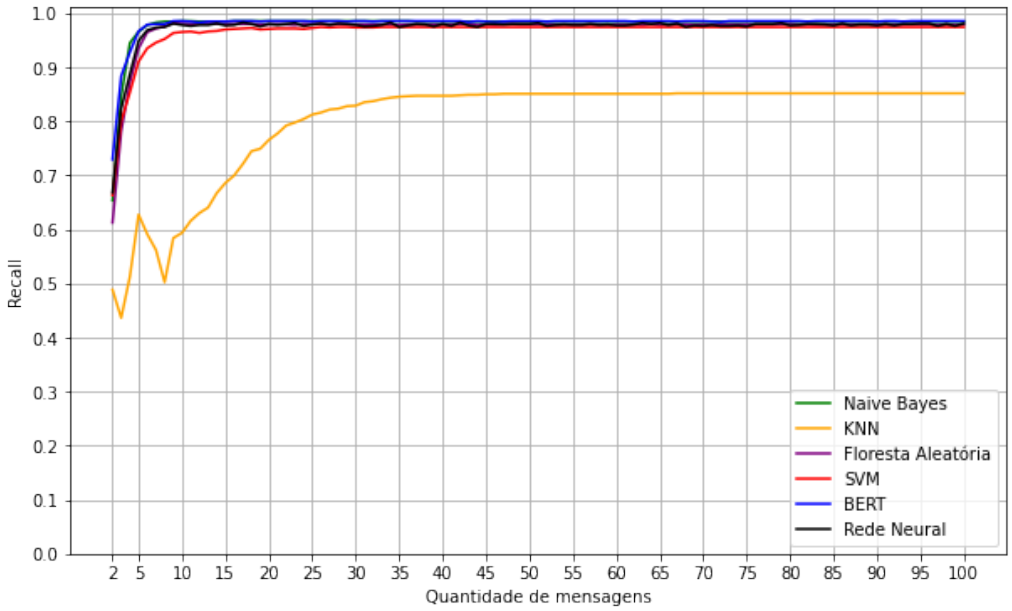


Figura 4.51: Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2.

Para as métricas F_1 (Figura 4.52) e $F_{0.5}$ (Figura 4.53) os resultados seguiram o mesmo padrão.

O KNN precisou de aproximadamente 37 mensagens para estabilizar os resultados, obtendo neste ponto $F_1 = 91,69\%$ e $F_{0.5} = 96,49\%$. O algoritmo SVM obteve $F_1 = 98,21\%$ e $F_{0.5} = 99,28\%$ para as primeiras 10 mensagens, e a Rede Neural MLP obteve $F_1 = 98,88\%$ e $F_{0.5} = 99,55\%$ para a mesma quantidade.

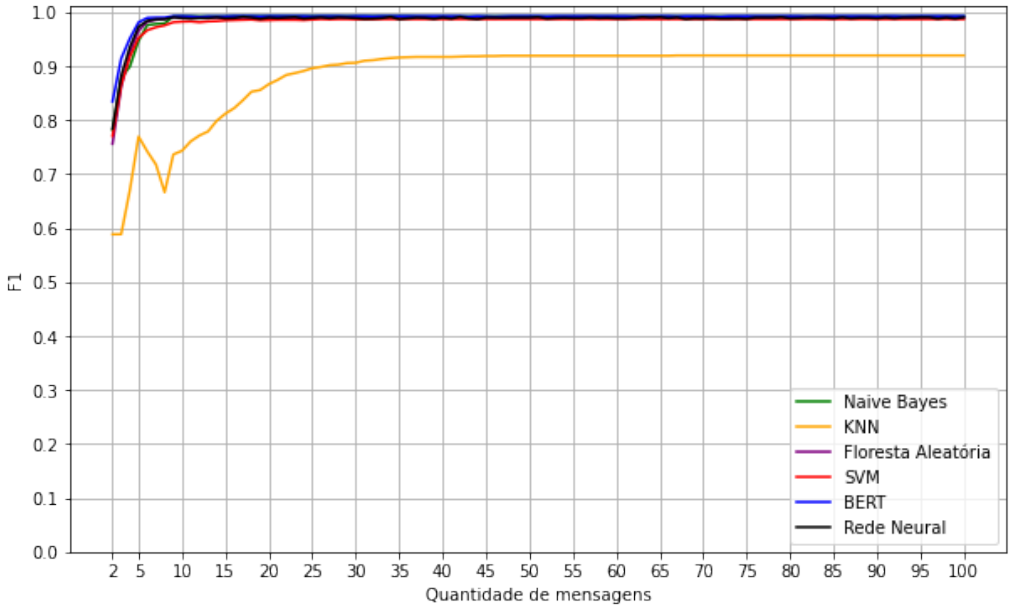


Figura 4.52: Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2.

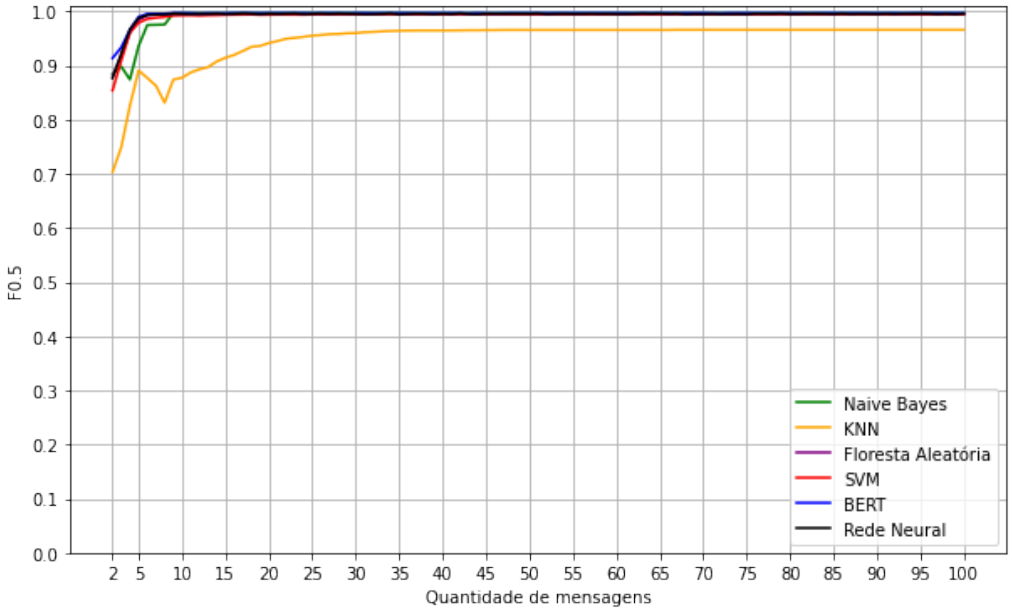


Figura 4.53: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 2.

Para a fase de testes, o algoritmo Rede Neural MLP foi selecionado por obter os melhores resultados de $F_{0.5}$, no geral.

Teste

Os resultados de acurácia e precisão podem ser visualizados na Figura 4.54 e Figura 4.55.

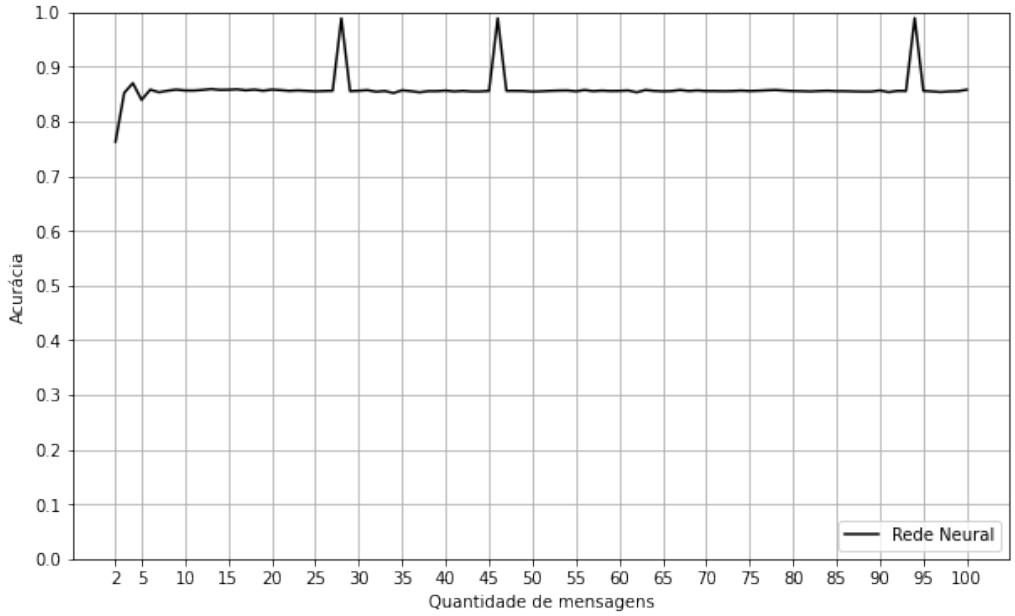


Figura 4.54: Resultado final da acurácia para o experimento com *undersampling* para a estratégia 2.

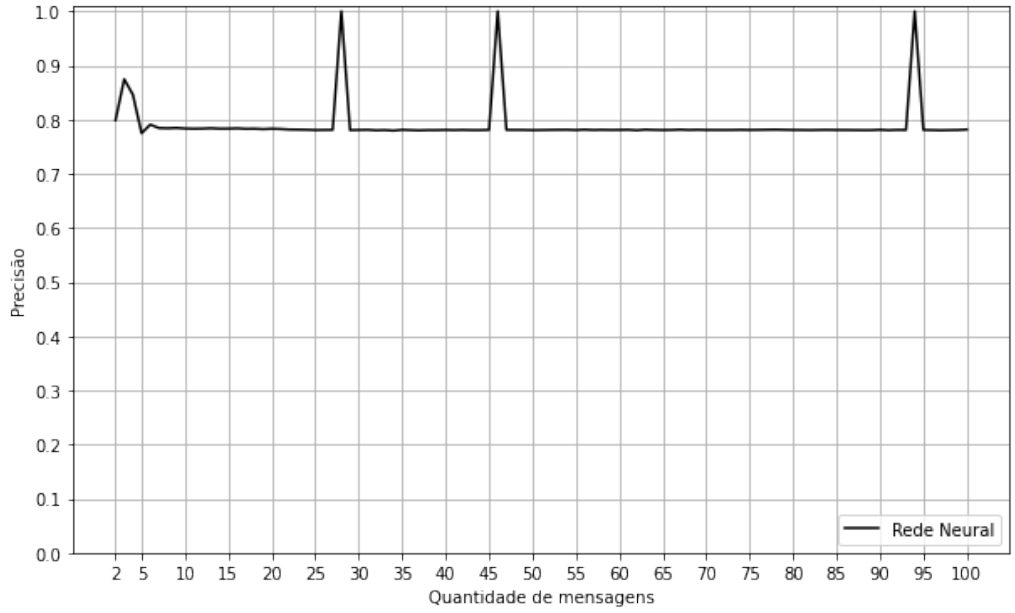


Figura 4.55: Resultado final da precisão para o experimento com *undersampling* para a estratégia 2.

Ambos apresentaram resultados semelhantes. Com 10 mensagens, o algoritmo obteve 85,66% de acurácia e 78,35% de precisão. Com 50 mensagens, obteve 85,46% de acurácia e 78,05% de precisão.

O *recall*, apresentado na Figura 4.56 teve valores superiores. Com 10 mensagens a Rede Neural MLP conseguiu 98,56% e atingiu a estabilidade.

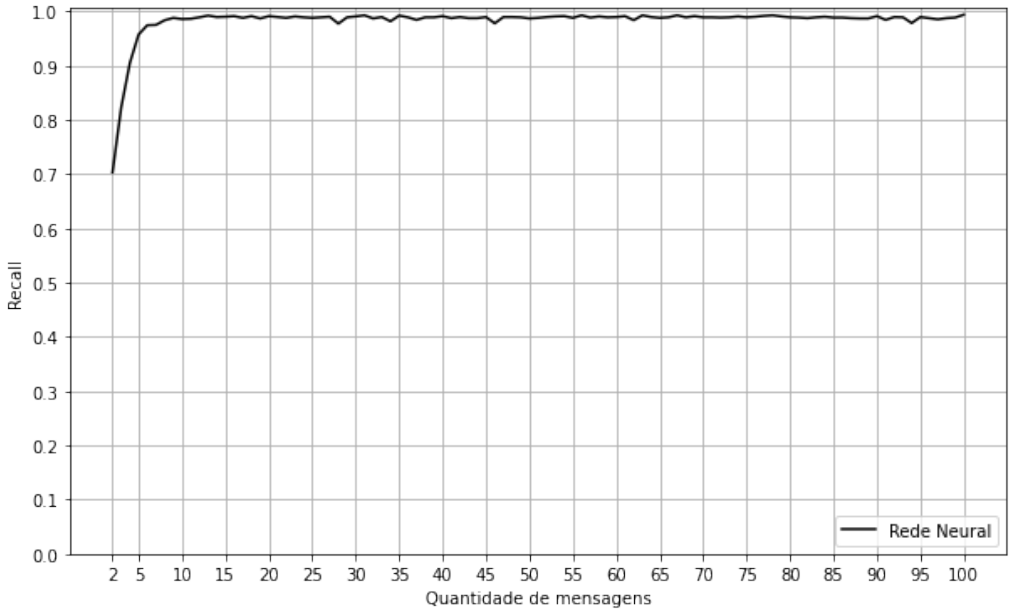


Figura 4.56: Resultado final do *recall* para o experimento com *undersampling* para a estratégia 2.

As métricas F_1 (Figura 4.57) e $F_{0.5}$ (Figura 4.58) também apresentaram resultados semelhantes.

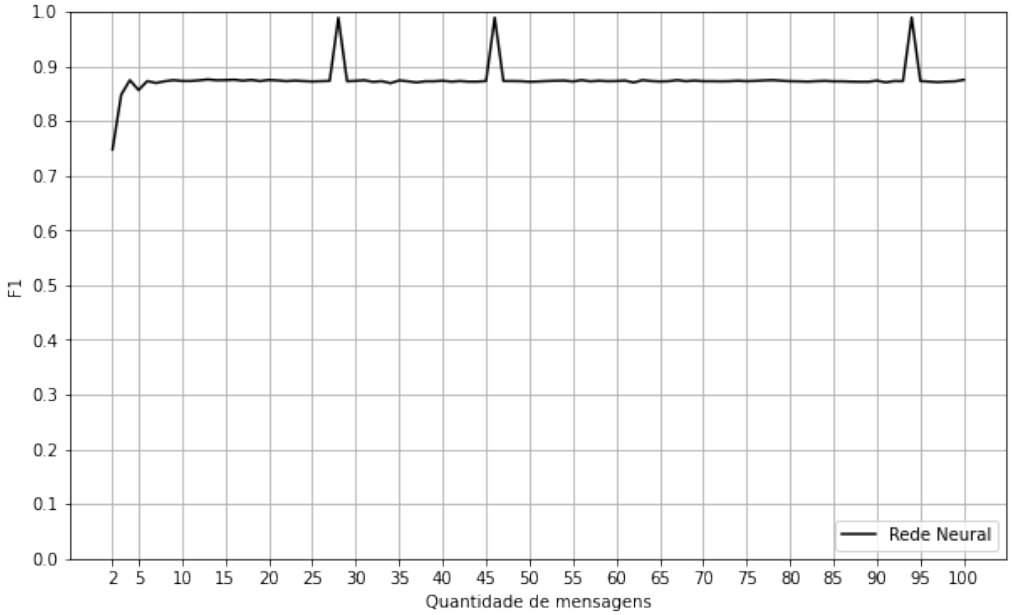


Figura 4.57: Resultado final do F_1 para o experimento com *undersampling* para a estratégia 2.

Com 10 mensagens, o algoritmo obteve $F_1 = 87,30\%$ e $F_{0.5} = 81,70\%$ e com 20 mensagens, o algoritmo obteve $F_1 = 87,49\%$ e $F_{0.5} = 81,74\%$.

Os resultados deste experimentos também superaram os resultados obtidos por KULSRUD (2019), tendo apresentado resultados superiores com menos mensagens.

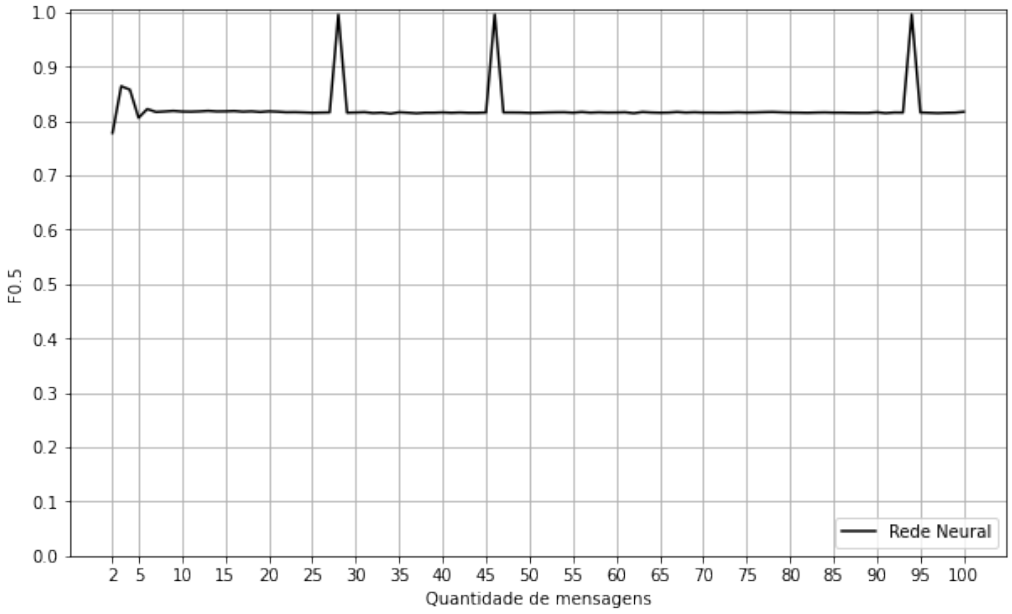


Figura 4.58: Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 2.

A Figura 4.59 exibe a quantidade de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem.

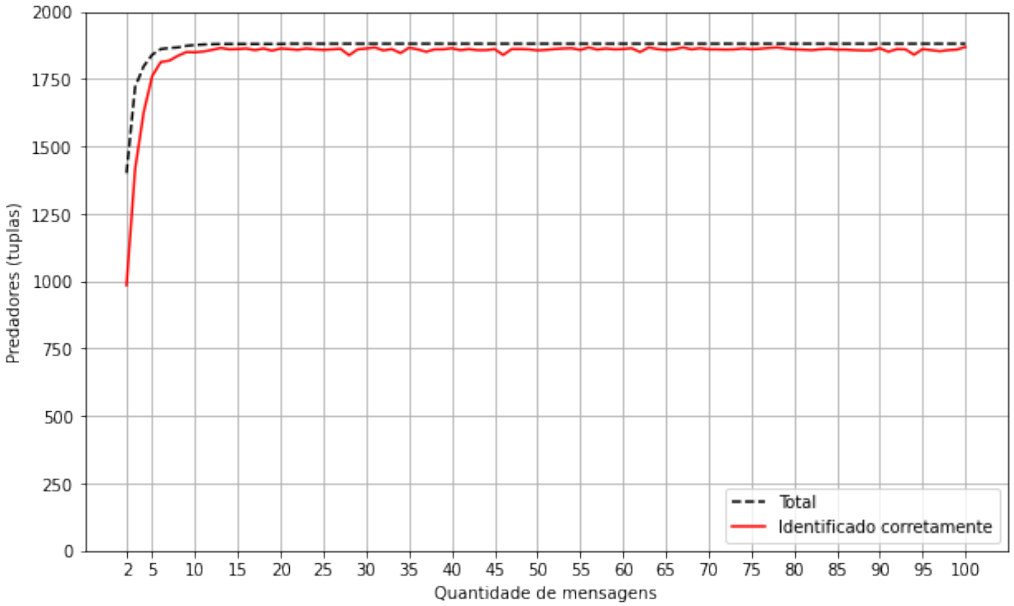


Figura 4.59: Quantidade de predadores (tuplas) corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 2.

Com 10 mensagens, o algoritmo classificou corretamente 1.849 autores como predadores sexuais, de um total de 1.876 possíveis de serem identificados.

Com 20 mensagens, o algoritmo classificou corretamente 1.863 autores como predadores sexuais, de um total de 1.880 tuplas possíveis de serem identificadas.

Com 50 mensagens foram classificados corretamente 1.856 autores como predadores sexuais, de um total de 1.881 tuplas possíveis de serem identificadas.

A Figura 4.60 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem. Após execução das etapas de pré-processamento e pré-filtro, restaram 236 predadores sexuais e são necessárias 6 mensagens para que hajam 236 predadores únicos identificáveis.

Com 9 mensagens, o algoritmo foi capaz de detectar 236 predadores dos 236 possíveis. Assim, em relação ao total de 236 predadores da base, 100% dos predadores puderam ser identificados com apenas 9 mensagens.

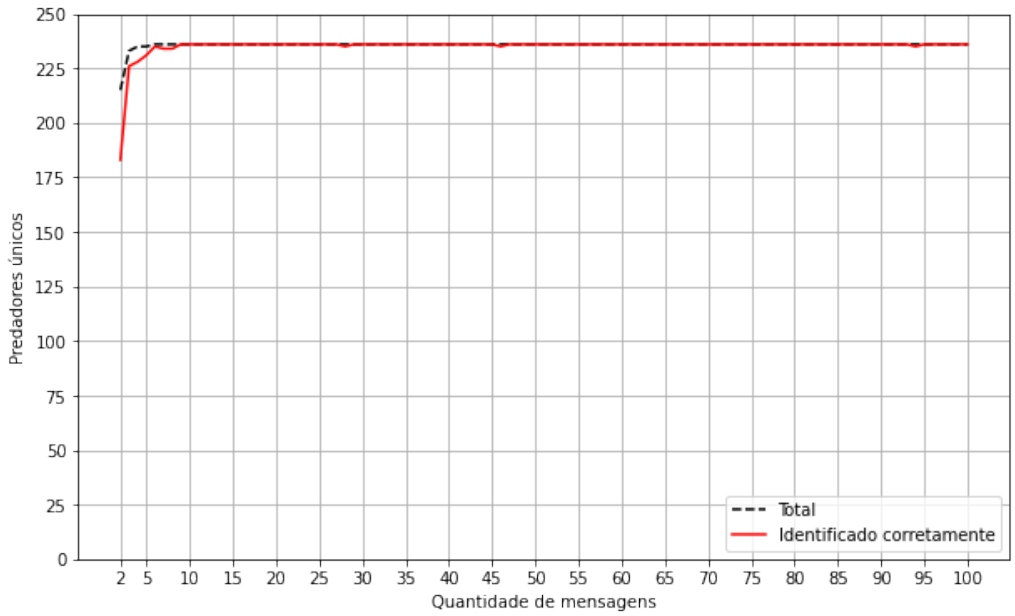


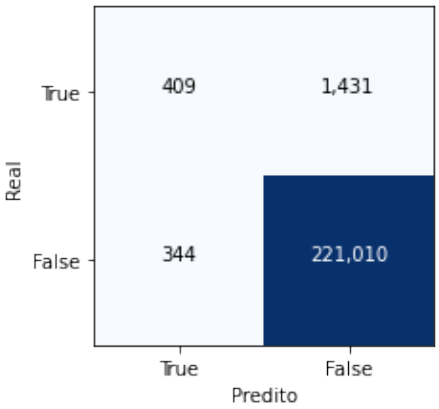
Figura 4.60: Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 2.

A Tabela 4.8 exibe os resultados detalhados para algumas quantidades de mensagens.

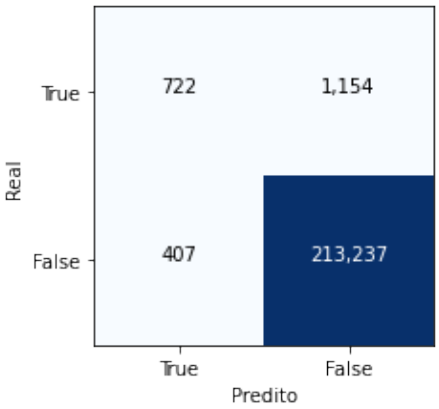
Tabela 4.8: Resultados detalhados das métricas para o experimento com *undersampling* dos dados para a estratégia 2.

Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,8394	0,7747	0,9571	0,8563	0,8054
10	0,8566	0,7835	0,9856	0,8730	0,8170
20	0,8582	0,7831	0,9910	0,8749	0,8174
24	0,8559	0,7812	0,9888	0,8728	0,8154
50	0,8546	0,7805	0,9867	0,8716	0,8145
100	0,8581	0,7817	0,9936	0,8750	0,8165

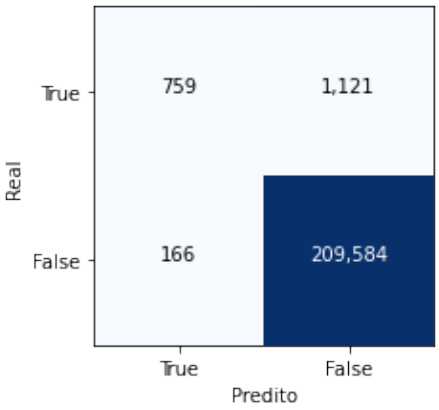
Por fim, a Figura 4.61 exibe as matrizes de confusão para algumas quantidades de mensagens.



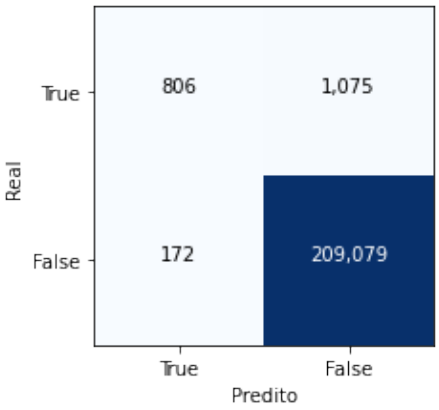
(a) 5 mensagens



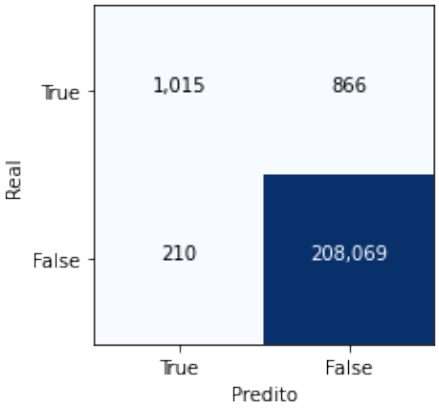
(b) 10 mensagens



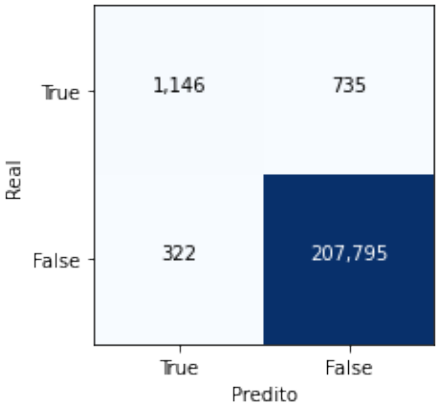
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.61: Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 2.

4.9 Estratégia 3 - Distinguir Conversas Predatórias e Gerais + Distinguir Predador e Vítima

Para implementar a terceira e última estratégia, as duas estratégias anteriores foram combinadas em sequência, ou seja, primeiro foram executados todos os métodos da estratégia 1 (Distinguir Conversas Predatórias e Gerais) para que se pudesse obter as conversas predatórias.

Em seguida, foram executados todos os métodos da estratégia 2 (Distinguir Predador e Vítima) apenas para as conversas predatórias identificadas pelo primeiro algoritmo.

Todos os métodos de agrupamento e remoção de conversas que não tivessem dois autores também foram utilizados nesta estratégia.

Foram realizados experimentos com e sem técnicas de balanceamento, que são detalhados a seguir, e, para todos os experimentos desta estratégia, o algoritmo Rede Neural MLP foi utilizado na primeira etapa, pois, como visto na seção 4.7 deste capítulo, este foi o algoritmo que apresentou melhores resultados para a métrica $F_{0.5}$.

4.9.1 Sem Balanceamento dos Dados

Treinamento

A Figura 4.62 apresenta os resultados para a métrica acurácia.

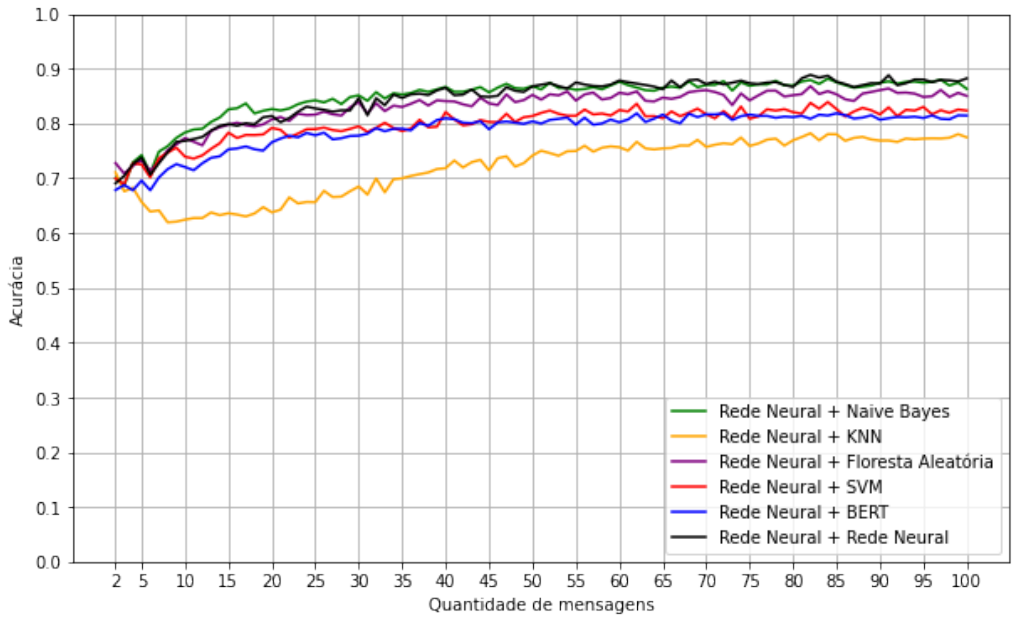


Figura 4.62: Resultado da acurácia para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3.

Com 30 mensagens, a combinação Rede Neural + *Naive Bayes* obteve acurácia de 85,13%. Já a combinação Rede Neural + Floresta Aleatória obteve 83,90% para a mesma quantidade de mensagens.

Os algoritmos obtiveram resultados muito parecidos de precisão para as primeiras 5 mensagens, começando a divergir a partir de 10 mensagens, conforme pode ser visualizado na Figura 4.63. Com 30 mensagens, a combinação Rede Neural + Rede Neural obteve a melhor precisão entre os algoritmos, 83,33%, enquanto que a combinação Rede Neural + KNN obteve o pior resultado, com 69,81%.

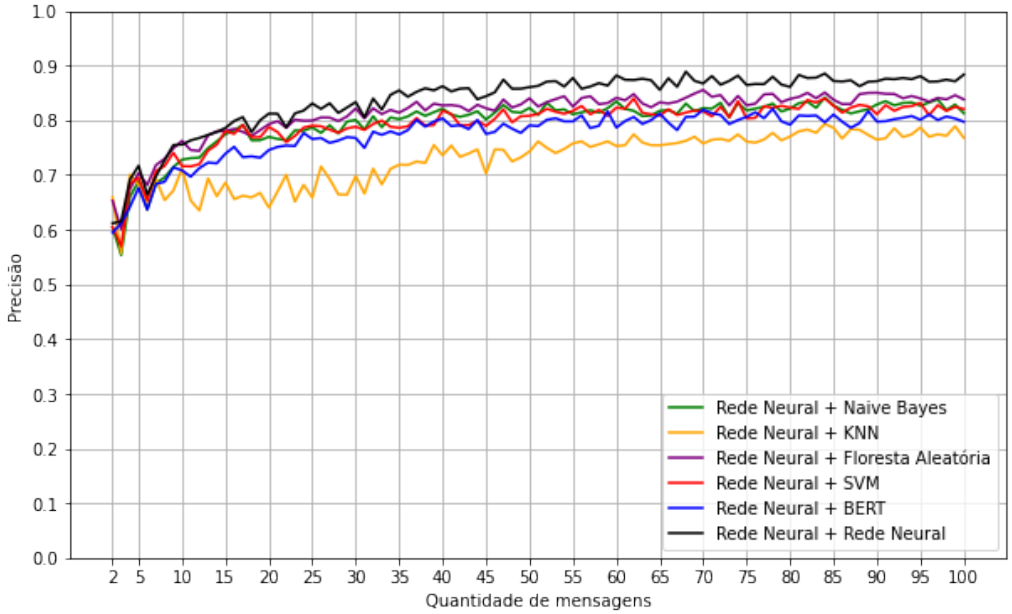


Figura 4.63: Resultado da precisão para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3.

Para o *recall*, apresentada na Figura 4.64, a combinação Rede Neural + *Naive Bayes* obteve melhores resultados. Com apenas 10 mensagens já foi possível obter um *recall* de 82,35%, com 20 mensagens, 87,90%, e com 40 mensagens, 91,88%.

As combinações Rede Neural + Rede Neural e Rede Neural + Floresta Aleatória tiveram resultados parecidos, obtendo respectivamente, 83,83% e 83,52% de *recall* para as primeiras 30 mensagens.

Os resultados das métricas F_1 e $F_{0.5}$ podem ser visualizados na Figura 4.65 e na Figura 4.66, respectivamente. Em ambas as métricas o melhores resultados foram obtidos pelas combinações Rede Neural + *Naive Bayes* e Rede Neural + Rede Neural.

A combinação Rede Neural + *Naive Bayes* obteve $F_1 = 77,18\%$ e $F_{0.5} = 74,47\%$ para as primeiras 10 mensagens, $F_1 = 82,01\%$ e $F_{0.5} = 78,90\%$ para as 20 primeiras e $F_1 = 86,80\%$ e $F_{0.5} = 84,01\%$ para as 50 primeiras.

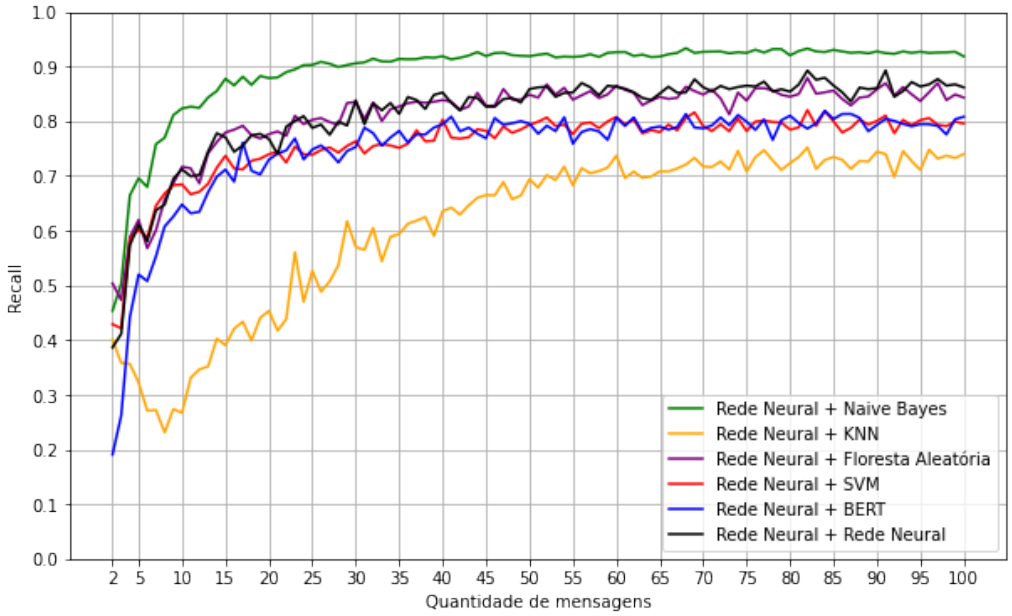


Figura 4.64: Resultado do *recall* para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3.

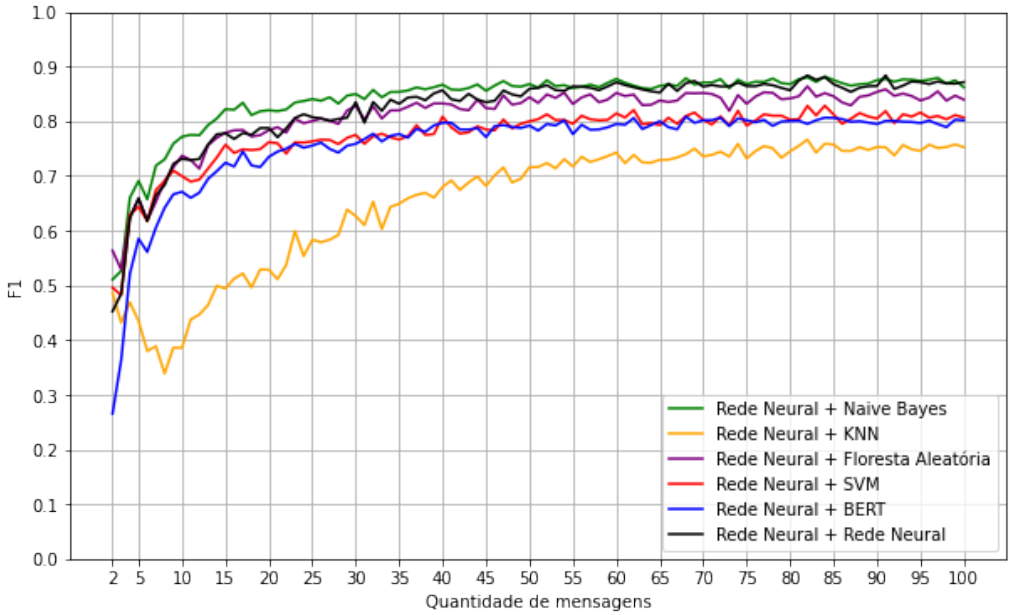


Figura 4.65: Resultado do F_1 para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3.

A combinação Rede Neural + Rede Neural obteve $F_1 = 73,14\%$ e $F_{0.5} = 74,52\%$ para as primeiras 10 mensagens, $F_1 = 78,76\%$ e $F_{0.5} = 80,17\%$ para as primeiras 20 mensagens e $F_1 = 85,94\%$ e $F_{0.5} = 86\%$ para as primeiras 50 mensagens.

A combinação Rede Neural + KNN apresentou os piores resultados, não chegando a atingir 80% de F_1 ou $F_{0.5}$.

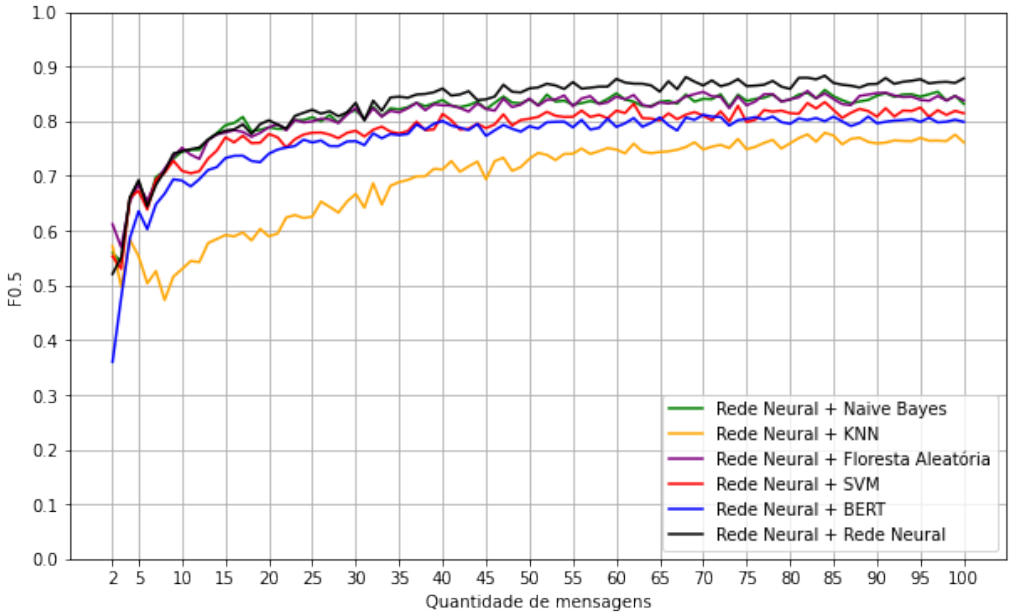


Figura 4.66: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento sem balanceamento dos dados para a estratégia 3.

Para a fase de testes, a combinação Rede Neural + Rede Neural foi selecionada por apresentar melhores resultados de $F_{0.5}$ em relação aos demais algoritmos.

Teste

O resultado da acurácia na base de testes pode ser observado na Figura 4.67. Com 20 mensagens, foi obtido 77,34%.

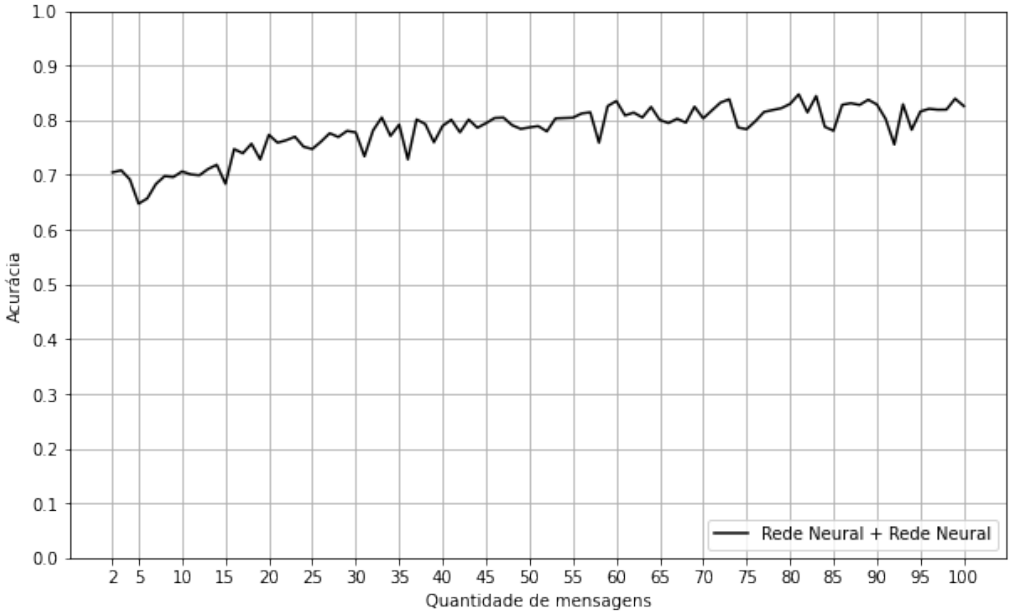


Figura 4.67: Resultado final da acurácia para o experimento sem balanceamento dos dados para a estratégia 3.

O resultado da precisão pode ser visualizado na Figura 4.68. Inicialmente, a precisão apresentou valores mais baixos, com 73,40% para as primeiras 10 conversas, chegando a 89,03% para 100 mensagens.

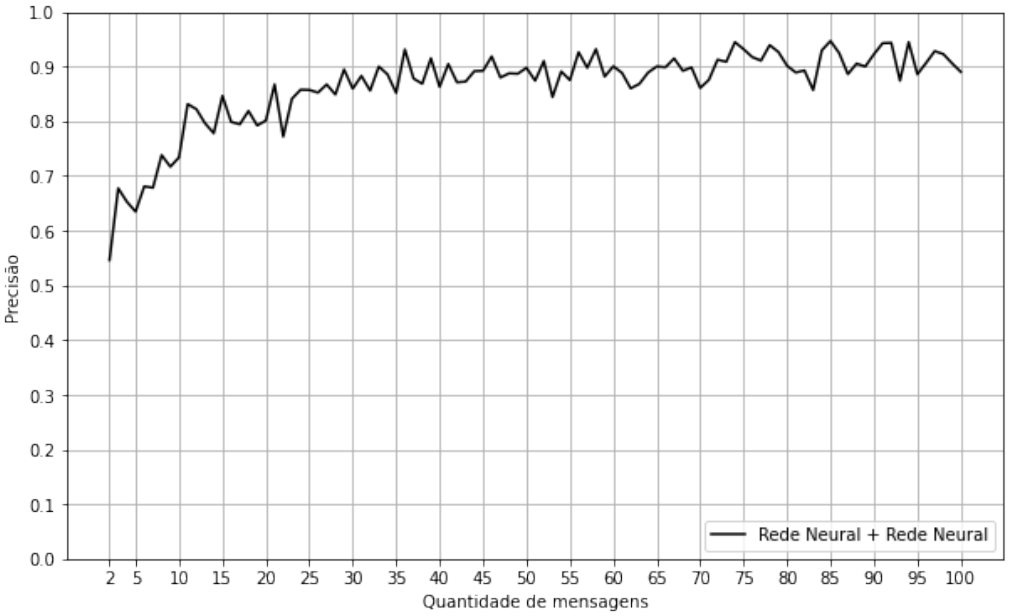


Figura 4.68: Resultado final da precisão para o experimento sem balanceamento dos dados para a estratégia 3.

O *recall* (Figura 4.69) também apresentou resultados muito variáveis, com valores mais baixos nas primeiras mensagens. Com 10 mensagens obteve-se *recall* de 48,01%, enquanto que o maior valor encontrado para a métrica foi 81,47% com 83 mensagens.

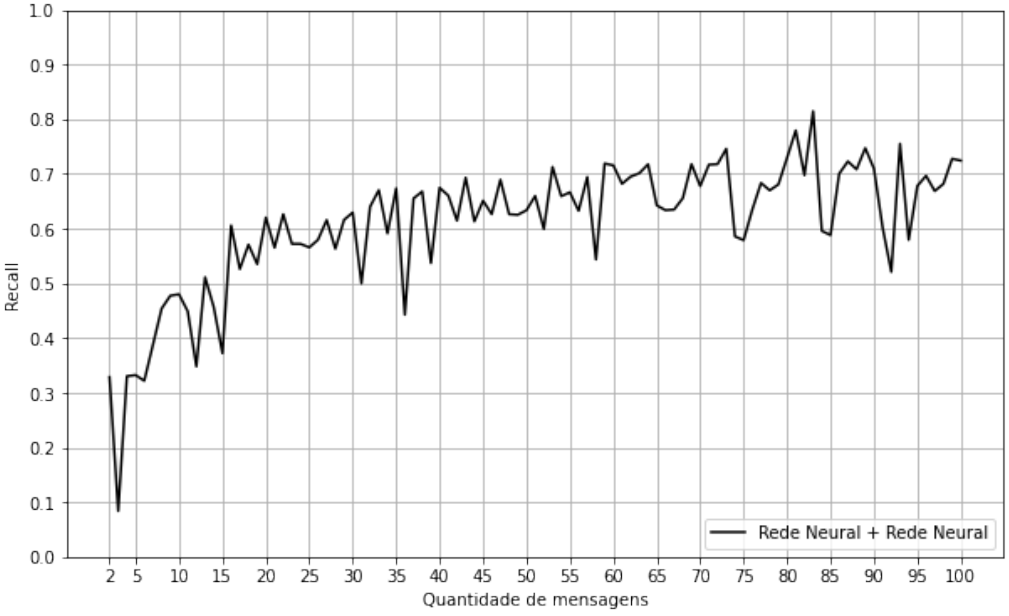


Figura 4.69: Resultado final do *recall* para o experimento sem balanceamento dos dados para a estratégia 3.

O resultado para a métrica F_1 pode ser observado no gráfico da Figura 4.70. Com 10 mensagens, o classificador obteve $F_1 = 58,05\%$. Com 20 mensagens, obteve $F_1 = 69,93\%$ e com 30 mensagens, obteve $F_1 = 72,68\%$.

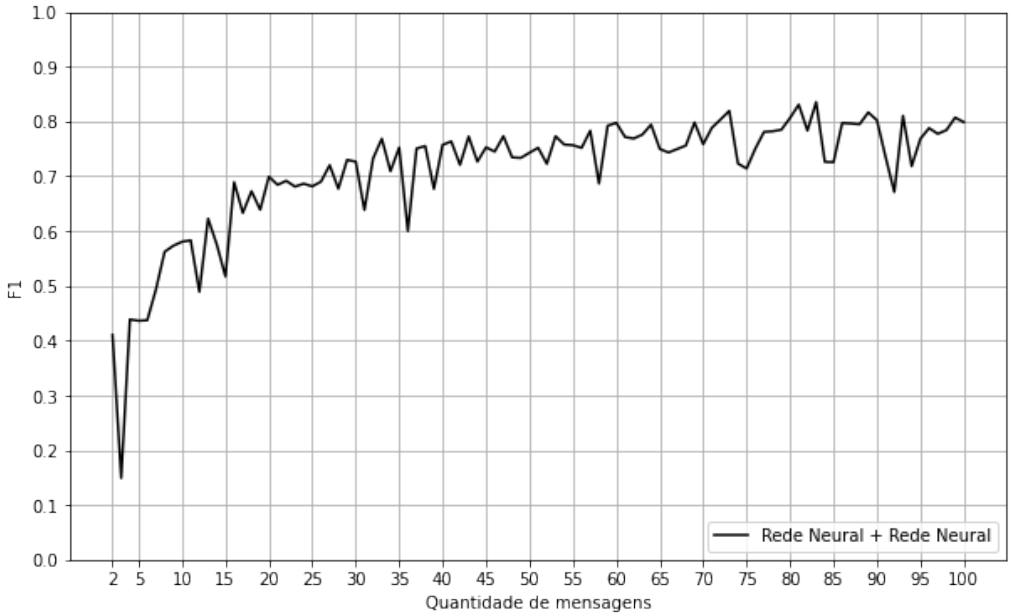


Figura 4.70: Resultado final do F_1 para o experimento sem balanceamento dos dados para a estratégia 3.

Para a métrica $F_{0.5}$, foi obtido o resultado da Figura 4.71.

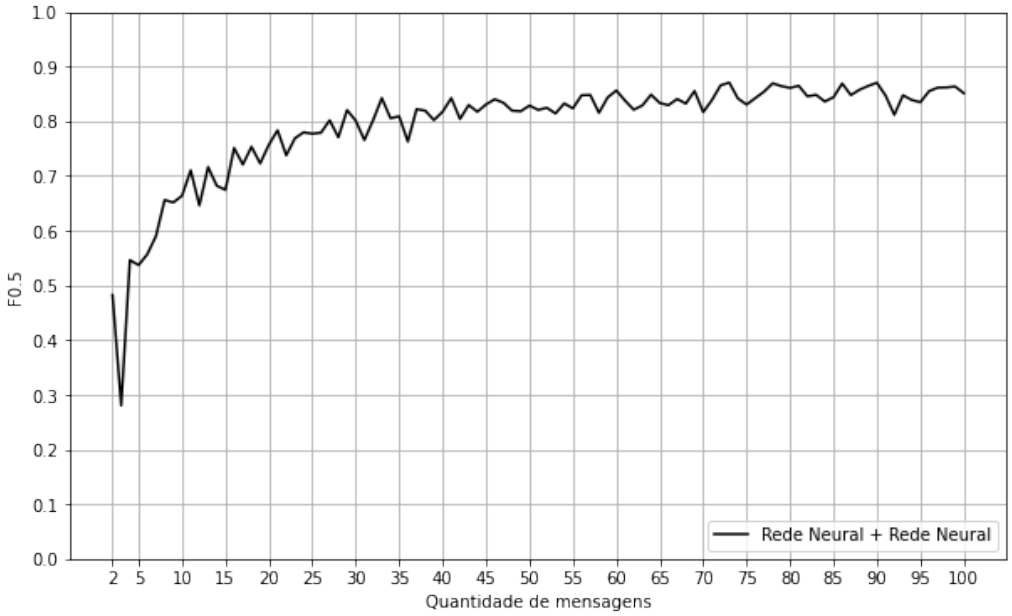


Figura 4.71: Resultado final do $F_{0.5}$ para para o experimento sem balanceamento dos dados para a estratégia 3.

Com 10 mensagens, a combinação Rede Neural + Rede Neural obteve $F_{0.5} = 66,38\%$. Com 20 mensagens obteve $F_{0.5} = 75,71\%$, com 24 mensagens obteve

$F_{0.5} = 77,98\%$, e com 30 mensagens obteve $F_{0.5} = 80,10\%$.

Este experimento não superou o resultado de KULSRUD (2019), necessitando também de mais 24 mensagens para apresentar resultados superiores a 80% de $F_{0.5}$.

Foram gerados gráficos para as quantidades de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem, e para as quantidades de predadores únicos corretamente identificados para cada quantidade de mensagem. Como informado anteriormente, a quantidade de tuplas de predadores não é igual a quantidade de predadores únicos.

A Figura 4.72 exibe a quantidade de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem. Na linha tracejada estão as quantidades totais de predadores (tuplas) existentes para cada quantidade de mensagem, representando assim, as quantidades máximas que poderiam ser alcançadas pelo algoritmo para cada mensagem. Neste experimento, a quantidade de tuplas de predadores possíveis de serem identificadas depende da quantidade de conversas classificadas como predatórias pelo primeiro classificador. Assim, a quantidade de tuplas possíveis de serem identificadas é igual a quantidade de conversas predatórias classificadas corretamente pelo primeiro classificador.

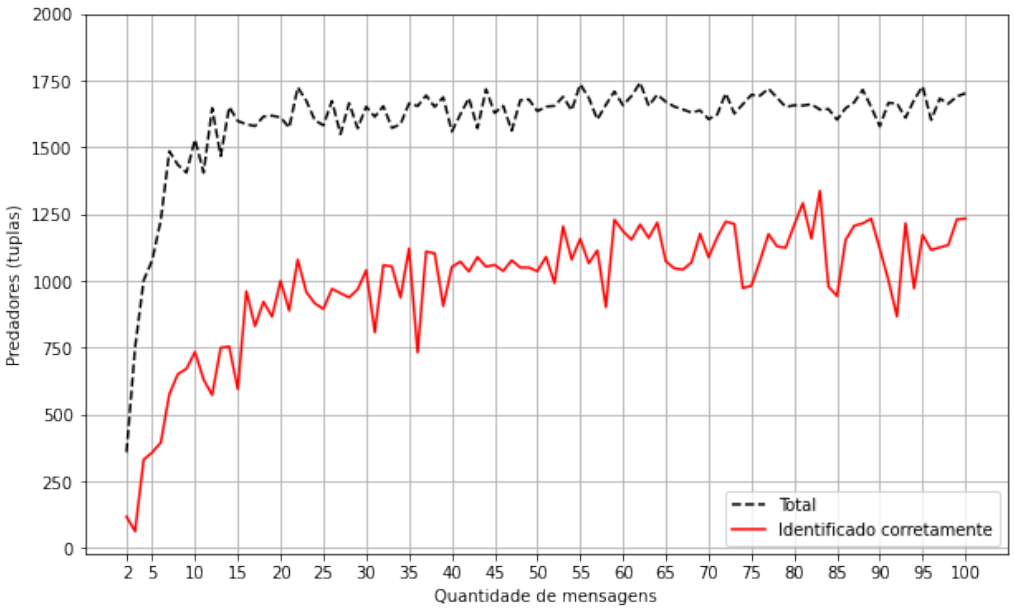


Figura 4.72: Quantidade de predadores (tuplas) corretamente identificadas para o experimento sem balanceamento dos dados para a estratégia 3.

Com 10 mensagens, o algoritmo classificou corretamente 734 autores como predadores sexuais, de um total de 1.529 possíveis de serem identificados.

Com 20 mensagens, o algoritmo classificou corretamente 1.000 autores como predadores sexuais, de um total de 1.612 tuplas possíveis de serem identificadas. Com 50 mensagens foram classificados corretamente 1.036 autores como predadores sexuais, de um total de 1.635 tuplas possíveis de serem identificadas.

A Figura 4.73 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem. Após execução das etapas de pré-processamento e pré-filtro, restaram 236 predadores sexuais. Como esta estratégia depende das conversas que foram classificadas como predatórias pelo primeiro classificador, a quantidade de predadores únicos totais pode ser menor. Neste experimento, com 100 mensagens, só é possível identificar até 230 predadores sexuais.

Com 10 mensagens, o algoritmo foi capaz de detectar 173 predadores dos 221 possíveis. Com 20 mensagens, foram detectados 197 predadores de 227 possíveis, e com 50 mensagens, 202 predadores de 224 possíveis.

Em relação ao total de 236 predadores da base, 73,31% dos predadores puderam ser identificados com apenas 10 mensagens. Com 20 mensagens, foram detectados 83,47% dos predadores, e com 50 mensagens, foram detectados 85,59% dos predadores sexuais.

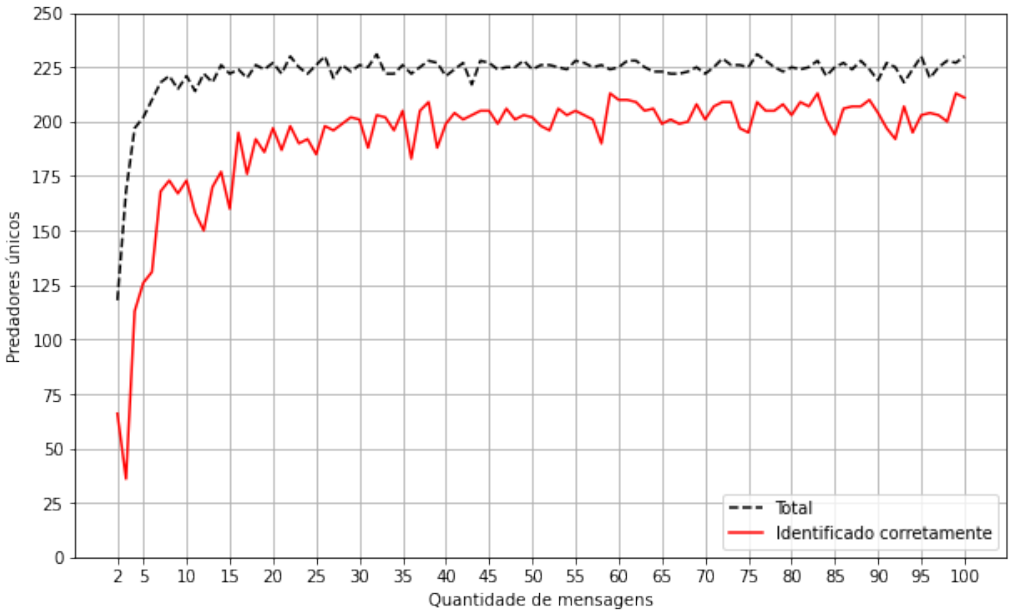


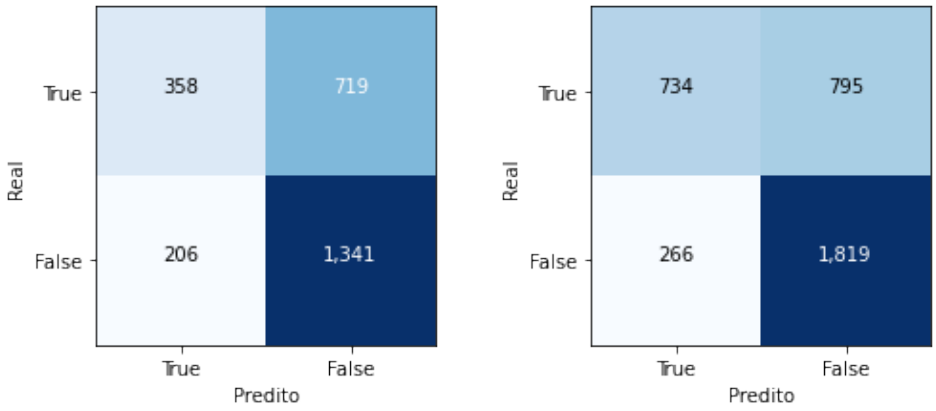
Figura 4.73: Quantidade de predadores únicos corretamente identificados para o experimento sem balanceamento dos dados para a estratégia 3.

A Tabela 4.9 exibe os resultados detalhados para algumas quantidades de mensagens.

Tabela 4.9: Resultados detalhados das métricas para o experimento sem balanceamento dos dados para a estratégia 3.

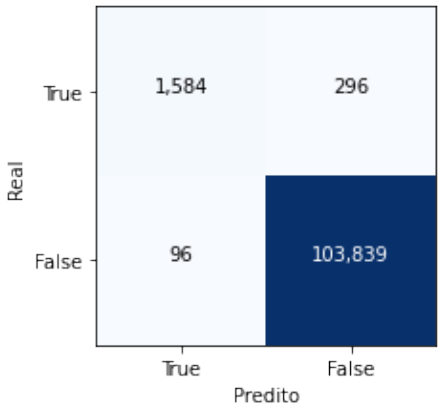
Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,6475	0,6348	0,3324	0,4363	0,5371
10	0,7064	0,7340	0,4801	0,5805	0,6638
20	0,7734	0,8013	0,6203	0,6993	0,7571
24	0,7516	0,8577	0,5721	0,6864	0,7798
50	0,7870	0,8977	0,6336	0,7429	0,8287
100	0,8260	0,8903	0,7244	0,7988	0,8513

Por fim, a Figura 4.74 exibe as matrizes de confusão para algumas quantidades de mensagens.

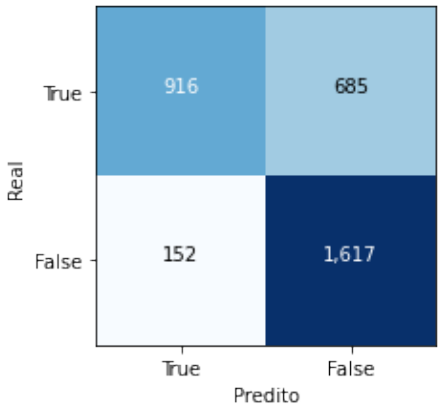


(a) 5 mensagens

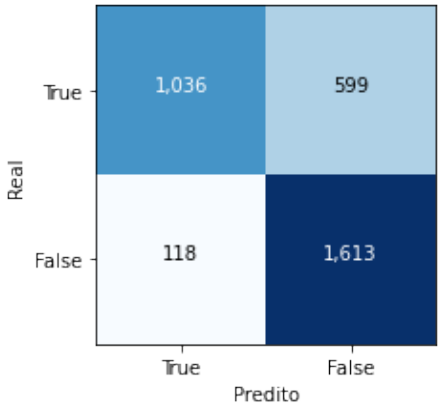
(b) 10 mensagens



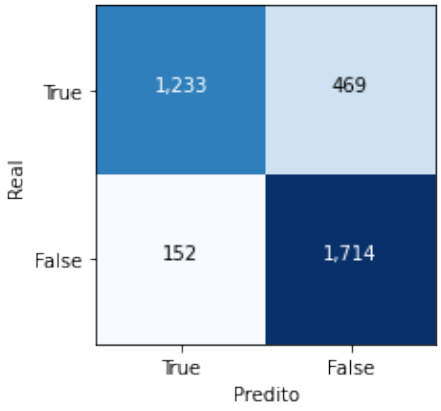
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.74: Matrizes de confusão para o experimento sem balanceamento dos dados para a estratégia 3.

4.9.2 Com *Undersampling*

Treinamento

Os resultados de acurácia e precisão do treinamento para o experimento com *undersampling* para a estratégia 3 podem ser visualizados na Figura 4.75 e Figura 4.76.

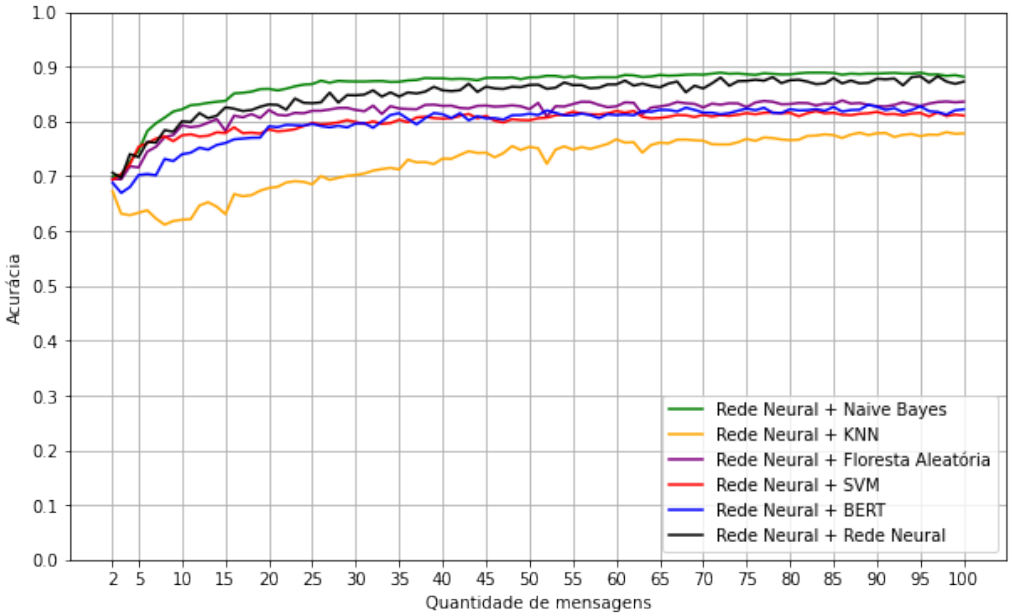


Figura 4.75: Resultado da acurácia para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3.

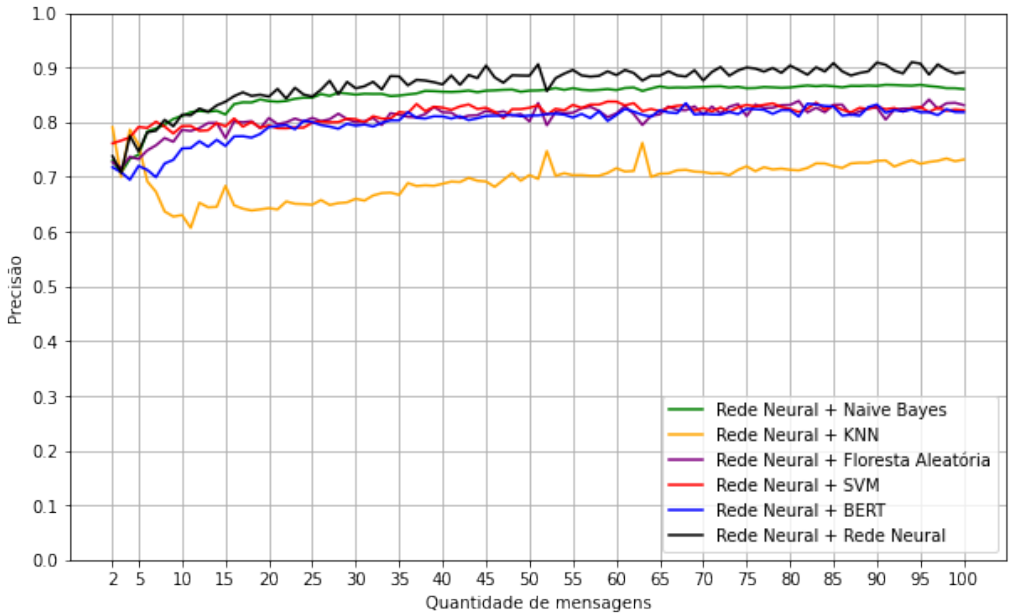


Figura 4.76: Resultado da precisão para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3.

A combinação Rede Neural + *Naive Bayes* apresentou os melhores resultados para a métrica acurácia, obtendo 87,29% para as primeiras 30 mensagens, e 88,16% para as primeiras 100 mensagens.

Já para a precisão, a combinação Rede Neural + Rede Neural obteve os melhores resultados, obtendo 86,19% de precisão com 30 mensagens, e 89,13% de precisão com 100 mensagens.

A combinação Rede Neural + SVM obteve 77,47% de acurácia e 79,30% de precisão para as primeiras 10 mensagens e 80,22% de acurácia e 82,65% de precisão para 50 mensagens.

Os resultados mais baixos de acurácia e precisão foram da combinação Rede Neural + KNN. Com 10 mensagens, o algoritmo obteve 62,07% de acurácia e 63,04% de precisão, e com 50 mensagens obteve 75,37% de acurácia e 70,36% de precisão.

A combinação Rede Neural + *Naive Bayes* também apresentou os melhores resultados para o *recall*, conforme pode ser observado na Figura 4.77. Com 10 mensagens, o classificador obteve 84,15% de *recall* e manteve os resultados acima de 90% de *recall* a partir de 24 mensagens.

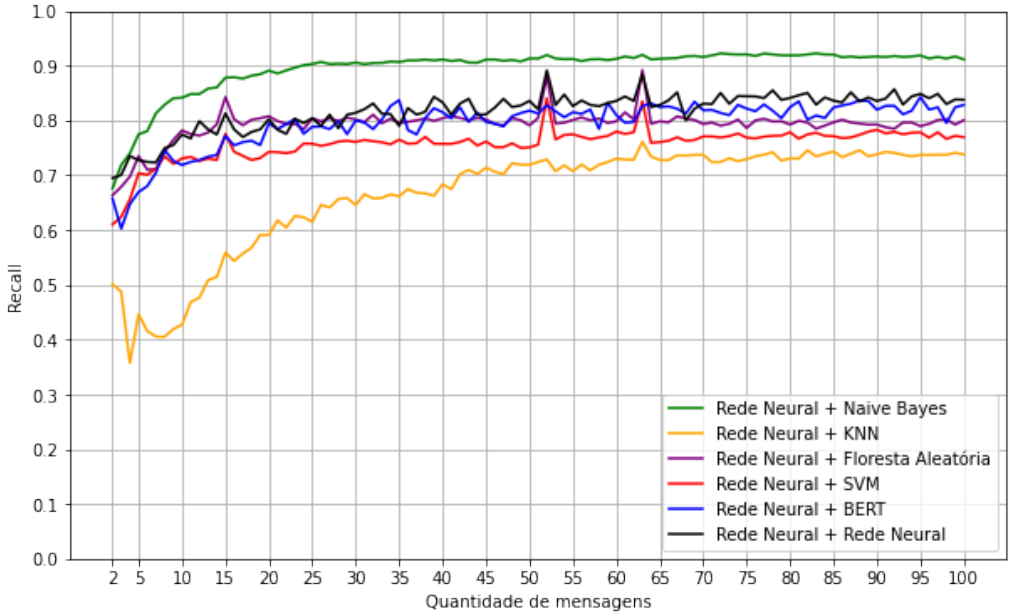


Figura 4.77: Resultado do *recall* para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3.

Os resultados de F_1 e $F_{0.5}$ podem ser visualizados na Figura 4.78 e Figura 4.79, respectivamente.

A combinação Rede Neural + *Naive Bayes* conseguiu os melhores resultados de F_1 e $F_{0.5}$. Com 10 mensagens, o algoritmo obteve $F_1 = 82,45\%$ e $F_{0.5} = 81,62\%$, com 25 mensagens obteve $F_1 = 87,19\%$ e $F_{0.5} = 85,53\%$, e com 50 mensagens obteve $F_1 = 88,32\%$ e $F_{0.5} = 86,74\%$.

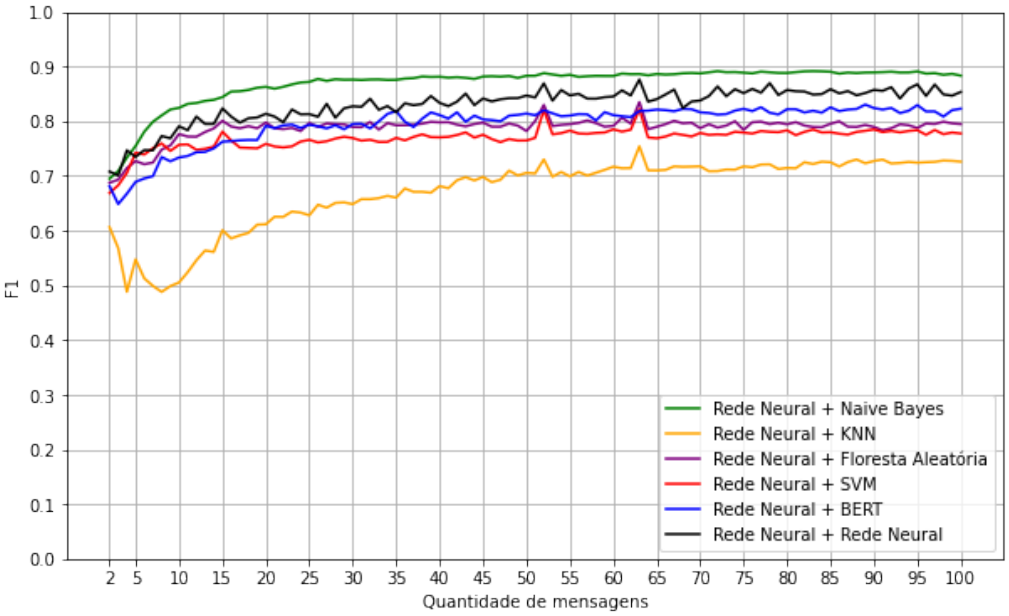


Figura 4.78: Resultado do F_1 para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3.

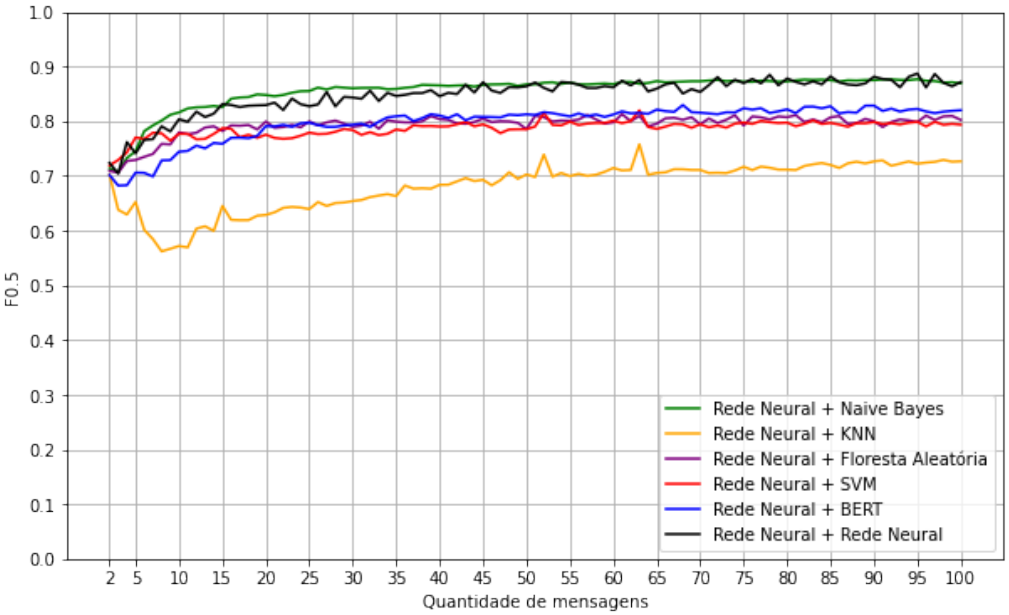


Figura 4.79: Resultado do $F_{0.5}$ para diferentes quantidades de mensagens para o experimento com *undersampling* para a estratégia 3.

A combinação Rede Neural + Rede Neural obteve $F_1 = 79,08\%$ e $F_{0.5} = 80,38\%$ para as primeiras 10 mensagens, enquanto que a combinação Rede Neural + BERT obteve $F_1 = 73,38\%$ e $F_{0.5} = 74,44\%$ para as primeiras 10 mensagens.

Por fim, a combinação Rede Neural + *Naive Bayes* foi escolhida para ir para a fase de testes por apresentar os melhores resultados para a métrica $F_{0.5}$ na maioria das mensagens.

Teste

O resultado da acurácia dpode ser visualizado na Figura 4.80. Com 10 mensagens os resultados estabilizaram por volta de 50% de acurácia, obtendo neste ponto 49,97%.

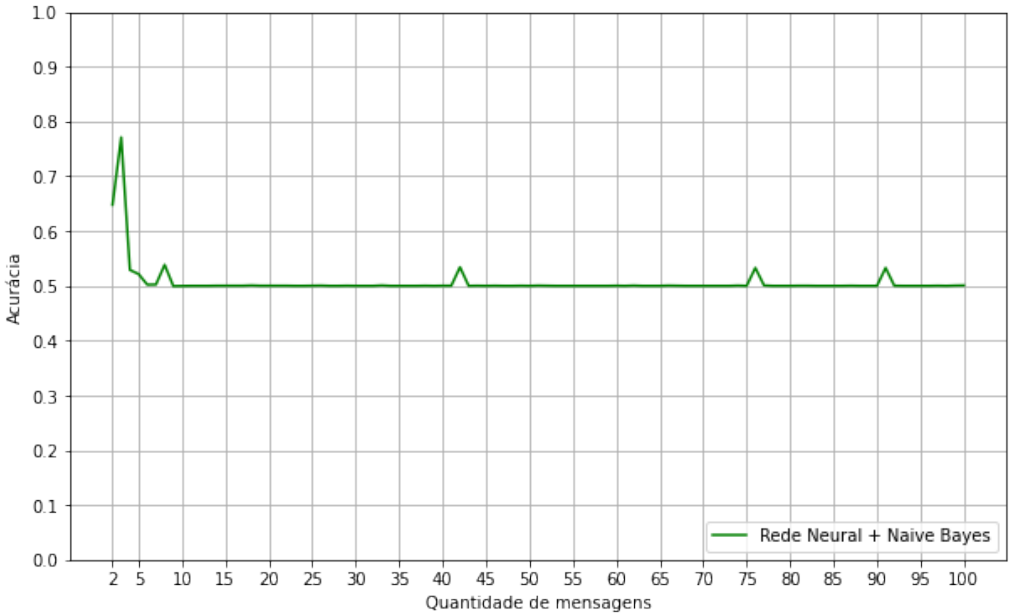


Figura 4.80: Resultado final da acurácia para o experimento com *undersampling* para a estratégia 3.

A precisão, representada pela Figura 4.81, apresentou o mesmo padrão. Com 10 mensagens, a combinação Rede Neural + *Naive Bayes* obteve 49,99% de precisão.

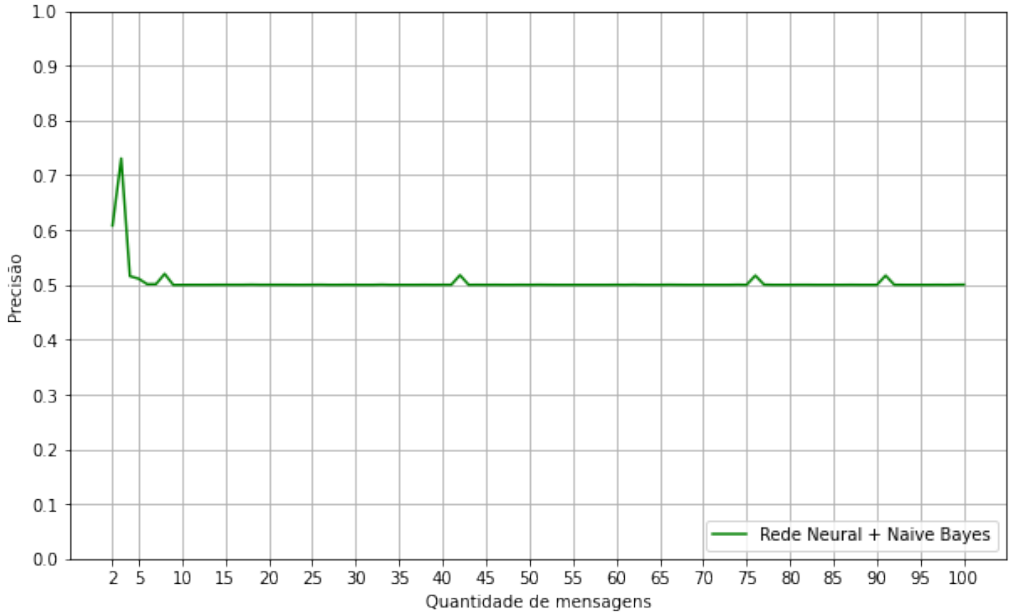


Figura 4.81: Resultado final da precisão para o experimento com *undersampling* para a estratégia 3.

Os resultados para o *recall* foram mais altos. Com 10 mensagens o algoritmo obteve 99,46% de *recall*, e com 20 mensagens obteve 99,57%, conforme pode ser observado na Figura 4.82.

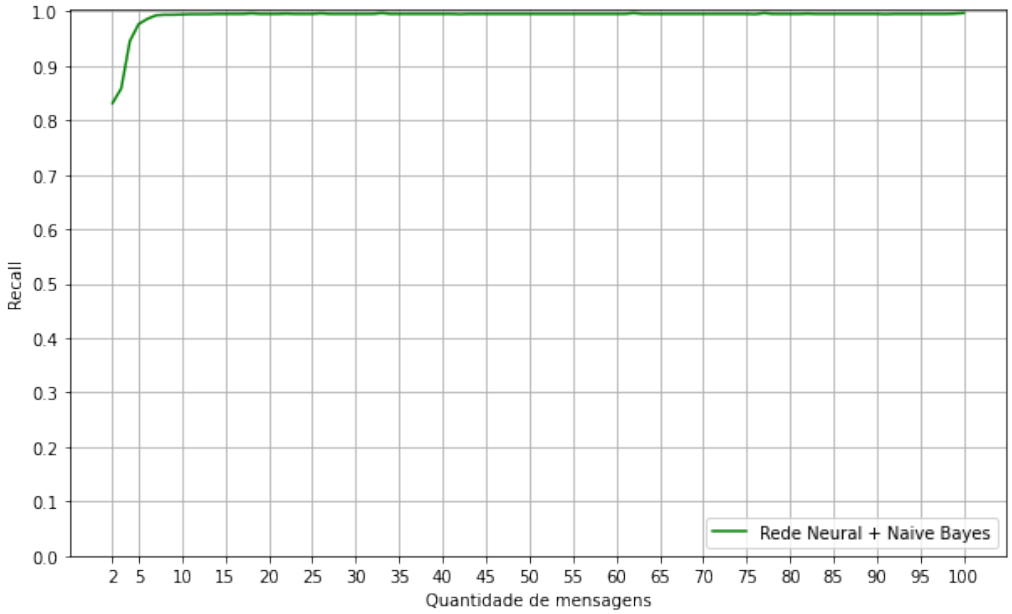


Figura 4.82: Resultado final do *recall* para o experimento com *undersampling* para a estratégia 3.

Para a métrica F_1 (Figura 4.83), foi obtido $F_1 = 66,54\%$ para as primeiras 10 mensagens, $F_1 = 66,58\%$ para 20 mensagens, e $F_1 = 66,57\%$ para 30 mensagens.

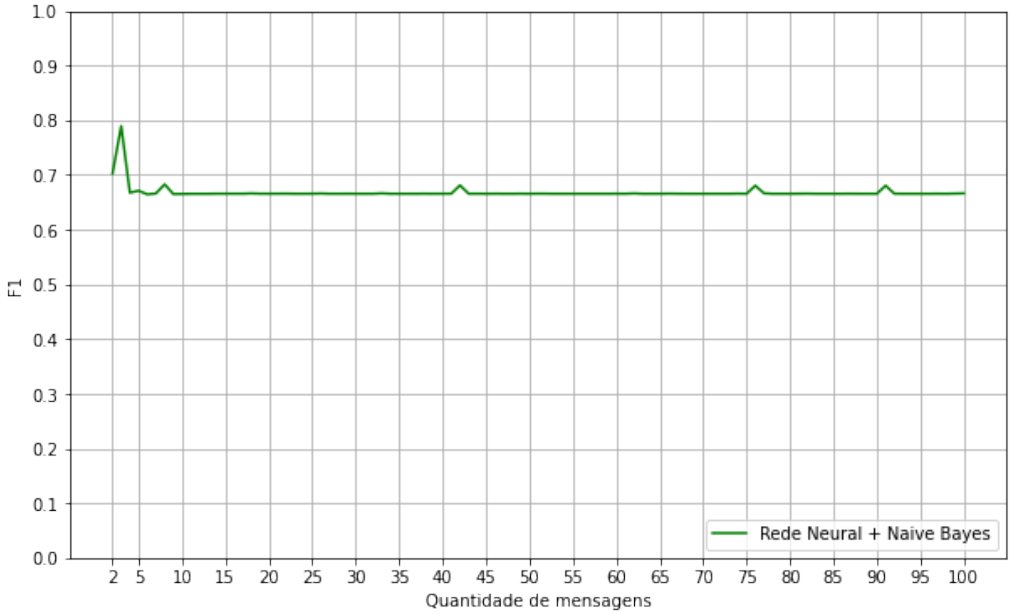


Figura 4.83: Resultado final do F_1 para o experimento com *undersampling* para a estratégia 3.

Por fim, o resultado para a métrica $F_{0.5}$ pode ser visualizado na Figura 4.84. Com 10 mensagens, o algoritmo obteve $F_{0.5} = 55,51\%$, com 20 mensagens obteve $F_{0.5} = 55,54\%$, e com 30 mensagens obteve $F_{0.5} = 55,53\%$.

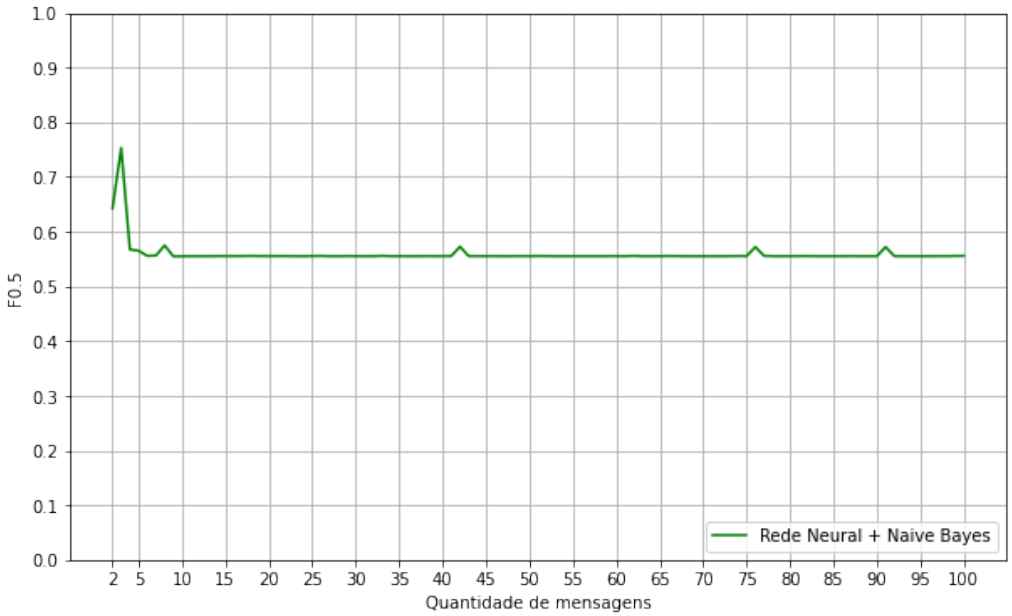


Figura 4.84: Resultado final do $F_{0.5}$ para o experimento com *undersampling* para a estratégia 3.

A Figura 4.85 exibe a quantidade de predadores (tuplas) corretamente identificadas para cada quantidade de mensagem.

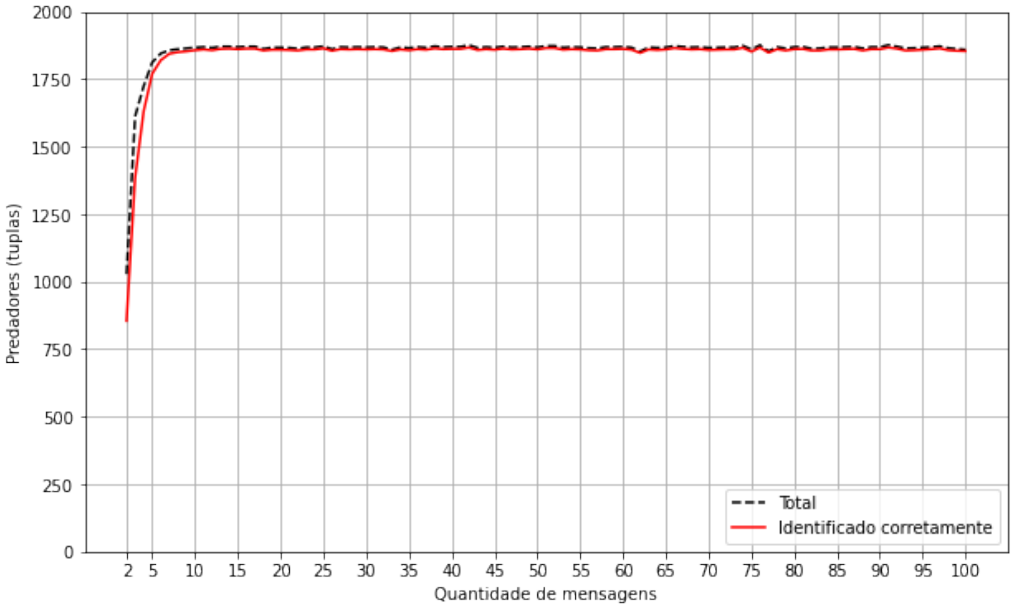


Figura 4.85: Quantidade de predadores (tuplas) corretamente identificadas para o experimento com *undersampling* dos dados para a estratégia 3.

Neste experimento, a quantidade de tuplas de predadores possíveis de serem iden-

tificadas também depende da quantidade de conversas classificadas como predatórias pelo primeiro classificador. Assim, a quantidade de tuplas possíveis de serem identificadas é igual a quantidade de conversas predatórias classificadas corretamente pelo primeiro classificador.

Com 10 mensagens, o algoritmo classificou corretamente 1.857 autores como predadores sexuais, de um total de 1.867 possíveis de serem identificados.

Com 20 mensagens, o algoritmo classificou corretamente 1.859 autores como predadores sexuais, de um total de 1.867 tuplas possíveis de serem identificadas. Com 50 mensagens foram classificados corretamente 1.860 autores como predadores sexuais, de um total de 1.868 tuplas possíveis de serem identificadas.

A Figura 4.86 exibe a quantidade de predadores únicos identificados para cada quantidade de mensagem. Após execução das etapas de pré-processamento e pré-filtro, restaram 236 predadores sexuais, porém, como esta estratégia depende das conversas que foram classificadas como predatórias pelo primeiro classificador, a quantidade de predadores únicos totais pode ser menor. Neste experimento, a partir de 8 mensagens é possível identificar os 236 predadores sexuais.

Já com 8 mensagens o algoritmo foi capaz de detectar 236 predadores dos 236 possíveis. Assim, em relação ao total de 236 predadores da base, 100% dos predadores sexuais puderam ser identificados com apenas 8 mensagens.

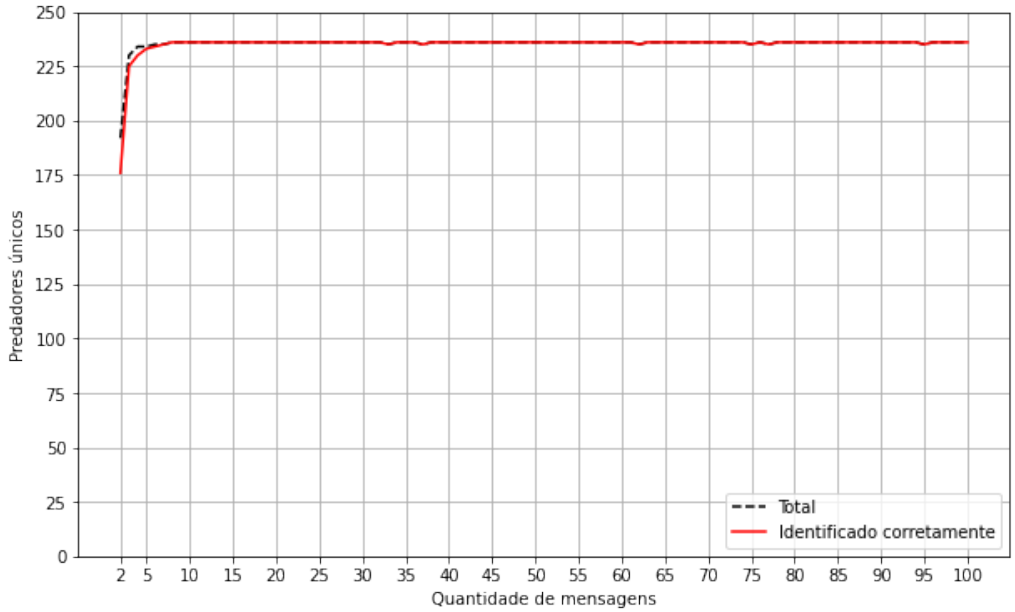


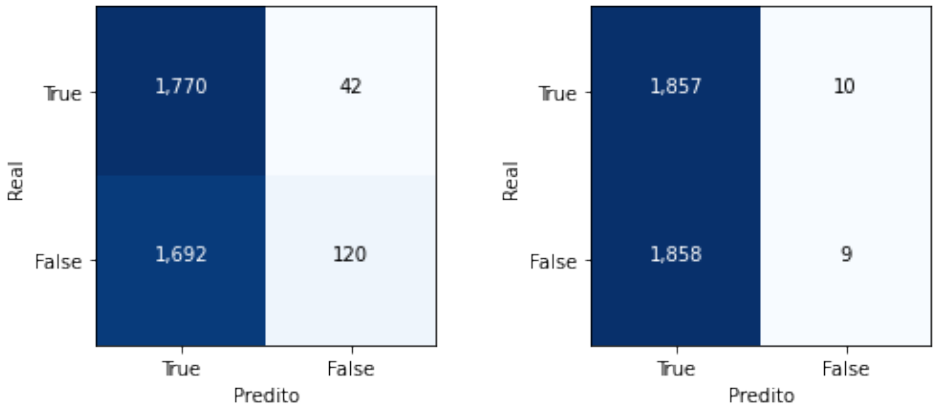
Figura 4.86: Quantidade de predadores únicos corretamente identificados para o experimento com *undersampling* dos dados para a estratégia 3.

A Tabela 4.10 exibe os resultados detalhados para algumas quantidades de mensagens.

Tabela 4.10: Resultados detalhados das métricas para o experimento com *under-sampling* dos dados para a estratégia 3.

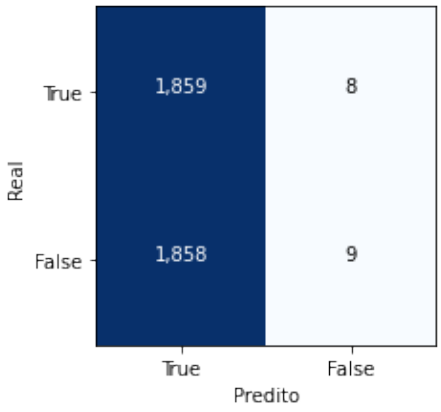
Mensagens	**Acurácia**	**Precisão**	***Recall***	F_1	$F_{0.5}$
5	0,5215	0,5113	0,9768	0,6712	0,5651
10	0,4997	0,4999	0,9946	0,6654	0,5551
20	0,5003	0,5001	0,9957	0,6658	0,5554
24	0,5000	0,5000	0,9957	0,6657	0,5553
50	0,5000	0,5000	0,9957	0,6657	0,5553
100	0,5008	0,5004	0,9973	0,6664	0,5558

Por fim, a Figura 4.87 exibe as matrizes de confusão para algumas quantidades de mensagens.

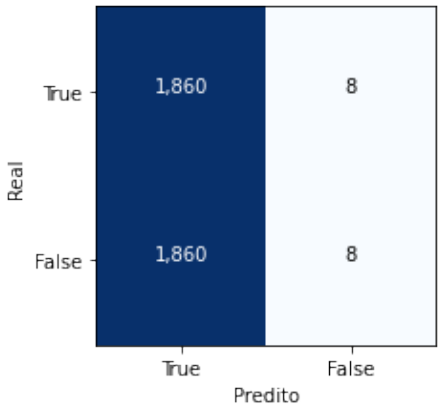


(a) 5 mensagens

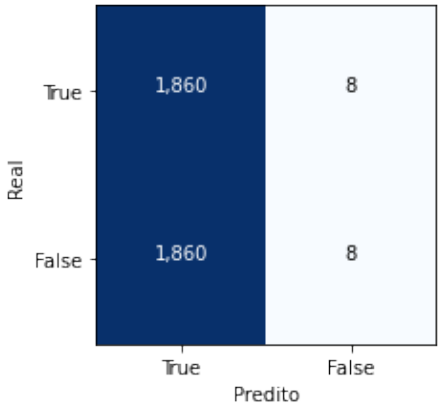
(b) 10 mensagens



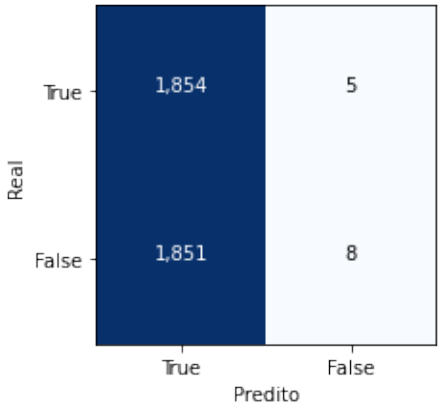
(c) 20 mensagens



(d) 24 mensagens



(e) 50 mensagens



(f) 100 mensagens

Figura 4.87: Matrizes de confusão para o experimento com *undersampling* dos dados para a estratégia 3.

4.10 Considerações Finais

Tendo em vista os resultados apresentados na seção anterior, conclui-se que as estratégias 1 e 2 são eficientes para detectar precocemente predadores sexuais.

Estratégia 1

Com o experimento sem balanceamento foi possível detectar 218 predadores sexuais com 10 mensagens, representando 92,37% dos predadores que poderiam ser identificados com 10 mensagens. Os resultados foram superiores ao estado da arte. Com 10 mensagens, o algoritmo obteve $F_{0.5} = 85,96\%$ e com 24 mensagens, obteve $F_{0.5} = 90,60\%$.

O trabalho considerado estado da arte, KULSRUD (2019) conseguiu detectar 200 predadores com até 30 mensagens e obteve $F_{0.5}$ maior que 80% após 24 mensagens.

Na base de testes, no total, haviam 254 predadores sexuais. Porém, 18 predadores foram perdidos por conta do pré-filtro, que exclui conversas que não tem apenas dois autores. Restaram então, 236 predadores sexuais.

Com o experimento com *undersampling*, 100% dos predadores sexuais possíveis de serem identificados foram detectados a partir de 8 mensagens (236 predadores). Este experimento obteve os resultados mais altos desta dissertação. Com 10 mensagens, o algoritmo obteve $F_{0.5} = 99,89\%$, e com 24 mensagens obteve $F_{0.5} = 99,82\%$, superando o estado da arte.

Estratégia 2

O experimento sem balanceamento de dados não superou o estado da arte. Com 10 mensagens, o algoritmo conseguiu detectar 171 predadores e obteve $F_{0.5} = 56,48\%$. Com 24 mensagens, o algoritmo obteve $F_{0.5} = 69,57\%$.

No experimento com *undersampling*, o algoritmo conseguiu detectar 100% dos predadores sexuais possíveis de serem identificados a partir de 9 mensagens (236 predadores). Com 10 mensagens, o algoritmo obteve $F_{0.5} = 81,70\%$, e com 24 mensagens obteve $F_{0.5} = 81,54\%$, também superando o estado da arte.

Estratégia 3

O experimento sem balanceamento dos dados não superou o estado da arte, necessitando também de mais de 24 mensagens para obter resultados superiores a 80% de $F_{0.5}$. Com 10 mensagens, o algoritmo obteve $F_{0.5} = 66,68\%$ e conseguiu detectar 173 predadores. Com 24 mensagens, o algoritmo obteve $F_{0.5} = 77,98\%$.

O experimento com *undersampling* também não superou o estado da arte, obtendo $F_{0.5} = 55,53\%$ para 24 mensagens. Porém, a partir de 8 mensagens, o classifi-

cador foi capaz de detectar os 236 predadores sexuais possíveis de serem identificados. Analisando o detalhamento das métricas e as matrizes de confusão apresentados na seção anterior, é possível entender que isto ocorreu porque o segundo classificador classificou quase todos os autores como sendo predadores sexuais. Isto aumentou o *recall*, tornando possível a identificação de todos os predadores sexuais, mas diminuiu a precisão, causando uma redução nas métricas F_1 e $F_{0.5}$.

Capítulo 5

Conclusão

Esta dissertação teve por objetivo detectar precocemente predadores sexuais em conversas realizadas na internet entre duas pessoas através do desenvolvimento de três estratégias distintas utilizando algoritmos de classificação de textos.

Como visto ao longo deste trabalho, *chats* de redes sociais e jogos *online* podem ser ambientes propícios para pedófilos (CUNHA, 2017), que podem utilizá-los para abordar crianças e adolescentes, que, em sua maioria, não tem maturidade para entender os riscos que correm *online*, chegando até mesmo a fornecer informações pessoais a estranhos (NIC.BR, 2020).

A partir da premissa de que predadores sexuais executam certo comportamento padrão para abordar suas vítimas antes do contato sexual (OLSON *et al.*, 2007), foi possível inferir que, identificando precocemente que existe um predador sexual em uma conversa, é possível evitar o encontro físico, e por conseguinte, o abuso sexual.

Desta forma, conclui-se que detectar precocemente predadores sexuais em conversas virtuais é relevante para a sociedade, uma vez que pode ser capaz de auxiliar a reduzir o número de casos de abuso sexual infantil.

Uma das maiores dificuldades encontradas na área foi a falta de dados verídicos entre predadores sexuais e vítimas reais. Grande parte do problema deve-se a própria natureza dos dados, que são sigilosos e envolvem questões legais. Na revisão da literatura, apresentada no capítulo 2, nota-se que a base mais utilizada pelos trabalhos realizados a partir de 2012 é a base do desafio *Sexual Predator Identification* da competição do PAN 2012, que contém dados do site PJ, que são conversas reais entre predadores sexuais condenados e voluntários se passando por crianças.

Além disso, identificar predadores sexuais é um problema naturalmente de classes desbalanceadas, já que, em um cenário real, a quantidade de conversas predatórias é muito inferior a quantidade de conversas normais (INCHES e CRESTANI, 2012).

Nas seções seguintes são apresentadas as contribuições do trabalho e os trabalhos futuros.

5.1 Contribuições

O problema de detecção precoce de predadores sexuais é relativamente novo, não tendo muitos trabalhos relacionados. Assim, nota-se como primeira contribuição o desenvolvimento de três estratégias distintas para detectar precocemente predadores sexuais a fim de auxiliar a reduzir o número de casos de abuso sexual infantil.

As três estratégias foram capazes de detectar precocemente predadores sexuais, com alguns experimentos tendo resultados superiores ao estado da arte, o que demonstra que algoritmos de classificação de textos são eficientes para a tarefa.

A estratégia 1, em particular, denominada Distinguir Conversas Predatórias e Gerais, apresentou os melhores resultados entre as três estratégias e conseguiu superar o estado da arte para os dois experimentos realizados, mostrando que para o problema de detecção precoce de predadores sexuais esta estratégia é mais eficiente que a estratégia de dois classificadores em sequência.

Para o experimento sem balanceamento dos dados da estratégia 1, foi obtido $F_{0.5} = 85,96\%$ para as primeiras 10 mensagens e $F_{0.5} = 90,60\%$ para as primeiras 24 mensagens. E para o experimento com *undersampling* da estratégia 1, foi obtido $F_{0.5} = 99,89\%$ para as primeiras 10 mensagens e $F_{0.5} = 99,82\%$ para as primeiras 24 mensagens, sendo o melhor resultado desta dissertação.

O experimento com *undersampling* da estratégia 2, denominada Distinguir Predador e Vítima, também superou o estado da arte. Com 10 mensagens, o classificador obteve $F_{0.5} = 81,70\%$, e com 24 mensagens obteve $F_{0.5} = 81,54\%$.

Além disso, os experimentos com *undersampling* das três estratégias conseguiram detectar 100% dos predadores sexuais possíveis de serem detectados, com menos de 10 mensagens.

Como contribuição também pode ser citado o artigo realizado sobre o tema desta dissertação intitulado "Uma Abordagem para Detecção Precoce de Predadores Sexuais em Conversas Virtuais" (PANZARIELLO e XEXÉO, 2021).

5.2 Limitações e Trabalhos Futuros

Como trabalhos futuros, primeiramente sugere-se a aplicação das estratégias deste trabalho em bases que contenham dados verídicos entre predadores sexuais e vítimas.

É interessante também a aplicação de técnicas de *machine learning* para avaliar imagens e vídeos. A base do PAN 2012 utilizada neste trabalho não contém este tipo de informação, mas é possível que elas existam em conversas reais entre predadores e vítimas. Inclusive, este tipo de estratégia pode ser ser aplicado também ao problema de identificação de pornografia infantil.

Em relação à base do PAN 2012, poderia tentar-se extrair algum significado das

URLs presentes da base, investigando se é possível correlacioná-las com as mensagens predatórias. Neste trabalho, todas as URLs foram excluídas.

Como foram realizados experimentos com *undersampling*, pode ser interessante também avaliar a aplicação de técnicas de *oversampling*.

Por fim, sugere-se a aplicação das estratégias em cenários reais, como *chats* de jogos e redes sociais.

Referências Bibliográficas

ANDRIJAUSKAS, A., SHIMABUKURO, A., 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*.

BEECH, A. R., ELLIOTT, I. A., BIRGDEN, A., et al., 2008, "The Internet and child sexual offending: A criminological review", *Aggression and Violent Behavior*, v. 13, n. 3, pp. 216–228.

BILCHES, W., 2020. "Alerta aos pais: pedofilia virtual aumenta no Brasil em meio à pandemia". Gazeta do Povo. Disponível em: <`https://www.gazetadopovo.com.br/vida-e-cidadania/alerta-aos-pais-pedofilia-virtual-aumenta-no-brasil-em-meio-a-pandemia/`>. Acesso em: 20 abr. 2021.

BORJ, P. R., BOURS, P., 2019, "Predatory Conversation Detection". In: *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*, pp. 1–6. IEEE.

BORJ, P. R., RAJA, K., BOURS, P., 2020, "On Preprocessing the Data for Improving Sexual Predator Detection : Anonymous for review". In: *SMAP 2020 - 15th International Workshop on Semantic and Social Media Adaptation and Personalization*.

CARDEI, C., REBEDEA, T., 2017, "Detecting Sexual Predators in Chats using Behavioural Features and Imbalanced Learning", *Natural Language Engineering*, v. 23, n. 4, pp. 589–616.

CORDEIRO, A. M., DE OLIVEIRA, G. M., RENTERÍA, J. M., et al., 2007, "Revisão sistemática: uma revisão narrativa", *Revista do Colégio Brasileiro de Cirurgiões*, v. 34, pp. 428–431.

CRAVEN, S., BROWN, S., GILCHRIST, E., 2006, "Sexual grooming of children: Review of literature and theoretical considerations", *Journal of Sexual Aggression*, v. 12, n. 3, pp. 287–299.

CUNHA, J., 2017. "Aliciamento Sexual Infantil Online". Safernet. Disponível em: <https://new.safernet.org.br/content/aliciamento-sexual-infantil-online>. Acesso em: 01 maio 2022.

DEVLIN, J., CHANG, M.-W., LEE, K., et al., 2018, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*.

DOS SANTOS, L., GUEDES, G., 2019, "Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na Internet". In: *Brazilian Workshop on Social Network Analysis and Mining*, pp. 143–154.

DOS SANTOS, L. F., 2021, *Identificação Automática de Atividade Predatória Sexual em Conversas Virtuais no Brasil*. Tese de Mestrado, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca.

DOS SANTOS, L. F., GUEDES, G. P., 2018, "Detecção de traços de narcisismo em conversas com predadores sexuais". In: *Brazilian Workshop on Social Network Analysisand Mining*.

EBRAHIMI, M., SUEN, C. Y., ORMANDJIEVA, O., 2016a, "Detecting predatory conversations in social media by deep Convolutional Neural Networks", *Digital Investigation*, v. 18, pp. 33–49.

EBRAHIMI, M., SUEN, C. Y., ORMANDJIEVA, O., et al., 2016b, "Recognizing Predatory Chat Documents usingSemi-supervised Anomaly Detection", *Electronic Imaging*, v. 2016, n. 17, pp. 1–9.

ESCALANTE, H. J., VILLATORO-TELLO, E., JUÁREZ, A., et al., 2013, "Sexual predator detection in chats with chained classifiers". In: *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 46–54.

ESCALANTE, H. J., VILLATORO-TELLO, E., GARZA, S. E., et al., 2017, "Early detection of deception and aggressiveness using profile-based representations", *Expert Systems with Applications*, v. 89, pp. 99–111.

FAUZI, M. A., BOURS, P., 2020, "Ensemble Method for Sexual Predators Identification in Online Chats". In: *2020 8th International Workshop on Biometrics and Forensics, IWBF 2020 - Proceedings*.

GROZEA, C., POPESCU, M., 2012, "Encoplot - Tuned for High Recall (also proposing a new plagiarism detection score)". In: *CLEF*.

HILLMAN, H., HOOPER, C., CHOO, K.-K. R., 2014, "Online child exploitation: Challenges and future research directions", *Computer Law & Security Review*, v. 30, n. 6, pp. 687–698.

INCHES, G., CRESTANI, F., 2012, "Overview of the International Sexual Predator Identification Competition at PAN-2012". In: *CLEF*.

ITU, 2022. "Individuals using the Internet". International Telecommunication Union. Disponível em: <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>. Acesso em: 01 maio 2022.

KONTOSTATHIS, A., EDWARDS, L., LEATHERMAN, A., 2009, "ChatCoder: Toward the Tracking and Categorization of Internet Predators". In: *Proc. Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining*.

KONTOSTATHIS, A., EDWARDS, L., LEATHERMAN, A., 2010, "Text Mining and Cybercrime". pp. 149–164.

KULSRUD, H. B., 2019, *Detection of cyber grooming during an online conversation*. Tese de Mestrado, Norwegian University of Science and Technology.

LIU, D., SUEN, C. Y., ORMANDJIEVA, O., 2017, "A Novel Way of Identifying Cyber Predators", .

LIVINGSTONE, S., HADDON, L., GÖRZIG, A., et al., 2011. "EU Kids Online: final report". EU Kids Online. Disponível em: <http://eprints.lse.ac.uk/45490/1/EU%20Kids%20Online%20final%20report%202011%28lsero%29.pdf>. Acesso em: 12 mar. 2018.

MCGHEE, I., BAYZICK, J., KONTOSTATHIS, A., et al., 2011, "Learning to Identify Internet Sexual Predation", *International Journal of Electronic Commerce*, v. 15, n. 3, pp. 103–122.

MISRA, K., DEVARAPALLI, H., RINGENBERG, T. R., et al., 2019, "Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 623–628.

MOURA, J., CANGUÇU, P., 2022. "Suspeito de pedofilia contra 65 crianças fez vítimas em jogo on-line". R7. Disponível em: <https://noticias.r7.com/brasilia/suspeito-de-pedofilia-

contra-65-criancas-fez-vitimas-em-jogo-on-line-26012022>. Acesso em: 01 maio 2022.

NIC.BR, N. D. I. E. C. D. P. B., 2020. "Pesquisa sobre o uso da Internet por crianças e adolescentes no Brasil: TIC Kids Online Brasil". Cetic.br. Disponível em: <http://cetic.br/pt/arquivos/kidsonline/2019/pais>. Acesso em: 29 abr. 2021.

OLSON, L. N., DAGGS, J. L., ELLEVOLD, B. L., et al., 2007, "Entrapping the Innocent: Toward a Theory of Child Sexual Predators' Luring Communication", *Communication Theory*, v. 17, n. 3, pp. 231–251.

PANZARIELLO, M., XEXÉO, G., 2021, "Uma Abordagem para Detecção Precoce de Predadores Sexuais em Conversas Virtuais". In: *Anais da VII Escola Regional de Sistemas de Informação do Rio de Janeiro*, pp. 136–139. SBC.

PARREIRAS, M., VASCONCELLOS, A., MANGELI, E., et al., 2022, "Inteligência artificial aplicada para o aumento da produtividade no atendimento de intimações". In: *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pp. 180–191. SBC.

PEERSMAN, C., VAASSEN, F., ASCH, V. V., et al., 2012, "Conversation Level Constraints on Pedophile Detection in Chat Rooms". In: *CLEF*.

PENDAR, N., 2007, "Toward Spotting the Pedophile Telling victim from predator in text chats". In: *International Conference on Semantic Computing (ICSC 2007)*, pp. 235–241.

RAHMANMIAH, M. W., YEARWOOD, J., KULKARNI, S., 2011, "Detection of child exploiting chats from a mixed chat dataset as a text classification task", *Proceedings of Australasian Language Technology Association Workshop*, pp. 157–165.

RBA, R., 2017. "Crescem denúncias de abuso sexual de crianças e adolescentes". Rede Brasil Atual. Disponível em: <https://www.redebrasilatual.com.br/cidadania/2017/05/crescem-denuncias-de-abuso-e-sexual-de-criancas-e-adolescentes>. Acesso em: 12 mar. 2018.

SCIKIT-LEARN, 2022. "Cross-validation: evaluating estimator performance". Scikit-learn. Disponível em: <https://scikit-learn.org/stable/modules/cross_validation.html>. Acesso em: 13 ago. 2022.

UN, U. N., 2020. “Online predators put millions of children at risk during COVID-19 pandemic lockdown”. United Nations News. Disponível em: <https://news.un.org/en/story/2020/04/1061742>. Acesso em: 30 abr. 2021.

UNICEF, 2019. “UNICEF poll: More than a third of young people in 30 countries report being a victim of online bullying”. UNICEF. Disponível em: <https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying>. Acesso em: 30 abr. 2021.

VILLATORO-TELLO, E., JUÁREZ-GONZÁLEZ, A., ESCALANTE, H. J., et al., 2012, “A Two-step Approach for Effective Detection of Misbehaving Users in Chats”. In: *CLEF*.