ISSN on-line: 2238-4170
http://estacio.periodicoscientificos.com.br/index.php/gestaocontemporanea
Gestão Contemporânea, v.14, n.1, p. 22-44, jun. 2024.

ARTIGO ORIGINAL

PROCESSAMENTO DE LINGUAGEM NATURAL APLICADO A POSTAGENS EM PORTUGUÊS NO TWITTER

ORIGINAL ARTICLE

NATURAL LANGUAGE PROCESSING APPLIED TO POSTS IN PORTUGUESE ON TWITTER

Henrique de Menezes Alves Junior[1]
Murilo Henrique Tank Fortunato[2]
Centro de Conhecimento e Pesquisa - PECEGE, Brasil

Resumo

Todos os dias, devido à grande utilização de redes sociais por parte da sociedade, milhões de terabytes de dados não estruturados são gerados. Com isso, empresas e pesquisadores na área de dados encontram neste volume de informação uma possibilidade de realizar estudos, para compreender padrões no comportamento dos usuários, entre outras diversas análises possíveis. Com isso, neste trabalho tivemos como objetivo construir um sistema capaz de realizar análises assertivas e de fácil compreensão, de postagens na rede social Twitter, sobre qualquer assunto desejado. Para cumprir com essa meta, foi desenvolvido um pipeline de tarefas, onde realizamos a captura dessas postagens, em seguida, um modelo de Processamento de Linguagem Natural (NLP) foi construído para classificar os tweets, alcançando 80% de acurácia e F1-Score médio de 79,66%. Após essas etapas, foram realizadas análises com as informações disponíveis, como a criação de Nuvens de Palavras e um estudo de Correlação entre os termos mais utilizados nos tweets e a classificação realizada pelo modelo, com o intuito de auxiliar os usuários finais do sistema na análise comportamental dos usuários na rede social. Por fim, com este projeto buscamos contribuir com trabalhos desenvolvidos no Brasil na área de dados, visto que nosso sistema foi desenvolvido com a proposta de analisar textos em português, utilizando técnicas modernas de NLP, como o modelo BERT.

Palavras-chave: NLP; BERT; Redes Sociais.

Abstract

Every day, due to the extensive use of social media by society, millions of terabytes of unstructured data are generated. As a result, companies and researchers in the data field find in this volume of information a possibility to conduct studies to understand patterns in user behavior, among other various possible analyses. With this in mind, our objective in this work was to build a system capable of performing assertive and easily understandable analyses of posts on the social media platform Twitter, on any desired subject. To achieve this goal, a task pipeline was developed, where we captured these posts, and then a Natural Language Processing (NLP) model was built to classify the tweets, achieving 80% accuracy and an average F1-Score of 79.66%. After these steps, analyses were performed with the available information, such as creating Word Clouds and a study of the correlation between the most used terms in the tweets and the classification performed by the model, with the aim of assisting end users of the system in the behavioral analysis of users on the social network. Finally, with this project, we aim to contribute to works developed in Brazil in the data field, since our system was developed with the purpose of analyzing texts in Portuguese, using modern NLP techniques such as the BERT model.

[1] Especialista em Data Science e Analytics pela USP/ESALQ. E-mail: henrique.menezesjr@gmail.com.
[2] Mestre em Ciências Ambientais pela UNIFAL; Doutor em Agricultura Sustentável pela UNIFENAS. E-mail: mtank@live.com.

Submetido em 23/01/2024
Aceito em 06/05/2024

Keywords: NLP; BERT; Social Network.

INTRODUÇÃO

É inegável como o mundo está sendo transformado pelos dados. Prova disso, é que segundo a International Business Machines Corporation (2020), até 2024 o volume de dados globais que são processados, transformados e armazenados em datacenters e serviços de nuvens, chegará a 143 zetabytes (143 sextilhões de bytes). Outra informação relevante, um relatório divulgado pela página Data Reportal (2023) mostra que 4,76 bilhões de pessoas no mundo utilizam redes sociais todos os dias, passando em média 2 horas e 31 minutos conectados.

Atualmente, de acordo com o site Internet Live Stats, que calcula os números atuais de usuários da internet, uma das principais redes sociais é o Twitter, contendo mais de 1,3 bilhão de contas e 500 milhões de tweets (postagens por usuário) por dia. Em outras palavras, o Twitter é uma grande fonte de dados, que podem ser explorados para compreender o comportamento dos usuários sobre qualquer assunto (Mazumdar, 2020). Um exemplo disso ocorre no dia de lançamento de um filme muito esperado, milhões de postagens são feitas pelos usuários desejando expressar suas opiniões. Com isso, é possível que seja monitorado pela produção do filme, se ele está sendo bem aceito pelo público, ou não. Essas informações mostram um fato: um volume grande de dados é trafegado na internet a cada minuto, gerando interesse em diversas áreas da sociedade, buscando entender o comportamento dos usuários na rede. Nesse sentido, é importante destacar que essas postagens nas redes sociais são monitoradas para compreender a reação do público ao lançamento do filme, analisando opiniões, comentários e sentimentos expressos pelos usuários sobre o filme.

Aproveitando este crescimento da quantidade de dados à disposição, áreas como a ciência de dados buscam tirar valor dessas informações, trazendo análises relevantes sobre as informações disponíveis. Um dos problemas enfrentados por quem trabalha neste ramo é a quantidade de dados de forma não estruturada, isto é, dados que não estão no padrão de linha por coluna, como por exemplo, as tabelas de

Bancos de Dados (Eberendu, 2016). Segundo o Blog da IBM (2021), 80% dos dados à disposição de empresas são não estruturados. Alguns exemplos são imagens, vídeos, publicações em blogs e redes sociais.

Para auxiliar o processo de análise dos dados em formato de texto, existem as técnicas de Natural Language Processing (“Processamento de linguagem natural”, traduzindo do inglês), ou NLP, uma vertente dentro da Inteligência Artificial dedicada a desenvolver modelos capazes de compreender, gerar e manipular textos, o que acaba se tornando uma grande ferramenta para se trabalhar com essa grande quantidade de dados não estruturados disponíveis (Vajjala et al, 2020). Uma das técnicas NLP existentes, considerada estado da arte nesta área, se chama BERT (acrônimo para “Bidirectional Encoder Representations from Transformers”). Desenvolvido pela Google e utilizado em sua ferramenta de busca, o BERT é um modelo poderoso para compreensão de contextos, apresentando inovações na forma de realizar o processamento dos textos apresentados a ele.

Contudo, mesmo que fosse apresentado um resultado do modelo satisfatório, é necessário que exista uma aplicação prática para a sua utilização, além de uma interpretação e análises de seu resultado de forma a auxiliar o propósito pelo qual foi construído. Com isso, este projeto tem como objetivo apresentar para empresas e organizações, que é possível o desenvolvimento de uma ferramenta capaz de auxiliar na interpretação do comportamento dos usuários na rede social Twitter, baseando-se nas suas postagens sobre assuntos desejados. Importante destacar que também é um objetivo utilizar apenas postagens no idioma português, com o intuito de contribuir com trabalhos realizados no Brasil com a língua nacional.

Para cumprir com este objetivo, foi proposto um algoritmo capaz de realizar a captura de uma amostragem de tweets da plataforma, levando em consideração o assunto que se deseja estudar e o período que foram feitas as publicações. Após isso, uma análise foi feita para entender se o assunto escolhido está sendo bem avaliado ou não pelos usuários. Além do resultado do modelo, também serão apresentadas algumas formas de visualização do resultado da análise, que auxiliem o usuário final a interpretar e tomar as melhores decisões possíveis a respeito do que se deseja estudar.

MATERIAL E MÉTODOS

Para dar início ao desenvolvimento do projeto, primeiramente o dividimos em etapas. Temos primeiro a etapa de captura dos dados que serão utilizados durante os processos do projeto. Após isso, temos o pré-processamento destes dados, para que eles estejam no padrão exigido pelo modelo. Temos também a fase de treinamento do modelo, e complementando esta etapa temos a sua validação, para que fosse possível analisar os seus parâmetros e realizar alterações caso necessário. Por fim, temos o teste do modelo com um conjunto de dados específicos para esta fase, que não foi utilizado em nenhuma etapa antes, além da criação de formas visuais para facilitar a análise final do usuário.

Todo código desenvolvido para este projeto foi realizado na linguagem de programação Python, com o auxílio de bibliotecas específicas (serão mencionadas em seus respectivos tópicos) para a realização de cada função do sistema proposto. Abaixo será explicada cada etapa separadamente.

CAPTURA DOS DADOS

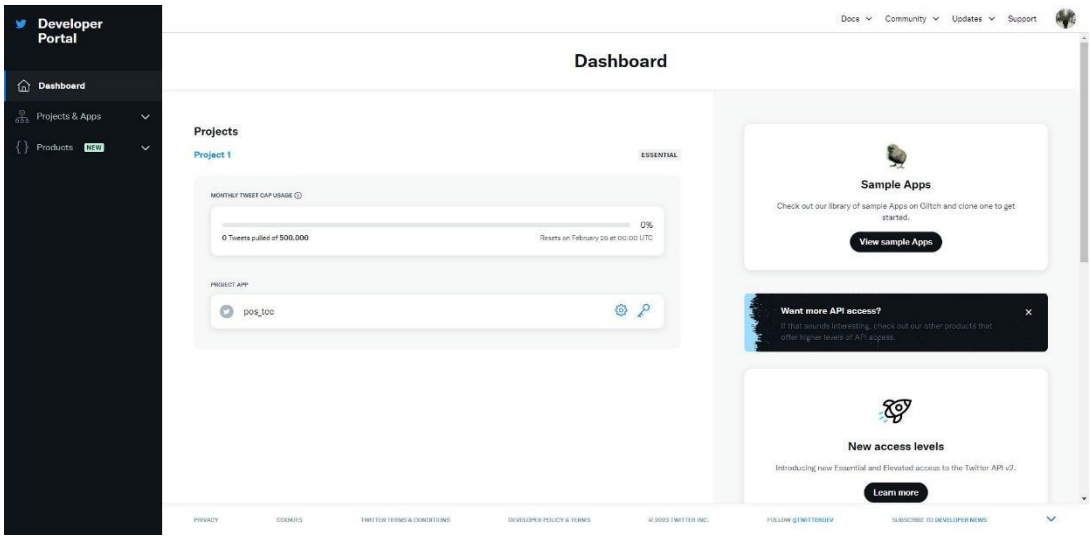
Neste projeto, uma dificuldade foi encontrar uma base de dados robusta e confiável com tweets em português, e que ainda estivesse classificada para realizar a análise de sentimentos. Devido à exigência de dados bem específicos, foi necessária a construção de um dataset com fontes diferentes, para que fosse possível a realização das etapas de treinamento, validação e teste do modelo. Com isso, para compor a base de dados deste projeto, teremos textos extraídos diretamente do Twitter, e também postagens de avaliação de aplicativos para celular Android, extraídos diretamente da Google Play Store. Optamos por essa outra fonte de dados, pois identificamos que a linguagem utilizada nas duas fontes é similar, além do fato da finalidade dos textos serem de avaliação, facilitando o trabalho de classificação dos respectivos textos. Abaixo, é demonstrado como foi feita a extração dos dados para cada fonte.

TWITTER

Para extração de informações retiradas do Twitter, utilizamos uma API disponibilizada pela própria empresa, chamada Tweepy. Para utilizar esta ferramenta, faz-se necessário criar uma conta na rede social, e se cadastrar também na plataforma para desenvolvedores. Nesta plataforma, os desenvolvedores têm acesso de forma gratuita a captura de 500.000 tweets mensalmente, porém, com algumas pequenas restrições no momento da extração, como por exemplo, só podem ser enviadas requisições para a API que capturem 100 tweets de uma só vez.

Na Figura 1 é possível visualizar a página inicial da área de desenvolvedor.

Figura 1: Página inicial da área de desenvolvedor do Twitter



Fonte: Twitter Develop Platform. Disponível em: <https://developer.twitter.com/>. Acesso em: 07 fev. 2023.

Na plataforma de desenvolvedores, faz-se necessária a criação de tokens de acesso, para que possa ser feita a autenticação do seu usuário no momento da requisição via código. Criadas as credenciais, é possível enviar requisições para a API via script Python, sendo possível também, escrever consultas especificando o assunto desejado, além de outros parâmetros.

Utilizando esta ferramenta, um script de captura de tweets foi desenvolvido, usando como parâmetros o idioma ser sempre o português, foi excluído os retweets, além de filtrar as datas desejadas para cada assunto. Com isso, capturamos cerca de

20.000 tweets envolvendo assuntos diferentes que foram bastante comentados durante o desenvolvimento deste projeto.

Na Tabela 1, é apresentado alguns exemplos de assuntos buscados para compor a base de dados, além do período que foi realizada a consulta.

Tabela 1: Exemplos de assuntos que foram buscados no Twitter para captura de postagens durante os períodos indicados

Assunto	Período da Consulta (Data Início – Data Fim)
#ApuracaoRJ	22/02/2023 - 22/02/2023
#ChampionsLeague	15/02/2023 - 15/02/2023
#BBB23	24/01/2023 - 25/01/2023

Fonte: Dados originais da pesquisa

AVALIAÇÃO DE APLICATIVOS

Com o auxílio da biblioteca Python “google_play_scraper”, é possível realizar extração dos comentários e avaliações que os usuários da Google Play Store indicam para os aplicativos disponíveis. É possível realizar filtros na extração como o idioma dos comentários, e também o país de origem do comentário.

Escolheu-se 5 aplicativos dentre os mais populares para a área de alimentação da plataforma, por se tratarem de aplicativos bastante utilizados no dia a dia. Eles podem ser avaliados com uma nota entre 1 (sendo a pior nota) até 5 (a melhor nota) na plataforma. Foi desenvolvido um algoritmo que realiza a captura dos comentários mais relevantes e recentes dos usuários. Com isso, foram coletadas 12.000 amostras de comentários no dia 02/02/2023, sendo 4.000 amostras de notas entre 1 e 2 (consideradas negativas), 4.000 amostras para a nota 3 (considerada neutras) e 4.000 amostras de notas entre 4 e 5 (consideradas positivas). Essa amostragem foi definida, pois o volume total de comentários presentes na Google Play Store para esses aplicativos se encontra em torno de 15 milhões, e como o objetivo do projeto é uma prova de conceito, foi identificado que essa amostragem iria compor de forma satisfatória a nossa base de dados.

PRÉ-PROCESSAMENTO

Para o pré-processamento também foi necessária uma divisão para as diferentes fontes, pois os dados provenientes da Google Play Store já estão classificados. Abaixo, vamos descrever o processo de pré-processamento para cada um dos conjuntos de dados.

Twitter

O primeiro passo para preparar os textos oriundos do Twitter é retirar caracteres e partes do texto que podem prejudicar o treinamento do modelo, seja por ser um caractere especial, ou um texto que não acrescenta informação relevante. Com isso, cada tweet da nossa base é processado por um script de tratamento, onde são retiradas menções a outros usuários (caracterizadas pelo caractere “@”), links para outras páginas na internet e números. Também são retiradas palavras consideradas “stopwords”, que são palavras que podem ser removidas do texto sem perda considerável na sua compreensão, por exemplo: “as”, “e”, “os”, “de”, “para”. Para realizar esta tarefa, utilizamos a biblioteca Python “nltk”, que possui uma lista de stopwords em português disponível para utilização, além da possibilidade de incluir novas palavras que achamos necessárias, como por exemplo, xingamentos e abreviações. Por fim, deixamos todas as letras em minúsculo.

A próxima etapa é a categorização destes tweets, para que seja possível o treinamento e a validação do modelo posteriormente. Esta etapa se faz necessária, pois, como informado anteriormente, não encontramos uma base de dados já classificada de postagens no Twitter em português, que atendesse à demanda deste projeto. Portanto, para que fosse possível a classificação da quantidade de textos capturados em tempo hábil, foi necessária a automatização do processo.

Nesta etapa, um script foi desenvolvido dentro da plataforma Google Cloud, onde é possível utilizar a API de Linguagem Natural disponibilizada pela empresa. Foi criada uma conta na plataforma de forma gratuita, a qual a empresa disponibiliza por 90 dias um crédito de 300 dólares para que usuários novos possam conhecer suas ferramentas. Utilizando a ferramenta Vertex AI, construímos um ambiente capaz de executar scripts Python voltados para Machine Learning. Com o auxílio da biblioteca “language_v1”, é possível definir uma pontuação (podendo ser chamado também de

"score") ente -1 e 1 para cada texto apresentado para a API, sendo -1 o grau mais "negativo" para a sentença, e 1 o grau mais positivo. Foi desenvolvido um algoritmo que utilizou o modelo do Google para definir uma pontuação para cada tweet presente na base. Após isso, classificamos os textos em: "negativo" para score abaixo de -0,3; "positivo" para score acima de 0,3; e "neutro" para score entre -0,3 e 0,3. Na Tabela 2, podemos ver exemplos de sentenças em cada categoria.

Tabela 2: Exemplos de *tweets* categorizados pelo script utilizando a biblioteca language_v1

Tweet	*Score*	Classificação
"não entra na minha cabeça a viola davis não ter sido indicada ao oscar por a mulher rei"	-0.600	Negativo
"será que vem aí menor audiência do Oscar"	0.0	Neutro
"sério, gostei desse filme muito, ele é tão cativante. a história é legal, as personagens são legais, a animação é boa. em geral, adorei, se ganhar o oscar agora, vai ser merecido."	0.899	Positivo

Fonte: Dados originais da pesquisa

AVALIAÇÃO DE APLICATIVOS

Conforme demonstrado na seção anterior, os textos de avaliação de aplicativos já foram extraídos com uma nota de 1 até 5 vinculada em cada um. O pré-processamento para estes dados é a criação de uma nova classificação para os textos, baseado nessas notas: Notas 1 e 2 são considerados textos negativos; notas 3 são considerados textos neutros; e notas 4 e 5 são textos considerados positivos. Além disso, foi realizado o processo de tratamento destes textos para padronização da base, como retirada de "stopwords" e deixando as palavras em minúsculo.

Após todas as etapas de pré-processamentos para ambas as fontes serem concluídas, os conjuntos de dados são agrupados e salvos em um único arquivo, para facilitar no manuseio durante o processo de treinamento do modelo, que vem a seguir.

TREINAMENTO E VALIDAÇÃO DO MODELO

De posse da base de dados classificada, é possível realizar o treinamento do modelo. Neste projeto, desejamos utilizar uma técnica considerada Estado da Arte no que tange a projetos de NLP (Ravichandiran, 2021). O BERT é um modelo de aprendizado profundo desenvolvido pelo Google para a área de Processamento de Linguagem Natural. Ele é utilizado na ferramenta de busca do Google e seu artigo foi recentemente publicado, disponibilizando seu uso para a comunidade (Devlin et al., 2018).

De forma geral, as técnicas de NLP trabalham com “tokens”, que é a atribuição de números às palavras para que o computador possa interpretar o texto. Com isso, os modelos convencionais buscavam compreender o contexto destes tokens realizando varreduras no texto da esquerda para a direita, ou o contrário, como explicado no artigo do próprio BERT. A principal inovação do BERT é o treinamento do modelo realizado nas duas direções do texto, buscando melhorar o entendimento do contexto de cada palavra presente. Por exemplo, nas frases “Vou seguir este caminho para casa”, e “Eu caminho na praia todos os dias” a palavra “caminho” seria interpretada de uma maneira similar por modelos tradicionais como o word2vec e GloVe, mas com o BERT, ele assume significados diferentes para ambas as frases (Khurana, D et al, 2020).

Como foi dito anteriormente, as técnicas de NLP trabalham utilizando tokens, e para ser possível a realização de um trabalho de análise textual, é necessário que exista um dicionário de tokens para cada palavra desejada. Em uma pesquisa realizada em 2018 por Jorge Filho e outros colaboradores, foi desenvolvida uma base de dados contendo tokens para as palavras do idioma português (Filho et al, 2018). O repositório brWac utilizou cerca de 145 milhões de sentenças para a construção de 2,7 bilhões de tokens, e se encontra disponível de forma gratuita para todos os projetos que desejam utilizá-lo. Em um artigo diferente, realizado por Fábio Souza e colaboradores (2020), temos um modelo BERT sendo treinado utilizando a base de dados brWac, em que os autores realizam a comparação com o modelo Multilingual BERT, e conseguem melhores resultados para as tarefas como Sentence Similarity (determinar o quanto um texto é similar) e Textual entailment (capacidade do modelo de reconhecer a relação entre partes de um texto). O modelo pré-treinado deste

projeto, denominado BERTimbau, também se encontra disponível para a utilização da comunidade.

Uma técnica muito utilizada em projetos envolvendo trabalhos com redes neurais, é chamada de Transfer Learning (Transferência de Aprendizado, na tradução direta do idioma inglês). Com ela, é possível realizar a reutilização de um modelo já treinado utilizando uma base de dados genérica, em um problema novo e mais específico. O que justifica a popularidade do Transfer Learning é o fato de que treinar novos modelos do início pode ser uma tarefa complicada, pois pode envolver um custo alto envolvendo hardware capaz de processar o modelo, além do tempo necessário para o treinamento, que, em alguns casos, chegam a demorar dias para terminar (Weiss et al, 2016). Além desses aspectos, há também o desafio de construir uma base de dados robusta para o treinamento, validação e teste do modelo. O Transfer Learning oferece uma solução ao permitir que aproveitemos modelos disponíveis para a comunidade, podendo adicionar etapas para direcionar o modelo para o problema específico que estamos tentando resolver.

Neste trabalho, propomos a utilização da técnica de Transfer Learning no modelo denominado BERTimbau, adicionando novas camadas de processamento na saída da rede pré-treinada para classificação dos textos retirados do Twitter. A utilização desta técnica traz grandes benefícios para este projeto, pois temos disponível para utilização um modelo que já foi treinado com milhões de sentenças, ou seja, a parte mais custosa e demorada do processamento já ocorreu. Com isso, podemos focar na realização do ajuste fino do modelo pré-treinado, realizando um novo treinamento com a base de dados contendo os textos do Twitter, e ajustando os parâmetros deste treinamento em busca do melhor resultado.

Para realizar o treinamento de um modelo BERT, a primeira etapa é gerar os tokens das palavras em cada texto presente na base. Para isso, foi utilizado um objeto tokenizer da biblioteca Python “transformers”, com o modelo BERTimbau já treinado. Este processo atribui números únicos para cada token, o que permite ao modelo realizar os cálculos necessários para entendimento do relacionamento entre as palavras. Ainda em relação à base de dados, foi realizada a separação das amostras

para cada etapa do processo, seguindo a proporção na separação: 80% para treinamento, 10% para validação do modelo e 10% para o teste final.

Após esta etapa, precisamos definir o nosso classificador. Ele será composto pelo modelo já treinado do BERTimbau Básico (12 camadas escondidas, com tamanho 768 e 12 attention heads), conectado a sua saída teremos mais uma camada de Dropout, que elimina neurônios da nossa rede de forma aleatória, o que auxilia na regularização do modelo durante seu treinamento, e adicionaremos uma última camada linear, para nos retornar o valor calculado pela rede. A escolha pelo BERTimbau básico ocorreu devido à falta de hardware disponível para utilizar o modelo “Large” (24 camadas escondidas, com tamanho 1024 e 16 attention heads), sendo necessário mais de 64 GB de memória RAM disponível.

Outra etapa importante para o treinamento de um modelo NLP é a escolha do otimizador. O otimizador é uma função matemática que auxilia durante o treinamento de um modelo de Machine Learning na diminuição do erro da função e na maximização de sua eficiência (Choi; Dami et al., 2019).

Para este projeto, foi utilizado o otimizador Adam (Kingma e Ba, 2014), muito comum em projetos de Deep Learning. Abaixo, a Tabela 3 contém os parâmetros do nosso modelo que foram variados durante o processo de treinamento e validação em busca do melhor resultado, onde os valores utilizados no modelo final foram: 16 para Tamanho do Batch; 3 e-5 para Taxa de aprendizado; e 3 para número de Épocas.

Tabela 3: Parâmetros definidos para execução do treinamento do modelo NLP proposto

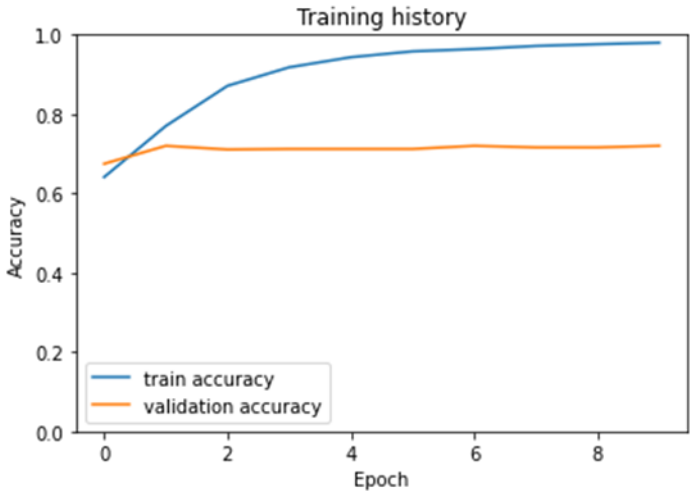
Parâmetro	Descrição	Valor
Tamanho do Batch	Número de exemplos utilizado por iteração.	16, 32
Taxa de Aprendizado	Taxa de ajuste para cada passo dado em direção ao mínimo da função de perda.	1e-4, 1e-5, 3e-5, 5e-5
Épocas	Quantidade de iterações para o treinamento do modelo.	3, 5, 10

Fonte: Dados originais da pesquisa

Com isso, o treinamento foi realizado contendo 7 mil amostras de cada classe, e podemos ver na Figura 2 abaixo o comportamento dele ao longo de 10 épocas. É possível identificar que logo entre as épocas 2 e 4 temos valores próximos do máximo

atingido de acurácia. No próprio artigo do BERT, os autores realizam o ajuste fino do modelo variando entre 2 e 3 épocas, para evitar o overfitting (quando o modelo se torna especialista apenas no conjunto de treinamento, e não se torna genérico o suficiente para funcionar em novas situações).

Figura 2: Demonstração das curvas do treinamento e validação do modelo ao longo das épocas



Fonte: Dados originais da pesquisa

MÉTRICAS PARA VALIDAÇÃO DE MODELOS

Para compreendermos melhor os resultados do modelo desenvolvido, faz-se necessária a definição de alguns conceitos de validação de modelos de Aprendizado de Máquina. Quando o modelo realiza classificações ele pode cometer erros, com isso, existem algumas definições de acordo como a classificação foi realizada. Abaixo está cada definição:

Verdadeiro Positivo (VP): Quando o modelo acerta a classificação da amostra.

Verdadeiro Negativo (VN): Quando o modelo acerta ao não classificar a amostra.

Falso Positivo (FP): Quando o modelo erra a classificação da amostra.

Falso Negativo (FN): Quando o modelo erra ao não classificar a amostra.

Com base nestas definições, algumas métricas foram criadas para validação de modelos de aprendizado de máquina. A métrica de Acurácia indica o desempenho geral do modelo, isto é, de todas as classificações do modelo, qual a sua taxa de acerto. Ela pode ser definida pela eq. (1).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

Onde, VP: Número de classificações consideradas verdadeiras positivas; VN: Número de classificações consideradas verdadeiras negativas; FP: Número de classificações consideradas falsas positivas; FN: Número de classificações consideradas falsas negativas.

Precisão é a variável que mede a relação entre a quantidade de classificações corretas feitas pelo modelo, pela quantidade total de classificações feitas (corretas e incorretas) para a classe desejada. Pode ser expressa pela eq. (2).

$$Precisão = \frac{VP}{VP + FP} \quad (2)$$

onde, VP: Número de classificações consideradas verdadeiras positivas; FP: Número de classificações consideradas falsas positivas.

Sensibilidade (em inglês, recall) mede o nível de sensibilidade do modelo a uma classe, comparando a quantidade de classificações corretas pelo sistema, pela quantidade real da classe desejada presente na base de dados. Esta métrica é definida pela eq. (3).

$$Sensibilidade = \frac{VP}{VP + FN} \quad (3)$$

onde, VP: Número de classificações consideradas verdadeiras positivas; FN: Número de classificações consideradas falsas negativas.

Já o F1-Score é uma métrica que calcula a média harmônica entre as duas variáveis anteriores, em que o valor mais alto possível representa que o modelo estaria atuando bem, pois estaria sendo preciso nas suas classificações, e ao mesmo tempo sensitivo para cada exemplo presente na base de dados. Podemos expressar o F1-Score pela eq. (4).

$$F1\ Score = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (4)$$

onde, Precisão: valor calculado da métrica de Precisão para o modelo; Sensibilidade: valor calculado para métrica Sensibilidade do modelo.

RESULTADOS E DISCUSSÃO

Para demonstrar os resultados deste projeto, vamos separar em dois tópicos nesta seção. No primeiro, iremos discutir os resultados do modelo de classificação de tweets. Após isso, iremos demonstrar algumas análises realizadas utilizando os dados extraídos e o resultado do modelo.

MODELO

A Tabela 4 demonstra os valores das variáveis definidas anteriormente neste texto, para o modelo sendo aplicado na base de dados de teste, com textos que não tiveram contato durante seu treinamento. No geral, o modelo teve acurácia total de 80,1%.

Tabela 4: Valores de Precisão, Sensitividade e F1-*Score* para cada classe na base de dados de teste.

Classes	Precisão (%)	Sensitividade (%)	F1-*Score* (%)
Negativo	78%	84%	81%
Neutro	79%	74%	76%
Positivo	82%	81%	82%

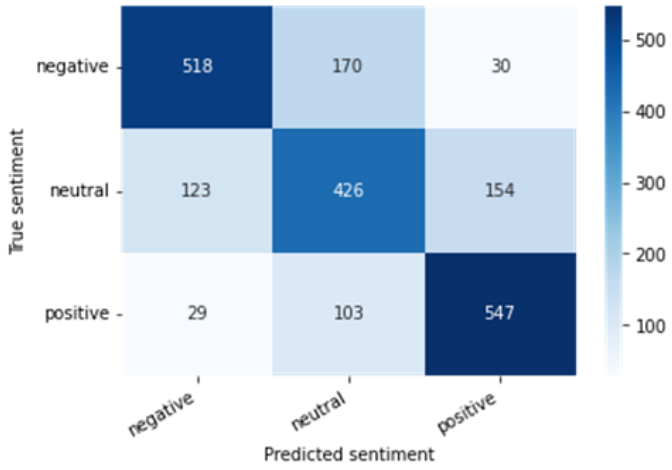
Fonte: Dados originais da pesquisa.

Podemos ver também que o modelo teve mais dificuldades em classificar corretamente exemplos da classe Neutro. Isto pode ocorrer pelo fato de que textos meramente informativos (muitas vezes com tom neutro) podem conter palavras que possuem significado próprio de algo positivo ou negativo, fazendo o modelo ser tendencioso a classificar toda a sentença como algo que não seja neutro. Inclusive, durante a realização da validação do modelo, identificamos a necessidade de excluir da base de treinamento os textos classificados como “Neutros” oriundos das classificações de aplicativos da Google Play Store. O motivo para esta ação se deve ao fato de grande parte dos comentários com Nota 3 terem uma tendência a serem positivos, com palavras como “bom”, “útil”, “satisfeito”, o que estava prejudicando o

treinamento do modelo. Após essa ação, o modelo ganhou cerca de 5% de acurácia geral.

Na Figura 3 pode-se ver uma matriz de confusão, que ilustra as classificações do modelo para cada classe. Nesta imagem, a diagonal principal desta matriz representa as classificações corretas feitas pelo modelo. É possível notar que as classificações que envolvem a classe Neutro realmente é a que possui mais erros de classificação, por exemplo, foram classificados 170 textos como Neutros, porém, são textos considerados Negativos.

Figura 3: Matriz de Confusão demonstrando as classificações que o modelo realizou para cada classe



Fonte: Dados originais da pesquisa.

Diante do que foi apresentado, consideramos que o modelo está com um bom desempenho, mantendo sua acurácia próxima aos de trabalhos publicados com o mesmo tema. Por exemplo, Anupama B. S. e outros pesquisadores (2020) desenvolveram um modelo utilizando técnicas de Nayve Bayes para classificar os tweets em “Positivos” e “Negativos”, alcançando 83% de acurácia. Outro trabalho interessante de se mencionar foi o realizado por Lihua Yao e demais pesquisadores (2020), onde utilizaram técnicas diferentes para classificarem os textos em “Positivos”, “Negativos” e “Neutros”.

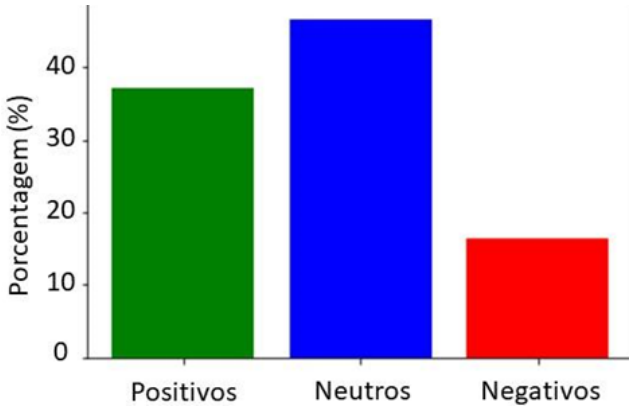
O melhor resultado que obtiveram foi utilizando o BERT para extração de características dos tweets, alcançando 75% de acurácia. Nosso projeto alcançou 80% de acurácia utilizando um classificador BERT para textos em português, tendo um desempenho menor do que o primeiro trabalho apresentado por conta da presença da classe “Neutro” em nosso projeto, já que esta é a classe onde temos os resultados menos satisfatórios, mas que ainda são considerados relevantes.

ANÁLISES DOS RESULTADOS

Além de apresentar um modelo que classifique de forma assertiva os textos apresentados a ele, o objetivo deste projeto também foi desenvolver análises que possam contribuir para a criação de um sistema, onde usuários comuns possam compreender como um determinado assunto está sendo visto pelas redes sociais. Para ilustrar melhor as análises e demonstrar o funcionamento do projeto, fizemos uma requisição ao nosso sistema que capturasse 1000 tweets que utilizaram a tag “#Oscars2023”, entre os dias 12/03/2023 (noite da cerimônia de premiação do Oscars) e 13/03/2023, e as seguintes análises foram realizadas utilizando estas postagens.

Primeira informação relevante que podemos mostrar é a porcentagem de tweets para cada classe, pois pode ser relevante saber como está a proporção de aceitação do público, no lugar de saber apenas se, de forma geral, está sendo “positivo”, por exemplo. Na Figura 4 podemos ver o gráfico demonstrando como estavam os tweets referentes à cerimônia do Oscar.

Figura 4: Proporção de tweets referentes a cerimônia do Oscar, entre os dias 12/03/2023 e 13/03/2023, para cada classe definida no projeto



Fonte: Dados originais da pesquisa

Analisando a Figura 4, podemos dizer que, para os tweets capturados durante este período, nós tivemos um baixo índice de pessoas comentando de forma negativa, chegando a somente 16,3%. A classe com textos considerados pelo modelo “neutros” foi a maior, com 46,33%, o que faz sentido, pois na noite da cerimônia do Oscar temos muitas postagens apenas com cunho informativo, por exemplo, “elenco de six of crows acaba de chegar no tapete vermelho”.

Outra análise interessante seria mostrar para os usuários os termos que estão mais vinculados ao assunto que se deseja estudar, para cada tipo de classificação. Com isso, utilizamos o recurso de Nuvens de Palavras, muito comum em análises de grandes volumes de textos, por ser uma ferramenta intuitiva e visualmente agradável para os usuários (Depaolo et al, 2014).

Foi criada uma Nuvem de Palavras para cada classe, onde pode ser visto os termos mais ou menos utilizados se referindo à Cerimônia do Oscar, baseado em seus tamanhos e em sua respectiva classificação. Nas Figuras 5, 6 e 7 são ilustradas essas Nuvens de Palavras.

Figura 5: Nuvem de Palavras com os termos relacionados à cerimônia do Oscar 2023, para a classe de tweets “Positivo”

Fonte: Dados originais da pesquisa

Figura 6: Nuvem de Palavras com os termos relacionados à cerimônia do Oscar 2023, para a classe de tweets “Neutro”

Fonte: Dados originais da pesquisa

Figura 7: Nuvem de Palavras com os termos relacionados à cerimônia do Oscar 2023, para a classe de tweets “Negativo”

Fonte: Dados originais da pesquisa

É possível realizar análises diferentes para cada Nuvem de Palavra apresentada. Na Figura 5, para a classe “Positivo”, podemos ver que os principais termos estão se referindo ao filme “Tudo em Todo Lugar ao Mesmo Tempo”, e à atriz Michelle Yeoh, vencedores de prêmios ao longo da noite. Na Figura 6, para a classe

"Neutro", temos muitas palavras relacionadas à Cerimônia e ao Tapete Vermelho (momento de chegada ao evento dos artistas), onde os nomes mais citados foram os do ator Pedro Pascal e da atriz e cantora Lady Gaga. Para a Figura 7, classe "Negativo", podemos ver que os usuários comentaram sobre a transmissão brasileira do evento.

Por fim, a última análise realizada neste projeto, foi o estudo de correlação entre as palavras mais utilizadas para cada classe, e a classificação que o modelo atribuiu para cada tweet. Como se trata de uma análise qualitativa, ou seja, sobre variáveis que não são números, é necessário realizar um processo de criação de variáveis binárias. Foi desenvolvido um processo de captura das principais palavras utilizadas nos textos, e posteriormente a criação de colunas na nossa base de dados, onde seus conteúdos podem ser "1" ou "0", indicando respectivamente se aquele termo estava ou não presente na sentença.

Após este processo, foi possível a realização do cálculo estatístico de Correlação de Pearson, onde é feito uma estimativa do quanto cada variável influencia em outra. Na Tabela 5 podemos ver como a presença de cada um dos principais termos utilizados em cada classe, pode ter influenciado na classificação do modelo.

Tabela 5: Estudo de Correlação de Pearson, para cada um dos principais termos utilizados durante a Cerimônia do Oscar 2023, e os tipos de classes do modelo

Principais Palavras	Positivo (%)	Neutro (%)	Negativo (%)
bom	0%	-2%	3%
horrível	-5%	-7%	16%
lady	-4%	9%	-7%
lindo	17%	-11%	-6%
maravilhoso	10%	-8%	-3%
mesmo	5%	-5%	-1%
red	-2%	4%	-2%
ruim	-5%	-6%	14%
vermelho	3%	-1%	-3%
linda	20%	-14%	-6%
melhor	19%	-11%	-9%

Fonte: Dados originais da pesquisa

Podemos observar na Tabela 5 que as palavras apresentam correlações baixas com as classes do modelo, não passando de 20%. Como nossa base de dados contém textos com quantidade de palavras diferentes, e os usuários podendo escrever

como quiserem, seria difícil um termo específico ter uma influencia grande na classificação do nosso modelo. Contudo, podemos analisar alguns termos chaves, como por exemplo “linda”, que foi muito utilizado nas postagens classificadas como “Positivo”, apresenta uma correlação de 20% com esta classe, e se comparada com as demais, mostra-se uma correlação “alta” para o nosso caso. Temos também a palavra “horrível”, com 16% de correlação com a classe “Negativo”, o que faz sentido já que esta palavra carrega um significado “negativo”.

Com isso, podemos mostrar como é possível realizar diversas análises mais intuitivas e relevantes para os usuários finais de um sistema de Análise de Dados, onde o foco do estudo seriam os comentários de usuários do Twitter sobre qualquer assunto. Outras análises podem ser realizadas utilizando como base o resultado do nosso modelo e nossos primeiros estudos, contudo, para o objetivo deste projeto estamos satisfeitos com as análises já demonstradas.

CONCLUSÃO

Neste trabalho, foi apresentado como pode ser favorável para uma organização a utilização de técnicas NLP para realizar análises em textos de redes sociais. É possível identificar padrões de comportamento dos usuários, observar se aceitaram ou não um novo produto no mercado, ou estão reagindo bem a um evento, como no exemplo demonstrado ao longo do texto, entre outras diversas análises possíveis.

Foi utilizado o BERT, uma técnica de processamento de linguagem natural moderna e robusta. Com ela, foi possível chegar a 80% de acurácia na tarefa de classificação dos textos nas classes “positivo”, “negativo” e “neutro”. Como mencionado neste trabalho, devido à falta de projetos e publicações realizados na área de NLP utilizando o idioma português, houve dificuldade em encontrar bases de dados com texto neste idioma, classificados baseados no sentimento de cada sentença. Com isso, também houve a contribuição com a criação de uma base de dados contendo cerca de 20 mil tweets e 12 mil avaliações de usuários na Google Play Store, todos os textos classificados baseados no sentimento predominante da sentença. Além disso, também se demonstrou como tal tarefa foi realizada, desde a

extração dos textos do Twitter e da Google Play Store, até a utilização da ferramenta de processamento de linguagem natural presente na plataforma Google Cloud.

Como sugestões de trabalhos futuros, é possível enriquecer a base de dados com mais sentenças de assuntos diferentes, o que iria garantir mais diversidade para o treinamento do modelo, e que pode ocasionar em uma melhora da acurácia na classificação de textos novos. Também seria importante a realização do processo de classificação dos tweets ser feito manualmente por um humano, pois o nosso modelo foi treinado com textos classificados pelo modelo da Google disponível no seu serviço de Nuvem, que pode errar suas classificações. Além disso, podemos aperfeiçoar o classificador adicionando mais camadas escondidas na saída do modelo BERTimbau pré-treinado, ou combinar outras técnicas de Aprendizado de Máquina buscando melhorar seu desempenho. Uma opção seria utilizar os modernos Large Language Models, como por exemplo o ChatGPT, pois devido a sua alta capacidade de interpretação de textos e de gerar respostas com detalhes e assertivas, torna-se um modelo com capacidade para ser utilizado em diversas tarefas diferentes (Liu et al, 2023). Por fim, novas análises podem ser propostas utilizando a saída do modelo, além da criação de uma interface com o usuário para a utilização real deste sistema.

REFERÊNCIAS

B, Anupama; D, Rakshith; Kumar, Rahul; M, Navaneeth. (2020). Real Time Twitter Sentiment Analysis using Natural Language Processing. International Journal of Engineering Research & Technology (IJERT).

Choi, Dami; Shallue, Christopher J.; Nado, Zachary; Lee, Jaehoon; Maddison, Chris J.; Dahl, George E. (2019). On Empirical Comparisons of Optimizers for Deep Learning.

Data Reportal. (2023). DIGITAL 2023: Global Overview Report. Disponível em: <https://datareportal.com/reports/digital-2023-global-overview-report>. Acesso em: 25 mar. 2023.

Depaolo, Concetta; Wilkinson, Kelly. (2014). Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. TechTrends.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Eberendu, Adanma. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology (IJCTT).

Filho, Jorge; Wilkens, Rodrigo; Idiart, Marco; Villavicencio, Aline. (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese.

Google Cloud Platform. (2008-2023). Disponível em: <https://cloud.google.com/>. Acesso em: 07 fev. 2023.

Google Play Store. (2008-2023). Disponível em: <https://play.google.com/store/>. Acesso em: 07 fev. 2023.

Internet Live Stats. [INTERNET LIVE STATS]. (2018). Twitter Usage Statistics. Disponível em: <https://www.internetlivestats.com>. Acesso em: 07 out. 2022.

IBM. [IBM]. (2020). Por que uma estratégia de armazenamento e proteção de dados?; Disponível em: <https://www.ibm.com/blog/>. Acesso em: 23 ago. 2023.

IBM. [IBM]. (2021). Structured vs. Unstructured Data: What's the Difference?; Disponível em: <https://www.ibm.com/blog/>. Acesso em: 23 ago. 2023.

Khurana, D.; Koli, A.; Khatter, K. et al. (2022). Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. Disponível em: <https://doi.org/10.1007/s11042-022-13428-4>.

Kingma, Diederik P.; Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization.

Mazumdar, S.; Thakker, D. (2020). Citizen Science on Twitter: Using Data Analytics to Understand Conversations and Networks. Future Internet.

Python. (2001-2023). Disponível em: <https://www.python.org/>. Acesso em: 07 fev. 2023.

Ravichandiran, Sudharsan. (2021). Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. 1ed. Packt Publishing.

Souza, F.; Nogueira, R.; Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Proceedings of the Brazilian Conference on Intelligent Systems, pages 403–417. Springer.

Towards Data Science. (2020). A Look at Precision, Recall, and F1-Score. Disponível em:<https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec>. Acesso em: 07 fev. 2023.

Twitter. (2006-2023). Disponível em: <https://twitter.com/>. Acesso em: 07 fev. 2023.

Vajjala, Sowmya; Majumder, Bodhisattwa; Gupta, Anuj; Surana, Harshit. (2020). Practical Natural Language Processing. O'Reilly Media.

Weiss, K.; Khoshgoftaar, T. M.; Wang, D. (2016). A survey of transfer learning.

Yao, Lihua; Jerry, Li; Alam, Hassan; Melnokov, Oleg. (2020). An Evaluation of Tweet Sentiment Classification Methods. 2020 International Conference on Computational Science and Computational Intelligence (CSCI).

Liu, Yiheng; Han, Tianle; Ma, Siyuan; Zhang, Jiayue; Yang, Yuanyuan; Tian, Jiaming; He, Hao; Li, Antong; He, Mengshen; Liu, Zhengliang; Wu, Zihao; Zhao, Lin; Zhu, Dajiang; Li, Xiang; Qiang, Ning; Shen, Dingang; Liu, Tianming; Ge, Bao. (2023). Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models.