DOI: https://doi.org/10.23925/ddem.v.3.n.12.68378

IDENTIFICAÇÃO E CLASSIFICAÇÃO DE DADOS SENSÍVEIS USANDO TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

SENSITIVE DATA IDENTIFICATION AND CLASSIFICATION USING NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES

Eric Henrique Da Silva Passos[1]
Lisleandra Machado[2]
Domingos Sávio da Cunha Garcia[3]
Leonardo Amorim de Araújo[4]
Samuel Alves de Freitas[5]
Ana Paula Lima dos Santos[6]
Gustavo José Santiago Rosseti[7]
Silvana Rodrigues Pires Moreira[8]

RESUMO
Este estudo investiga a aplicação de técnicas de Processamento de Linguagem Natural (PLN) e Machine Learning (ML) na identificação e classificação de dados sensíveis, com ênfase na conformidade com a Lei Geral de Proteção de Dados (LGPD). O processo inclui o pré-processamento de dados textuais, a vetorização com TF-IDF, e a implementação dos algoritmos Naive Bayes e Random Forest, com otimização de hiperparâmetros utilizando Grid Search. O desempenho dos modelos é avaliado por análises como acurácia, matriz de confusão e curva ROC. A abordagem proposta tem como objetivo auxiliar as empresas na proteção e gerenciamento de dados, garantindo o cumprimento das exigências de privacidade e segurança determinadas pela legislação.

1 Pós Graduado do curso em DATA SCIENSE e ANALYTICS – 2024 do MBA USP/ESALQ. Bacharel em direito. advericpassos@hotmail.com. https://orcid.org/0009-0002-4683-9111
2 Professora e Pesquisadora pelo CNPQ, FAPEMIG, FUNDEP e CAPES. Possui graduação em Direito, Administração de Empresas, Engenharia de Produção, Pedagogia Doutora em Engenharia de Produção pela UNIMEP e mestra em Engenharia de Produção pela UFSC - Universidade Federal de Santa Catarina. Atualmente é coordenadora do Curso de graduação em Engenharia Ferroviária e Metroviária. Tem profundos conhecimentos em Data Science y Analytics, Digital Business (Business Intelligence). Professora no Instituto Federal de Educação Ciência e Tecnologia do Sudeste de Minas Gerais - Juiz de Fora, MG. Desde 2002, é avaliadora ad hoc de cursos de graduação (INEP/MEC). lisleandra.machado@ifsudestemg.edu.br. https://orcid.org/0000-0002-7761-8023.
3 Doutor pelo Curso de História Econômica da UNICAMP. domingos.garcia@unemat.br. https://orcid.org/0000-0002-8754-6780.
4 Doutor pelo Curso de Engenharia de Transportes da UFRJ. leonardo.araujo@ifsudestemg.edu.br. https://orcid.org/0000-0003-2722-7531.
5 Mestre pelo Curso de Engenharia de Mecânica da UFSJ. samuel.freitas@ifsudestemg.edu.br. https://orcid.org/0000-0001-8985-8975.
6 Mestre pelo Curso de Engenharia Elétrica pela UFSJ. ana.santos@ifsudestemg.edu.br. https://orcid.org/0000-0002-0061-4243.
7 Doutor em Engenharia Elétrica pela UFJF. gustavo.rosseti@ifsudestemg.edu.br. https://orcid.org/0000-0002-3945-9746.
8 Doutora em Bioquímica Agrícola pela UFV. silvana.moreira@ifsudestemg.edu.br. https://orcid.org/0000-0001-7514-7216.

Palavras Chave: Machine Learning (ML); Classificação de Dados; Naive Bayes; Random Forest; Matriz de Confusão.

ABSTRACT
This study investigates the application of Natural Language Processing (NLP) and Machine Learning (ML) techniques in the identification and classification of sensitive data, with an emphasis on compliance with the General Data Protection Law (LGPD). The process includes the preprocessing of textual data, vectorization with TF-IDF, and the implementation of Naive Bayes and Random Forest algorithms, with hyperparameter optimization using Grid Search. The performance of the models is evaluated by analyses such as accuracy, confusion matrix and ROC curve. The proposed approach aims to assist companies in data protection and management, ensuring compliance with the privacy and security requirements determined by the legislation.

Keywords: Machine Learning (ML); Data Classification; Naive Bayes; Random Forest; Confusion Matrix.

INTRODUÇÃO

Nos últimos anos, o avanço das tecnologias digitais e a tecnologia de dados envolveram mudanças profundas na forma como indivíduos e organizações interagem. O ambiente em rede possibilitou um fluxo massivo de informações, que, embora tenha facilitado o desenvolvimento econômico e a globalização, também aumentou a exposição de dados pessoais e sensíveis. Essa realidade criou uma demanda urgente por regulamentações externas à proteção da privacidade dos indivíduos, gerando novas responsabilidades para as organizações empresariais.

No Brasil, a Lei Geral de Proteção de Dados (LGPD), sancionada em 2018, segue a tendência de legislações internacionais, como o Regulamento Geral de Proteção de Dados (GDPR) da União Europeia, estabelecendo um marco regulatório que visa garantir a privacidade e a segurança das informações. A LGPD tem como objetivo principal regular o tratamento de dados pessoais, seja no meio digital ou físico, e responsabilizar as empresas pelo uso inadequado ou inseguro dessas informações (Lei nº 13.709, 2018). Não se trata de uma jurisdição ao uso dos dados, mas sim de uma regulação que busca promover a transparência, o consentimento informado e a proteção dos direitos fundamentais dos titulares de informações críticas.

A aplicação da LGPD abrange um amplo escopo de entidades, incluindo pessoas jurídicas de direito público e privado, bem como indivíduos que atuam no exercício de atividades econômicas (Teixeira, 2021). Essa abrangência impõe às organizações a necessidade de se adequarem, criando políticas e mecanismos robustos de proteção de dados. Isso inclui não

apenas o tratamento de dados pessoais comuns, mas, em especial, dos chamados dados sensíveis, que podem revelar origem racial ou étnica, convicções religiosas, opiniões políticas, dados de saúde, entre outros. A proteção desses dados é particularmente importante, pois a manipulação convencional pode resultar em impactos severos nas liberdades e direitos fundamentais garantidos pela Constituição Federal de 1988 (Lima, 2021).

A complexidade da LGPD não se limita apenas à distinção entre dados pessoais e sensíveis. Existe também o desafio de identificar quando uma combinação de dados pessoais aparentemente inofensivos pode gerar informações sensíveis, exigindo um nível de tratamento mais específico. Essa questão é particularmente relevante em um cenário de big data, no qual grandes volumes de informações são processados e cruzados constantemente (Lima, 2021). Nessa perspectiva, a identificação e classificação automatizada de dados, por meio de ferramentas tecnológicas avançadas, torna-se uma estratégia fundamental para garantir a conformidade com a lei e mitigar os riscos de exposição de informações.

O Processamento de Linguagem Natural (PLN) e o Machine Learning (ML) surgem como tecnologias-chave para lidar com essa complexidade. O PLN permite que sistemas computacionais compreendam e analisem grandes volumes de dados textuais de forma eficiente, enquanto o AM possibilita que os algoritmos aprendam padrões e realizem predições com base em grandes conjuntos de dados. Essas técnicas, quando combinadas, podem automatizar a identificação de dados pessoais e sensíveis, oferecendo uma solução eficaz para empresas que buscam se adequar às exigências da LGPD. Além disso, uma aplicação de algoritmos como Naive Bayes e Random Forest, amplamente utilizada no aprendizado supervisionado, contribui para a construção de modelos preditivos robustos.

Considerando esses desafios e oportunidades, esta pesquisa visa investigar a aplicação de técnicas de PLN e ML para identificar e classificar dados pessoais, não pessoais e sensíveis. O objetivo principal é desenvolver uma abordagem automatizada que auxilie as empresas a atender aos requisitos da LGPD, implementando medidas de proteção adequadas e minimizando os riscos de vazamento e violação de dados.

A relevância deste estudo encontra a interseção entre a necessidade crescente de proteção de dados e a aplicação de tecnologias avançadas para facilitar esse processo. À medida que as empresas enfrentam uma pressão crescente para se adequarem à LGPD, a automatização da identificação e classificação de dados pessoais e sensíveis pode oferecer uma solução eficiente, ao mesmo tempo em que garanta o cumprimento das regulamentações e preserve os direitos dos indivíduos.

1. MATERIAL E MÉTODOS

Este trabalho adota uma abordagem aplicada, utilizando técnicas de PLN e ML para a identificação e classificação de dados sensíveis, pessoais e não pessoais, com foco em garantir a conformidade com a Lei Geral de Proteção de Dados (LGPD). O estudo foi conduzido com dados de uma empresa do setor de seguros, com mais de 100 anos de atuação, que conta com um portfólio de mais de sete milhões de clientes e cerca de cinco mil funcionários.

1.1. Conjunto de Dados

A base possui 12.126 (doze mil cento e vinte e seis) linhas, sendo que 1.299 (um mil duzentos e noventa e nove) são da classe sensível, 3.849 (três mil oitocentos e quarenta e nove) são da classe não pessoal, e 7.474 (sete mil quatrocentos e setenta e quatro) são da classe pessoal.

A Tabela 1 representa uma amostra desses dados:

Conjunto de Tabelas	Tabela	Coluna	Tipo	Descrição	Rótulo
owlandmdmprd	bureautransunionpfans	cpf	string	número cpf	Pessoal
owlandmdmprd	bureautransunionpfans	nme completo	string	nome completo	Pessoal
owlandmdmprd	bureautransunionpfans	cns	string	nome cartão nacional saúde	Pessoal
owlandmdmprd	bureautransunionpfans	nme operadora sauda ans	string	número operadora nacional saúde ans	Não Pessoal
owlandmdmprd	bureautransunionpfans	plano	string	número plano	Não Pessoal

Tabela 1. Amostra do conjunto de dados
Fonte: Dados originais da pesquisa

Nesses dados, a variável “Conjunto de Tabelas” representa o conjunto global de informações. A variável “Tabela” designa o nome de uma tabela derivada desse conjunto de dados. A variável “Coluna” especifica o nome da coluna contida na tabela. A variável “Tipo” especifica o formato de dados presente na respectiva coluna. A “Descrição” contém uma breve

explicação do conteúdo daquela coluna. Finalmente, a variável “Rótulo” fornece informações sobre a natureza do conteúdo da coluna, classificando-a com pessoal, não pessoal ou sensível.

1.2. Pré-processamento

Para as etapas de pré-processamento, análise e modelagem, foi utilizada a plataforma Databricks, compatível com a Linguagem Python 3.9.5, em um cluster de 14 GB de memória e 4 núcleos.

1.3. Modelos de Classificação

Para realizar as classificações foram empregados dois modelos de ML, o Naive Bayes e o Random Forest.

Naive Bayes: é um método que funciona com base no teorema de Bayes, fórmula utilizada para calcular a probabilidade condicional de um evento acontecer com base em informações prévias (Vajjala et al., 2020).

Random Forest: é um algoritmo baseado em árvore de decisão que cria um conjunto, uma floresta de árvores de decisão (Louppe, 2014). Na onde a estrutura da árvore é semelhante a um fluxograma, onde cada nó representa uma decisão e as ramificações indicam os possíveis caminhos a seguir com base nas respostas às perguntas feitas em cada nó.

1.4. Métricas de Avaliação

Para avaliar o desempenho de ambos os modelos, foi adotada a matriz de confusão, examinada a acurácia, e também o gráfico e os valores dada curva ROC (Receiver Operating Characteristic), ferramentas utilizadas em contextos de classificação.

Conforme descrito por Jonhson e Kuhn (2016), a matriz de confusão é uma representação simples das classes observadas e previstas em um conjunto de dados. Em contraste, a acurácia é uma medida da precisão global do modelo, indicando a correspondência entre as classes observadas e as previstas.

A acurácia pode ser calculada usando a seguinte equação eq.1:

$$Acurácia = \frac{Número\ de\ Predições\ Corretas}{Número\ Total\ de\ Predições} \quad (1)$$

Ainda, com base nos ensinamentos de Johnson e Kuhn (2016), essa métrica não faz distinção entre os tipos de erros cometidos, por exemplo a filtragem de spam, onde o custo de excluir erroneamente um e-mail importante é provavelmente maior do que permitir que um e-mail de spam passe pelo filtro de forma incorreta.

De outro modo, a curva é ROC um método que integra duas medidas condicionais, a sensibilidade e a especificidade, em um único valor. A sensibilidade representou a precisão apenas para a população de eventos, enquanto a especificidade foi aplicada aos não eventos (Johnson e Kuhn, 2016).

A fórmula a seguir ilustra o cálculo da taxa de verdadeiro positivo eq.2:

$$TVP = \frac{Sensibilidade\ x\ Prevalência}{(Sensibilidde\ x\ Prevalência)+((1-Especificidade)\ x\ (1-Prevalências))} \quad (2)$$

onde a especificidade representa a probabilidade de diagnosticar corretamente um evento quando ele realmente ocorre, a prevalência é a proporção de verdadeiros eventos, e a especificidade é a probabilidade de diagnosticar corretamente um não evento quando ele realmente não ocorre.

Da mesma forma, a equação a seguir calcula a taxa de verdadeiros negativos eq.3:

$$TVN = \frac{Especificidade\ x\ (1-Prevalência)}{(Prevalência\ x\ (1-Sensibilidade))+(Especificidade\ x\ (1-Prevalência))} \quad (3)$$

2. RESULTADOS E DISCUSSÃO

As variáveis e as observações utilizadas para o pré-processamento, treinamento e teste dos modelos de ML foram apresentadas na Tabela 1 que, com exceção da variável Rótulo, teve todas as suas variáveis (strings) concatenadas em apenas uma coluna.

Isso porque o conteúdo dessas variáveis tem no máximo duas palavras, e se o TF-DF fosse aplicado a cada variável individualmente, o resultado das previsões seria prejudicado devido à dificuldade de identificar palavras relevantes em observações com apenas uma ou duas palavras.

No início do pré-processamento, foi realizada a limpeza dos dados textuais para garantir a qualidade das informações, eliminando caracteres especiais e reduzindo espaços em branco.

Após a conclusão do pré-processamento, foi realizada a etapa de vetorização dos dados, por meio da técnica TF-IDF que, em tradução livre para o português, significa Frequência do Temo – Frequência Inversa nos Documentos (Vajjala et al., 2020). Essa vetorização permitiu transformar os dados de texto em um formato numérico compreensível pelos algoritmos de ML.

O TF-IDF atribuiu um valor ponderado a cada palavra, considerando a relação entre sua frequência em um documento específico e sua frequência em todo o conjunto de documentos, realçou palavras-chave relevantes, enquanto reduziu o peso dos termos comuns que não acrescentaram um impacto significativo (Vajjala et al., 2020).

O TF (frequência do termo) mediu com que frequência um termo (palavra) aparece em um determinado documento. Como os documentos em um corpus podem variar em comprimento, um termo pode ocorrer mais frequentemente em um documento mais longo em comparação com um documento mais curto. Para ajustar essas contagens, a técnica TF-IDF dividiu o número de ocorrências pelo comprimento do documento (Vajjala et al., 2020).

O cálculo do TF é expresso na seguinte equação eq.4:

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{(\text{Número de ocorrências do termo t no documento d})}{(\text{Total de termos no documento d})} \qquad (4)$$

onde o termo t: é a palavra; e o documento d: o texto na onde a palavra está inserida.

Por outro lado, o IDF (frequência inversa do documento), avaliou a relevância de um termo em um corpus, penalizando termos muitos comuns em prol dos mais raros, ou seja, aqueles que ocorrem com menor frequência (Vajjala et al., 2020).

Quanto mais raro foi o termo, maior foi o valor IDF e, consequentemente, maior a sua importância quando utilizado no cálculo do TF-IDF para um documento específico. Esse cálculo é expresso na seguinte equação eq.5:

$$\mathrm{IDF}(\mathrm{t}) = \log_{\mathrm{e}} \frac{(\text{Número total de documentos no corpus})}{(\text{Número de documentos que contêm o termo t})} \qquad (5)$$

O resultado foi uma matriz numérica, em que cada linha corresponde a um documento, e cada coluna representa uma palavra única presente nos documentos.

A partir dos dados da Tabela 1, o algoritmo Naive Bayes estimou a probabilidade condicional de cada característica de um texto para cada classe, com base na ocorrência dessa característica naquela classe específica (Vajjala et al., 2020).

De acordo com Witten et al. (2021), a aplicação do Naive Bayes no contexto de classificação é representada pela seguinte fórmula eq. 6:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)} \quad (6)$$

onde Pr representa a probabilidade condicional de que a variável resposta Y seja k, dado um valor específico da variável preditora X, em outras palavras, é a probabilidade de que uma observação pertença à classe k dado um valor particular para as características x; π_k representa a probabilidade de uma observação aleatória pertencer à classe k; $f_k(x)$ é a função de densidade condicional de X para a classe k, representa a probabilidade de observar um terminado valor (x) das características preditoras (X) dado que a observação pertence à classe k, descreve como os valores de X estão distribuídos na classe k; $\sum_{l=1}^{k} \pi_l$ representa o somatório para l começando de 1 até k que é o número total de classes, π_l é a probabilidade de uma observação pertencer a classe l; $f_l(x)$ representa a probabilidade de observar um determinado valor x das características preditoras X dado que a observação pertence à classe l, descreve como que os valores de X estão distribuídos na classe l.

De outro modo, o Random Forest, algoritmo baseado em árvore de decisão, criou um conjunto, uma floresta de árvores de decisão (Louppe, 2014). A estrutura dessa árvore é semelhante a um fluxograma, onde cada nó representou uma decisão e as ramificações indicaram os possíveis caminhos a seguir com base nas respostas às perguntas feitas em cada nó.

Melhor explicando, segundo Witten et al. (2021), ao construir uma árvore de decisão, uma amostra de m preditores é selecionada para cada divisão a partir do conjunto completo de p preditores. Nesse contexto, a divisão foi restrita a usar apenas um dos m preditores selecionados. A cada divisão, uma nova amostra de m preditores foi escolhida, sendo que geralmente m é aproximadamente igual a raiz quadrada de p. Esse método introduziu uma dose

de aleatoriedade, assegurando que as árvores fossem mais diversas e, consequentemente, menos correlacionadas entre si, provocando uma melhoria significativa na precisão do modelo.

A modelagem do Random Forest para a tarefa de classificação é formalizada pela equação eq. 7:

$$\{h(x, \theta_k)k = 1, 2 \dots\} \quad (7)$$

onde h refere-se a uma coleção de arvores, k é um indexador variando a partir de 1, e θ_k representa os parâmetros específicos de cada árvores na coleção.

No Random Forest foi aplicada a técnica de otimização conhecida como Grid Search, procedimento que testa as combinações dos valores de hiperparâmetro fornecidos para encontrar o melhor modelo.[9]

Antes de iniciar o processo de Grid Search, os hiperparâmetros foram estruturados em um dicionário utilizando a linguagem Python do seguinte modo: para os tuned_params_v2, os valores foram definidos como variando entre 100, 200, 300, 400 e 500. Para min_samples_split, foram considerados os valores: 2, 5 e 10, enquanto para min_samples_leaf, os valores utilizados foram 1, 2 e 3.

O parâmetro 'n_estimators' representa o número de árvores na floresta que serão criadas durante o treinamento. O 'min_samples_split' indica o número mínimo de amostras que um nó deve ter para ser dividido em sub-nós, enquanto o 'min_samples_leaf' indica o número mínimo de amostras necessárias para um nó se tornar uma folha, ou seja, sem sub-nós.

Por meio desse dicionário, o Grid Search explorou todas as combinações possíveis desses valores, treinando um modelo para cada configuração e avaliando o desempenho resultante.

Ao final do processo, identificou-se a configuração ideal que levou à criação do modelo mais eficaz para classificar os dados de texto.

Tanto o algoritmo Random Forest e o Nave Bayes quanto o Grid Search foram acessados por meio da biblioteca de código aberto scikit-learn, versão 1.3.1, e sua documentação e os exemplos de valores para os hiperparâmetros estão disponíveis em:

9 . Disponível em: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em 11/11/2024.

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Para avaliar o desempenho de ambos os modelos, foi adotada a matriz de confusão, examinada a acurácia, e também o gráfico e os valores da curva ROC.

A acurácia do modelo Random Forest foi de 0.93 enquanto a do Nave Bayes foi de 0.89. Ressalta-se que a variável "Rótulo" no conjunto de dados ilustrado na Tabela 1 não possui classes balanceadas, portanto, a acurácia foi analisada com cuidado e em conjunto com as outras métricas. Isso se deve ao fato de que essa métrica não lida bem com cenários desse tipo.

As Figuras 1 e 2 a seguir ilustram o resultado da matriz de confusão para ambos os modelos e detalham as previsões para cada classe:

Matrix de Confusão			
Não Pessoal	1769	45	4
Pessoal	36	911	15
Sensível	38	58	229

Figura 1. Matriz de Confusão do Modelo Random Forest
Fonte: Dados originais da pesquisa

Matrix de Confusão			
Não Pessoal	1819	45	4
Pessoal	97	843	22
Sensível	59	97	169

Figura 2. Matriz de Confusão do Modelo Naive Bayes
Fonte: Dados originais da pesquisa

Sobre a curva ROC, a taxa de verdadeiros positivos e a taxa de falsos positivos são representadas em um gráfico, onde uma é plotada em relação à outra (Johnson e Kuhn, 2016). Os gráficos das Figuras 3 e 4, ilustram essa relação e os resultados obtidos:

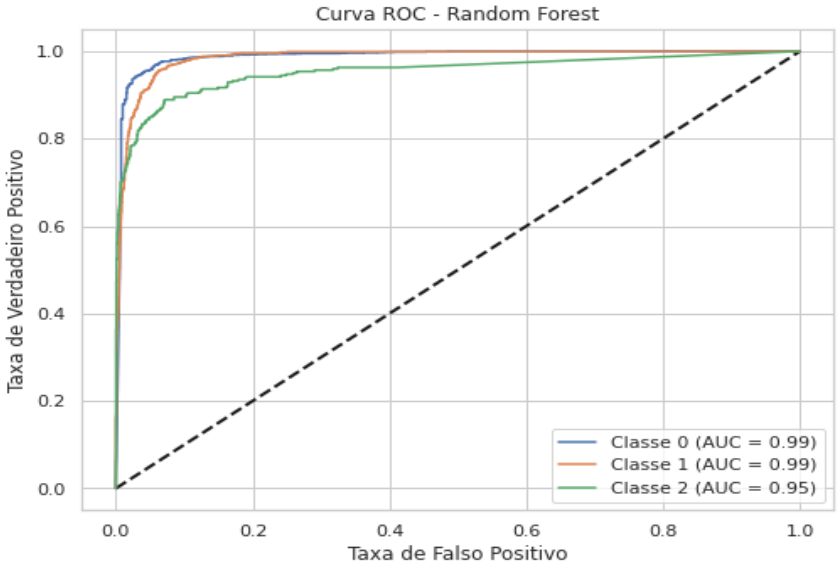

Imagem 3. Gráfico da Curva ROC do Modelo Random Forest
Fonte: Dados originais da pesquisa

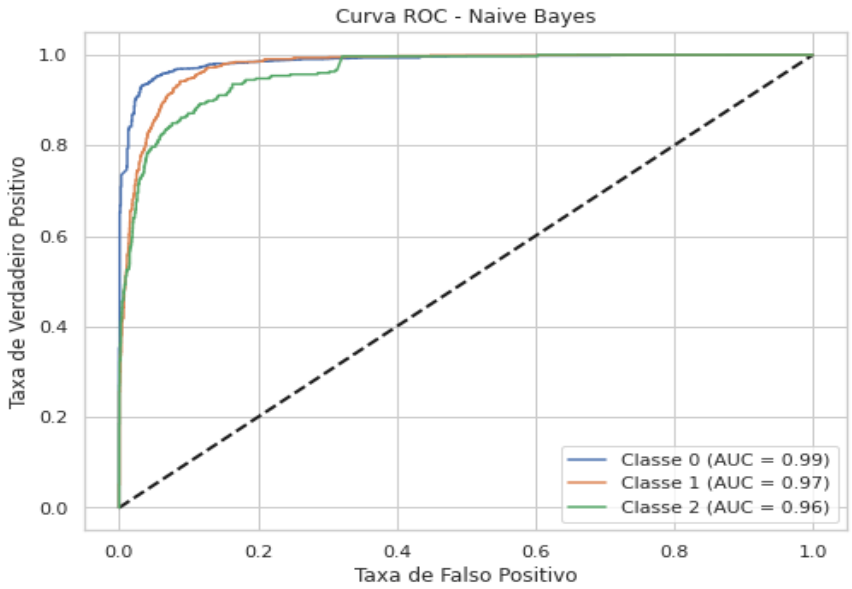

Imagem 4. Gráfico da Curva ROC do Modelo Naive Bayes
Fonte: Dados originais da pesquisa

Quanto mais perto das extremidades esquerda e superior estiverem as curvas de cada classe, melhor o modelo está. No entanto, assim como a acurácia, essa análise foi combinada com a matriz de confusão para uma avaliação mais completa e confiável.

CONSIDERAÇÕES FINAIS

Os resultados obtidos foram positivos, refletindo a eficácia dos algoritmos de Machine Learning aplicados a um conjunto de dados com características particulares. A similaridade dentro de cada classe, combinada com distinções marcantes entre as diferentes classes, foi um fator crucial para o bom desempenho dos modelos. O Random Forest, em particular, demonstrou uma capacidade superior de capturar essas diferenças, resultando em maior acurácia e melhores métricas de classificação comparado ao Naive Bayes.

No entanto, é importante considerar que o conjunto de dados apresentava classes desbalanceadas, o que influenciou as métricas de desempenho, como a acurácia. Para trabalhos futuros, seria interessante explorar técnicas de balanceamento de classes, como oversampling ou undersampling, para melhorar a performance, especialmente nas classes minoritárias.

Além disso, embora o TF-IDF tenha sido eficaz para a vetorização dos dados textuais, o uso de técnicas mais avançadas, como Word Embeddings (por exemplo, Word2Vec ou BERT), poderia fornecer representações semânticas mais ricas, potencialmente melhorando ainda mais os resultados.

Por fim, a aplicação do Grid Search para otimização dos hiperparâmetros demonstrou ser uma abordagem eficiente para melhorar a performance dos modelos, mas há espaço para expandir essa análise com técnicas mais avançadas, como Random Search ou Bayesian Optimization, que poderiam refinar ainda mais a escolha dos parâmetros.

Os resultados sugerem que, com o uso de técnicas de pré-processamento adequadas e modelos bem ajustados, é possível alcançar bons resultados em tarefas de classificação textual. A continuação desse estudo, com dados mais variados e ajustes adicionais, pode trazer ainda mais insights sobre a aplicabilidade dessas técnicas em diferentes contextos.

REFERÊNCIAS

BRASIL. 1988. **Constituição da República Federativa do Brasil. Estabelece a Constituição Federal do Brasil**. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 29 de agosto de 2023.

BRASIL. 2018. **Lei nº 13.709, de Agosto de 2018**. Dispõe de maneira geral sobre a proteção de dados no âmbito nacional. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 29/08/2023.

JOHNSON, Kjell; KUHN, Max. 2016. **Applied Predictive Modeling**.1ed. Editora Springer. New York, USA. Disponível em: https://www.ic.unicamp.br/~wainer/cursos/1s2021/432/2013_Book_AppliedPredictiveModeling.pdf. Acesso em: 04 de outubro de 2023.

LIMA, Ana P. M. C.; CRESPO, Marcelo; PINHEIRO, Patricia P.2020. **LGPD aplicada.** Editora Atlas, São Paulo, SP, Brasil. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788597026931/epubcfi/6/10[%3Bvnd.vst.idref%3Dcopyright]!/4/12/4/1:0[%2Cby. Acesso em: 29 de setembro de 2023.

LIMA, Rosa. P. Lima. 2021. ANPD e LGPD: **Desafios e perspectivas.** Editora Almedina, São Paulo, SP, Brasil. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9786556272764/pageid/50. Acesso em: 29 de setembro de 2023.

LOUPPE, Gilles. 2014. **Understanding Random Forest: From Theory to Practice.** PhD Dissertation. University Of Liége. Liége, Bélgica. Disponível em: https://arxiv.org/pdf/1407.7502.pdf. Acesso em: 02 de setembro de 2023.

OLIVEIRA, Adrielly L. S.; SANTOS, Alessandra, P. B.; LIRA, Bruno B. L.; ABRÃO, Bianca, B.; CAMARGO, Caio P. F. 2022. **LGPD e a Proteção de dados pessoais na SOCIEDADE EM REDE**: Dados de Criança e Adolescentes na Internet; Tratamento de Proteção de Dados no Comércio Eletrônico; Proteção de Dados Falecidos; Violação de Direitos da Personalidade e Responsabilidade Civil.1ed. Editora Almedina, São Paulo, SP, Brasil. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9786556276373/pageid/3. Acesso em: 28 de setembro de 2023.

TEIXEIRA, Tarcisio. 2021. **LGPD e E-commerce**. 2ed. Editora Saraiva, São Paulo, SP, Brasil. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9786555598155/epubcfi/6/36[%3Bvnd.vst.idref%3Dmiolo15.xhtml]!/4. Acesso em: 20 de setembro de 2023.

VAJJALA, Sowmya. MAJUMDER, Bodhisattwa. GUPTA, Anuj. SURANA Harshit. 2020. **Practical Natural Language Processing:** A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media, Inc., Sebastopol, CA, USA. Disponível em: https://www.oreilly.com/library/view/practical-natural-language/9781492054047/. Acesso em: 08 de setembro de 2023.

WITTEN, Daniela; JAMES Gareth; TIBSHIRANI. 2021. **An introduction to statistical learning**: Witch Application in R. Editora Springer. New York, USA. Disponível em: https://www.stat.berkeley.edu/users/rabbee/s154/ISLR_First_Printing.pdf. Acesso em: 01 de outubro de 2023.

Recebido – 18/09/2024
Aprovado – 31/10/2024