Uma Abordagem para Detecção Automática de Fraudes em Aplicativos de Mensagens Instantâneas

Alexsandro Nascimento[1], **Thiago Gadelha**[1],
José Maria Monteiro[1], **Javam Machado**[1]

[1]Universidade Federal do Ceará (UFC)
Av. Humberto Monte, s/n, Pici - CEP 60440-593 – Fortaleza – CE – Brasil

{alexsandro.nascimento, thiago.gadelha}@lsbd.ufc.br

{jose.monteiro, javam.machado}@lsbd.ufc.br

Abstract. *Instant messaging applications have enabled easy and efficient communication. However, they have also facilitated the widespread dissemination of cyber threats, such as financial fraud. In this context, the rapid and effective detection of fraud conveyed in texts shared on instant messaging applications becomes of paramount importance to prevent financial losses. This work presents two publicly available labeled datasets consisting of Brazilian Portuguese (PT-BR) messages collected from public groups on WhatsApp and Telegram, respectively, containing fraudulent messages, which were named FraudWhatsApp.Br and FraudTelegram.Br. Additionally, we conducted a series of text classification experiments, combining two different feature extraction methods, three distinct token generation strategies, two forms of preprocessing, and nine classification algorithms to discriminate texts into two categories: fraudulent and non-fraudulent texts. Our best results achieved an F1-score of 0.99 for both FraudTelegram.Br and FraudWhatsApp.Br datasets, showing the feasibility of the proposed approach.*

Resumo. *Os aplicativos de troca de mensagens instantâneas possibilitam uma comunicação simples e eficiente. Contudo, eles também propiciam a disseminação em massa de ameaças cibernéticas, tais como as fraudes financeiras. Neste contexto, a rápida e efetiva detecção de fraudes veiculadas em textos compartilhados nesses aplicativos torna-se de fundamental importância para evitar prejuízos financeiros. Este trabalho apresenta dois conjuntos de dados rotulados e disponíveis publicamente, formado por mensagens em português do Brasil (PT-BR) coletadas de grupos públicos do WhatsApp e Telegram, respectivamente, contendo mensagens fraudulentas, os quais foram denominados FraudWhatsApp.Br e FraudTelegram.Br. Adicionalmente, conduzimos uma série de experimentos de classificação de texto, combinando dois diferentes métodos de extração de atributos, três diferentes estratégia para geração de tokens, duas formas de pré-processamento e nove algoritmos de classificação para discriminar os textos em duas categorias: textos fraudulentos e não fraudulentos. Nossos melhores resultados alcançaram uma pontuação F1 de 0,99 tanto para FraudTelegram.Br quanto para o FraudWhatsApp.Br, comprovando a viabilidade da abordagem proposta.*

1. Introdução

Nos últimos anos, a ascensão dos aplicativos de troca de mensagens instantâneas alterou significativamente o modo como produzimos, compartilhamos e consumimos informação. O aplicativo WhatsApp é muito popular no Brasil, com mais de 165 milhões de usuários [de Sá et al. 2023b]. Por outro lado, em apenas um ano, a proporção de *smartphones* com Telegram instalado cresceu no Brasil de 45% para 60% em 2022 [de Sá et al. 2023a]. A popularidade desses aplicativos se deve à sua versatilidade e facilidade de uso. Eles possibilitam o compartilhamento instantâneo de diferentes tipos de mídia, como textos, links, imagens, áudios e vídeos. Além disso, eles fornecem um recurso particularmente importante: grupos de bate-papo públicos. Esses grupos públicos são acessíveis por meio de *links* de convite e, geralmente, possuem temas específicos para discussão, como política, esporte, finanças ou educação, por exemplo. Tanto o WhatsApp quanto o Telegram permitem que um usuário participe de centenas de grupos, conectando-se assim a milhares de outros usuários simultaneamente.

Contudo, ao mesmo tempo que esses aplicativos permitem uma comunicação rápida e eficiente, a ausência de um controle centralizado e de uma regulamentação apropriada torna-os propícios para a disseminação em massa de ameaças cibernéticas, tais como: fraudes financeiras, *phishing*, *smishing* (um tipo especial de *phishing* enviado via SMS - *Short Message Service*), *malware*, roubo de identidade, *stalking* (perseguição insistente em meio digital), roubo de dados pessoais, engenharia social, dentre outras [Apruzzese et al. 2023]. Estes riscos tornam-se ainda mais graves quando os usuários são crianças ou idosos.

A quantidade de fraudes financeiras tem crescido gradualmente [1]. Uma fraude bastante comum começa com uma mensagem de um número desconhecido ou até mesmo de uma pessoa próxima, pedindo para que o usuário pague uma conta ou faça um depósito bancário para salvá-la de uma situação difícil. Alguns golpes se apresentam como uma oferta de trabalho ou promoções, incentivando o usuário a clicar em um *link* fraudulento e fornecer dados pessoais, os quais podem ser utilizados para abrir contas bancárias e solicitar empréstimos, por exemplo. Propostas de investimento financeiro onde o investidor recebe o dobro do valor investido em apenas 30 dias também são fraudes corriqueiras. Esses são apenas alguns exemplos de fraudes financeiras comumente encontradas em grupos públicos do WhatsApp e Telegram.

Neste contexto, a rápida e efetiva detecção de fraudes veiculadas em textos compartilhados em aplicativos de mensagens instantâneas torna-se de fundamental importância para evitar prejuízos financeiros. Contudo, apesar deste cenário, existem poucos métodos de detecção de fraudes desenvolvidos especificamente para essas plataformas. Além disso, para que métodos eficientes sejam desenvolvidos é necessário que existam conjuntos de dados rotulados com mensagens fraudulentas que tenham sido veiculadas nesses aplicativos, uma vez que a maneira como seus usuários se expressam variam significativamente em comparação com redes sociais públicas como Facebook e Twitter[Rosenfeld et al. 2018]. Todavia, não encontramos nenhum conjunto de dados contendo mensagens fraudulentas extraídas do WhatsApp ou Telegram.

[1] https://valor.globo.com/financas/noticia/2022/06/15/cresce-numero-de-vitimas-ou-tentativas-de-golpes-bancarios-e-financeiros-aponta-pesquisa-da-febraban.ghtml

Assim, para preencher essa lacuna, foram construídos dois conjuntos de dados de grande escala, rotulados, anônimos e disponíveis publicamente, formado por mensagens em português do Brasil (PT-BR) coletadas de grupos públicos do WhatsApp e Telegram, respectivamente, contendo mensagens fraudulentas, os quais foram denominados FraudWhatsApp.Br e FraudTelegram.Br. Em seguida, foi realizada uma série de experimentos de classificação de texto, combinando dois diferentes métodos de extração de atributos (*Bag of Words or BoW* e *TF-IDF or Term Frequency – Inverse Document Frequency*), três diferentes estratégia para geração de *tokens* (unigramas, bigramas e trigramas), duas formas de pré-processamento (sem pré-processamento e com remoção de *stop words* mais lematização) e nove algoritmos de classificação para discriminar os textos em duas categorias: textos fraudulentos (relacionados a algum tipo de fraude) e não fraudulentos (não relacionados a fraudes). Esses 108 experimentos foram executados tanto utilizando o FraudWhatsApp.Br quanto o FraudTelegram.Br.

Os resultados mostram que é possível identificar de forma eficiente uma mensagem fraudulenta. Os melhores resultados alcançaram uma pontuação F1 de 0,99 tanto para o conjunto de dados FraudTelegram.Br quanto para o FraudWhatsApp.Br, comprovando a viabilidade da abordagem proposta. Até onde sabemos, não há nenhum trabalho anterior que tenha disponibilizado conjuntos de dados rotulados de mensagens do WhatsApp e Telegram contendo golpes financeiros e nem avaliado estratégias para detecção automática de fraudes neste contexto. Os experimentos realizados estabelecem assim uma base para análise de desempenho (*baseline*) e indica quais abordagens dentre as testadas são mais adequadas, provendo informações sobre esse problema ainda não-explorado.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 são discutidos os principais trabalhos relacionados. Os conjuntos de dados FraudWhatsApp.Br e FraudTelegram.Br são apresentados na Seção 3. A Seção 4 detalha os experimentos realizados com a finalidade de avaliar diferentes modelos preditivos para a detecção de fraudes em aplicativos de troca de mensagens instantâneas. Na Seção 5 discutem-se os resultados obtidos. Por fim, na Seção 6 são apresentadas as conclusões obtidas e as possibilidades para trabalhos futuros.

2. Trabalhos Relacionados

Em [Apruzzese et al. 2023], os autores apresentam uma visão ampla e de alto nível dos benefícios, problemas e desafios envolvidos na aplicação das técnicas de aprendizado automático (*Machine Learning - ML*) na área de segurança cibernética. Um método baseado em aprendizado de máquina para detecção de URLs maliciosas brasileiras foi apresentado em [Ayres et al. 2019]. Os autores avaliaram quatro classificadores distintos: *Naive Bayes* (NB), *K-Nearest Neighbors* KNN, *Support Vector Machine* (SVM) e Árvore de Decisão. A avaliação foi realizada com dados reais extraídos do catálogo de fraudes da rede acadêmica brasileira e outras fontes. O melhor resultado foi alcançado pelo algoritmo KNN, que obteve um *F1-score* (F1) de 0,96. Um sistema especialista para a detecção de páginas maliciosas, denominado Xphide, foi proposto em [Barros et al. 2020]. A base da construção do sistema foi realizada por meio de uma análise aprofundada a respeito de atributos relevantes para descrição de páginas web. Esta análise serviu de insumo para a elaboração das regras do processo decisório do Xphide. O sistema proposto foi avaliado em três diferentes bases de dados e os resultados mostraram que o Xphide superou algoritmos de classificação tradicionais em termos de precisão e revocação.

Uma ferramenta para detecção de *SMS Phishing* baseada em aprendizado automático foi apresentada em [Boukari et al. 2021]. Os autores avaliaram dois classificadores: *Random Forest* (RF) e *Naive Bayes* (NB). O primeiro obteve um valor de *F1-score* (F1) de 0,92, enquanto o segundo alcançou um F1 de 0,72. Nesta avaliação, eles utilizaram um conjunto de dados altamente desbalanceado, contendo 5.000 instâncias, das quais apenas 246 eram da classe positiva, ou seja, *SMS Phishing*. Já em [Mishra and Soni 2023], os autores apresentam um modelo de detecção de *smishing* (*Short Text Messages Phishing* ou *SMS Phishing*) composto por duas fases: Fase de Verificação de Domínio e Fase de Classificação de SMS. A Fase de Verificação de Domínio examina a autenticidade da URL. A fase de classificação do SMS examina o conteúdo de texto das mensagens e extrai um conjunto de atributos. Por fim, o sistema classifica as mensagens usando o algoritmo *Backpropagation*. Um protótipo do sistema foi desenvolvido e avaliado utilizando conjuntos de dados público. Os resultados da avaliação alcançaram uma precisão de 97,93%, o que mostra que o método proposto é bastante eficiente para a detecção de *smishing*.

Em [Prabhu Kavin et al. 2022], os autores propõem uma abordagem baseada na inteligência artificial para detecção automática de perfis falsos (ou *Spamers*) no Twitter. Eles exploraram três algoritmos de classificação diferentes: *Support Vector Machine* (SVM), *Random Forest* (RF) e *Multilayer Perceptron* (MLP). O melhor resultado foi alcançado pelo algoritmo SVM, que obteve um *F1-score* (F1) de 0,94. Uma análise comparativa entre algoritmos preditivos para detecção de *cyberbullying* foi apresentada em [Kumar and Bhat 2022]. Os resultados indicaram que técnicas de aprendizado de máquina, aprendizado profundo e processamento de linguagem natural são fundamentais para detectar ataques de *cyberbullying* de forma eficaz.

Diferentemente dos trabalhos anteriores, a abordagem proposta neste artigo tem por finalidade detectar a ocorrência de fraudes em textos compartilhados em aplicativos de mensagens instantâneas, tais como WhatsApp e Telegram. Tais fraudes não necessariamente envolvem o roubo de informações pessoais ou financeiras das vítimas, como no caso do *Phishing* e do *SMS Phishing*. Muitas vezes, por exemplo, elas utilizam engenharia social para obter transferências de valores via PIX, realizadas pelo próprio proprietário da conta bancária. Adicionalmente, disponibilizamos dois conjuntos de dados rotulados de mensagens do WhatsApp e Telegram, respectivamente, contendo golpes financeiros. Por fim, nossa avaliação experimental explorou nove algoritmos de classificação para discriminar os textos em duas categorias: textos fraudulentos e não fraudulentos. A Tabela 1 apresenta as principais características dos trabalhos relacionados, incluindo a "Tarefa" ou problema investigado, se o trabalho disponibilizada conjuntos de dados, além da quantidade de classificadores avaliados nos experimentos.

Tabela 1. Principais Características dos Trabalhos Relacionados

Trabalho	Tarefa	Disponibiliza Conjuntos de Dados	Qtd. de Classificadores
[Ayres et al., 2019]	Detecção de URLs Maliciosas	Não	4
[Barros et al., 2020]	Detecção de URLs Maliciosas	Não	3
[Boukari et al., 2021]	Detecção de SMS Phishing	Não	2
[Mishra et al., 2023]	Detecção de SMS Phishing	Não	1
[Prabhu et al., 2022]	Detecção de Perfis Falsos	Não	3
[Kumar et al., 2022]	Detecção de *Cyberbullying*	Não	3
[Este Trabalho]	Detecção de Fraudes em AMI	Sim	9

3. Os Conjuntos de Dados FraudWhatsApp.Br e FraudTelegram.Br

Para que seja possível desenvolver um detector automático de potenciais fraudes no contexto de aplicativos de mensagens instantâneas, é crucial a utilização de um conjunto de dados rotulados de larga escala, formado por mensagens em português do Brasil (PT-BR), que tenham circulado nessas plataformas. Entretanto, não encontramos um corpus com essas características. Assim, para preencher essa lacuna, construímos dois conjuntos de dados, denominados FraudWhatsApp.Br e FraudTelegram.Br, formados por mensagens coletadas de grupos públicos do WhatsApp e do Telegram, respectivamente. Neste trabalho, utilizamos as diretrizes metodológicas propostas em [Rubin et al. 2015] para a construção de um corpus textual voltado para problemas de classificação.

3.1. Coleta dos Dados

A coleta de mensagens do Whatsapp e do Telegram foi realizada utilizando-se plataforma [de Sá et al. 2023b], no período de 01 de agosto de 2022 a 31 de dezembro de 2022. No WhatsApp foram coletadas 813.106 mensagens únicas (ou seja, sem repetição), a partir de 179 grupos públicos. Já no Telegram foram capturadas 767.847 mensagens únicas, a partir de Telegram 150 grupos e/ou canais. Logicamente, torna-se inviável rotular manualmente uma quantidade tão grande de mensagens. Neste contexto, alguma estratégia para reduzir o número de mensagens a serem rotuladas precisa ser aplicada. Neste trabalho, utilizamos uma estratégia baseada no filtro de palavras-chave, a qual é descrita a seguir:

1. **Critério de Inclusão**: Incluir mensagens em que haja a ocorrência de algum dos termos a seguir: retorno, imediato, *bitcoin*, transfiro, consult, investimento, esquema, oportunidade, empréstimo, emprego, banco, conta suspensa, problema de segurança, dinheiro, oportunidade única, verificação de conta, atualização de segurança, prêmio e sorte.
2. **Critério de Exclusão**: Excluir mensagens em que haja a ocorrência de algum dos termos a seguir: estados unidos, notícias do dia, guerra, stf, Moraes, congresso, senado, *qanon*, *Trump*, globalista e esquerdista.
3. **Critério de Não Repetição**: Excluir mensagens repetidas.

As palavras utilizadas no filtro foram escolhidas a partir de uma análise exploratória sobre os dados coletados. Por exemplo, embora existam muitas mensagens com fraudes envolvendo transferências eletrônicas, a palavra "PIX" não se mostrou útil como critério de inclusão, uma vez que o conjunto de dados continha uma grande quantidade de mensagens de propaganda eleitoral que incluíam esse termo e não se caracterizavam como fraude. Após a aplicação do filtro o conjunto de dados FraudTelegram.Br passou de 767.847 mensagens para 4.215 mensagens únicas. Já o conjunto de dados FraudWhatsapp.Br passou de 813.106 mensagens para 2.475 mensagens únicas. Essa redução na quantidade de mensagens viabilizou a rotulagem manual dos textos. Vale ressaltar que os grupos públicos utilizados na coleta das mensagens possuem como tema principal conteúdos políticos de extrema direita. Esta escolha foi motivada pela disponibilidade dos dados. Porém, acreditamos que esta seleção não influencia os resultados obtidos uma vez que o filtro aplicado seleciona um conjunto de mensagens que envolvem temas financeiros.

3.2. Anonimização dos Dados

A fim de assegurar a privacidade dos usuários, dados pessoais como nomes e números de telefone foram anonimizados. Adicionalmente, utilizamos uma função *hash* para criar um identificador único e anônimo para cada usuário a partir do seu número de telefone. Por fim, utilizamos uma função *hash* para criar um identificador único e anônimo para cada grupo a partir do seu nome. Como esses grupos são publicamente disponíveis, nossa abordagem não viola a política de privacidade do WhatsApp[2] e nem a Lei Geral de Proteção de Dados (LGPD).

3.3. Rotulagem dos Textos

A rotulagem dos dados é mais um desafio complexo, pois temos que especificar se o texto é ou não fraudulento, ou seja, está ou não relacionado a alguma fraude. A seguir, descrevemos o processo de rotulagem manual do conteúdo textual das mensagens obtidas após a aplicação do filtro de palavras-chave. Vale ressaltar que o processo de rotulação foi inteiramente manual para garantir que o corpus textual seja de alta qualidade. Três anotadores conduziram o processo de rotulagem. Resolvemos divergências de rotulagem realizando uma revisão coletiva. O processo de rotulagem utilizado baseou-se nos seguintes itens:

1. Se o texto contém indícios de fraude, nós o rotulamos como passível de fraude. Para esse propósito, utilizamos buscas na Web e fazemos uso de plataformas brasileiras que noticiam golpes, como *Uol*[3], *Agência Lupa*[4] e *Banco Central*[5].
   Exemplo: PIX INVESTIMENTO IMEDIATO CAI NA HORA FAZEMOS DE QUALQUER BANCO FÍSICO OU DIGITAL O ROBÔ INVESTIMENTO PIX INVESTI EM AÇÕES DE RETORNO IMEDIATO FAZEMOS DE TODOS BANCO FÍSICO E DIGITAL R 50,00 transfiro 500,00 REAIS R 100,00 transfiro 1000,00 REAIS R 150,00 transfiro 1500,00 REAIS R 200,00 transfiro 3000,00 REAIS R 250,00 transfiro 4000,00 REAIS R 300,00 transfiro 5000,00 REAIS R 350,00 transfiro 6500,00 REAIS R 400,00 transfiro 8000,00 REAIS R 450,00 transfiro 9500,00 REAIS R 500,00 transfiro 10.000,00 REAIS R 1.000,00 transfiro 20.000,00 R 5.000,00 transfiro 50.000,00 R 10.000,00 transfiro 100.000,00 REAIS R 20.000,00 transfiro 200.000,00 REAIS R 30.000,00 300.000,00 REAIS O valor cai na conta imediatamente via pix TODAS CONTAS DIGITAIS E BANCO FÍSICO COM O PIX DA CERTO qualquer conta que tenha pix
2. Se o texto contiver alegações que não podem ser verificadas como fraude e que são imprecisas, tendenciosas, alarmistas, rotulamos como passível de fraude.
   Exemplo: INVISTA CONOSCO E RECEBA SEU LUCRO ESTÁ SEGURO E RETORNO 100% GARANTIDO ... MUITO FÁCIL DE COMEÇAR A GANHAR DINHEIRO COM BITCOIN. (CONTACTE-ME AGORA) COMO INVESTIR COMO FAÇO SOBRE ISSO COMO POSSO COMEÇAR COMO FUNCIONA NÃO SEI NADA SOBRE ISSO E QUERO QUE VOCÊ ME DÊ INFORMAÇÕES SOBRE ISSO QUERO INVESTIR CONTACTE-ME E VOU APRESENTAR COMO COMEÇAR A GANHAR CONOSCO

[2]https://www.whatsapp.com/legal/privacy-policy
[3]`https://www.uol.com.br/`
[4]`http://piaui.folha.uol.com.br/lupa/`
[5]`https://www.bcb.gov.br/`

3. Se nenhuma das indicações anteriores for encontrada no texto, rotulamos que o texto não contém fraude. Adicionalmente, quando o texto contém uma opinião em vez de uma afirmação ou é humorístico, ele também foi rotulado como não fraudulento.
   Exemplo: OS NÚMEROS NÃO MENTEM. O caso dos analistas de noticiário se repete em relação às previsões sobre as taxas de desemprego, inflação, confiança etc.

Após o processo de rotulagem, o conjunto de dados FraudTelegram.Br ficou com 2.924 mensagens únicas rotuladas como fraude (rótulo 1) e 1.291 mensagens únicas classificadas como não fraude (rótulo 0). Já o conjunto de dados FraudWhatsApp.Br ficou com apenas 84 mensagens únicas classificadas como fraude (rótulo 1) e 2.556 mensagens únicas rotuladas como não fraude (rótulo 0). Assim, o FraudWhatsApp.Br apresenta um número pequeno de exemplos de fraudes e um grande desbalanceamento entre as classes, o que pode comprometer o desempenho dos modelos preditivos construídos a partir deste conjunto de dados. Por fim, aplicamos a seguinte estratégia para computar a interseção entre os dois conjuntos de dados. Cada mensagem fraudulenta "w" do conjunto FraudWhatsApp.Br é comparada com cada mensagem fraudulenta "t" de FraudTelegram.Br. Caso as mensagens 'w" e "t" difiram em mais de 30 caracteres, elas são consideradas distintas. Caso contrário, calcula-se a métrica de Jaro-Winkler (com *threshold* de 0,7) a fim de medir a similaridade entre 'w" e "t". Se o valor da métrica de Jaro-Winkler for maior que 0,85, assume-se que as mensagens 'w" e "t" são as mesmas. Se não, considera-se que as mensagens 'w" e "t" são diferentes. Ao final deste processo, observou-se que apenas 185 mensagens fraudulentas estavam presentes nos dois conjuntos de dados. Desta forma, podemos concluir que as mensagens contendo fraudes que circulam no WhatsApp e no Telegram possuem características particulares, sendo dependentes da plataforma, o que ilustra a importância de explorarmos os dois conjuntos de dados.

4. Avaliação Experimental

Com o intuito de prover uma *baseline* para o problema de detecção de fraudes em mensagens de texto do Telegram e Whatsapp em português, foram realizados uma série de experimentos utilizando os conjuntos de dados FraudTelegram.Br e FraudWhatsapp.Br.

4.1. Atributos e Algoritmos de Classificação

Como mencionado anteriormente, foram avaliados dois métodos de extração de atributos distintos (*BoW* e *TF-IDF*). Decidiu-se não utilizar vetores de *embedding* pré-treinados devido à grande quantidade de palavras com erros de ortografia, *emoticons* e neologismos no corpus. Nesse contexto, as características dos métodos *BoW* e *TF-IDF* são interessantes devido à sua simplicidade, velocidade e ampla utilização na classificação de textos.

Antes de aplicar os métodos *BoW* e *TF-IDF*, o texto foi convertido para letras minúsculas. É importante mencionar que os *emojis* estão muito presentes nos textos e desempenham um papel importante na linguagem utilizada nos aplicativos de troca de mensagens. Por essa razão, decidiu-se mantê-los. No entanto, como combinações de *emojis* podem gerar diferentes tipos de *tokens*, optou-se por separá-los por espaços em branco, criando assim um *token* específico para cada *emoji*. Além disso, realizou-se a normalização de URLs, mantendo apenas o nome do domínio. Devido à diversidade lexical do corpus, os vetores resultantes são esparsos e possuem alta dimensionalidade.

Três diferentes estratégia para geração de *tokens* foram avaliadas: unigramas, bigramas e trigramas. Embora isso resulte em vetores de alta dimensionalidade, acreditamos que essa abordagem pode revelar padrões distintos presentes nas mensagens, uma vez que bigramas e trigramas podem capturar uma quantidade maior de informações relacionadas ao contexto. Além disso, para avaliar o impacto de técnicas de pré-processamento, duas possibilidades foram consideradas: i) sem pré-processamento e ii) com remoção de *stop words* e lematização. Essas técnicas visam eliminar ruídos, permitindo uma representação mais precisa das características relevantes presentes nas mensagens.

Dessa forma, foram criados 12 cenários de execução distintos, combinando dois métodos de extração de atributos (*BoW* e *TF-IDF*), três estratégia para geração de *tokens* (unigramas, bigramas e trigramas) e duas formas de pré-processamento (sem e com pré-processamento). Para cada um desses cenários, avaliamos 9 algoritmos clássicos para tarefas de classificação [Pranckevičius and Marcinkevičius 2017], abrangendo diferentes categorias: modelos lineares (Logistic Regression - LR), modelos generativos (*Bernoulli Naive-Bayes - BNB* e *Multinomial Naive Bayes - MNB*), aprendizado baseado em instâncias (*K-Nearest Neighbors - KNN*), máquinas de vetores de suporte (*Linear Support Vector Machine - LSVM* e *Stochastic Gradient Descent - SGD*), *ensemble* de algoritmos (*Random Forest - RF* e *Gradient Boosting - GB*), além de redes neurais (*Multilayer Perceptron - MLP*). Assim, para cada conjunto de dados desenvolvidos realizamos um 108 experimentos. Os algoritmos de classificação foram implementados utilizando a biblioteca Python *scikit-learn* [Pedregosa et al. 2011].

No MLP, foram utilizados um *batch size* de 64 e uma estratégia de treinamento com parada antecipada, em que reservou-se 10% dos dados de treinamento para validação e interrompeu-se o treinamento quando o desempenho da validação não melhorava em pelo menos 0,001 em 5 épocas consecutivas. Os demais hiperparâmetros foram mantidos com valor padrão para todos os algoritmos. Embora não tenhamos realizado uma seleção sistemática de hiperparâmetros, a diversidade das abordagens testadas nos permite obter informações sobre quais estratégias de aprendizado podem ser mais adequadas para o problema investigado, estabelecendo assim um *baseline*. Todos os dados e códigos utilizados nos experimentos estão disponíveis em nosso repositório *online*[6].

4.2. Métricas de Desempenho

Como mencionado anteriormente, o problema investigado consiste em uma tarefa de classificação binária em que fraude representa a classe positiva (esta também é nossa classe de interesse) e não fraude representa a classe negativa. Para avaliar o desempenho de cada modelo gerado, as seguintes métricas foram utilizadas: *False positive rate* (FPR), *Precision* (PRE), *Recall* (REC) e *F1-score* (F1). Uma vez que utilizamos uma validação cruzada com *k-fold* ($k = 5$), iremos apresentar a média e o desvio padrão de cada métrica obtida, em cada experimento realizado.

5. Resultados

Nesta seção, os resultados obtidos para os dois conjuntos de dados avaliados, FraudTelegram.Br e FraudWhatsApp.Br, serão apresentados e discutidos. Cada uma das tabelas mostradas a seguir contém informações sobre seis cenários diferentes. Assim, cada tabela

[6]https://github.com/jmmfilho/sec-fraudimabr

Tabela 2. Resultados do FraudTelegram.Br com o Método *BoW*

(a) BoW-1. Atributos: 18.539

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.12	0.98±0.13	**0.98±0.13**
BNB	0.97	0.92±0.26	0.83±0.34	0.87±0.33
MNB	0.99	0.98±0.13	0.98±0.12	0.98±0.13
LSVM	0.99	0.98±0.14	0.97±0.16	0.98±0.15
KNN	0.97	0.91±0.28	0.92±0.27	0.91±0.27
SGD	0.99	0.98±0.14	0.98±0.14	0.98±0.14
RF	0.99	0.98±0.13	0.98±0.13	0.98±0.13
GB	0.99	0.97±0.17	0.97±0.16	0.97±0.17
MLP	0.99	0.98±0.13	0.98±0.13	0.98±0.13

(b) BoW-1 C/Pré. Atributos: 17.835

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.12	0.99±0.11	**0.99±0.12**
BNB	0.98	0.93±0.25	0.83±0.34	0.87±0.33
MNB	0.99	0.97±0.15	0.98±0.14	0.98±0.14
LSVM	0.99	0.98±0.13	0.98±0.13	0.98±0.13
KNN	0.96	0.90±0.30	0.90±0.30	0.90±0.30
SGD	0.99	0.98±0.15	0.98±0.14	0.98±0.14
RF	0.99	0.98±0.13	0.98±0.13	0.98±0.13
GB	0.99	0.95±0.20	0.97±0.18	0.96±0.19
MLP	0.99	0.98±0.11	0.98±0.12	0.98±0.12

(c) BoW-1,2. Atributos: 102.085

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.14	0.98±0.14	0.98±0.14
BNB	0.79	0.91±0.27	0.74±0.35	0.77±0.40
MNB	0.99	0.99±0.11	0.98±0.12	0.98±0.12
LSVM	0.99	0.97±0.15	0.98±0.15	0.98±0.15
KNN	0.97	0.88±0.30	0.92±0.26	0.90±0.30
SGD	0.99	0.98±0.12	0.98±0.13	0.98±0.12
RF	0.99	0.98±0.13	0.98±0.12	0.98±0.13
GB	0.99	0.97±0.16	0.98±0.15	0.97±0.15
MLP	0.99	0.99±0.10	0.99±0.12	**0.99±0.11**

(d) BoW-1,2 C/Pré. Atributos: 85.824

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.13	0.98±0.12	**0.98±0.12**
BNB	0.79	0.91±0.27	0.72±0.35	0.75±0.40
MNB	0.99	0.98±0.12	0.98±0.11	0.98±0.12
LSVM	0.99	0.98±0.13	0.98±0.12	0.98±0.13
KNN	0.96	0.87±0.32	0.90±0.29	0.88±0.31
SGD	0.99	0.98±0.12	0.99±0.11	0.98±0.12
RF	0.99	0.97±0.17	0.98±0.14	0.97±0.16
GB	0.99	0.95±0.20	0.97±0.16	0.96±0.19
MLP	0.99	0.98±0.11	0.98±0.12	0.98±0.12

(e) BoW-1,2,3. Atributos: 215.272

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.14	0.98±0.14	0.98±0.14
BNB	0.71	0.90±0.28	0.68±0.34	0.71±0.42
MNB	0.99	0.99±0.09	0.99±0.10	**0.99±0.10**
LSVM	0.99	0.98±0.15	0.97±0.15	0.97±0.15
KNN	0.97	0.86±0.32	0.92±0.27	0.88±0.32
SGD	0.99	0.98±0.14	0.98±0.14	0.98±0.14
RF	0.99	0.97±0.15	0.98±0.13	0.98±0.14
GB	0.99	0.97±0.17	0.98±0.15	0.97±0.16
MLP	0.99	0.99±0.11	0.98±0.12	0.98±0.12

(f) BoW-1,2,3 C/Pré. Atributos: 165.975

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.98±0.14	0.98±0.13	0.98±0.14
BNB	0.69	0.90±0.28	0.67±0.33	0.70±0.42
MNB	0.99	0.99±0.10	0.99±0.11	**0.99±0.10**
LSVM	0.99	0.98±0.13	0.98±0.13	0.98±0.13
KNN	0.96	0.86±0.32	0.91±0.28	0.88±0.32
SGD	0.99	0.98±0.13	0.98±0.12	0.98±0.12
RF	0.99	0.97±0.17	0.98±0.14	0.97±0.16
GB	0.99	0.94±0.22	0.96±0.18	0.95±0.21
MLP	0.99	0.99±0.11	0.98±0.12	0.98±0.12

possui seis sub-tabelas, uma para cada cenário avaliado, e estão organizadas da seguinte forma: a) apenas unigramas, sem pré-processamento; b) apenas unigramas, com remoção de *stopwords* e lematização; c) unigramas e bigramas, sem pré-processamento; d) unigramas e bigramas, com remoção de *stopwords* e lematização; e) unigramas, bigramas e trigramas, sem pré-processamento; f) unigramas, bigramas e trigramas, com remoção de *stopwords* e lematização. Cada sub-tabela informa também a quantidade de atributos gerados. Para cada sub-tabela destacamos os resultados proporcionados por cada um dos 9 classificadores avaliados

5.1. Resultados Obtidos para o Telegram

Os resultados obtidos nos experimentos realizados com o conjunto de dados FraudTelegram.Br estão resumidos nas Tabelas 2 e 3. A Tabela 2 apresenta os resultados para o método de extração de atributos *BoW*. Já a tabela 3 ilustra os resultados para o método de extração de atributos *TF-IDF*. A partir das Tabelas 2 e 3, podemos notar que os classificadores MNB e MLP apresentaram, em geral, melhores resultados, considerando o *F1-score*. BNB e KNN, em média, obtiveram os piores resultados. Os demais classificadores apresentaram um bom resultado em todos os cenários. Além disso, percebemos que a estratégia *TF-IDF* apresentou resultados melhores.

Tabela 3. Resultados do FraudTelegram.Br com o Método *TF-IDF*

(a) TF-IDF-1. Atributos: 18.539

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.97±0.16	0.98±0.13	0.98±0.15
BNB	0.98	0.92±0.26	0.84±0.34	0.87±0.33
MNB	0.99	0.98±0.14	0.98±0.14	0.98±0.14
LSVM	0.99	0.99±0.12	0.98±0.14	0.98±0.14
KNN	0.99	0.96±0.19	0.96±0.18	0.96±0.19
SGD	0.99	0.99±0.11	0.98±0.12	0.98±0.12
RF	0.99	0.98±0.13	0.98±0.13	0.98±0.13
GB	0.99	0.96±0.20	0.97±0.18	0.96±0.19
MLP	0.99	0.99±0.09	0.99±0.11	**0.99±0.10**

(b) TF-IDF-1 C/Pré. Atributos: 17.835

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.96±0.19	0.98±0.15	0.97±0.17
BNB	0.98	0.93±0.25	0.83±0.34	0.87±0.33
MNB	0.99	0.97±0.15	0.98±0.14	0.98±0.15
LSVM	0.99	0.98±0.14	0.98±0.13	0.98±0.13
KNN	0.98	0.97±0.18	0.95±0.22	0.96±0.20
SGD	0.99	0.99±0.11	0.98±0.12	0.98±0.12
RF	0.99	0.98±0.14	0.98±0.14	0.98±0.14
GB	0.99	0.96±0.20	0.96±0.19	0.96±0.19
MLP	0.99	0.98±0.12	0.99±0.11	**0.99±0.12**

(c) TF-IDF-1,2. Atributos: 102.085

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.97±0.16	0.98±0.13	0.98±0.15
BNB	0.79	0.91±0.27	0.74±0.35	0.77±0.40
MNB	0.99	0.99±0.11	0.98±0.12	0.98±0.12
LSVM	0.99	0.98±0.12	0.98±0.13	0.98±0.12
KNN	0.98	0.96±0.19	0.95±0.20	0.96±0.20
SGD	0.99	0.99±0.11	0.98±0.12	0.98±0.12
RF	0.99	0.99±0.11	0.99±0.10	**0.99±0.11**
GB	0.99	0.97±0.18	0.97±0.16	0.97±0.17
MLP	0.99	0.99±0.11	0.99±0.11	**0.99±0.11**

(d) TF-IDF-1,2 C/Pré. Atributos: 85.824

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.96±0.19	0.98±0.15	0.97±0.17
BNB	0.79	0.91±0.27	0.72±0.35	0.75±0.40
MNB	0.99	0.98±0.15	0.98±0.13	0.98±0.14
LSVM	0.99	0.99±0.11	0.99±0.11	0.99±0.11
KNN	0.98	0.97±0.18	0.96±0.19	0.96±0.18
SGD	0.99	0.99±0.10	0.99±0.11	**0.99±0.10**
RF	0.99	0.98±0.13	0.98±0.12	0.98±0.12
GB	0.99	0.96±0.20	0.96±0.18	0.96±0.19
MLP	0.99	0.98±0.13	0.98±0.12	0.98±0.13

(e) TF-IDF-1,2,3. Atributos: 215.272

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.97±0.17	0.98±0.14	0.97±0.15
BNB	0.71	0.90±0.28	0.68±0.34	0.71±0.42
MNB	0.99	0.98±0.13	0.98±0.12	0.98±0.13
LSVM	0.99	0.98±0.12	0.98±0.12	**0.98±0.12**
KNN	0.98	0.97±0.18	0.97±0.17	0.97±0.18
SGD	0.99	0.99±0.11	0.98±0.13	0.98±0.12
RF	0.99	0.99±0.11	0.98±0.12	0.98±0.12
GB	0.99	0.97±0.16	0.98±0.12	0.98±0.12
MLP	0.99	0.98±0.13	0.98±0.13	0.98±0.13

(f) TF-IDF-1,2,3 C/Pré. Atributos: 165.975

Método	Auc Score	Precision	Recall	F1-score
LR	0.99	0.96±0.18	0.97±0.15	0.97±0.17
BNB	0.69	0.90±0.28	0.67±0.33	0.70±0.42
MNB	0.99	0.97±0.16	0.98±0.13	0.98±0.14
LSVM	0.99	0.99±0.11	0.99±0.11	0.99±0.11
KNN	0.98	0.97±0.16	0.96±0.19	0.97±0.18
SGD	0.99	0.99±0.11	0.98±0.12	0.98±0.12
RF	0.99	0.98±0.12	0.98±0.14	0.98±0.13
GB	0.99	0.95±0.21	0.95±0.20	0.95±0.21
MLP	0.99	0.99±0.10	0.99±0.10	**0.99±0.10**

5.2. Resultados Obtidos para o WhatsApp

Os resultados obtidos nos experimentos realizados com o conjunto de dados FraudWhatsApp.Br estão resumidos nas Tabelas 4 e 5. A Tabela 4 apresenta os resultados para o método de extração de atributos *BoW*. Já a Tabela 5 ilustra os resultados para o método de extração de atributos *TF-IDF*. A partir das Tabelas 4 e 5, observa-se que os classificadores BNB e LSVM tiveram, em geral, melhores resultados em termos de *F1-score*. Por outro lado, o classificador KNN demonstrou consistentemente os piores resultados em todas as configurações consideradas. Isso sugere que o KNN teve dificuldades em lidar com a tarefa de classificação em questão. Os demais classificadores apresentaram um bom resultado em todos os cenários, mesmo em um conjunto de dados tão desbalanceado.

Vale destacar ainda que, em geral, os algoritmos avaliados apresentaram desempenhos superiores no conjunto de dados FraudTelegram.Br, em relação ao FraudWhatsApp.Br. Provavelmente, essa divergência decorre da pequena quantidade de exemplos de mensagens fraudulentas e do grande desbalanceamento presentes no FraudWhatsApp.Br.

Tabela 4. Resultados do FraudWhatsApp.Br com o Método *BoW*

(a) BoW-1. Atributos: 17.949

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.91±0.26	0.82±0.34	0.86±0.31
BNB	0.94	0.90±0.27	0.90±0.27	**0.90±0.27**
MNB	0.94	0.81±0.34	0.90±0.28	0.85±0.32
LSVM	0.95	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.76	0.94±0.22	0.76±0.35	0.82±0.33
SGD	0.95	0.92±0.24	0.88±0.30	0.90±0.28
RF	0.93	0.95±0.20	0.82±0.33	0.87±0.30
GB	0.95	0.88±0.29	0.90±0.27	0.89±0.28
MLP	0.91	0.89±0.29	0.85±0.32	0.87±0.31

(b) BoW-1 C/Pré. Atributos: 18.627

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.92±0.24	0.88±0.30	0.90±0.28
BNB	0.96	0.88±0.29	0.90±0.27	0.89±0.278
MNB	0.95	0.86±0.31	0.90±0.27	0.88±0.30
LSVM	0.97	0.95±0.20	0.85±0.32	0.89±0.28
KNN	0.79	0.90±0.27	0.79±0.35	0.84±0.33
SGD	0.97	0.96±0.19	0.88±0.30	**0.91±0.26**
RF	0.95	0.95±0.20	0.82±0.33	0.87±0.30
GB	0.96	0.89±0.29	0.85±0.32	0.87±0.31
MLP	0.94	0.95±0.20	0.85±0.32	0.89±0.28

(c) BoW-1,2. Atributos: 108.871

Método	Auc Score	Precision	Recall	F1-score
LR	0.97	0.95±0.20	0.82±0.33	0.87±0.30
BNB	0.97	0.86±0.31	0.90±0.27	0.88±0.30
MNB	0.96	0.86±0.31	0.90±0.27	0.88±0.30
LSVM	0.97	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.73	0.89±0.29	0.73±0.35	0.79±0.35
SGD	0.97	0.96±0.19	0.88±0.30	**0.91±0.26**
RF	0.95	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.96	0.92±0.24	0.88±0.30	0.90±0.28
MLP	0.90	0.94±0.21	0.79±0.34	0.85±0.32

(d) BoW-1,2 C/Pré. Atributos: 92.268

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.95±0.20	0.82±0.33	0.78±0.30
BNB	0.95	0.93±0.24	0.90±0.27	**0.92±0.25**
MNB	0.95	0.84±0.32	0.90±0.28	0.87±0.31
LSVM	0.97	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.73	0.93±0.24	0.73±0.35	0.80±0.35
SGD	0.95	0.95±0.20	0.82±0.33	0.87±0.30
RF	0.95	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.95	0.85±0.32	0.87±0.30	0.86±0.31
MLP	0.88	0.94±0.21	0.79±0.34	0.85±0.22

(e) BoW-1,2,3. Atributos: 233.164

Método	Auc Score	Precision	Recall	F1-score
LR	0.97	0.95±0.20	0.82±0.33	0.87±0.30
BNB	0.97	0.93±0.23	0.93±0.23	**0.93±0.23**
MNB	0.97	0.87±0.30	0.93±0.23	0.90±0.28
LSVM	0.98	0.94±0.21	0.79±0.34	0.85±0.32
KNN	0.73	0.89±0.29	0.73±0.35	0.79±0.35
SGD	0.98	0.92±0.24	0.88±0.30	0.90±0.28
RF	0.95	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.96	0.90±0.28	0.87±0.30	0.89±0.29
MLP	0.91	0.94±0.22	0.76±0.35	0.82±0.33

(f) BoW-1,2,3 C/Pré. Atributos: 176.130

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.95±0.20	0.82±0.33	0.87±0.30
BNB	0.95	0.96±0.18	0.91±0.27	**0.93±0.23**
MNB	0.95	0.84±0.32	0.90±0.28	0.87±0.31
LSVM	0.97	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.73	0.93±0.24	0.73±0.35	0.80±0.35
SGD	0.95	0.93±0.24	0.90±0.27	0.92±0.25
RF	0.91	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.95	0.90±0.27	0.90±0.27	0.90±0.27
MLP	0.91	0.91±0.26	0.82±0.34	0.86±0.31

Tabela 5. Resultados do FraudWhatsapp.Br com o Método *TF-IDF*

(a) TF-IDF-1. Atributos: 18.627

Método	Auc Score	Precision	Recall	F1-score
LR	0.95	0.92±0.25	0.85±0.32	0.88±0.30
BNB	0.96	0.88±0.29	0.90±0.27	0.89±0.28
MNB	0.97	0.75±0.35	0.89±0.29	0.81±0.35
LSVM	0.94	0.91±0.26	0.82±0.34	0.86±0.31
KNN	0.90	0.77±0.35	0.90±0.28	0.82±0.34
SGD	0.94	0.91±0.26	0.82±0.34	0.86±0.31
RF	0.95	0.99±0.08	0.79±0.34	0.86±0.31
GB	0.96	0.93±0.24	0.90±0.27	**0.92±0.25**
MLP	0.97	0.87±0.30	0.87±0.30	0.87±0.30

(b) TF-IDF-1 C/Pré. Atributos: 17.949

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.88±0.29	0.82±0.34	0.85±0.32
BNB	0.94	0.90±0.27	0.90±0.27	**0.90±0.27**
MNB	0.98	0.75±0.35	0.89±0.29	0.81±0.35
LSVM	0.95	0.88±0.29	0.82±0.34	0.85±0.32
KNN	0.87	0.73±0.35	0.86±0.32	0.78±0.36
SGD	0.95	0.91±0.26	0.82±0.34	0.86±0.31
RF	0.94	0.99±0.08	0.76±0.35	0.84±0.33
GB	0.94	0.92±0.25	0.85±0.32	0.88±0.30
MLP	0.96	0.90±0.28	0.87±0.30	0.89±0.29

(c) TF-IDF-1,2. Atributos: 108.871

Método	Auc Score	Precision	Recall	F1-score
LR	0.95	0.95±0.20	0.82±0.33	0.87±0.30
BNB	0.97	0.86±0.31	0.90±0.27	0.88±0.30
MNB	0.98	0.74±0.35	0.92±0.25	0.81±0.35
LSVM	0.95	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.93	0.77±0.35	0.92±0.25	0.83±0.33
SGD	0.95	0.95±0.20	0.82±0.33	0.87±0.30
RF	0.95	0.99±0.09	0.73±0.35	0.81±0.34
GB	0.95	0.96±0.19	0.88±0.30	0.91±0.26
MLP	0.98	0.95±0.20	0.85±0.32	**0.99±0.28**

(d) TF-IDF-1,2 C/Pré. Atributos: 92.268

Método	Auc Score	Precision	Recall	F1-score
LR	0.95	0.88±0.29	0.82±0.34	0.85±0.32
BNB	0.95	0.93±0.24	0.90±0.27	0.92±0.25
MNB	0.98	0.74±0.35	0.89±0.29	0.80±0.35
LSVM	0.95	0.88±0.29	0.82±0.34	0.85±0.32
KNN	0.90	0.76±0.35	0.90±0.29	0.81±0.34
SGD	0.95	0.95±0.20	0.82±0.33	0.87±0.30
RF	0.91	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.94	0.89±0.29	0.85±0.32	0.87±0.31
MLP	0.96	0.93±0.24	0.90±0.27	**0.92±0.25**

(e) TF-IDF-1,2,3. Atributos: 233.164

Método	Auc Score	Precision	Recall	F1-score
LR	0.95	0.95±0.20	0.82±0.33	0.87±0.30
BNB	0.97	0.93±0.23	0.93±0.23	**0.93±0.23**
MNB	0.98	0.74±0.35	0.95±0.20	0.81±0.34
LSVM	0.95	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.96	0.76±0.35	0.95±0.20	0.83±0.34
SGD	0.95	0.94±0.21	0.79±0.34	0.85±0.32
RF	0.95	0.99±0.09	0.73±0.35	0.81±0.34
GB	0.96	0.95±0.20	0.85±0.32	0.89±0.28
MLP	0.98	0.92±0.25	0.85±0.32	0.88±0.30

(f) TF-IDF-1,2,3 C/Pré. Atributos: 176.130

Método	Auc Score	Precision	Recall	F1-score
LR	0.96	0.88±0.29	0.82±0.34	0.85±0.32
BNB	0.95	0.96±0.18	0.91±0.27	**0.93±0.23**
MNB	0.98	0.74±0.35	0.89±0.29	0.80±0.35
LSVM	0.95	0.95±0.20	0.82±0.33	0.87±0.30
KNN	0.90	0.74±0.35	0.89±0.29	0.81±0.35
SGD	0.95	0.95±0.20	0.82±0.33	0.87±0.30
RF	0.94	0.93±0.24	0.73±0.35	0.80±0.35
GB	0.94	0.87±0.30	0.87±0.30	0.87±0.30
MLP	0.95	0.92±0.25	0.85±0.32	0.88±0.30

5.3. Ameaças à Validade

Na execução dos experimentos foram identificadas algumas ameaças à validade, as quais podem ser classificadas em: validade interna, validade externa, validade de construção e validade de conclusão [Wohlin et al. 2012]. A ***Validade Interna*** ameaça os fatores que podem influenciar as observações do estudo. A coleta das mensagens foi realizada no período de 01 de agosto de 2022 a 31 de dezembro de 2022, um momento de intensos debates políticos, o que pode ter aumentado a quantidade de mensagens não fraudulentas, por exemplo. Coletas realizadas em períodos distintos podem gerar conjuntos de dados diferentes, que, consequentemente, irão influenciar os resultados dos classificadores. Contudo, esses são desafios típicos dos problemas de classificação de texto. A ***Validade Externa*** ameaça a generalização das descobertas do estudo. As mensagens foram coletadas a partir de 179 grupos públicos do WhatsApp e 150 grupos/canais do Telegram, centrados no debate político. Esse recorte pode não capturar o comportamento geral dos grupos públicos brasileiros. A ***Validade de Construção*** ameaça o que diz respeito ao relacionamento entre teoria e observação. A primeira dificuldade consiste em definir teoricamente o que seria uma fraude no contexto das mensagens compartilhadas no WhatsApp e Telegram. O segundo problema consiste na rotulagem manual das mensagens, uma vez que estas podem ser rotuladas de maneira inadequada, devido à complexidade de assinalar se um texto envolve ou não uma fraude. A ***Validade de Conclusão*** ameaça a validade das conclusões. As conclusões obtidas neste trabalho somente podem ser consideradas válidas se as mensagens tiverem sido rotuladas corretamente.

6. Conclusões e Trabalhos Futuros

Neste trabalho, foram apresentados dois conjuntos de dados, denominados FraudWhatsApp.Br e FraudTelegram.Br, contendo mensagens fraudulentas em português do Brasil (PT-BR) que circularam em grupos públicos do WhatsApp e Telegram, respectivamente. Além disso, uma série de experimentos buscando construir uma solução eficiente para o problema de identificação de fraudes em textos compartilhados por meio de aplicativos de mensagens instantâneas foi discutida. O melhor resultado alcançou um escore F1 de 0,99 tanto para FraudWhatsApp.Br quanto para o FraudTelegram.Br. Como trabalho futuro pretendemos desenvolver *chatbots* que analisem os textos que trafegam em um determinado grupo e automaticamente alertem os usuários sobre a possibilidade de fraude.

Agradecimentos

Este trabalho foi parcialmente financiado pela Lenovo, como parte de seu investimento em P&D pela lei de informática. Os autores agradecem ao CNPq (316729/2021-3) e ao LSBD/UFC pelo financiamento parcial deste trabalho.

Referências

Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., and Di Franco, F. (2023). The role of machine learning in cybersecurity. *Digital Threats*, 4(1).

Ayres, L., Brito, I. V. S., and e Souza, R. G. (2019). Utilizando aprendizado de máquina para detecção automática de urls maliciosas brasileiras. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 972–985, Porto Alegre, RS, Brasil. SBC.

Barros, M., Silva, C., and Miranda, P. (2020). Xphide: Um sistema especialista para a detecção de phishing. In *Anais do XX Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 161–174, Porto Alegre, RS, Brasil. SBC.

Boukari, B. E., Ravi, A., and Msahli, M. (2021). Machine learning detection for smishing frauds. In *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC)*, pages 1–2.

de Sá, I. C., Gadelha, T., Vinuto, T., da Silva, J. W. F., Monteiro, J. M., and Machado, J. C. (2023a). A real-time platform to monitoring misinformation on telegram. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 271–278. SCITEPRESS.

de Sá, I. C., Galic, L., Franco, W., Gadelha, T., Monteiro, J. M., and Machado, J. C. (2023b). BATMAN: A big data platform for misinformation monitoring. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 237–246. SCITEPRESS.

Kumar, R. and Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *Int. J. Inf. Secur.*, 21(6):1409–1431.

Mishra, S. and Soni, D. (2023). Dsmishsms-a system to detect smishing SMS. *Neural Comput. Appl.*, 35(7):4975–4992.

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

Prabhu Kavin, B., Karki, S., Hemalatha, S., Singh, D., Vijayalakshmi, R., Thangamani, M., Haleem, S. L. A., Jose, D., Tirth, V., Kshirsagar, P. R., Adigo, A. G., and Jain, D. K. (2022). Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks. *Wirel. Commun. Mob. Comput.*, 2022.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., and Kraus, S. (2018). A study of whatsapp usage patterns and prediction models without message content. *arXiv preprint arXiv:1802.03393*.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.