MODELO PARA A ENTREGA DAS ATIVIDADES

COMPONENTE CURRICULAR:	PROJETO APLICADO II
NOME COMPLETO DO ALUNO:	Clayton dos Santos Lira Lorena Vaz Cord Tiago Clemente Rodrigues Lucas Vaz de Castro Oliveira
RA:	10416054 10424700 10423746 10424623

Atenção: Toda atividade deverá ser feita com fonte Arial, tamanho 11, espaço de 1,5 entre as linhas e alinhamento justificado entre as margens.

UNIVERSIDADE PRESBITERIANA MACKENZIE

Tecnologia em ciência de dados

RELATÓRIO TÉCNICO

Análise de Sentimento em Comentários de E-commerce Usando Processamento de Linguagem Natural e Aprendizado de Máquina

São Paulo

2024

1. Introdução

O comércio eletrônico, ou *e-commerce*, tem se consolidado como uma das formas mais relevantes e dinâmicas de comércio em todo o mundo, promovendo o crescimento da economia digital e transformando os padrões de consumo. Em consequência ao crescimento deste setor, é gerada uma grande quantidade de dados, dentre estes, comentários e avaliações sobre produtos, a partir dos quais as empresas podem extrair valor e usar como subsídio para tomada de decisões (LIU; YE, 2022).

O Brasil é um dos maiores mercados de e-commerce da América Latina, e utilizar os dados provenientes das avaliações dos produtos é essencial para entender as opiniões dos clientes e, com isso, manter a competitividade e oferecer experiências excepcionais. Nesse sentido, a análise de sentimentos é uma técnica de grande utilidade, a qual utiliza o processamento de linguagem natural (NLP) aliado à linguística computacional e mineração de dados para automatizar a análise de grandes volumes de dados textuais, avaliando as opiniões dos clientes permitindo respostas rápidas a feedbacks negativos. Tal estratégia pode reduzir inclusive a necessidade de pesquisas de satisfação com os clientes, processo que gera custos adicionais com recursos humanos além de levar tempo (SINGH; SARRAF, 2020).

Por se tratar de um problema de classificação, a análise de sentimentos pode ser feita com o uso de algoritmos de aprendizado de máquina supervisionado. Entre eles, o algoritmo Random Forest se destaca como uma das técnicas aplicadas, conforme discutido por Manpreet Kaur (2020). Segundo Sumatti et al. (2020), o Random Forest é um tipo de algoritmo de ensemble, que, a partir do conjunto de treinamento, gera um conjunto de árvores de decisão. Em seguida, os resultados dessas árvores são agregados para determinar a classe final, sendo o resultado uma saída combinada de cada árvore no conjunto. É um algoritmo flexível e robusto, composto por um grande número de árvores e à medida que o número de árvores aumenta, a robustez do classificador também tende a aumentar.

Diante do exposto e levando em consideração a aplicabilidade da análise de sentimentos no contexto do comércio eletrônico, o presente estudo propõe a criação de um modelo de aprendizado de máquina supervisionado com o intuito de classificar avaliações e resenhas deixadas por clientes, usando para isso os comentários de compras realizadas na Olist no período de 2016 a 2018. O objetivo é identificar áreas de melhoria na experiência dos clientes, promovendo uma compreensão mais profunda de suas emoções e contribuindo para a melhoria contínua dos serviços e produtos oferecidos pela empresa.

2. Objetivos

2.1. Objetivo geral

Construir um modelo, por meio da linguagem python, voltado para a análise de sentimento em comentários dos clientes a partir das avaliações e comentários feitos pelos clientes em uma base com dados de e-*commerce*, utilizando técnicas de processamento de linguagem natural e aprendizado de máquina supervisionado para identificar padrões de opinião e gerar insights que possam apoiar a tomada de decisões estratégicas.

2.2. Objetivos específicos

1.1.1. Realizar o pré-processamento linguístico dos textos, por meio de processos de tokenização, remoção de stopwords e lematização para transformá-los em inputs adequados para modelagem;

1.1.2. Utilizar as técnicas POS-tagging e a representação vetorial TF-IDF para converter os textos em representações numéricas para uso no modelo proposto;

1.1.3. Treinar e avaliar um modelo de aprendizado de máquina, utilizando algoritmo Random Forest com o intuito de classificar sentimentos e avaliar seu desempenho utilizando as métricas de acurácia, precisão, revocação e F1-score.

3. Metodologia

3.1. Aquisição dos dados

O conjunto de dados utilizado neste projeto foi disponibilizado pela empresa Olist, proprietária dos dados, por meio da plataforma Kaggle, renomada e conhecida comunidade online voltada para competições e troca de conhecimentos entre profissionais e entusiastas de ciência de dados.

A Olist é uma empresa que oferece um conjunto completo de soluções voltadas para negócios. Entre os serviços oferecidos, destaca-se a Olist Store, uma plataforma de vendas online que opera em diversas lojas e marketplaces populares no Brasil, como Amazon, Shopee, Magalu, Americanas, entre outros. Através da Olist, empresas têm a oportunidade de expor seus produtos nos sites de grandes redes de varejo, contando com suporte em áreas como gestão, logística e atendimento ao cliente. A Olist não comercializa produtos próprios, mas possibilita que lojistas e fabricantes utilizem sua plataforma para impulsionar suas vendas. Isso é possível porque a empresa possui grande relevância nesses marketplaces, o que resulta em maior visibilidade para os parceiros. Dessa forma, a Olist recebe uma comissão sobre as vendas, enquanto seus parceiros conseguem ampliar suas operações e aumentar o volume de vendas mensais. Para mais informações, é possível acessar o site oficial da Olist.

O conjunto de dados utilizado pode ser acessado através do link https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Esse dataset foi publicado sob a licença pública Creative Commons CC BY-NC-SA 4.0 (Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional), o que permite a reprodução e o compartilhamento dos dados originais ou modificados, desde que seja para fins não comerciais.

3.2. Descrição do dataset e metadados

De acordo com o proprietário dos dados, o dataset fornecido contém informações detalhadas sobre clientes, sua localização, vendedores, valores de frete e produtos, além das avaliações e comentários de clientes a respeito das compras realizadas através da plataforma. O conjunto abrange aproximadamente 100.000 pedidos realizados entre 2016 e 2018 em diversos marketplaces no Brasil. Embora os dados sejam reais, a identidade dos clientes e parceiros comerciais foi preservada, com os nomes substituídos por referências da série de televisão Game of Thrones, garantindo o anonimato.

O conjunto de dados é composto por nove tabelas no formato CSV. Contudo, este estudo se concentra especificamente na quinta tabela, denominada "olist_order_reviews_dataset.csv". A seguir, são apresentadas a descrição dos atributos desta tabela.

- olist_order_reviews_dataset.csv

  Descritivo: Dados com avaliações e comentários feitos pelos clientes.

  Linhas: 99.224

  Atributos:

  - review_id (string)
  - order_id (string)
  - review_score (integer)
  - review_comment_title (string)
  - review_comment_message (string)
  - review_creation_date (datetime)
  - review_answer_timestamp (datetime)

3.3. Ferramentas, linguagem e bibliotecas utilizadas

O conjunto de dados "olist_order_reviews_dataset" foi analisado com uso da linguagem Python. A análise foi conduzida no Google Colab, e os resultados foram carregados ao fim de cada etapa em um repositório do GitHub, onde todo o estudo, incluindo explicações, foi documentado nos próprios códigos e notas do Colab.

Para importação e tratamento dos dados foi utilizada a biblioteca pandas, que se trata de uma ferramenta de código aberto amplamente utilizada para análise e manipulação de dados na linguagem Python. Ela fornece estruturas de dados flexíveis e eficientes, como *DataFrame* e *Series*. Sendo assim, a tabela "olist_order_reviews_dataset" será referida de agora em diante como *DataFrame*, visto que esta foi importada como este tipo de objeto para a presente análise.

Além disso, quando necessária, a visualização dos dados foi através de gráficos de coluna, com auxílio da biblioteca Matplotlib, a qual foi projetada para criar gráficos de maneira personalizável, sendo amplamente usada em ciência de dados, análise de dados e visualização para comunicação de resultados. Já a confecção, treinamento e avaliação de métricas do modelo de classificação foram executados por meio de módulos da biblioteca de aprendizado de máquina Scikit-learn, a qual oferece uma ampla gama de ferramentas para tarefas como classificação, regressão, agrupamento, redução de dimensionalidade e validação de modelos. Mais especificamente, foram usados os módulos *resample, TfidfVectorizer, train_test_split, RandomForestClassifier* e *classification_report*.

4. Resultados

4.1. Análise exploratória

Após importação dos dados como dataframe, foram realizadas as etapas de tratamento e pré-processamento (os códigos e detalhes podem ser encontrados no GitHub). Durante o tratamento, foi realizada a verificação de dimensões, ou seja, colunas e linhas do dataframe utilizando os comandos head e shape.

Foi verificado que o dataframe possui um total de 99224 linhas e 7 colunas. Nesse sentido, as colunas que apresentaram relevância para a análise foram review_score, review_comment_title e review_comment_message. Desses atributos, o review_score não apresentou nenhum valor nulo, com 99224 registros, enquanto review_comment_message figurou com 40977 registros (58247 nulos) e review_comment_title teve o menor número de registros (1156 válidos e 87656 nulos).

Foi possível constatar que atributo **review_comment_message** possui uma cobertura maior e com mais informações detalhadas que **review_comment_title**, sendo uma boa fonte de dados para a análise proposta. No entanto, embora o atributo **review_comment_title**, contenha menor quantidade de dados válidos, oferece informações importantes de forma mais concisa e representativa do sentimento expresso pelo cliente. Em outras palavras, o título tende a refletir o sentimento mais proeminente e imediato (positivo, negativo ou neutro). Por isso, optou-se pela combinação do título, quando disponível, com o conteúdo do comentário, de forma a enriquecer a análise, na forma do novo atributo nomeado **combined_text**, usando como artifício a concatenação em cada registro do conteúdo da coluna **review_comment_title** somado ao conteúdo de **review_comment_message**, intercalados por uma string de espaço ‘ ‘ (o código e detalhes podem ser encontrados no GitHub).

Após este procedimento, os valores ausentes da nova coluna foram removidos através do método dropna() que está definido dentro do objeto dataframe do pandas. Com isso, restaram como valores válidos 40977 registros.

Por utilizar o aprendizado de máquina supervisionado, a estratégia usada neste estudo foi gerar a variável objetivo a partir das informações presentes no atributo **review_score.** Este atributo corresponde a avaliações que são representadas por números inteiros variando de 1 a 5, para representar o grau de satisfação do cliente que vai de muito insatisfeito até muito satisfeito, respectivamente. Para facilitar a análise, optou-se pela eliminação das avaliações de nota 3 (consideradas neutras e correspondentes a 3557 dos registros válidos - 8,7%) e dicotomização das restantes entre positivas (notas 4 e 5) e negativas (notas 1 e 2) conforme esquematizado na tabela abaixo. O rótulo gerado foi

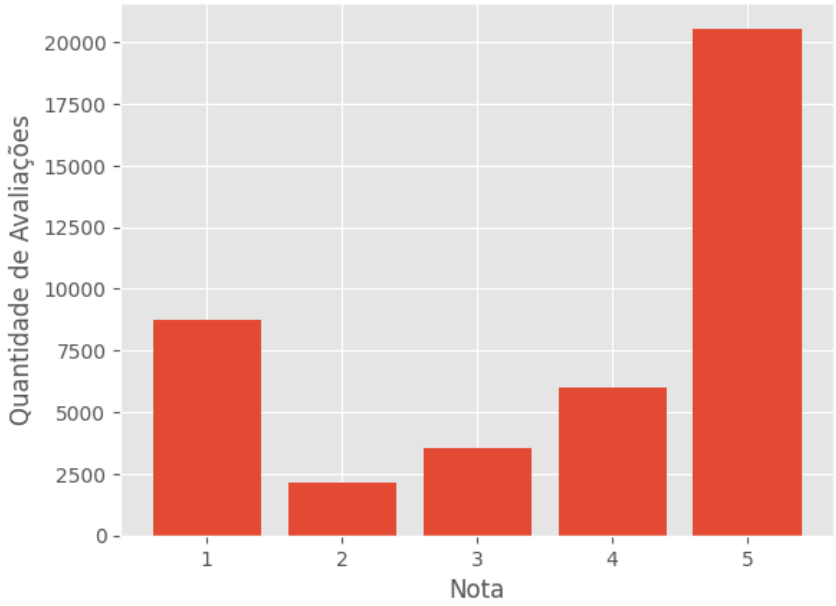
armazenado em um novo atributo denominado sentimento. Ademais, a figura abaixo possibilita uma melhor avaliação da distribuição das avaliações.

Tabela 1 - Quantidade de avaliações por classe do atributo review_score e rótulo atribuído.

Review_score	Quantidade	Proporção	Rótulo atribuído
1	8745	21,3%	Negativo
2	2145	5,2%	Negativo
3	3557	8,7%	-
4	5976	14,6%	Positivo
5	20554	50,2%	Positivo
Total	40977	100%	-

Fonte: o autor.

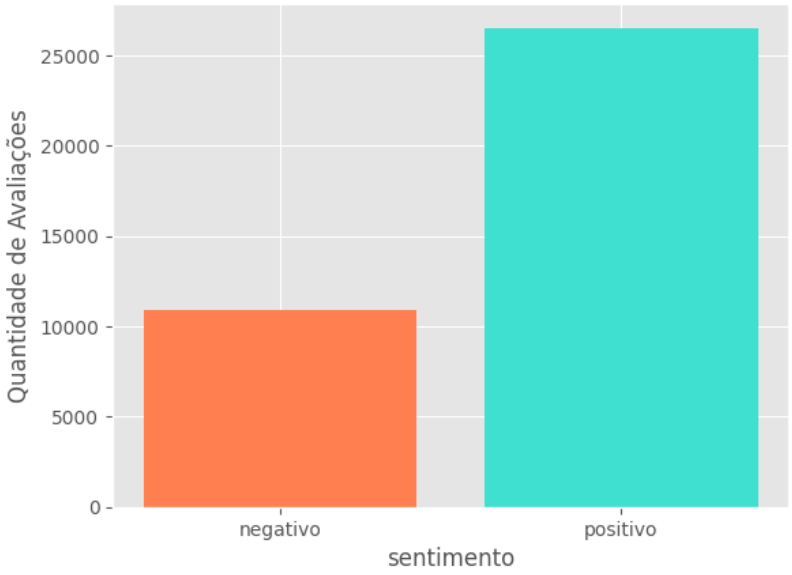
Figura 1 - Distribuição do atributo review_score



Fonte: o autor.

A figura 2 mostra a distribuição após a dicotomização das variáveis. Nesse caso, foi possível observar um desequilíbrio entre as avaliações positivas (mais frequentes) e avaliações negativas (menos frequentes), o que poderia levar ao enviesamento da análise.

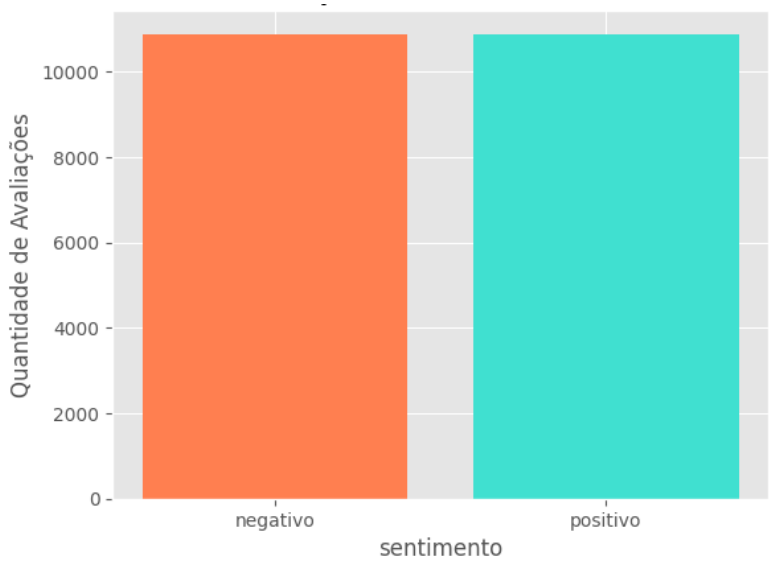
Figura 2 - Distribuição do atributo sentimento rotulado de acordo com as classes do review_score



Fonte: o autor.

Diante do desequilíbrio entre as classes da variável objetivo, foi utilizada a função resample para fazer uma subamostragem das avaliações positivas. Após este procedimento, foi obtido um *dataframe* com 10890 registros para cada uma das classes, totalizando 21780 registros válidos para o prosseguimento do projeto. A figura 3 representa a distribuição das classes do novo atributo sentimento após o procedimento supracitado para endereçar a predomínio de avaliações positivas.

Figura 3 - distribuição das classes positivo e negativo após subamostragem



Fonte: o autor.

4.2. Pré-processamento do atributo preditor combined_text com técnicas de processamento de linguagem natural (NLP)

Para gerar o input adequado para o modelo proposto, foi realizada inicialmente uma limpeza de ruídos com o auxílio das bibliotecas unidecode, emoji e re. Além disso, foram executados diversos procedimentos referentes ao processamento de linguagem natural (NLP) com o auxílio da biblioteca NLTK (Natural Language Toolkit). Esta é amplamente utilizada para o desenvolvimento de aplicações de NLP., oferecendo um conjunto de ferramentas que facilitam as tarefas de tokenização, remoção de stopwords, lematização, POS-tagging entre outras.

Nesse contexto, a tokenização pode ser definida como o processo de dividir um texto em unidades menores chamadas tokens, como palavras ou frases, permitindo a análise individual de cada elemento. Já a remoção de *stopwords*, consiste em eliminar palavras muito comuns, como artigos e preposições, que não agregam valor significativo ao texto. A lematização, por sua vez, reduz as palavras a sua forma canônica, preservando o significado gramatical. Por fim, o POS-tagging (Part-of-Speech Tagging) atribui categorias gramaticais, como substantivo, verbo ou adjetivo, a cada token, permitindo a análise contextual. Essas etapas são essenciais para estruturar e compreender dados textuais em projetos de processamento de linguagem natural (JAKHOTIYA et al., 2022).

Dessa forma, foi realizado o tratamento dos dados textuais combinados, removendo a capitalização das letras, emojis, acentuação, URLs e espaços em branco indesejados nos comentários. Em seguida, foram executadas etapas de tokenização, remoção de stopwords, lematização, POS-tagging.

Por fim, foi feita representação vetorial utilizando o modelo de Term Frequency — Inverse Document Frequency (TF-IDF), a qual consiste em uma técnica usada no NLP para converter textos em vetores numéricos, destacando a relevância de palavras em um documento em relação a um conjunto de documentos. Essa abordagem ajuda a reduzir a influência de palavras muito frequentes, mas pouco informativas (como *stopwords*), atribuindo maior peso às palavras que aparecem com frequência em um documento, mas são raras no corpus, o que torna o TF-IDF eficaz para tarefas como classificação de texto e recuperação de informações (ALLAHYARI et al. 2017).

Após esse processamento, foi feita a separação do conjunto de dados entre variável preditora e variável objetivo. Logo após, foi utilizada a função *train_test_split* para separação dos conjuntos de treinamento e teste para dar prosseguimento à construção do modelo de aprendizado de máquina.

4.3. Treinamento do modelo

Neste estudo, optou-se pelo modelo de classificação Random Forest, que é um algoritmo amplamente utilizado em tarefas de classificação devido à sua capacidade de lidar com dados complexos e à resistência ao overfitting. Para isso, foi utilizada a biblioteca Sci-kit learn e os parâmetros para o modelo foram os seguintes: max_features='log2', class_weight = 'balanced'.

4.4. Avaliação das métricas

A avaliação do desempenho do modelo treinado foi feita por meio de métricas de precisão, recall, F1-Score e acurácia. Observando os resultados obtidos foi possível observar que o modelo apresenta um bom desempenho, com alta precisão, recall e F1-Score para ambas as classes.

As métricas de avaliação incluem uma acurácia geral de 92% indicando que o modelo classificou corretamente a maior parte das avaliações como positivas ou negativas. Para a classe "negativo", o modelo alcançou uma precisão de 89% e um recall de 95%, sugerindo que, embora tenha cometido alguns erros ao classificar avaliações negativas, foi altamente eficaz em identificar corretamente as avaliações dessa classe.

Já para a classe "positivo", a precisão foi de 95% e o recall de 89%, o que mostra que o modelo foi mais assertivo ao prever corretamente avaliações positivas, mas deixou de identificar algumas. O F1-score para ambas as classes foi de 0,92, refletindo um bom equilíbrio entre precisão e recall. A média ponderada e a macro média das métricas também corroboram a performance consistente do modelo.

5. Conclusão

A análise de sentimentos tem se tornado uma ferramenta essencial para empresas que desejam entender melhor a percepção de seus clientes, e a Olist, uma plataforma de e-commerce que conecta vendedores a grandes marketplaces, é um exemplo claro dessa aplicação.

A Olist não apenas facilita a venda de produtos através de sua plataforma, mas também coleta uma vasta quantidade de dados sobre as experiências dos consumidores. Com a crescente concorrência no setor de e-commerce, compreender as emoções e opiniões dos clientes se tornou crucial para manter a competitividade. A análise de sentimentos permite que a Olist identifique rapidamente áreas que necessitam de melhorias, respondendo de forma proativa a feedbacks negativos e aprimorando a experiência do usuário.

Utilizando técnicas avançadas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, a Olist pode analisar comentários e avaliações feitas por clientes. Por exemplo, ao aplicar um modelo de floresta aleatória para classificar sentimentos em avaliações, a empresa consegue não apenas determinar se uma opinião é positiva ou negativa, mas também extrair insights valiosos sobre o que os clientes realmente pensam sobre seus produtos e serviços.

O impacto da análise de sentimentos na Olist pode ser significativo. Com uma acurácia geral do modelo em torno de 92%, a empresa pode confiar nas classificações geradas para tomar decisões informadas sobre melhorias em seus serviços. Isso não apenas melhora a satisfação do cliente, mas também fortalece a reputação da marca no mercado. Ao entender melhor as emoções dos consumidores, a Olist pode não apenas melhorar sua oferta, mas também se posicionar como uma líder no competitivo cenário do e-commerce brasileiro.

6. Melhorias e perspectivas futuras

Apesar do modelo apresentar bom desempenho, algumas melhorias podem ser implementadas nesse estudo. Inicialmente, vale lembrar que os registros com nota 3 foram eliminados da análise devido a baixa frequência. Nesse sentido, a ausência do rótulo para a classe neutro pode limitar a aplicabilidade do modelo, já que ao manter as três classes (positivo, neutro e negativo), o modelo seria mais completo, representando melhor a diversidade das opiniões. Além disso, é importante destacar que nem sempre a avaliação com notas de 1 a 5 é coerente com a resenha dos comentários. Por exemplo, um usuário pode dar a nota 1 em uma avaliação, porém escrever um comentário positivo ou o oposto. Nesse sentido, poderiam ser utilizadas outras estratégias como a de *Emotional word tagging* utilizada no estudo de Liu e Ye (2022). Outro ponto de melhoria diz respeito ao processo de seleção de modelos e hiperparâmetros. Para isso, poderia ser utilizada a ferramenta *GridSearchCV* também presente na biblioteca Scikit-Learn para melhoria das métricas e otimização do modelo.

7. Repositório do Github, apresentação com storytelling e vídeo da apresentação

Endereço para o repositório: https://github.com/lucasvazcastro/Projeto_Aplicado_II

Endereço para a apresentação (storytelling): https://docs.google.com/presentation/d/1aZ-AmbbitcTsmrcO_uN6QDp6HCelj-DZ/edit?usp=sharing&ouid=105411754165150340934&rtpof=true&sd=true

Endereço para o vídeo da apresentação: https://youtu.be/uLBK1RBgJts

Referências

LIU, Yanqiu; YE, Fuming. Sentiment Analysis of Online Catering User Comments Based on Random Forest Feature Extraction. *In*: **2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)**. Zhangye, China: IEEE, 2022, p. 667–670.

SINGH, Shailendra Narayan; SARRAF, Twinkle. Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm. In: **2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**. Noida, India: IEEE, 2020, p. 112–116.

KAUR, Manpreet. An Approach for Sentiment Analysis Using Gini Index with Random Forest Classification. In: SMYS, S.; TAVARES, João Manuel R. S.; BALAS, Valentina Emilia; et al (Orgs.). **Computational Vision and Bio-Inspired Computing**. Cham: Springer International Publishing, 2020, v. 1108, p. 541–554.

SUMATHI, B.; SIJI GEORGE C.; et al. Classification of Sentiment on Business Data for Decision Making using Supervised Machine Learning Methods. **International Journal of Engineering and Advanced Technology**, v. 9, n. 3, p. 3595–3600, 2020.

ALLAHYARI, Mehdi; POURIYEH, Seyedamin; ASSEFI, Mehdi; et al. Text Summarization Techniques: A Brief Survey. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 10, 2017.

JAKHOTIYA, Aachal; JAIN, Harshada; JAIN, Bhavik; CHANIYARA, Charmi. Text Pre-Processing Techniques in Natural Language Processing: A Review. I**nternational Research Journal of Engineering and Technology (IRJET)**, v. 9, n. 2, p. 878–880, Feb. 2022.