risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 15/03/2024
Aceitação/Acceptance: 20/05/2024

Uma Aplicacão para Explicabilidade de Predições de um SVM em Tweets de COVID-19

Ivo de Abreu Araújo[1], Renato Hidaka Torres[2], Nelson Cruz Sampaio Neto[3]

ivoabreu94@gmail.com; renatohidaka@ufpa.br; nelsonneto@ufpa.br

[1] Universidade Federal do Sul e Sudeste do Pará, 68638-000, Rondon do Pará – PA - Brasil

[2,3] Universidade Federal do Pará - Faculdade de Computação, 66075-110, Belém – PA – Brasil

DOI: 10.17013/risti.54.121–137

Resumo: Este trabalho propõe uma aplicação web que usa um modelo de caixa preta *Support Vector Machine (SVM)* com 79% de acurácia para classificar o sentimento de *tweets* sobre a COVID-19 integrando o *framework LIME* de forma interativa para explicar decisões sobre previsões. Além do ganho de transparência em relação à avaliação de amostras falso-positivas, notou-se também que o modelo SVM tende a falhar ao associar um teste positivo de COVID-19 a um bom sentimento e se confunde em previso˜es envolvendo palavras sobre a COVID-19, como *Omicron*, que indica falta de representatividade na base de dados. Além disso, a partir dos resultados do *LIME*, foi possível melhorar a acura´cia do modelo para 81% ao incluir as *stopwords* "*not*" e "*no*".

Palavras-chave: Processamento Natural de Linguagem; Framework de Explicabilidade; Explicabilidade de Modelos de Caixas Pretas; Aprendizado de Máquina.

An Application for Explicability of SVM Predictions on COVID-19 Tweets

Abstract: The present work proposes a web application that uses an Support Vector Machine (SVM) black box model with 79% accuracy to classify sentiment from tweets about COVID-19 integrating the LIME framework in an interactive way to explain decisions about predictions. Besides the gain in transparency in relation to the evaluation of false positive samples, it was also noted that the SVM model tends to fail when associating a positive COVID-19 test with a good sentiment and gets confused in specific predictions involving words related to COVID-19 variants such as Omicron, which indicate lack of representativeness in the database. In addition, from the LIME results, it was possible to improve the model accuracy to 81% by including the stopwords "not" and "no".

Keywords: Natural Language Processing; Explicability framework; Explicability of black box models; Machine learning.

1. Introdução

A pandemia causada pelo vírus SARS-CoV2 afetou o mundo e obrigou os países a tomarem decisões com o intuito de reduzir os consequentes impactos sociais, econômicos, culturais e políticos. Dentre as medidas tomadas, pode-se citar o isolamento social, que gerou como consequência uma maior interação das pessoas por meio das mídias sociais. Além disso, mídias sociais, como o Twitter, tornaram-se relevantes veículos de notícias e opiniões sobre a doença, bem como o uso de tecnologias para o apoio de processos estratégicos em setores da sociedade que gerenciam pessoas como turismo (Khan et al., 2020; López et al., 2023).

Com a imposição das restrições físicas, o uso de mídias sociais potencializou um envolvimento mais ativo de usuários com manifestações públicas por meio de mensagens de texto, que levou áreas como a de ensino a adaptarem seus profissionais a metodologias de ensino remotas, gerando assim uma base de dados onde é possível constatar experiências vividas pela população mundial em decorrência da COVID-19 (Aguayo et al., 2021).

Nessa perspectiva, Meena & Bai (2019) definem que as redes sociais são plataformas que propagam informações volumosas que permitem explorar e compreender sentimentos de indivíduos de uma sociedade. A análise dessas informações pode ajudar no desenvolvimento de ações que amenizem os problemas de saúde mental provocados pelas restrições da pandemia. De acordo com Recuero (2009), os termos rede social e mídia social são conceitualmente diferentes. Embora os sites de redes sociais atuem como suporte para as interações que constituirão as redes sociais, eles não são, por si, redes sociais. Assim, sistemas como o *Twitter* serão considerados mídias sociais, ou seja, um meio capaz de manter uma rede social ativa.

À vista disso, tem sido desafiador o uso de aplicações robustas de aprendizado de máquina na área de *NLP (Natural Language Processing)* para mapear opiniões e sentimentos diante de cenários inesperados, como o provocado pela COVID-19, com o intuito de adequar soluções inteligentes no âmbito de problemas sociais e psicológicos (Glowacki et al., 2021; Kolluri & Murthy, 2021).

Um dos atuais desafios em aprendizado de máquina é o uso de algoritmos com alta capacidade preditiva que são inerentemente não transparentes em suas decisões, uma vez que são considerados caixas pretas devido a complexidade que envolve suas funções internas (Molnar, 2019). Já o termo caixa branca é o oposto de caixa preta, pois, se refere a algoritmos que funcionam com base em estruturas mais simples de serem compreendidas pelo ser humano, por exemplo, a profundidade de uma árvore de decisão.

Dessa maneira, algoritmos de caixa branca garantem uma maior confiabilidade aos resultados preditivos. Nesse contexto, Nielsen (2020) define que a interpretabilidade de um algoritmo está diretamente relacionada com a compreensão do usuário quanto ao que o modelo decidiu em uma predição, por exemplo, os pesos de uma regressão linear, enquanto que a explicabilidade busca justificar como o modelo tomou uma decisão. Já Ahmad (2020) afirma que a explicabilidade de um algoritmo é desenvolvida pela capacidade qualitativa de identificar features que estão direta ou indiretamente

relacionadas a uma predição. Dado os conceitos apresentados, o presente artigo visa explorar a explicabilidade de um algoritmo de caixa preta.

Uma vez que medidas de tendência central, como *score* e *recall*, não são suficientes para um entendimento minucioso e explicativo de algoritmos de caixa preta (Ribeiro et al., 2016), *frameworks* de explicabilidade, como *LIME (Local Interpretable Model-Agnostic Explanations)* e *SHAP (Shapley Additive Explanations)*, têm se popularizado. Esses *softwares* tentam explicar predições de modelos de aprendizado de máquina em contextos locais, explorando *features* com a finalidade de entender decisões internas de algoritmos de caixa preta (Lundberg & Lee, 2017).

Tendo em vista a importância de se compreender modelos de aprendizado de máquina mais robustos, principalmente aqueles que entregam soluções pouco transparentes, este artigo propõe estudar a explicabilidade do algoritmo de caixa preta *SVM (Support Vector Machine)* na tarefa de classificar sentimentos em textos coletados de uma mídia social. Uma análise sobre a ação de palavras vazias (ou *stopwords*, em inglês) na acurácia do modelo também é apresentada. Para isso, foi desenvolvida uma aplicação *web* no intuito de facilitar a interação com o *framework LIME*. A revisão bibliográfica realizada não encontrou trabalhos que explorassem o uso do *LIME* em tarefas de proces samento de língua no *Twitter*.

Assim, o objetivo principal é classificar sentimentos em mensagens de usuários no Twitter sobre a COVID-19 e compreender a influência de certas palavra nas predições. O presente trabalho também busca responder as seguintes questões: O estudo da explicabilidade ajuda a determinar falta de representatividade na base de dados? A identificação de palavras mais ou menos influentes pode reduzir a ocorrência de predições falso-positivas? O uso de determinadas stopwords é capaz de melhorar a acurácia do modelo preditivo?

Os resultados obtidos demonstraram que a classificação de sentimentos de *tweets* sobre a COVID-19, agregada com a explicação da predição, possibilita a compreensão das palavras mais relevantes, mesmo quando o modelo treinado falha em classificar determinadas amostras. Outro destaque é a melhora de desempenho do modelo *SVM* após o uso de *stopwords*. Ao serem mantidas as palavras de negação "*not*" e "*no*" durante um novo processo de limpeza e treinamento da base de dados, a acurácia do modelo incrementou de 79% para 81%.

O restante do artigo encontra-se estruturado nas seguintes seções. A seção 2 apresenta trabalhos relacionados a pesquisa desenvolvida. A seção 3 trata do referencial teórico. A seção 4 descreve os materiais e métodos empregados. Na seção 5, os resultados são apresentados e discutidos. Por fim, a seção 6 traz as considerações finais e os trabalhos futuros.

2. Trabalhos relacionados

Esta seção apresenta uma descrição de estudos relacionados com análise de sentimentos e explicabilidade em aprendizado de máquina. É importante destacar que não foram encontrados artigos que seguiram a linha de pesquisa do presente trabalho, ou seja,

explicabilidade com o uso do *LIME* em *NLP*. O contexto escolhido foi a COVID-19 na mídia social *Twitter*.

Em Silva et al., (2021), é realizado um estudo na área de *NLP* no contexto de mensagens do *Twitter* relacionadas ao uso do serviço de saúde brasileiro durante a pandemia de COVID-19. A coleta dos dados ocorreu no período de dezembro de 2019 a outubro de 2020 por meio da ferramenta *Twint*. Para a classificação das mensagens, os autores utilizaram o classificador *NRC sentimen*t com três classes: Positiva, Negativa e Neutra. Os resultados obtidos apresentaram um mapeamento de sentimentos durante cada mês do período de coleta dos dados, por exemplo, em fevereiro de 2019, houve uma predominância de sentimentos positivos, por conta das menções ao programa "Previne Brasil". É interessante destacar que não houve o uso de um *framework* de explicabilidade para buscar entender quais termos eram mais importantes para a decisão da classificação.

Nessa mesma perspectiva, Garcia & Berton (2021) fizeram uma análise de sentimentos no *Twitter* em mensagens de texto do Brasil e dos EUA visando identificar opiniões sobre a pandemia de COVID-19. Classificadores de Regressão Logística, *Random Forest* e *SVM* foram comparados e obtiveram desempenhos similares na maioria dos testes realizados em uma base de dados coletada entre abril e agosto de 2020. Os resultados mostraram uma predominância de sentimentos negativos nas duas línguas no assunto de "cuidados com a proliferação do vírus". Esse trabalho também não fez um estudo de explicabilidade sobre os modelos de caixa preta utilizados.

O artigo de Ayoub et al., (2021) apresentou um estudo de explicabilidade em mensagens falsas (ou *fakenews*, em inglês) coletadas na Internet e relacionadas a COVID-19. Para o desenvolvimento da pesquisa, os autores utilizaram o *framework SHAP* com o objetivo de entender o comportamento de um classificador *DistilBert*. O artigo reporta a identificação das palavras mais relevantes para a classificação de uma nova notícia e, por conseguinte, um melhor entendimento da lógica de decisão seguida pelo modelo, principalmente nos equívocos.

Seguindo na linha da explicabilidade, Liu et al., (2021) propõem um estudo para classificar sentimentos em mensagens do *Twitter* com temática plural e entender predições equivocadas de uma rede neural utilizando o *LIME*. A aplicação do *framework* permitiu ranquear as palavras mais influentes para a resposta do modelo e, assim, analisar as classificações que falharam. Além da explicabilidade local, os autores também consideraram a análise da explicabilidade global no seu estudo.

Em Behl et al., (2021), os autores realizaram uma análise de dados sobre situações emergenciais de desastres naturais. O artigo explorou a explicabilidade de um modelo *MLP (Multilayer Perceptron)* treinado em bases de dados referentes a terremotos no Nepal e na Itália com o objetivo de classificar as mensagens nas seguintes classes: "necessidade de ajuda", "oferta de ajuda" e "outros". Além dos testes realizados nessas bases de dados, o modelo também foi aplicado em mensagens de situações emergenciais da pandemia de COVID-19 no *Twitter*, resultando em 83% de acurácia para predições nas três classes mencionadas. Com relação a explicabilidade, o uso do framework *LIME* permitiu identificar as palavras mais importantes para a classificaão do *MLP*.

Diferentemente dos artigos citados acima, que não dispõem de uma aplicação integrada, Kouvela et al., (2020) propõe um sistema *web* que usa o algoritmo de caixa preta *Random*

Forest para identificar usuários *bots* no *Twitter*. Para isso, o algoritmo foi treinado com uma base de dados contendo mensagens coletadas com a API do *Twitter*. Por fim, o *framework LIME* foi utilizado para gerar explicações em 20 *tweets* de cada perfil de usuário classificado pela caixa preta como legítimo ou *bot*.

3. Revisão da Literatura

Nos últimos anos, a explicabilidade tornou-se um fator importante quando o propósito é explorar comportamentos de predições em aprendizado de máquina, principalmente quando se trata do desafio de entender decisões de algoritmos de caixa preta em contextos em que a compreensão de resultados preditivos é essencial (Hall et al., 2021).

Nesse cenário, a inteligência artificial explicável, também conhecida como *XAI (Explainable Artificial Intelligence)*, complementa o conhecimento sobre decisões de algoritmos de uma maneira transparente com explicações compreensíveis no nível humano (Rothman 2020, Turek 2021).

Com relação à explicabilidade de aprendizado de máquina, Ahmad (2020) destaca dois tipos de abordagens importantes de serem compreendidas:

Explicabilidade local: Visa prover explicações mais minuciosas em pequenas regiões de um modelo, onde cada região diz respeito a uma amostra do *dataset*. Por exemplo, essa é a proposta do *framework LIME*, que gera explicabilidade apenas para uma amostra específica, haja vista que amostras individuais de um modelo são mais suscetíveis ao uso de regressões lineares do que toda a base de dados (Ribeiro et al., 2016).

Explicabilidade global: É interessante quando se busca uma visão completa do comportamento do modelo em predições. No entanto, esse tipo de explicação é mais complexa, considerando que a tarefa de encontrar relações entre as *features* com base no comportamento geral do modelo é mais desafiadora. Árvores de decisão e regressão linear são exemplos de algoritmos que inerentemente permitem serem compreendidos na sua totalidade.

3.1. Framework LIME

O *LIME* é uma solução de *framework* que tem se popularizado na área de aprendizado de máquina por explicar predições de qualquer modelo de aprendizado supervisionado. Conforme Ribeiro et al., (2016), o *LIME* tem como objetivo auxiliar desenvolvedores e usuários finais a compreenderem predições por meio de explicações baseadas em um modelo explicável.

O processo de explicação do *LIME* em *NLP* acontece localmente, ou seja, a partir de uma amostra específica escolhida pelo usuário. O *LIME* gera, aleatoriamente, amostras artificiais com o propósito de entender qual é o impacto de cada palavra presente no texto da amostra original. Para isso, cada nova amostra artificial tem como base o texto original, no entanto, com a remoção de algumas palavras.

Em seguida, o modelo de caixa preta (alvo da explicabilidade) realiza uma classificação para cada amostra artificial, verificando quais foram classificadas na mesma classe da amostra original e atribui um peso para cada palavra, baseado no seu impacto para a

saída do modelo. Por fim, o *LIME* aplica uma regressão linear (modelo substituto) para buscar explicar de forma aproximada a relação dos pesos de cada palavra com a predição da amostra original.

A função de explicação do *LIME* é dada por (1), otimizada para gerar a explicação mais próxima da predição original:

$$\xi(x) = \frac{argmin}{g \in G} L(f, g, \pi x) + \Omega(g) \quad (1)$$

A explicação ξ dada pelo LIME para uma amostra x é obtida por um modelo substituto $g \in G$, onde G representa um conjunto de algoritmos explicáveis. Já $\Omega(g)$ diz respeito a complexidade do modelo substituto, ou seja, quanto mais baixa, maior a capacidade de explicabilidade a um nível humano. Por convenção, utiliza-se como modelo substituto uma regressão linear otimizada com uma função de erro $L(f, g, \pi x)$ que busca reduzir a complexidade da explicação, onde L é a função que minimiza o erro, por exemplo, uma regularização *Lasso*; f representa o modelo original, por exemplo, um *Random Forest*, que será explicado pelo *LIME*; e πx é a métrica que mede a proximidade entre as amostras artificias geradas próximo da amostra original.

3.2. Métricas de Avaliação

Sobre a avaliação de desempenho dos algoritmos classificadores em aprendizado de máquina, o uso de alguns indicadores possibilita medir o nível de efetividade em predições. A seguir, são apresentadas as métricas citadas em [Kelleher et al. 2015]:

Acurácia: É considerada uma das métricas mais utilizadas, visto que é o resultado, em porcentagem, da divisão do número de amostras acertadas pelo total de amostras presentes no conjunto de teste.

Precisão: Avalia a quantidade de amostras positivas que o modelo acertou, considerando todas as amostras positivas e falso-positivas. À vista disso, há uma perspectiva e quão eficiente encontra-se o modelo para classificações que realmente eram positivas.

Revocação: Está relacionada ao nível de confiança que se pode obter do modelo em que, para o total de amostras positivas existentes (verdadeiramente positivas e falso-negativas), quantas realmente foram corretamente classificadas.

F-1 score: É a média harmônica entre a precisão e a revocação. Assim, quanto maiores forem os valores das métricas precisão e revocação, maior será o valor da F-1 score.

4. Materiais e Métodos

Nesta seção, são mostrados os recursos metodológicos utilizados no desenvolvimento de uma aplicação *web* que almeja facilitar o entendimento de um processo de categorização automático de sentimentos em *tweets* sobre a COVID-19. Todos os códigos desenvolvidos podem ser livremente acessados em (Araújo 2022ª).

4.1. Base de Dados

Este trabalho usou a base de dados desbalanceada disponibilizada por Miglani, (2020) contendo um total de 41.157 mensagens na língua inglesa, relacionadas a COVID-19, que foram extraídas do *Twitter* até dezembro de 2020 no formato *CSV (Comma Separated Values)*. O domínio de classificação é do tipo multiclasses com cinco sentimentos: Positivo; Extremamente Positivo; Negativo; Extremamente Negativo; e Neutro.

Já que a base de dados possui uma quantidade significativa de mensagens, optou-se por não usar validação cruzada e nem particionamento dos dados. Segundo Shalev-Shwartz & Ben-David, (2014), as técnicas de validação cruzada são recomendadas principalmente quando há escassez de dados rotulados, fato que pode resultar em conjuntos de teste insuficientes para certificar a capacidade de generalização do modelo. Assim, a base de dados foi dividida em conjuntos de treino e teste, com 30.867 (75%) e 10.290 (25%) instâncias, respectivamente.

Das seis *features* originais (*username, screenname, local, tweetat*, tweetoriginal e sentimento), foram mantidas apenas as consideradas mais relevantes para o estudo: A tweetoriginal com o conteúdo da mensagem; e a sentimento, que é a classe alvo, com o rótulo do sentimento associado à mensagem.

4.2. Pré-processamento dos dados

Devido à similaridade entre as palavras relacionadas aos sentimentos positivo e extremamente positivo, buscou-se estabelecer fronteiras de explicação bem definidas entre as classes preditivas no intuito de facilitar o processo de explicabilidade. Dessa forma, todas as amostras positivas foram concentradas em uma "nova" classe chamada Positiva, ligada a expressões como felicidade, esperançaa e interesse. O mesmo procedimento foi replicado para os exemplos negativos, que foram juntados na classe Negativa, associada a expressões como tristeza, caos e fraqueza. Assim, tem-se três classes para uma predição: Positiva (18.046 amostras); Negativa (15.398 amostras); e Neutra (7.713 amostras).

Na tentativa de reduzir inconsistências, os seguintes procedimentos de *NLP* foram realizados no conteúdo das mensagens: Remoção de URLs, HTMLs, caracteres especiais, espaços e números; e conversão de todas as palavras para minúsculo. Nessa etapa, *scripts* da biblioteca *Scikit-learn*, versão 0.24.1, foram usados para realizar as ações.

Em seguida, aplicou-se o método de *stopwords*, que consiste na remoção de palavras irrelevantes, como artigos e conjunções, as quais são frequentes, mas não implicam na compreensão da semântica de uma sentença (Bonaccorso, 2017). Para isso, este trabalho fez uso da biblioteca *NLTK (Natural Language Toolkit*), versão 3.4.5, com uma lista de *stopwords* (Araújo, 2022c) contendo 179 palavras que, por convenção, são tratadas como irrelevantes na língua inglesa.

A etapa final de preparação dos dados tratou da tokenização, ou seja, processo que divide um texto em palavras transformado-as em um vetor de inteiros, com números que representam a frequência de cada palavra (Hapke et al., 2019).

4.3. Classificação de sentimentos

Para a etapa de classificação, foram utilizados três algoritmos: *SVM*, *Random Forest*; e *RNAR (Rede Neural Artificial Rasa)*. Esses métodos foram escolhidos por serem caracterizados na literatura como caixas pretas e não transparentes. Dessa maneira, o objetivo foi comparar o desempenho de cada um e escolher o mais preciso.

O *Random Forest* foi configurado com base no conjunto padrão de hiperparâmetros com 100 árvores. Com relação ao *SVM*, também optou-se pela utilização das especificações padrões, com a definição de uma função de *kernel* do tipo linear. Sobre a *RNAR*, empregou-se uma arquitetura baseada em Lad, (2020), sendo uma rede do tipo *feed-foward*, configurada com cinco camadas, totalizando 127 neurônios e 381.265 parâmetros treinados com 50 épocas.

Ao comparar os resultados obtidos pelos três algoritmos no conjunto de teste, optou-se pelo *SVM*, em função desse algoritmo ter se mostrado superior em todas as métricas convencionais: acurácia (79%), precisão (77%), *F-1 score* (78%) e revocação (78%). Assim, o modelo *SVM* foi utilizado na aplicação *web* e nos experimentos descritos a seguir.

4.4. Aplicação web

O desenvolvimento de uma aplicação *web* integrada com o *framework LIME* se justifica ao agilizar o processo de predição e explicação de novas amostras. A ferramenta proposta oferece acessibilidade online para estudos de explicabilidade e exploração do comportamento de algoritmos de caixa preta, tanto por especialistas em aprendizado de máquina, quanto por usuários de aplicações em produção. É destacável também que, para eventuais atualizaçõees de modelo ou aplicação, há um ganho no reuso da arquitetura criada.

Dito isto, para a construção da aplicação *web*, denominada *Twitter COVID Explainer*, utilizou-se uma arquitetura de *software* cliente-servidor por meio da *API do Flask*, que dispõe de bibliotecas que permitem criar soluções integradas com *Python*. A arquitetura da ferramenta é ilustrada na Figura 1.

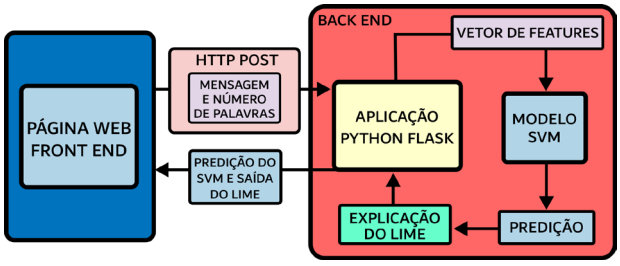

Figura 1 – Arquitetura da aplicação *web*.

Inicialmente, o usuário fornece como entrada a mensagem que deseja categorizar e a quantidade de palavras (20 termos no máximo) a serem ranqueadas de acordo com o seu nível de importância para o processo preditivo. Em seguida, a mensagem é pré-

processada, transformada em um vetor de inteiros e classificada pelo modelo *SVM*. A última etapa consiste na explicação da predição por meio do *framework LIME*, mais especificamente, usou-se o módulo *LimeTextExplainer*, versão 0.2.0.1. A página inicial da aplicação, que encontra-se publicamente disponível em (Araújo 2022b), pode ser visualizada na Figura 2.

O resultado da predição de um *tweet* classificado corretamente pelo modelo *SVM* como um sentimento positivo pode ser visto na Figura 3, evidenciando as probabilidades.

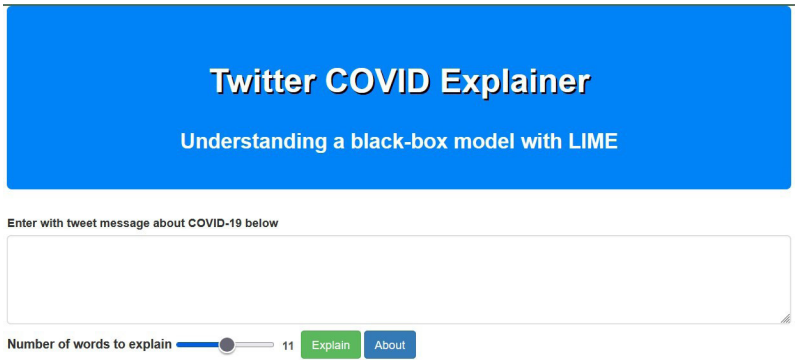

Figura 2 – Tela inicial da aplicação *web*. O usuário insere um *tweet* e define a quantidade de palavras a serem explicadas. Fonte: Elaborada pelo autor.

de cada classe e o graáfico de explicação do *LIME* configurado para mostrar as 10 palavras mais relevantes para a predição do *tweet*. Nesse exemplo, apenas três palavras (destacadas na cor verde) influenciaram positivamente a predição. Já as outras expressões foram consideradas influentes para as demais classes. É possível observar que a palavra "*free*" foi considerada a mais importante pelo modelo, adicionando uma probabilidade de 49% para o sentimento positivo.

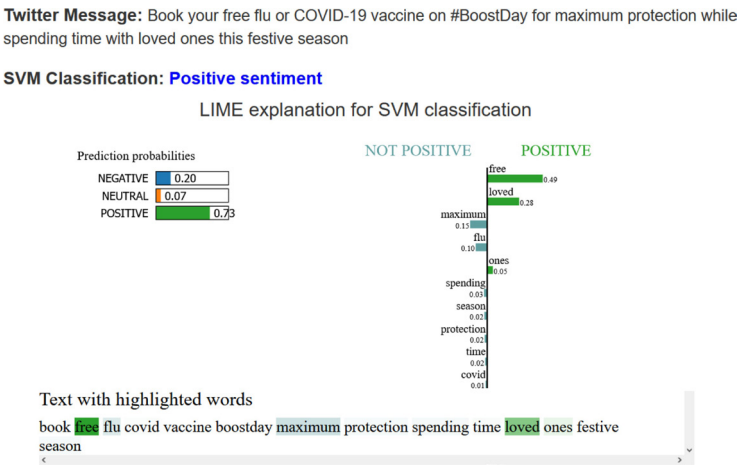

Figura 3 – Resultado da classificação correta de um tweet com sentimento positivo.

5. Resultados

Esta seção descreve os testes realizados com a aplicação *Twitter COVID Explainer*. A ideia central dos experimentos foi de avaliar a explicabilidade de forma qualitativa, observando o efeito de certas palavras nas predições.

5.1. Predições falso-positivas

Inicialmente, 30 mensagens do *Twitter* na língua inglesa relacionadas ao tema COVID-19 foram aleatoriamente selecionadas entre os anos de 2021 e 2022. Em seguida, esses *tweets* foram manualmente rotulados: 15 negativas, 11 positivas e 4 neutras. Lembrando que a base de dados usada no treinamento dos modelos foi coletada até o ano de 2020.

A explicação de uma predição falso-positiva para uma mensagem que na realidade corresponde a um sentimento negativo, devido ao aumento de pessoas que testaram positivo para COVID-19, pode ser vista na Figura 4. Observa-se, claramente, que a palavra "positive" influenciou fortemente o modelo a classificar a mensagem como sentimento positivo.

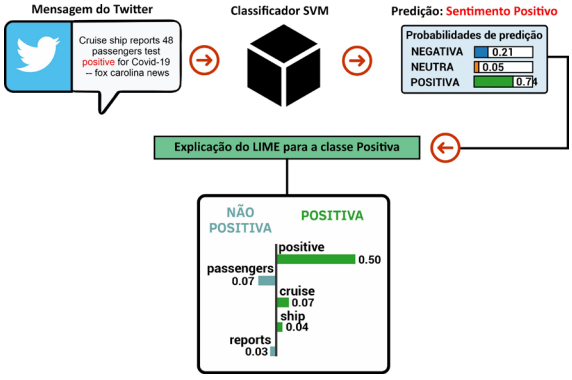

Figura 4 – Resultado da aplicação *web* configurada para explicar cinco palavras de um tweet classificado equivocadamente como sentimento positivo.

No intuito de validar o entendimento, a palavra "*positive*" foi removida da mensagem, a predição refeita e, como esperado, a classificação foi corretamente corrigida para negativa. É evidente que a remoção da palavra citada reduz a semântica da mensagem, no entanto, o objetivo do experimento foi atestar que o modelo diminui a certeza para uma mensagem falso-positiva e classifica corretamente o *tweet*. A base de dados contempla poucos exemplos (menos de 1% das mensagens) envolvendo testes de COVID-19. Tendo acesso a esse tipo de informação, o especialista pode otimizar a base de dados na tentativa de mitigar essa falta de representatividade.

Outro exemplo de limitação encontrado refere-se a palavra "vacina". Como a co- leta dos dados ocorreu em 2020, período que antecede a popularizacão de expressões como vacinas e variantes, foram encontradas apenas 139 amostras com esses termos, o que

corresponde a menos de 1% do total. Por exemplo, na Figura 5, em que a mensagem do *Twitter* relata que o reforço de vacinas é importante (algo que é positivo), o modelo classificou o *tweet* como sentimento negativo. Nota-se que a palavra "vacina" foi considerada pouco influente na predição e a palavra "*Omicron*", que corresponde a uma variante de COVID-19, sequer foi considerada.

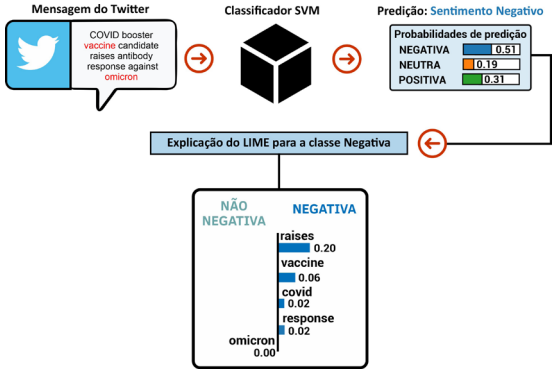

Figura 5 – Resultado da aplicação *web* configurada para explicar cinco palavras de um tweet classificado equivocadamente a respeito do termo vacina.

5.2. Stopwords

Esta etapa dos experimentos estudou o efeito que a retirada das *stopwords*, ou seja, palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido, tem sobre o processo de classificação. Para isso, se escolheu as *stopwords* "*no*" e "*not*" com a hipótese que elas podem ajudar a influenciar negativamente uma sentença. A palavra "*no*" ocorre em 2.300 amostras (5,5%) e a palavra "*not*" em 5.489 amostras (13,3%). Um exemplo de amostra formatada e corretamente classificada pode ser visto na Figura 6.

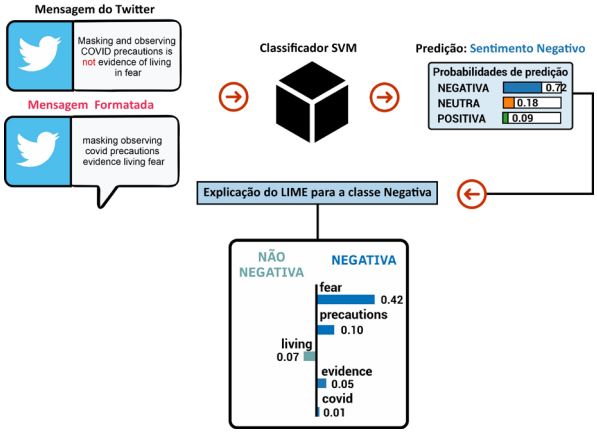

Figura 6 – Resultado da aplicação *web* para a classificação de uma mensagem com sentimento negativo explicada pelo *LIME*.

As palavras "*not*" e "*no*" foram, então, mantidas em todas as amostras e os modelos *SVM, RNAR* e *Random Forest (RF)* retreinados. Os resultados das duas avaliações, com e sem as stopwords analisadas, podem ser comparados na Tabela 1. Nota-se um ganho em todas as métricas, por exemplo, a acurácia do classificador *SVM* que passou de 79% para 81%. Contudo, o teste estatístico não paramétrico de *Kolmogorov-Smirnov* mostrou que os modelos com e sem *stopwords* não possuem diferença significativa (p-value = 0,92). Outras métricas de desempenho, como precisão e revocação, que tratam da assertividade com relação às amostras falso-positivas e falso-negativas, também melhoraram, mas não significativamente.

Treino	Modelo	Stopwords "no" e "not"	Acurácia	*Precisão*	*F1-Score*	*Revocação*
1	SVM	Removidas	79%	77%	78%	78%
2	**SVM**	**Incluídas**	**81%**	**79%**	**79%**	**80%**
1	RF	Removidas	74%	73%	74%	73%
2	**RF**	**Incluídas**	**77%**	**75%**	**76%**	**76%**
1	RNAR	Removidas	72%	71%	69%	70%

Tabela 1 – Resultados obtidos com os algoritmos considerando a remoção e a inclusão das palavras "no" e "not".

A mesma sentença usada na Figura 6 foi novamente classificada, mas, dessa vez, usando o modelo SVM retreinado e preservando a palavra "not". A explicação do *framework LIME* pode ser vista na Figura 7. Nota-se que o classificador passou a considerar a palavra "*not*" no processo preditivo, sendo a segunda mais influente, e ratificou o sentimento negativo da sentença, no entanto, com mais convicção.

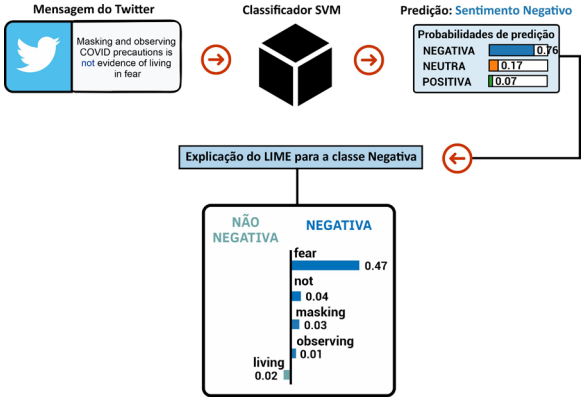

Figura 7 – Resultado da aplicação *web* para a classificação de uma mensagem com sentimento negativo com a palavra "*not* " tendo influência na predição.

Um exemplo de classificação onde a palavra "*no*" influenciou fortemente a predição para a classe negativa pode ser vista na Figura 8. Antes do retreino, essa mesma sentença

(sem a *stopword* em questão) havia sido classificada equivocadamente como neutra. Evidentemente, as palavras "*not*" e "*no*" são irrelevantes em determinados contextos da língua inglesa e uma explicação local não implica em um comportamento semelhante para todas as amostras.

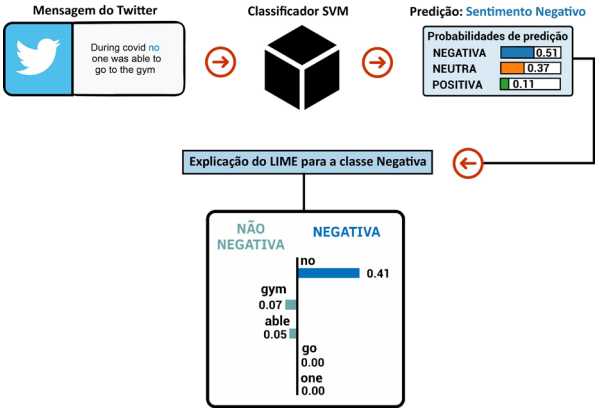

Figura 8 – Resultado da aplicação web para a classificação de uma mensagem com sentimento negativo com a palavra "*no*" tendo influência na predição.

O uso de um classificador inteligente acompanhado de um *framework* de explicabilidade, como o *LIME*, trouxe mais transparência e novas possibilidades de exploração para as predições realizadas. Em estudos na área de *NLP*, a validação das *stopwords* é uma etapa importante dentro do processo preditivo. Por fim, constatou-se que, em mensagens curtas, o *LIME* se mostrou pouco útil, talvez por não conseguir encontrar uma representação linear capaz de explicar adequadamente as relações entre as palavras.

6. Conclusões e trabalhos futuros

Este estudo permitiu constatar a importância de se compreender melhor o trabalho de algoritmos mais robustos como os do tipo caixa preta em soluções de aprendizado de máquina, onde há o dilema: transparência versus alto desempenho.

O uso da explicabilidade mostrou ausência de representatividade (desatualização) na base de dados. Em certas predições falso-positivas, notou-se que palavras julgadas relevantes e atuais não estavam sendo realcioadas pelo *LIME* em função do surgimento de novos termos com o avanço da pandemia da COVID-19. A análise da explicabilidade também possibilitou identificar palavras mais ou menos influentes, fato que contribuiu para ajustes no modelo com o objetivo de reduzir a ocorrência de predições falso-positivas.

Sobre a hipótese que o uso de *stopwords* pode aumentar a acurácia do modelo, observou-se que a predição das amostras com sentimento negativo melhorou com a adição das palavras "*no*" e "*not*"nas mensagens. A acurácia do classificador *SVM*, por exemplo,

passou de 79% para 81%. Embora essa diferença não seja estatisticamente significativa, em temas sensíveis qualquer melhora preditiva pode ser importante.

A integração de um *framework* de explicabilidade, como o *LIME*, a uma aplicação de aprendizado de máquina para tarefas de *NLP* agregou transparência e possibilitou alcançar resultados mais esclarecedores e qualitativos sobre o modelo, complementando as métricas de avaliação quantitativas convencionais. Dessa maneira, usuários comuns e profissionais da área têm a sua disposição um panorama da representatividade da base de dados e de como palavras-chave podem contribuir para uma predição, seja ela assertiva ou equivocada.

Com relação aos trabalhos futuros, sugere-se a inclusão de dados mais atuais sobre a COVID-19, de modo a melhorar a generalização do modelo *SVM* para novos vocabulários. Ainda na linha de pesquisa deste trabalho, comparar o processo preditivo de outros algoritmos de caixa preta, como *Random Forest* e *RNAR*, usando diferentes *frameworks* de explicabilidade, como o *SHAP*. Também seria interessante a realização de mais estudos sobre stopwords e explicabilidade em outros idiomas, por exemplo, a língua portuguesa.

Referências

Aguayo, R., Lizarraga, C., López-Bojórquez, M., Quiñonez, Y., & Cabrera, A. (2022). Implementación de plan de contingencia ante la pandemia covid-19 llamado rompiendo paradigmas docentes. RISTI - *Revista Ibérica de Sistemas e Tecnologias de Informação*, (45),48–63. https://doi.org/10.17013/risti.45.48-63

Ahmad, I. (2020). 40 Algorithms every programmer should know: Hone your problem-solving skills by learning different algorithms and their implementation in python.

Araújo,I. (2022a). Códigos de nlp. https://github.com/ivoaabreu/dissertacao-mestrado-ufpa-codigos-nlp

Araújo, I. (2022b). Endereço da aplicação web. https://twitter-explainer.onrender.com

Araújo,I. (2022c). Lista de stopwords. https://github.com/ivoaabreu/dissertacao-mestrado-ufpa-codigos-nlp- lime/blob/main/lista-stop-words.pdf

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4):102569. https://doi.org/10.1016/j.ipm.2021.10256

Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. (2021). Twitter for disaster relief through sentiment analysis for covid-19 and natural hazard crises. *International Journal of Disaster Risk Reduction*, 55,102101. https://doi.org/10.1016/j.ijdrr.2021.102101

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Garcia, K. & Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the USA. *Applied Soft Computing*, 101:107057. https://doi.org/10.1016/j.asoc.2020.107057

Glowacki, E. M., Wilcox, G. B., & Glowacki, J. B. (2021). Identifying# addiction concerns on twitter during the covid-19 pandemic: A text mining analysis. *Substance abuse*, 42(1),39–46. https://doi.org/10.1080/08897077.2020.1822489

Hall, P., Gill, N., & Cox, B. (2021). Responsible machine learning.

Hapke, H., Howard, C., & Lane, H. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster.

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine le- arning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, A. (2020). Social media analysis with ai: sentiment analysis techniques for the analysis of twitter covid-19 data. *J. Crit. Rev*, 7(9),2761–2774.

Kolluri, N. L. & Murthy, D. (2021). Coverifi: A covid-19 news verification system. *Online Social Networks and Media*, 22,100123. https://doi.org/10.1016/j.osnem.2021.10012

Kouvela, M., Dimitriadis, I., & Vakali, A. (2020). Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, pp. 55–63. https://doi.org/10.1145/3415958.3433075

Lad, R. (2020). *Parkinson's disease classification*. Disponível em https://www.kaggle.com/richalad/parkinsons-predictions

Liu, Z., Guo, Y., & Mahmud, J. (2021). When and why does a model fail? a human-in-the-loop error detection framework for sentiment analysis. https://doi.org/10.48550/arXiv.2106.00954

López, M. P. V., de Freitas, P. O., & Vargas, S. M. L. (2023). A relação entre a inovação tecnológica e o desempenho nos meios de hospedagem no contexto da pandemia da covid-19. RISTI - *Revista Ibérica de Sistemas e Tecnologias de Informação*, (52),45–60. https://doi.org/10.17013/risti.52.45-60

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774.

Meena, R. & Bai, V. T. (2019). Study on machine learning based social media and sentiment analysis for medical data applications. In *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 603–607. IEEE. https://doi.org/10.1109/I-SMAC47947.2019.9032580

Miglani, A. (2020). *Coronavirus tweets nlp - text classification*. https://www.kaggle.com/sagarkhambad/text-classification/data

Molnar, C. (2019). *Interpretable machine learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Nielsen, A. (2020). *Practical Fairness*. O'Reilly Media.

Recuero, R. (2009). Redes sociais na internet, difusão de informacção e jornalismo: elementos para discussão. *Metamorfoses jornalísticas*, 2,1–269.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Rothman, D. (2020). *Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps*. Packt Publishing Ltd.

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Silva, H., Andrade, E., Araújo D., & Dantas, J. (2021). Sentiment analysis of tweets related to sus before and during covid-19 pandemic. *IEEE Latin America Transactions*, 20(1),6–13. https://doi.org/10.1109/TLA.2022.9662168

Turek, M. (2021). Explainable artificial intelligence.

Redondo, A. M. F., & Cárdenas, F. de J. N. (2022). DevOps: Un vistazo rápido. Ciencia Huasteca Boletín Científico de la Escuela Superior de Huejutla, 10(19), 35-40.