THIAGO SALLES SANTOS

UMA ABORDAGEM BASEADA NO MODELO BERT PARA MINERAÇÃO DE OPINIÃO SOBRE A BEBIDA CACHAÇA

LAVRAS – MG

2023

THIAGO SALLES SANTOS

UMA ABORDAGEM BASEADA NO MODELO BERT PARA MINERAÇÃO DE OPINIÃO SOBRE A BEBIDA CACHAÇA

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Bacharelado em Ciência da Computação para a obtenção do título de Bacharel.

Prof. Denilson Alves Pereira, Ph.D.
Orientador

LAVRAS – MG

2023

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos

da Biblioteca Universitária da UFLA

Snatos, Thiago Salles

Uma Abordagem Baseada no Modelo BERT para Mineração de Opinião sobre a Bebida Cachaça / Thiago Salles Santos. 1^a ed. rev., atual. e ampl. – Lavras : UFLA, 2023.

36 p. : il.

TCC(graduação)–Universidade Federal de Lavras, 2023.

Orientador: Prof. Denilson Alves Pereira, Ph.D..

Bibliografia.

1. TCC. 2. Monografia. 3. Dissertação. 4. Tese. 5. Trabalho Científico – Normas. I. Universidade Federal de Lavras. II. Título.

CDD-808.066

THIAGO SALLES SANTOS

UMA ABORDAGEM BASEADA NO MODELO BERT PARA MINERAÇÃO DE OPINIÃO SOBRE A BEBIDA CACHAÇA

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Bacharelado em Ciência da Computação para a obtenção do título de Bacharel.

APROVADA em 06 de dezembro de 2023.

Prof. Dra. Marluce Rodrigues Pereira	UFLA
Prof. Dra. Paula Christina Figueira Cardoso	UFLA

gov.br
Documento assinado digitalmente
DENILSON ALVES PEREIRA
Data: 19/12/2023 20:55:31-0300
Verifique em https://validar.iti.gov.br

Prof. Denilson Alves Pereira, Ph.D.
Orientador

LAVRAS – MG
2023

OS SONHOS DAS PESSOAS... NÃO TÊM FIM! NÃO É VERDADE?!
(Marshall D. Teach)

RESUMO

A análise de sentimentos é uma técnica que permite identificar e quantificar opiniões, atitudes e emoções expressas em textos. Ela é uma ferramenta valiosa para aplicações como monitoramento de marca, pesquisa de mercado e tomada de decisões políticas. No entanto, a análise de sentimentos ainda apresenta alguns desafios, como a interpretação correta de nuances linguísticas e a aplicação em domínios específicos. Este trabalho aborda o desafio de um modelo de análise de sentimentos no domínio da bebida cachaça. O objetivo principal é apresentar uma abordagem para a mineração de opinião em textos, a partir de bases de dados extraídas de redes sociais, utilizando modelos baseados no modelo de linguagem *BERT*. Tais abordagens obtiveram resultados de 97,2% na métrica F1-Score para comentários extraídos da rede social *Facebook* e de 95,2% para *tweets* extraídos da plataforma *Twitter*

Palavras-chave: Análise de Sentimento. Cachaça. Redes Sociais

ABSTRACT

Sentiment analysis is a technique that allows identifying and quantifying opinions, attitudes and emotions expressed in texts. It is a valuable tool for applications such as brand monitoring, market research and political decision making. However, sentiment analysis still presents some challenges, such as correctly interpreting linguistic nuances and applying it to specific domains. This work addresses the challenge of a sentiment analysis model in the cachaça beverage domain. The main objective is to present an approach for opinion mining in texts, from datasets extracted from social networks, using models based on the *BERT* language model. Such approaches obtained results as 97.2% in the F1-Score metric for comments extracted from the social network *Facebook* and as 95.2% for tweets extracted from the *Twitter* platform.

Keywords: Sentiment Analysis. Cachaça. Social Media

SUMÁRIO

Parte I

PRIMEIRA PARTE - INTRODUÇÃO GERAL

1 INTRODUÇÃO

O presente trabalho apresenta o artigo desenvolvido intitulado como "Uma Abordagem Baseada no Modelo BERT para Mineração de Opinião sobre a Bebida Cachaça", que atua com o desenvolvimento de abordagens para a mineração de opinião na língua portuguesa para a bebida brasileira cachaça, utilizando-se de modelos baseados no modelo *BERT*(DEVLIN et al., 2019).

A mineração de opinião é uma divisão da análise de sentimento, que tem como objetivo a extração de opiniões em texto não estruturados, como por exemplo, extrair se o texto exprime uma opinião positiva, negativa ou neutra (SANTOS; BECKER; MOREIRA, 2014).

Já a bebida cachaça, é uma bebida alcoólica, tradicional do Brasil, que é produzia através da destilação do caldo da cana-de-açúcar, podendo ter o envelhecimento em barris de madeira ou aço-inox, além da adição de açúcares (ALCARDE, 2018).

O objetivo deste é trabalho apresentar o artigo "Uma Abordagem Baseada no Modelo BERT para Mineração de Opinião sobre a Bebida Cachaça", com o propósito de desenvolver abordagens para a tarefa de mineração de opinião sobre o domínio específico da bebida cachaça, utilizando-se de comentários coletados de redes sociais, acerca do assunto cachaça, além de desenvolver uma base de dados coletada de redes sociais.

A estrutura adotada para este documento, segue conforme descrito no manual de normalização e estrutura de trabalhos acadêmicos da Universidade Federal de Lavras (UFLA) para trabalhos no formato de artigo para a publicação em periódicos científicos, sendo dividido em duas partes, a primeira parte do documento trata-se de um resumo geral do presente trabalho, apresentando apenas uma introdução que descreve o trabalho como um todo, mostrando os assuntos abordados, os objetivos almejados, a finalidade do trabalho e também apresentar a estrutura geral do documento. Já a segunda parte do documento trata-se do artigo desenvol-

vido para a publicação em periódicos intitulado como "Uma Abordagem Baseada no Modelo BERT para Mineração de Opinião sobre a Bebida Cachaça", que segue a estrutura estabelecida pela Sociedade Brasileira de Computação (SBC), neste artigo os tópicos principais presente são: introdução, trabalhos relacionados, base de dados, avaliação experimental, conclusão e trabalhos futuros.

REFERÊNCIAS

ALCARDE, A. **Cachaça: ciência, tecnologia e arte**. 1st online. ed. Editora Blucher, 2018. Disponível em: https://books.google.com.br/books?id=4StdDwAAQBAJ. ISBN 9788521208457. Disponível em: <https://books.google.com.br/books?id=4StdDwAAQBAJ>.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>.

SANTOS, A.; BECKER, K.; MOREIRA, V. Um estudo de caso de mineração de emoções em textos multilíngues. In: **Anais do III Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brazil: SBC, 2014. p. 140–151. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/6810>.

Parte II

SEGUNDA PARTE - ARTIGOS

Uma Abordagem Baseada no Modelo BERT para Mineração de Opinião sobre a Bebida Cachaça

Thiago Salles Santos[1], **Denilson Alves Pereira**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Lavras – Caixa postal 3037, Lavras, 37.200-900, MG, Brasil.

thiago.santos6@estudante.ufla.br, denilsonpereira@ufla.br

Abstract. *Sentiment analysis is a technique that allows identifying and quantifying opinions, attitudes and emotions expressed in texts. It is a valuable tool for applications such as brand monitoring, market research and political decision making. However, sentiment analysis still presents some challenges, such as correctly interpreting linguistic nuances and applying it to specific domains. This work addresses the challenge of a sentiment analysis model in the cachaça beverage domain. The main objective is to present an approach for opinion mining in texts, from datasets extracted from social networks, using models based on the BERT language model. Such approaches obtained results as 97.2% in the F1-Score metric for comments extracted from the social network Facebook and as 95.2% for tweets extracted from the Twitter platform.*

Resumo. *A análise de sentimentos é uma técnica que permite identificar e quantificar opiniões, atitudes e emoções expressas em textos. Ela é uma ferramenta valiosa para aplicações como monitoramento de marca, pesquisa de mercado e tomada de decisões políticas. No entanto, a análise de sentimentos ainda apresenta alguns desafios, como a interpretação correta de nuances linguísticas e a aplicação em domínios específicos. Este trabalho aborda o desafio de um modelo de análise de sentimentos no domínio da bebida cachaça. O objetivo principal é apresentar uma abordagem para a mineração de opinião em textos, a partir de bases de dados extraídas de redes sociais, utilizando modelos baseados no modelo de linguagem BERT. Tais abordagens obtiveram resultados de 97,2% na métrica F1-Score para comentários extraídos da rede social Facebook e de 95,2% para tweets extraídos da plataforma Twitter.*

1. Introdução

Vivemos em um cenário no qual saber, compreender, analisar e mais do que tudo predizer as opiniões, atitudes e emoções expressas nas mensagens textuais das pessoas não é mais uma característica adicional, e sim uma atividade essencial para o mercado, para aplicações como: monitoramento de marca, pesquisa de mercado, tomada de decisões políticas, detecção de tendências, entre outros [Gomes et al. 2017].

A análise de sentimentos (AS) permite extrair informações valiosas a partir de dados em larga escala, que são gerados e compartilhados diariamente nas redes sociais, fóruns online e outras plataformas digitais [Gomes et al. 2017].

No entanto, apesar dos avanços alcançados, a análise de sentimentos ainda apresenta alguns desafios, como a interpretação correta de nuances linguísticas, como sarcasmo, ironia e ambiguidade, a aplicação em domínios específicos, dentre outros. Neste

trabalho iremos abordar o desafio de um modelo de análise de sentimentos no domínio específico da bebida cachaça.

1.1. Conceitos Básicos

A análise de sentimentos é uma ramificação da área de Processamento de Linguagem Natural (PLN), que tem como objetivo quantificar ou qualificar as opiniões ou emoções em dados não estruturados [Liu 2012].

A análise de sentimentos possui duas sub-áreas de atuação, que são: a mineração de opinião e a mineração de emoções. A mineração de opinião tem o foco na identificação de opiniões, como por exemplo, se o texto possui uma opinião positiva, negativa ou neutra. A mineração de emoções identifica o tipo de emoção expressa no texto, como por exemplo, felicidade, tristeza, raiva e entre outros [Santos et al. 2014].

Para a implementação das análises, existem diversas abordagens, desde abordagens simples baseadas em regras gramaticais, até abordagens mais complexas, como modelos de aprendizado de máquina (do inglês *machine learning*, da sigla ML). Neste trabalho, são tratadas apenas as abordagens de ML, mais especificamente abordagens que utilizam de redes neurais [Perreira 2021].

A AS pode ser utilizada em diferentes cenários, como por exemplo o marketing, para entender o que os clientes acham ou sentem sobre um determinado produto ou serviço. Outro exemplo é na política, para avaliar a popularidade e adesão de um determinado candidato ou partido. Compreender as opiniões e os sentimentos do seu público alvo, permite que as estratégias de comunicação e posicionamentos sejam desenvolvidos de acordo com as expectativas e desejos do grupo-alvo [Gomes et al. 2017].

No contexto de bebidas, a cachaça, ou também conhecida como aguardente, é uma bebida tradicional do Brasil, tendo suas primeiras produções em terra canarinha ainda na época de colônia, onde era um produto secundário produzido nos engenhos de cana-de-açúcar, mas hoje se tornou destaque no mercado de bebida nacional [Câmara 2018].

A bebida é produzida a partir da destilação do caldo da cana-de-açúcar, de alta graduação alcoólica, variando entre 38% a 48% de graduação, tendo a adição de açúcares e o envelhecimento em barris [Alcarde 2018].

A cachaça possui grande presença no mercado nacional, sendo a segunda bebida alcoólica mais consumida no país [SEBRAE 2015] além de empregar mais de 600 mil trabalhadores [SEBRAE 2022]. Já no âmbito de exportações, só em 2022 o valor de vendas do produto para fora do país foi de mais de US$ 18 milhões, sendo exportada para mais de 70 países, os principais sendo: Estados Unidos, Alemanha, Portugal e Itália. O volume exportado chega na casa dos 8 milhões de litros [Brandão 2022].

1.2. Motivação para o Trabalho

A cachaça é uma bebida altamente consumida e produzida no Brasil, tendo uma produção na casa de bilhões de litros [Alcarde 2018]. Segundo pesquisas do SEBRAE (Serviço Brasileiro de Apoio às Micro e Pequenas Empresas), o mercado de cachaças no Brasil se encontra em expansão [SEBRAE 2022]. Além de que, a cachaça é considerada um patrimônio cultural no Brasil [Dias 2014], carregando consigo uma parte da cultura brasileira.

Porém, mesmo com todas estas características, falta investimentos em pesquisas, de empresas e instituições locais, sobre a bebida [Martins et al. 2018], o que nos motiva a realizar este trabalho, com intuito de acrescentar conhecimento neste âmbito, tão importante para a cultura brasileira.

1.3. Lacunas

As lacunas que o presente trabalho almeja preencher ou reduzir são a pouca disponibilidade de estudos na tarefa de análise de sentimentos para a língua portuguesa [Perreira 2021], estudos acerca do produto cachaça [Martins et al. 2018] e a falta de recursos no domínio específico da cachaça para área de inteligência artificial.

1.4. Objetivos

Este trabalho tem como objetivo principal apresentar uma abordagem para a mineração de opinião em textos, no domínio específico da bebida cachaça, a partir de bases de dados extraídas de redes sociais, utilizando de modelos baseados no modelo *BERT* [Devlin et al. 2019].

Como objetivos secundários, o trabalho visa a coleta de uma base de dados para a tarefa de mineração de opinião, para o domínio da bebida cachaça, a rotulação amostral da base de dados coletada, o uso de bases de dados de terceiros para o ajuste fino dos modelos e a realização de um comparativo de abordagens.

1.5. Principais Resultados

Os principais resultados foram 97,2% na métrica F1-Score para o *dataset Facebook* e 95,2% para o *dataset Twitter*, demonstrando que as abordagens que estão descritas neste trabalho possuem potencial para a tarefa de mineração de opinião no domínio específico da bebida cachaça.

1.6. Organização deste Documento

O presente trabalho se encontra dividido nas seguintes seções: a primeira seção, introdutória, apresentou os conceitos básicos que norteiam o trabalho, além dos principais resultados obtidos, já a segunda seção, apresenta alguns trabalhos relacionados, a terceira seção descreve sobre as bases de dados utilizadas no trabalho, a quarta seção dissemina os procedimentos realizados e os resultados obtidos juntamente com sua discussão, e a quinta seção conclui sobre o trabalho apresentado e indica possíveis futuros trabalhos.

2. Trabalhos Relacionados

Trabalhar com o tema cachaça e o uso de inteligência artificial no meio acadêmico, são características presentes no trabalho [Silva et al. 2023], onde é apresentado uma base de dados desenvolvida para tarefa de reconhecimento de entidades nomeadas (da sigla em inglês NER), possuindo mais de 180.000 *tokens*, rotulados em 17 categorias de entidades nomeadas, para o idioma português, no domínio específico da bebida cachaça, além de uma avaliação experimental do *dataset* desenvolvido, utilizando o modelo *BERTimbau*, obtendo 0,933 de micro-F1.

Outro trabalho que também aborda essa relação entre a cachaça e a computação é [Rodrigues et al. 2015], que utiliza da visão computacional, juntamente com uso de quatro algoritmos de classificação: redes neurais, k-NN, máquinas de vetores de suporte (da

sigla em inglês SVM), Naive Bayes, e da técnica *ensemble AdaBoost* para a combinação dos classificadores, com o intuito de predizer a tipagem da bebida, como por exemplo se a cachaça é premium, extra premium ou envelhecida.

Ambos os trabalhos [Silva et al. 2023] e [Rodrigues et al. 2015], trabalham essa relação entre a cachaça e o uso de inteligência artificial assim como o presente trabalho, porém no trabalho [Rodrigues et al. 2015] essa relação é trabalhada em volto da área da visão computacional. Já o trabalho [Silva et al. 2023] possui uma relação mais próxima com o presente trabalho, visto que ambos os trabalhos utilizam de textos coletados no domínio da bebida cachaça, porém sendo o trabalho [Silva et al. 2023] para a tarefa de NER e o presente trabalho para a tarefa de mineração de opinião, e também ambos os trabalhos utilizam o modelo *BERTimbau* [Souza et al. 2020] para a realização dos experimentos com as base de dados desenvolvida.

Utilizar a análise de sentimentos para a identificação de opiniões e emoções no meio das redes sociais já é um aspecto que vem sendo estudado há algum tempo. O trabalho [Evangelista and Padilha 2014] apresenta uma ferramenta para a classificação de texto nas redes sociais *Facebook* e *Twitter*, para análise da reputação de empresas de comércio eletrônico. Para isso, o trabalho apresenta técnicas de extração de informação, juntamente com os métodos de classificação, para analisar se os comentários são positivos, negativos ou neutros.

[Araújo et al. 2013], também trabalha com análise de sentimentos em comentários nas redes sociais, vista a capacidade das plataformas de comunicação em gerar informações, apresentando oito métodos para a classificação de sentimentos para comentários em redes sociais, fóruns e revistas, além de desenvolver um novo método, que combina as abordagens exigentes, e por fim, apresenta a plataforma *iFeel*[1], que é uma ferramenta para comparação de diversos métodos de classificação de texto.

Assim como o presente trabalho, os trabalhos [Evangelista and Padilha 2014] e [Araújo et al. 2013] trabalham com a mineração de opinião na língua portuguesa, utilizando de comentários extraídos de redes sociais, com o intuito de avaliar as abordagens propostas, porém nos trabalhos [Evangelista and Padilha 2014] e [Araújo et al. 2013], as abordagens não apresentam o uso de redes neurais, ao contrário das abordagens propostas neste documento.

Outro trabalho que faz um comparativo de abordagens para análise de sentimentos em *tweets* é [Cardozo and Freitas 2021], que propõe o uso dos modelos de memória de longo prazo (da sigla em inglês LSTM) e o modelo SVM, para a base de dados *TweetSentBR* [Brum and Nunes 2017], com e sem pré-processamento, além de comparar resultados obtidos com os resultados apresentados no trabalho [Brum and Nunes 2017]. O trabalho [Cardozo and Freitas 2021] assim como o presente trabalho fazem um comparativo de resultados entre os resultados obtidos pelas as abordagens propostas com os resultados do trabalho [Brum and Nunes 2017], com diferencial que no presente trabalho, as abordagens utilizadas são diferentes, além de que a base de dados *TweetSentBR* foi utilizada com o intuito de realizar o ajuste fino do modelo *BERTimbau*.

[1] `http://www.ifeel.dcc.ufmg.br/`

3. Base de Dados

Nesta seção, iremos descrever os processos realizados para a coleta de dados, para o seu pré-processamento nos *datasets* e para a rotulação das amostras.

3.1. Coleta de Dados

A base de dados coletada de redes sociais se deu em conjunto com o trabalho CachaçaNER [Silva et al. 2023], a qual é constituída de um acervo de dados de comentários extraídos das redes sociais *Facebook* e *Twitter*, que estão inseridos ou abordam o assunto da bebida cachaça.

O intuito dessa coleta foi criar um conjunto de dados sobre a bebida cachaça, para a tarefa de análise de sentimento na língua portuguesa. Porém, devido ao alto custo, foi rotulada apenas uma amostra do conjunto de dados. Para treinar os modelos de aprendizagem de máquina, foram adicionalmente utilizadas outras bases de dados, conforme descrito na Seção 3.4.

A base de dados coletada está dividida em duas partes. Uma parte é referente a comentários extraídos da rede social *Facebook*, e a outra é referente a *tweets* extraídos da rede social *Twitter*.

3.1.1. Base de dados Facebook

Trata-se de comentários extraídos da rede social *Facebook*, com 36.015 comentários extraídos de 22.064 postagens, de 19 páginas relacionadas a cachaça. A maioria dessas páginas são referentes aos sites de vendas e fóruns utilizados para a extração de informação para o projeto CachaçaNER [Silva et al. 2023].

Para a extração dos comentários, foi utilizada a API *GraphAPI*[2], que é uma ferramenta disponibilizada pelo *Facebook*, a qual permite o desenvolvedor retirar ou inserir informações na rede social através de um protocolo HTTPS. Para a comunicação e requisição dos dados, foi utilizado por sua vez a biblioteca Request em Python para o manuseio da API.

A extração dos dados ocorreu nas páginas: casadabebida, mbcachacaria, amburanabr, Sanhacu, lojacachacaepinga, cachacavelhobarreirooficial, cachacaepresente, domtapparoengenho, cachacarianacional, CachacaCompanheira, blubeer.com.br, araraunacachacaria, brme.oficial, cachacariasalinas.com.br, magnificadefaria, cachacasapucaia, wibacachaca, bebidaonline, cachacasbrasileirasoficial. Para cada página extraída, foi gerado um *dataset* com todos os comentários encontrados, em todas as postagens desde a criação da página até setembro de 2021. Para cada comentário, foram extraídos: o texto do comentário, a sua data de publicação, seu identificador e seu link permanente.

Toda a base de dados foi unificada, gerando assim apenas um *dataset*, contendo todos os comentários, porém 1.395 instâncias foram removidas, por não possuírem textos, e 180 foram desconsiderados, por serem repetidas, resultando assim em um *dataset* com 34.440 instâncias válidas.

[2] `https://developers.facebook.com/docs/graph-api`

3.1.2. Base de dados Twitter

Trata-se de *tweets* extraídos da rede social *Twitter*, com 17.766 *tweets* extraídos de 12 páginas relacionadas a cachaça, e também *tweets* que possuem em seu escopo textual a *hastag* "#cachaca".

Para a extração dos *tweets*, foi utilizado a API *Twitter - API*[3], que é uma ferramenta disponibilizada pelo *Twitter*, ao qual permite via protocolo HTTPS retirar ou inserir informações na rede social. Para a comunicação e requisição dos dados, foi utilizada novamente a biblioteca Request em Python para o manuseio da API.

A extração dos dados ocorreu nas páginas: araraunacachaca, BebaCompanheira, cachacanacional, cachacariasp, cachacaslinas, CachacaSapucaia, cachacawiba, DomTapparo, EmporioCCEma, NaBebidaOnline, sanhacu. Juntamente com a pesquisa por *tweets* na rede social que possuem "#cachaca" em seu escopo textual. Para cada *tweet*, foram extraídos: o texto do *tweet*, seu identificador e a data de sua publicação.

Toda a base de dados foi unificada, gerando assim apenas um *dataset*, que após a exclusão de uma instância repetida resultou em 17.765 instâncias válidas.

3.2. Pré-Processamento

O pré-processamento realizado consistiu da remoção de todas as instâncias que possuíam texto vazio, ou seja, nenhum carácter ou contendo apenas espaço, tabulação ou quebra de linhas. Posteriormente, foi realizada a limpeza dos textos, removendo das suas extremidades o espaço, a tabulação e a quebra de linha.

Também foi realizado um pré-processamento nos rótulos, onde foram numerados para os valores de: 0, 1 e 2, representando, respectivamente, as classes: negativo, neutro, positivo.

3.3. Rotulação

A rotulação de dados é o processo de atribuir rótulos a dados brutos, podendo ser de natureza categórica, numérica ou textual. A rotulação pode ser feita de forma manual ou automática [Reips and Hara 2022].

Para este trabalho, foi realizada a rotulação manual, através da ferramenta *doccano*[4], que é uma ferramenta *open source* de rotulação de dados para documentos de texto ou imagens [Nakayama et al. 2018].

A Figura 1 mostra a interface da ferramenta *doccano*, na qual pode-se visualizar, ao centro, o dado que se deseja rotular, juntamente com os rótulos inseridos, na parte esquerda, o menu de navegação, e na parte direita, informações sobre o dado a ser rotulado.

Foi rotulada uma amostra das bases de dados escolhida de forma aleatória entre todas as instâncias coletadas. O processo de rotulação foi realizado pelo próprio autor, que rotulou cada instância em uma das seguintes classes: negativo, positivo ou neutro. Textos negativos que expressam insatisfações, revisões negativas e reclamações, por exemplo: *"To esperando chegar... comprei dia 13 e ainda não enviaram pra transportadora..."*.

[3] `https://developer.twitter.com/en/docs/twitter-api`

[4] `https://github.com/doccano/doccano`

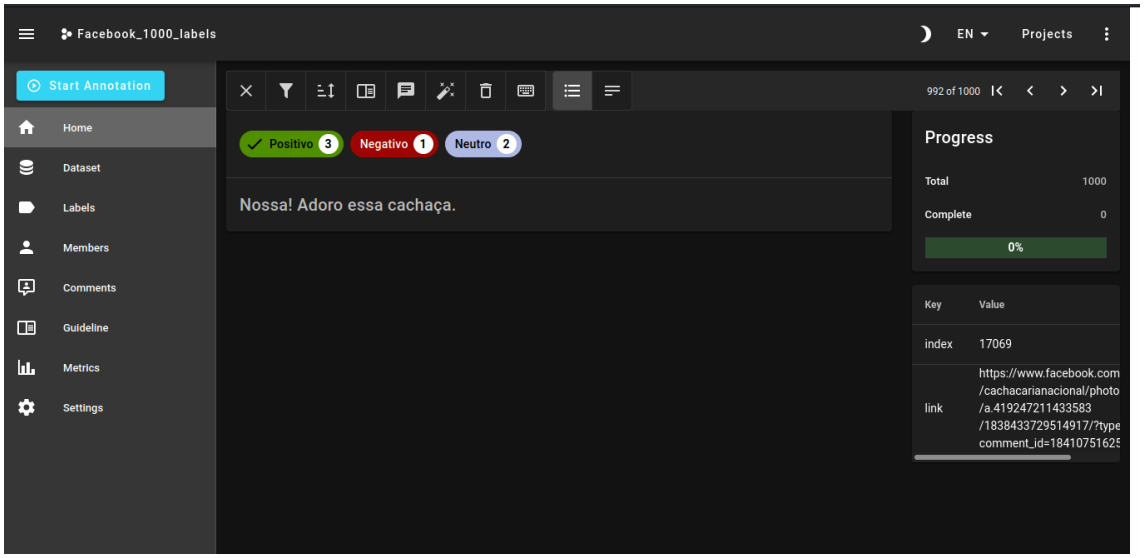

Figure 1. Interface da ferramenta doccano

Já o rótulo positivo é para textos que expressam elogios, revisões positivas, e desejo ao produto, por exemplo: *"Eu também quero muito"*. Por último, o rótulo neutro, para textos que expressam, tanto sentimento positivo como negativo, anúncios, textos com apenas marcações de outros usuários da plataforma, texto que não foi possível categorizar como positivo ou negativo, por exemplo: *"E boa sim , mas tem melhoras"*.

A base de dados rotulada foi denominada CachaçaOM (Cachaça Opinion Mining), a qual é composta de amostras dos dados coletados das redes sociais *Facebook* e *Twitter*.

3.3.1. Base de dados Facebook

Foram selecionados aleatoriamente do *dataset* 1.200 comentários, pertencentes ao conjunto de instâncias válidas, para a rotulação manual, entre as classes: negativo, neutro e positivo. O resultado é apresentado na Tabela 1.

Rótulo	Quantidade	Porcentagem
Negativo	31	2,58%
Neutro	593	49,42%
Positivo	576	48,0%

Table 1. Rótulos do dataset Facebook

A Tabela 2 mostra as informações dos textos de entrada, da base de dados *Facebook*, apresentando a média de palavras por instância, a mediana, a moda, o máximo de palavras em uma instância, o mínimo, e também o vocabulário, que mostra a quantidade de palavras únicas em todas as instâncias do *dataset*.

média	mediana	moda	máx	mín	Vocabulário
5	3	2	68	1	2.570

Table 2. Informações sobre a quantidade de palavras por instância do dataset Facebook

3.3.2. Base de dados Twitter

Foram selecionados 1.000 *tweets* aleatórios, pertencentes ao conjunto de instância válidas, para a rotulação manual, entre as classes: negativo, neutro e positivo. O resultado desta rotulação é apresentado na Tabela 3.

Rótulo	Quantidade	Porcentagem
Negativo	19	1,9%
Neutro	878	87,8%
Positivo	103	10,3%

Table 3. Rótulos do dataset Twitter

A Tabela 4 mostra as informações sobre a quantidade de palavras por instância da base de dados *Twitter*.

média	mediana	moda	máx	mín	Vocabulário
17	16	19	53	1	6.648

Table 4. Informações sobre a quantidade de palavras por instância do dataset Twitter

3.4. Outras Bases de Dados

As outras bases de dados utilizadas neste trabalho são compostas por acervos de dois trabalhos. São eles *Brazilian Portuguese Sentiment Analysis Datasets* [Souza and Filho 2021] e a *Building a Sentiment Corpus of Tweets in Brazilian Portuguese* [Brum and Nunes 2017], também conhecida com *TweetSentBR*. Ambas são bases de dados brasileiras públicas, extraídas de fontes digitais.

3.4.1. Base de dados Brazilian Portuguese Sentiment Analysis Datasets

A base de dados *Brazilian Portuguese Sentiment Analysis Datasets*, desenvolvida por [Souza and Filho 2021], composta por texto na língua portuguesa no domínio de revisões de produtos. O acervo se encontra disponível em `https://www.kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets`. A base de dados é composta por dados que foram retirados dos seguintes acervos públicos:

- **Brazilian E-Commerce Public Dataset by Olist**: É um acervo público desenvolvido pela Olist. Esse acervo é composto por análises de produtos

realizadas entre os anos de 2016 a 2018, por empresas parceiras da Olist [Olist and Sionek 2018].

- **B2W-Reviews01**: É um acervo desenvolvido pela equipe da B2W Digital, uma empresa latina de comércio digital. Esse acervo é composto por análises de produtos, realizadas entre os meses de janeiro a maio do ano de 2018, no site de vendas digital `americanas.com` [Real et al. 2021].
- **Corpus Bucapé**: É um corpus desenvolvido pelo projeto Opinando [Opinando 2020], composto por análises de produtos extraídos em setembro de 2013 [Hartmann et al. 2014].
- **UTLCorpus**: É um corpus desenvolvido pelo projeto Opinando [Opinando 2020], e foi dividido em dois corpora, um que trata de avaliações de filmes, que foram retirados do site `filmow.com`, no dia de 14 março de 2019. Já a outra parte do corpus trata de avaliações de aplicativos na plataforma *GooglePlay*, também no dia de 14 março de 2019 [Sousa et al. 2019].

Além dos *datasets* retirados de acervos públicos, o acervo de dados *Brazilian Portuguese Sentiment Analysis Datasets* fornece mais um *dataset*, composto da unificação dos *datasets* dos acervo públicos selecionados, chamado de *concatenate*. Porém os *datasets buscape* e *concatenate*, presentes neste acervo de dados não puderam ser empregados por falta de memória da placa de vídeo utilizada no experimento, uma vez que ao realizar os experimentos com estes *datasets* o erro OOM (Out of Memory) ocorria, devido a este fato ambos os *datasets* foram desconsiderados.

A Tabela 5 apresenta as informações estatísticas dos *datasets* considerando a quantidade de palavras por instâncias.

Dataset	qtd. instância	média	mediana	moda	máx	mín	Vocabulário
olist	37.953	7	6	2	35	1	14.601
b2w	115.977	14	10	7	618	1	48.211
utlc_apps	968.018	8	5	1	356	1	140.388
utlc_movies	1.188.497	21	10	2	4.515	1	265.990

Table 5. Informações sobre a quantidade de palavras por instância dos datasets do acervo de dados Brazilian Portuguese Sentiment Analysis Datasets

Já a Tabela 6 mostra o percentual de cada classe de rótulo para cada *dataset* do acervo, e entre parênteses a quantidade de instância.

Dataset	Negativo	Positivo
olist	29,97% (11.407)	70,03% (26.655)
b2w	30,81% (35.758)	69,19% (80.300)
utlc_movies	11,56% (137.539)	88,44% (1.052.003)
utlc_apps	22,51% (218.114)	77,49% (750.744)

Table 6. Percentual de classe do rótulo de cada dataset do acervo de dados Brazilian Portuguese Sentiment Analysis Datasets

A escolha do acervo *Brazilian Portuguese Sentiment Analysis Datasets* se deve ao fato de que, no geral, o conjunto de dados fornece uma base sólida para a análise

das opiniões e preferências dos consumidores em relação aos produtos, atrelando-se com nosso objetivo de analisar os sentimentos na esfera do produto cachaça.

3.4.2. Base de dados TweetSentBR

A base de dados *TweetSentBR* foi desenvolvida pelo trabalho que pertence ao projeto Opinando da USP [Brum and Nunes 2017]. Esse acervo possui enfoque em *tweets*, na esfera de programas de TV, possuindo um total de 15.047 *tweets* extraídos entre os meses de janeiro a junho de 2017. O acervo se encontra disponível publicamente em `https://bitbucket.org/HBrum/tweetsentbr/`.

A base de dados possui 212 instâncias de textos repetidos, as quais foram retiradas para realização do estudo, e também possui 47 instâncias não rotuladas, que também foram removidas, além de 4 instâncias removidas após o processo de pré-processamento do texto, contando com um total de 14.784 instâncias, rotuladas entre as classes: negativo, neutro e positivo, como mostra a Tabela 7.

Rótulo	Quantidade	Porcentagem
Negativo	4.403	29,78%
Neutro	3.783	25,59%
Positivo	6.598	44,63%

Table 7. Rótulo do dataset TweetSentBR

Já a Tabela 8, mostra informações sobre a quantidade de palavras por instância do texto de entrada utilizado da base de dados.

média	mediana	moda	máx	mín	Vocabulário
7	7	5	51	1	14.400

Table 8. Informações sobre a quantidade de palavras por instância do dataset TweetSentBR

A escolha dessa base de dados se deve ao fato de ser um acervo composto por *tweets*, carregando consigo o estilo e o comportamento do vocabulário utilizado em redes sociais, vinculando-se com as bases de dados que deseja-se avaliar, que também são extraídas de redes sociais.

4. Avaliação Experimental

Nesta seção, estão descritos os procedimentos e as configurações utilizadas na pesquisa, abordando os métodos, os parâmetros e a maneira de condução de cada experimento, pontuando as tecnologias utilizadas, como por exemplo, os modelos, as bibliotecas e a linguagem de programação. Também, são apresentados e discutidos os resultados obtidos.

4.1. Configuração Experimental

Os modelos de linguagem utilizados para a realização da pesquisa estão disponíveis no *Hugging Face*[5], que é uma plataforma online para o compartilhamento de modelos, *datasets* e ferramentas.

O primeiro modelo selecionado do repositório foi o *BERTimbau* [Souza et al. 2020], que é um modelo brasileiro, específico para a língua portuguesa, desenvolvido a partir do modelo *BERT* [Devlin et al. 2019]. O *BERT* por sua vez é um modelo bidirecional, ou seja, capaz de observar o contexto tanto à direita como à esquerda, proposto pela *Google* em 2018 e se tornou referência na comunidade de PLN.

O segundo modelo selecionado foi o *twitter-XLM-roBERTa*, que se trata de um modelo treinado com 198 milhões de *tweets* em diversos idiomas [Barbieri et al. 2022], incluindo o português, para a tarefa de análise de sentimento, podendo classificar entre três rótulos: positivo, negativo ou neutro. Este modelo utiliza como base o modelo *XLM-roBERTa* [Conneau et al. 2020], que é uma versão do modelo *roBERTa* [Liu et al. 2019] multilíngue.

Para a realização dos experimentos, foi utilizado um computador como servidor remoto, com as seguintes especificações: placa gráfica NVIDIA GeForce RTX 3090 contendo 24GB de memória dedicada, 128GB de memória RAM, processador Intel Core i7 da 10º geração, de 2.9GHz.

O servidor roda sobre o sistema operacional *Ubuntu* 22.04.3 LTS. Para a realização dos experimentos foi criado um ambiente virtual, através da ferramenta de gerenciador de pacotes *micromamba*[6] na versão 1.0.0. Para a execução dos experimentos, foi selecionada a linguagem de programação *Python* na versão 3.9.13, além das bibliotecas *Tensorflow* na versão 2.7.0, *Transformers* na versão 4.24.0 e a *Datasets* na versão 2.6.1.

Para a conexão com o servidor via SSH ou SFTP foi utilizada, a ferramenta *AnyDesk* [7] na versão 6.2.0, e também a IDE *Visual Studio Code*[8], com a extensão *Remote - SSH* na versão 0.106.4, que possibilitou a realização da codificação e da execução direta no servidor.

Para o controle de versionamento dos documentos foi utilizado a ferramenta git[9] na versão 2.34.1, juntamente com a plataforma github[10] para a hospedagem online do repositório. Repositório esse que se encontra disponível em `https://github.com/ThiagoSallesSantos/IC_AnaliseSentimento/`.

Os hiperparâmetros adotados no processo de ajuste fino do modelo *BERTimbau* foram os hiperparâmetros descritos no artigo [Devlin et al. 2019], são eles: número de épocas igual a 3, o otimizador *Adam* com a taxa de aprendizado igual a *2e-5*, com uma variação nos valores de *batchs*, sendo realizado um ajuste com *batch* igual a 16 e outro com valor igual a 32, devido ao tamanho variado dos *datasets*. Já para o modelo *twitter-*

[5] `https://huggingface.co/`
[6] `https://mamba.readthedocs.io/en/latest/user_guide/micromamba.html`
[7] `https://anydesk.com/pt`
[8] `https://code.visualstudio.com/`
[9] `https://git-scm.com/`
[10] `https://github.com/`

XLM-RoBERTa não foi adotado nenhum hiperparâmetro, visto que foi utilizado o modelo pré-treinado.

As partições adotadas para as bases de dados de terceiros foram as mesmas utilizadas nos trabalhos [Souza and Filho 2021] para a base de dados *Brazilian Portuguese Sentiment Analysis Datasets* e [Brum and Nunes 2017], para base de dados *TweetSentBR*.

Para ambos os modelos, foi utilizado o *tokenizador* correspondente ao modelo, que se encontram disponíveis junto à página dos modelos na plataforma *Hugging Face*.

Importante salientar que, para os experimentos de classificação binária com os *datasets Facebook* e *Twitter*, foram considerados apenas os dados rotulados como negativo e positivo, eliminando assim os dados rotulados como neutro, para que assim haja conformidade entre os modelos, visto que as bases do trabalho [Souza and Filho 2021] utilizadas para o ajuste fino do modelo *Bertimbau* não possui o rótulo neutro. Nos casos em que o modelo *twitter-XLM-RoBERTa* foi utilizado para predizer rótulos binários, as predições como neutro foram ignoradas.

4.2. Métricas de Avaliação

Nesta seção, apresentamos as métricas utilizadas nos experimentos. A Tabela 9 mostra o índice de siglas utilizada referente aos valores da matriz de confusão.

Sigla	Valor de Confusão
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo

Table 9. Índice das siglas da matriz de confusão

4.2.1. Acurácia

A Acurácia mede a proporção de predições corretas em relação ao total de predições feitas [Pedregosa et al. 2011]. Sua fórmula é:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN}$$

4.2.2. Revocação

O Revocação mede a proporção de verdadeiros positivos em relação ao total de instâncias que realmente pertencem à classe positiva [Pedregosa et al. 2011]. A fórmula para calcular o revocação é:

$$Revocacao = \frac{VP}{VP + VN}$$

4.2.3. Precisão

A Precisão avalia a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões positivas realizadas [Pedregosa et al. 2011]. A fórmula para calcular a Precisão é:

$$Precisao = \frac{VP}{VP + FP}$$

4.2.4. F1-Score

O F1-score combina tanto a precisão quanto revocação, oferecendo uma medida única [Pedregosa et al. 2011]. A fórmula para calcular o F1-score é:

$$F1 = 2 * \frac{Precisao * Revocacao}{Precisao + Revocacao}$$

A escolha do F1-score é devido ao fato de que ela oferece uma medida mais equilibrada e informativa do desempenho do modelo em cenários de desigualdade de classes, permitindo uma avaliação mais abrangente do modelo.

4.2.5. ROC-AUC

ROC-AUC é uma métrica utilizada para avaliar o desempenho de modelos de classificação binária. Ela mede a capacidade do modelo em discriminar corretamente entre as classes positiva e negativa em diferentes limiares de classificação. A curva ROC é um gráfico que representa a taxa de verdadeiros positivos em função da taxa de falsos positivos para diferentes limiares de classificação [Pedregosa et al. 2011].

A fórmula para calcular a Área sob a Curva ROC pode variar dependendo do método utilizado para construir a curva, um método comum é o cálculo da área usando a regra trapezoidal.

A escolha da métrica ROC-AUC se deve a conformidade com os *baselines* estabelecidos, para a realização da avaliação e consequentemente a escolha do modelo.

4.3. Baselines

Foram selecionados dois trabalhos para compor as *baselines*, são eles [Souza and Filho 2021] e [Brum and Nunes 2017].

4.3.1. Sentiment Analysis on Brazilian Portuguese User Reviews

O *Sentiment Analysis on Brazilian Portuguese User Reviews* trata-se de um trabalho desenvolvido pela UFRJ (Universidade Federal do Rio de Janeiro), com o propósito de realizar a tarefa de classificação de texto na língua portuguesa, utilizando de abordagens tradicionais [Souza and Filho 2021].

As abordagens tradicionais consistem em métodos mais básicos para classificação de textos, baseando-se em extrair características específicas dos documentos, e posteriormente, alimentar algum classificador com essas características extraídas.

No trabalho, foram utilizadas duas abordagens. A primeira consistiu em utilizar de *embeddings* de documentos, utilizando: *FastText*, *GloVe* e *Word2Vec*, todos treinados para a língua portuguesa, possuindo as seguintes variações em suas dimensões: 50, 100 e 300, e posteriormente foram utilizados os classificadores Logistic Regression, Random Forest e LightGBM. Já a segunda abordagem utilizou o esquema de pesos TF-IDF (*Term Frequency x Inverse Document Frequency*) para calcular os pesos da palavras, utilizando de variações do vocabulário a ser considerado para cada *dataset*, e posteriormente, foram utilizados os classificadores Logistic Regression e LightGBM. Porém para o presente trabalho apenas a primeira abordagem será utilizada para critérios comparativos, visto que, os resultados da segunda abordagem são apresentados apenas em formato de gráfico, sem os valores exatos, dificultando a comparação.

4.3.2. Building a Sentiment Corpus of Tweets in Brazilian Portuguese

O *Building a Sentiment Corpus of Tweets in Brazilian Portuguese* trata-se de um trabalho desenvolvido pela USP (Universidade de São Paulo), com o propósito de desenvolver uma base de dados de *tweets* rotulados para a tarefa de análise de sentimentos no contexto de programas de televisão brasileiros [Brum and Nunes 2017].

Para a avaliação experimental da base de dados desenvolvida, o trabalho, utilizou do saco de palavras (do inglês *bag of words*) para incorporação dos *tweets*, e também de três métodos de aprendizado de máquina: *SVM*, *Naive Bayes* e uma abordagem hibrida, que combina o *SVM* com um classificador léxico, esse último apenas para a avaliação binária da base de dados.

4.4. Resultados e Discussões

O primeiro conjunto de experimentos teve como objetivo avaliar os modelos baseados no *BERT* para a tarefa de classificação de sentimentos. Foi feita uma comparação com os dois trabalhos utilizados como *baseline*.

A Tabela 10 mostra os resultados pelos modelos gerados a partir do ajuste fino do modelo *BERTimbau* para a tarefa de classificação de texto, para cada *dataset*.

Para o modelo *BERTimbau*, foram mostrados apenas os resultados do ajuste fino com hiperparâmetro de *batch* igual a 16, visto que não houve diferença significativa entre os valores de 16 e 32 *batchs*.

O *dataset* TwitterASBR representa o acervo de dados do trabalho [Brum and Nunes 2017] e possui duas variações, uma com duas classes: positivo e negativo, e outra com três classes: positivo, negativo e neutro.

Já a Tabela 11 mostra os resultados após a predição do modelo *twitter-XLM-RoBERTa*, para cada *dataset*.

Vale ressaltar que tanto para o modelo *BERTimbau* como o modelo *twitter-XLM-RoBERTa*, o *dataset* TwitterASBR com três classes não possui a métrica ROC-AUC, visto

BERTimbau					
Dataset	Acurácia	Precisão	Revocação	F1	ROC-AUC
olist	0,944	0,954	0,967	0,960	0,929
b2w	0,970	0,984	0,972	0,978	0,969
utlc_apps	0,946	0,968	0,962	0,965	0,927
utlc _movies	0,952	0,965	0,981	0,973	0,856
TwitterASBR - 2 classes	0,890	0,936	0,878	0,906	0,893
TwitterASBR - 3 classes	0,747	0,749	0,747	0,748	-

Table 10. Resultados dos modelos gerados pelo ajuste fino do BERTimbau

twitter-XLM-RoBERTa					
Dataset	Acurácia	Precisão	Revocação	F1	ROC-AUC
olist	0,884	0,961	0,869	0,913	0,893
b2w	0,927	0,975	0,918	0,946	0,933
utlc_apps	0,861	0,979	0,838	0,903	0,889
utlc _movies	0,740	0,957	0,740	0,834	0,743
TwitterASBR - 2 classes	0,857	0,954	0,802	0,871	0,871
TwitterASBR - 3 classes	0,717	0,745	0,717	0,722	-

Table 11. Resultados do modelo twitter-XLM-RoBERTa

ser um métrica de classificação exclusivamente binária.

Observando as Tabelas 10 e 11, o modelo *BERTimbau* desempenhou melhor que o modelo *twitter-XLM-RoBERTa* de maneira geral. Isso se deve ao fato de que o modelo *BERTimbau* passou pelo processo de ajuste fino, utilizando-se das partições destinadas ao treino dos *datasets*, adquirindo assim conhecimento sobre o domínio, o que possibilitou um melhor resultado. Já o modelo *twitter-XLM-RoBERTa* vem pré-treinado para classificação de sentimentos, e não passou por um ajuste fino nos *datasets* utilizados neste trabalho. Além de que, tendencialmente modelos monolíngue desempenham melhores que modelos multilíngue [Conneau et al. 2020].

A Figura 2 mostra o gráfico dos comparativos entre os valores obtidos pela métrica ROC-AUC entre as abordagens para a tarefa de classificação de texto, para cada *dataset*. Foram selecionados os dois melhores resultados do trabalho [Souza and Filho 2021]. A métrica ROC-AUC foi a única utilizada no referido trabalho.

Os resultados mostram que o modelo *BERTimbau* conseguiu desempenhar melhor em alguns casos e em outros, o método clássico *FastText* com 300 dimensões utilizando o classificador LightGBM foi superior. Já o modelo *twitter-XLM-RoBERTa* foi inferior em todos os casos, provavelmente por ter sido treinado com base de dados em *tweets*, que possui geralmente um escopo de texto diferente.

As Figuras 3 e 4 mostram os gráficos dos comparativos entre os valores obtidos pelas métricas F1 e Acurácia entre as abordagens para a tarefa de classificação de texto, com 2 classes e 3 classes respectivamente, para o *dataset* TwitterASBR. Vale ressaltar que para a classificação de texto com 3 classes a abordagem *Hybrid Classifier* não foi utilizada no trabalho [Brum and Nunes 2017].

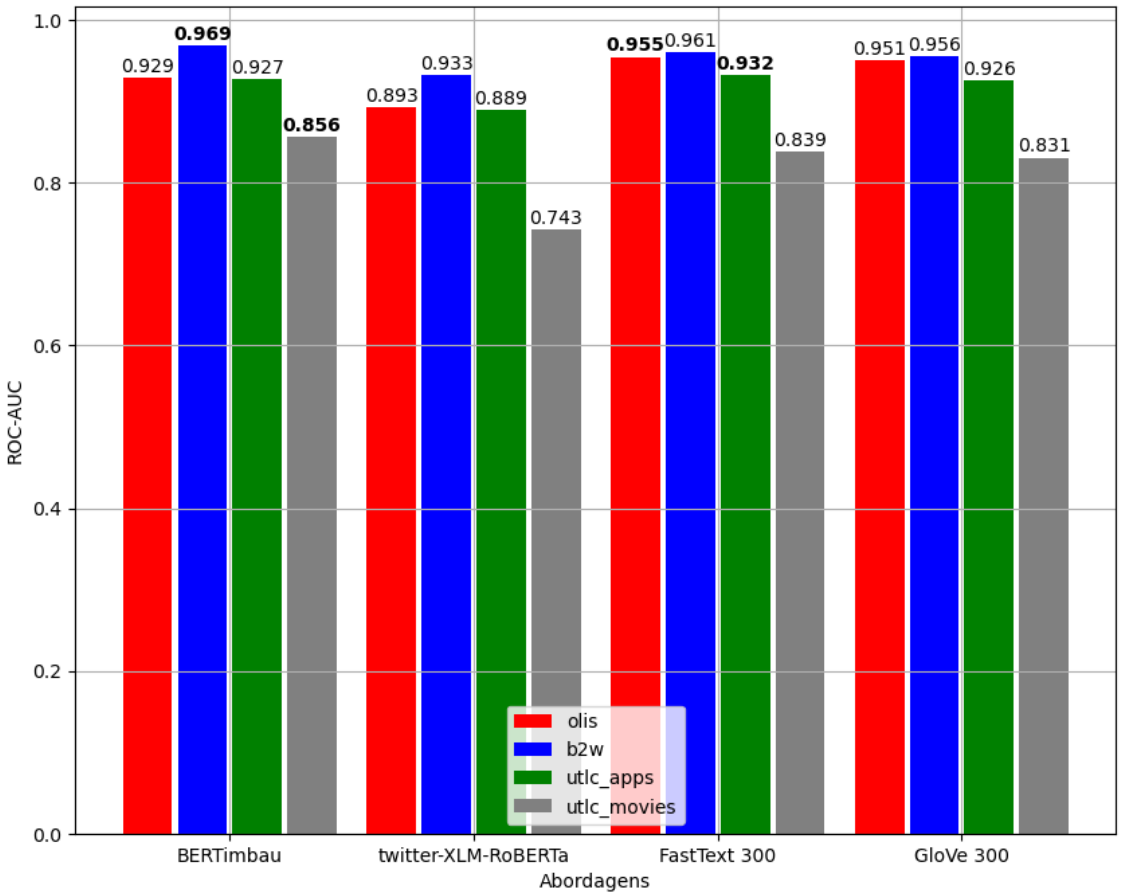

Figure 2. Gráfico do comparativo da métrica ROC-AUC entre as abordagens de classificação de texto

As Figuras 3 e 4 demonstram que os métodos propostos *BERTimbau* e *twitter-XLM-RoBERTa* foram melhores que métodos empregados no trabalho [Brum and Nunes 2017], tanto para a classificação de 2 classes como para classificação de 3 classes.

O segundo conjunto de experimentos teve como objetivo avaliar os modelos baseados no *BERT* para a tarefa de classificação de sentimentos nos *datasets* da bebida cachaça, coletadas das redes sociais. Como os datasets de cachaça são pequenos, eles foram utilizados apenas para o teste de predição, cujos resultados são mostrados a seguir. O treinamento dos modelos foi feito com cada um dos datasets de terceiros.

A Tabela 12 mostra os resultados ao predizer as 1.200 amostras rotuladas do *dataset* extraído da rede social *Facebook*, para as classes: positivo e negativo. A predição foi realizada utilizando dos modelos gerados, após o ajuste fino do modelo *BERTimbau*, e também da utilização do modelo pré-treinado *twitter-XLM-RoBERTa*.

Já a Tabela 13 mostra os resultados obtidos ao predizer as 1.000 amostras rotuladas do *dataset* extraído da rede social *Twitter*, para as classes: positivo e negativo.

Observando as Tabelas 12 e 13, a superioridade dos modelos *BERTimbau* em relação ao modelo *twitter-XLM-RoBERTa*, no que se trata do *dataset Facebook*, quando

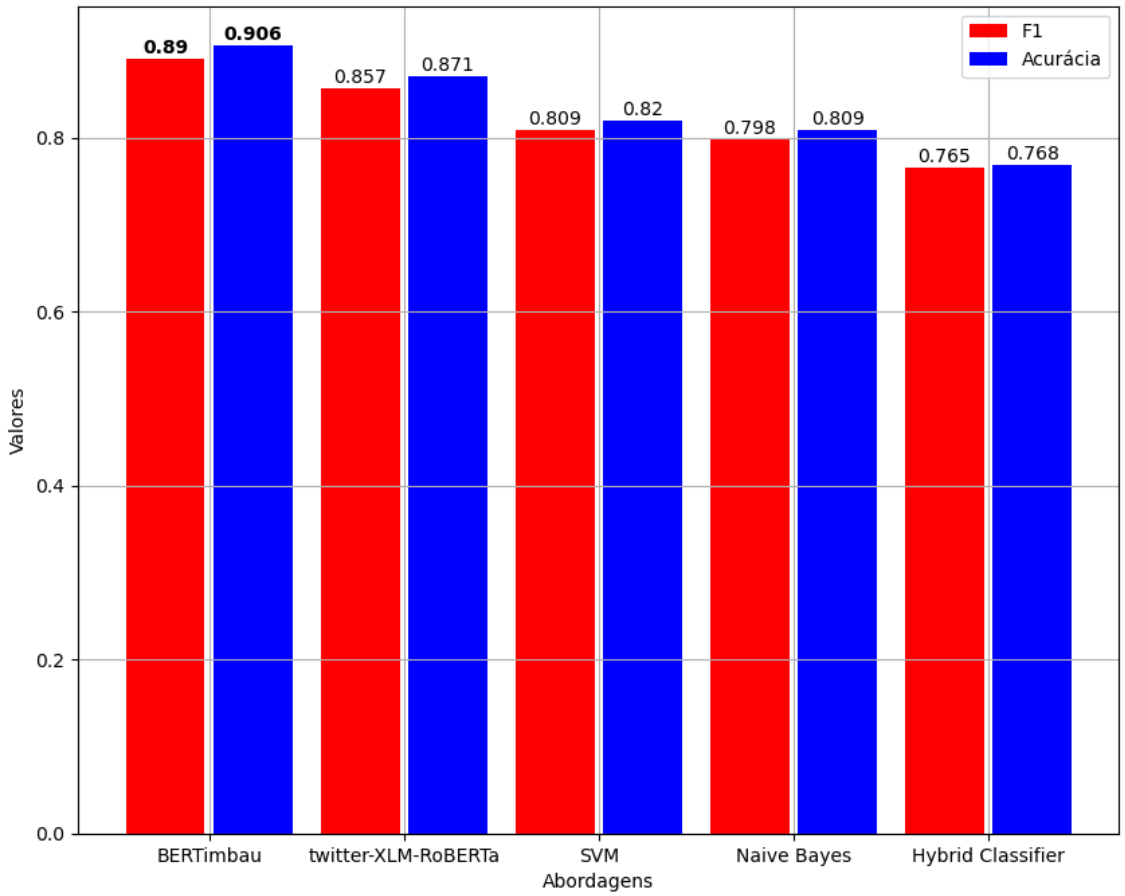

Figure 3. Gráfico do comparativo das métricas F1 e Acurácia entre as abordagens de classificação de texto para 2 classes

Modelo	Acurácia	Precisão	Revocação	F1-Score
BERTimbau - olist	0,945	0,977	0,965	0,971
BERTimbau - b2w	0,944	0,978	0,961	0,970
BERTimbau - utlc_apps	**0,947**	0,970	0,974	**0,972**
BERTimbau - utlc_movies	0,945	0,956	**0,987**	0,971
BERTimbau - twitterASBR	0,929	0,987	0,937	0,961
twitter-XLM-RoBERTa	0,866	**0,997**	0,861	0,924

Table 12. Resultados para o dataset Facebook - 2 classes

analisamos as métricas F1-Score e Acurácia. Porém, tal afirmação não pode ser dita para o *dataset Twitter*, devido ao fato que o modelo *twitter-XLM-RoBERTa* apresenta resultados de F1-Score e Acurácia superiores a alguns dos modelos *BERTimbau*, o que mostra a importância da similaridade entre os dados de treino com os dados reais. Isso fica mais evidente ao verificarmos que o modelo que obteve o melhor desempenho foi o *BERTimbau - twitterASBR*.

As Tabelas 14 e 15 mostram os resultados ao predizer os *datasets* do *Facebook* e *Twitter*, respectivamente, para as classes: positivo, negativo e neutro. Para essa classificação apenas os modelos *BERTimbau - twitterASBR* e *twitter-XLM-RoBERTa*

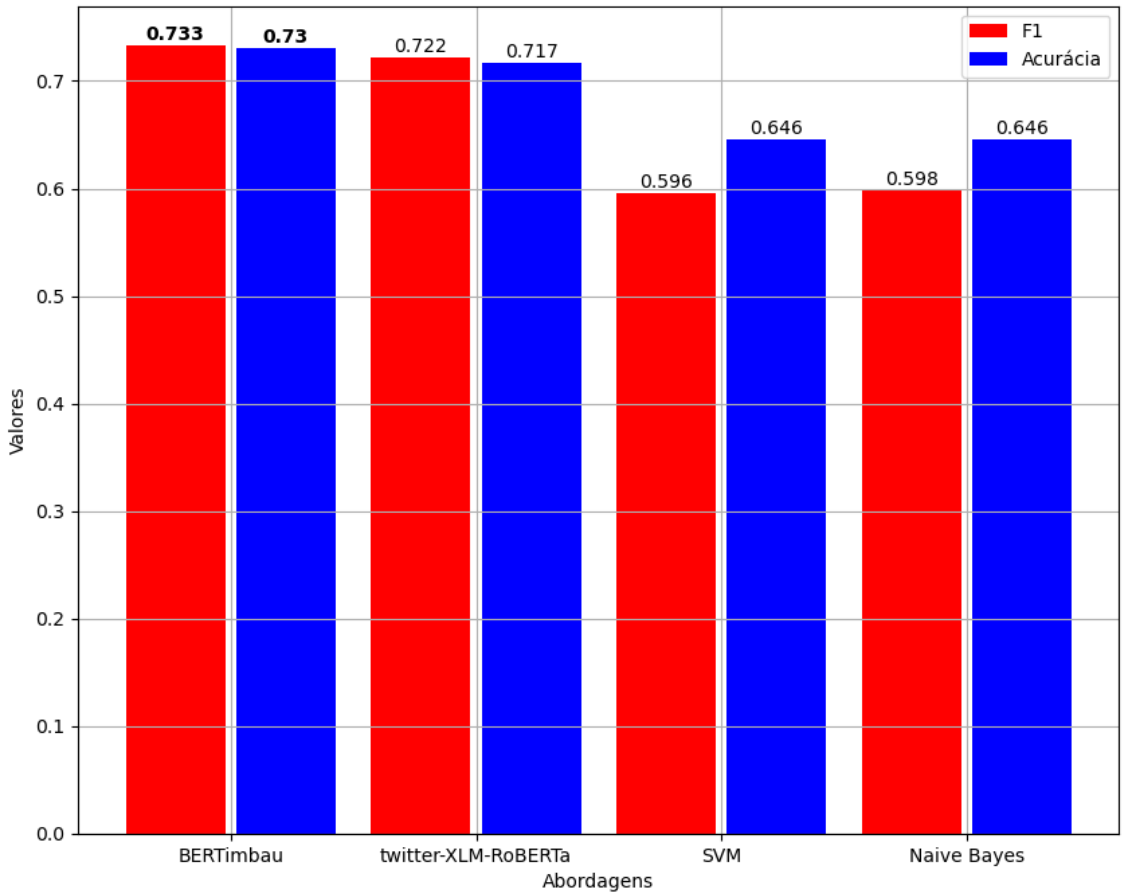

Figure 4. Gráfico do comparativo das métricas F1 e Acurácia entre as abordagens de classificação de texto para 3 classes

Modelo	Acurácia	Precisão	Revocação	F1-Score
BERTimbau - olist	0,901	0,909	0,980	0,943
BERTimbau - b2w	0,877	0,885	0,980	0,930
BERTimbau - utlc_apps	0,868	0,865	**1,0**	0,927
BERTimbau - utlc_movies	0,827	0,841	0,980	0,905
BERTimbau - twitterASBR	**0,918**	0,926	0,980	**0,952**
twitter-XLM-RoBERTa	0,877	**0,978**	0,873	0,923

Table 13. Resultados para o dataset Twitter - 2 classes

foram empregados, visto serem os únicos treinados para predizer para os três rótulos.

Modelo	Acurácia	Precisão	Revocação	F1-Score
BERTimbau - twitterASBR	**0,784**	**0,626**	0,730	**0,648**
twitter-XLM-RoBERTa	0,750	0,614	**0,767**	0,629

Table 14. Resultados para o dataset Facebook - 3 classes

O resultado apresentado na Tabela 14 demonstra que o modelo *BERTimbau - twitterASBR* desempenhou melhor, provavelmente por ter como base o modelo *BERTimbau*,

Modelo	Acurácia	Precisão	Revocação	F1-Score
BERTimbau - twitterASBR	0,755	0,487	0,638	0,522
twitter-XLM-RoBERTa	**0,780**	**0,498**	**0,742**	**0,542**

Table 15. Resultados para o dataset Twitter - 3 classes

visto que ambos os modelos não possuem similaridade com o *dataset Facebook*. Já a Tabela 15 apresenta um melhor desempenho do modelo *twitter-XLM-RoBERTa*, possivelmente devido ao fato de que o *twitter-XLM-RoBERTa* é um modelo pré-treinado em uma base de dados muito grande de *tweets*.

As Tabelas 16 e 17 apresentam os erros cometidos pelos modelos ao predizer os *datasets Facebook* e *Twitter* respectivamente, para as classes: positivo e negativo. A coluna "Erros Negativo" representa o erro do modelo ao predizer erroneamente uma instância negativa como positiva, e a "Erros Positivo" representa o erro do modelo ao predizer erroneamente uma instância positiva como negativa, e entre parênteses o percentual que o erro representa.

Modelo	Qtd Erros Negativo	Qtd Erros Positivo	Erros Totais
BERTimbau - olist	13 (41,93%)	20 (3,47%)	33 (5,43%)
BERTimbau - b2w	12 (38,70%)	22 (3,81%)	34 (5,60%)
BERTimbau - utlc_apps	17 (54,83%)	15 (2,60%)	32 (5,27%)
BERTimbau - utlc_movies	26 (83,38%)	7 (1,21%)	33 (5,43%)
BERTimbau - twitterASBR	7 (22,58%)	36 (6,25%)	43 (7,08%)
twitter-XLM-RoBERTa	1 (3,22%)	80 (13,88%)	81 (13,34%)

Table 16. Número de erros dos modelos no dataset Facebook

Modelo	Qtd Erros Negativo	Qtd Erros Positivo	Erros Totais
BERTimbau - olist	10 (52,63%)	2 (1,94%)	12 (9,83%)
BERTimbau - b2w	13 (68,42%)	2 (1,94%)	15 (12,29%)
BERTimbau - utlc_apps	16 (84,21%)	0 (0%)	16 (13,11%)
BERTimbau - utlc_movies	19 (100%)	2 (1,94%)	21 (17,21%)
BERTimbau - twitterASBR	8 (42,10%)	2 (1,94%)	10 (8,19%)
twitter-XLM-RoBERTa	2 (10,52%)	13 (12,62%)	15 (12,29%)

Table 17. Número de erros dos modelos no dataset Twitter

Analisando as Tabelas 16 e 17 pode-se observar que percentualmente a maioria dos modelos, com exceção do modelo *twitter-XLM-RoBERTa*, rotularam erroneamente as amostras negativas, tanto que o modelo *BERTimbau - utlc_movies*, rotulou erroneamente todas as amostras negativas do *dataset Twitter*. Analisando a Tabela 6, pode-se observar que houve menos amostras negativas em seu treinamento, o que pode influenciar no número de erros para os rótulos negativos.

Verificando as predições dos modelos no *dataset Facebook* houve erros em comentários, como em *"Excelente! Porém ainda estou à espera da minha do mês de Agosto. Está demorando, CN!..."*, que é uma amostra negativa, porém, cinco de seis modelos

predisseram de maneira errada essa instância, já que os modelos por serem baseados no modelo *BERT* analisam o contexto tanto a esquerda como a direita, para um melhor entendimento das relações entre as palavras presentes no texto analisado, portanto, provavelmente devido ao uso do termo ”Excelente” a relação entre as palavras obteve um cunho mais positivo, o que pode resultar em uma classificação errônea. Outro exemplo é o texto *”não está disponivel pra compra ainda??”*, que é uma amostra positiva, visto o interesse na aquisição do produto, porém todos os modelos, exceto o modelo *BERTimbau - utlc_movies*, rotularam erroneamente essa instância, possivelmente devido ao uso do termo de negação ”não”.

Já no *dataset Twitter*, foi verificado que todos os modelos rotularem erroneamente a instância *”#CACHAÇA e #PERFUME sao campeãs,em #IMPOSTOS. Cachaça,n posso confirmar.Mas #PERFUME,realmente tá bem salgado.Mas vale a pena,ser #cheirosa”*, que foi erroneamente rotulada por todos os modelos como amostras positivas, sendo uma amostra negativa, visto que trata-se de uma reclamação. Outro exemplo de instância é *”Fui tentar traduzir ”amor da minha vida”, a resposta foi CACHAÇA! https://t.co/JdMW5dIuI2.”*, que é uma amostra positiva, visto o apreço do cliente com a bebida, que foi erroneamente rotulada pelos modelos *BERTimbau - b2w* e *twitter-XLM-RoBERTa* como negativa.

Para concluir, ao observar os resultados de todas as abordagens utilizadas, que devido aos resultados apresentados pelo modelo *BERTimbau - utlc_apps* nas métricas Acurácia e F1-Score na Tabela 12, indica-se a utilização desse modelo para a predição de dados do *Facebook* para a classificação de duas classes. Já para a classificação de duas classes de dados do *Twitter*, indica-se a utilização do modelo *BERTimbau - TwitterASBR*, visto seus resultados nas métricas Acurácia e F1-Score na Tabela 13. Para as classificações de três classes de dados do *Facebook* e do *Twitter*, indica-se os modelos *BERTimbau - TwitterASBR* e *twitter-XLM-RoBERTa*, respectivamente, devido aos seus resultados nas Tabelas 14 e 15.

5. Conclusão e Trabalhos Futuros

Neste estudo, foram propostas abordagens para a tarefa de mineração de opinião, no domínio específico da bebida cachaça, utilizando modelos *BERT*, trazendo uma base de dados própria extraída de redes sociais. Porém, devido ao alto custo de rotular toda a base de dados, apenas uma amostra foi rotulada. Assim, foram utilizadas outras bases de dados de terceiros, que possuem similaridade com as bases de dados alvo, para realização do ajuste fino do modelo *BERTimbau* para a tarefa de mineração de opinião. Desta maneira, utilizou-se duas abordagens com as amostras rotuladas, uma com o modelo *BERTimbau* ajustado e outra com o modelo pronto *twitter-XLM-RoBERTa*.

Observando os resultados na Seção 4.4, pode-se concluir que as abordagens utilizadas neste trabalho obtiveram resultados consideráveis. Para o *dataset Facebook*, o melhor resultado foi de 97,2% na métrica F1-Score para o modelo *BERTimbau - utlc_apps*, já para o *dataset Twitter* o resultado foi de 95,2% com o modelo *BERTimbau - twitterASBR*. Portanto, é plausível afirmar que as abordagens utilizadas possuem potencial para a tarefa de mineração de sentimento no domínio específico da bebida cachaça.

O trabalho apresenta alguns pontos fracos, como o uso de uma pequena base dados rotulada para tarefa de mineração de opinião no domínio específico da bebida cachaça,

além de que a base foi rotulada por apenas uma pessoa, possuindo desta maneira um viés pessoal do rotulador, e também a quantidade de rótulos neutros presentes na rotulação amostral do *dataset Twitter*, visto que a remoção das amostra neutras, resultou em apenas 122 instâncias para a realização do experimento de classificação binária do *dataset Twitter*.

Como trabalhos futuros, propõe-se a avaliação de diferentes modelos, sejam eles baseados ou não no modelo *BERT*, a criação de uma equipe para a verificação e a finalização da rotulação de toda a base de dados coletada das redes sociais, e a realização dos mesmos experimentos, porém com uma infraestrutura de hardware superior, com memória e capacidade de processamento maior para processar bases maiores.

6. Agradecimentos

Gostaria de agradecer a FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) por financiarem a pesquisa. Agradecer ao professor Ph.D. Denilson Alves Perreira, por me orientar, agradecer também aos integrantes do LabRI, em especial o Arthur Franco e a Priscilla Silva, e também agradecer ao Me. Henrico Bertini Brum.

References

Alcarde, A. (2018). *Cachaça: ciência, tecnologia e arte*. Editora Blucher, 1st online edition. Disponível em: https://books.google.com.br/books?id=4StdDwAAQBAJ.

Araújo, M., Gonçalves, P., and Benevenuto, F. (2013). Measuring sentiments in online social networks. In *WebMedia '13: 19th Brazilian Symposium on Multimedia and the Web*, pages 97–105, New York, NY, USA. Association for Computing Machinery.

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Brandão, M. (2022). Mercado exportador de cachaça bate recorde em 2022. agênciaBrasil, https://agenciabrasil.ebc.com.br/economia/noticia/2022-12/mercado-exportador-de-cachaca-bate-recorde-em-2022. Acesso em 12 de Dezembro de 2023.

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv*. DOI: 10.48550/ARXIV.1712.08917.

Câmara, M. (2018). *Cachaça: Prazer Brasileiro*. Mauad Editora Ltda, 2nd edition. Disponível em: https://books.google.com.br/books?id=NYdiDwAAQBAJ.

Cardozo, L. and Freitas, L. (2021). Análise de sentimentos: Avaliando o desempenho de pré-processamento e de algoritmos de aprendizagem de máquina sobre o dataset tweetsentbr. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 169–174, Porto Alegre, RS, Brazil. SBC. DOI: 10.5753/brasnam.2021.16135.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J.,

editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dias, N. C. (2014). A cachaça é nossa: cultura e ideologia na construção da identidade nacional. In *Anais Brasileiros de Estudos Turísticos*, volume 4, pages 35—-44, Juiz de Foras, MG, Brazil. Disponível em: https://periodicos.ufjf.br/index.php/abet/article/view/3029.

Evangelista, T. R. and Padilha, T. P. P. (2014). Monitoramento de posts sobre empresas de e-commerce em redes sociais utilizando análise de sentimentos. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 152–163, Porto Alegre, RS, Brazil. SBC.

Gomes, F. P., Silva, E. M., Teixeira, I., and Brito, P. F. (2017). Análise de sentimentos: Uma revisão sistemática. In *XIX Encoinfo – Congresso de Computação e Tecnologias da Informação*, volume 19, pages 24–32, Palmas, TO, Brazil. Disponível em: https://ulbra-to.br/encoinfo/edicoes/2017/anais/.

Hartmann, N. S., Avanço, L. V., Balage, P. P., Duran, M. S., Nunes, M. d. G. V., Pardo, T. A. S., and Aluisio, S. M. (2014). A large corpus of product reviews in portuguese: tackling out-of-vocabulary words. In *International Conference on Language Resources and Evaluation - LREC*. Paris: ELRA. Disponível em: http://www.lrec-conf.org/proceedings/lrec2014/pdf/413_Paper.pdf.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*, pages 1–167. Springer Cham, Chicago, IL. Disponível em: https://doi.org/10.1007/978-3-031-02145-9.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Martins, M. d. F., Silva, N. C., Ouriques, R. A. d. B., and Cândido, G. A. (2018). Gestão e tecnologia em engenhos produtores de cachaça no brejo da paraíba-brazil. In *Revista Gestão Industrial*, volume 14, pages 209–230, Ponta Grossa, PR, Brazil. UTFPR.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Disponível em: https://github.com/doccano/doccano. Acesso em 10 de Novembro de 2023.

Olist and Sionek, A. (2018). Brazilian e-commerce public dataset by olist. Disponível em: https://www.kaggle.com/dsv/195341. Acesso em 10 de Novembro de 2023.

Opinando, P. (2020). Opinion mining for portuguese: Concept-based approaches and beyond. Universidade de São Paulo - USP, Disponível em: https://sites.google.com/icmc.usp.br/opinando/. Acesso em 10 de Novembro de 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(null):2825–2830. DOI: 10.5555/1953048.2078195.

Perreira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54:1087–1115. DOI: 10.1007/s10462-020-09870-1.

Real, L., Oshiro, M., and Mafra, A. (2021). B2w-reviews01: an open product reviews corpus. B2W Digital, Disponível em: https://github.com/americanas-tech/b2w-reviews01. Acesso em 10 de Novembro de 2023.

Reips, L. and Hara, C. (2022). Integração e rotulação automatizada de dados sobre o cnidário physalia physalis, usando a geolocalização como referência. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 105–111, Porto Alegre, RS, Brazil. SBC. DOI: 10.5753/sbbd_estendido.2022.21851.

Rodrigues, B. U., da Costa, R. M., and da Silva Soares, A. (2015). Reconhecimento do tipo de cachaça utilizando visão computacional e reconhecimento de padrões. Goiânia, GO, Brazil. Universidade Federal de Goiás.

Santos, A., Becker, K., and Moreira, V. (2014). Um estudo de caso de mineração de emoções em textos multilíngues. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 140–151, Porto Alegre, RS, Brazil. SBC.

SEBRAE (2015). Saiba mais sobre tendência do mercado de cachaça. SEBRAE, hhttps://sebrae.com.br/sites/PortalSebrae/artigos/saiba-mais-sobre-tendencia-do-mercado-de-cachaca,39aa6a2bd9ded410VgnVCM1000003b74010aRCRD. Acesso em 12 de Dezembro de 2023.

SEBRAE (2022). Produção de cachaça no brasil ainda tem muito potencial econômico. SEBRAE, https://sebrae.com.br/sites/PortalSebrae/artigos/producao-de-cachaca-no-brasil-ainda-tem-muito-potencial-economico,578ed967936ef710VgnVCM100000d701210aRCRD. Acesso em 10 de Novembro de 2023.

Silva, P., Franco, A., Santos, T., Brito, M., and Pereira, D. (2023). Cachacaner: a dataset for named entity recognition in texts about the cachaça beverage. *Language Resources and Evaluation*. DOI: 10.1007/s10579-023-09665-0.

Sousa, R. F. d., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Symposium in Information and Human Language Technology - STIL*. Porto Alegre: SBC. Disponível em: http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings-stil-2019-Final.pdf.

Souza, F. and Filho, J. (2021). Sentiment analysis on brazilian portuguese user reviews. *arXiv*. DOI: 10.48550/ARXIV.2112.05459.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.