



Universidade do Minho
Licenciatura em Ciências da Computação

Unidade Curricular Projeto

Ano Letivo de 2023/2024

Análise de Sentimentos a Comentários do TripAdvisor

Catarina Gonçalves (a101904)

Diogo Costa (a78133)

João Mirra (a100083)

Coordenador: Orlando Belo

maio, 2024

Data de Receção	
Responsável	
Avaliação	
Observações	

Análise Sentimentos a Comentários do TripAdvisor

**Catarina Gonçalves (a101904), Diogo Costa
(a78133), João Mirra (a100083)**

Maio, 2024

Dedicatória

Gostaríamos de expressar a enorme união e resiliência demonstrada por este grupo de trabalho. Em momentos de adversidade, mantivemo-nos unidos e determinados a alcançar os nossos objetivos. Um brinde à amizade e solidariedade que nos impulsionaram neste projeto!

“A verdadeira sabedoria está em reconhecer a própria ignorância.”

Sócrates

Resumo

O nosso objetivo neste estudo é realizar uma análise dos sentimentos presentes num conjunto específico de dados do TripAdvisor, com o propósito de compreender e aprimorar o nosso conhecimento acerca do processo de automatização dessa análise. Diante dos desafios inerentes a esta temática, como a complexidade do tema proposto e a escassez de ferramentas adequadas em língua portuguesa, o grupo de trabalho empenhou-se em abordar tais questões de forma eficaz e abrangente.

O principal intuito deste trabalho foi proporcionar um aprofundamento de conhecimento neste campo de estudo. A metodologia adotada baseou-se essencialmente na linguagem de programação Python e suas bibliotecas, o que permitiu a execução da tarefa de forma eficiente e precisa.

Contudo, é imprescindível ressaltar que o projeto não foi completamente desenvolvido, dada a complexidade desta área, que oferece múltiplos caminhos de investigação e evolução. Assim, mesmo que o trabalho não tenha sido realizado na sua plenitude, as bases foram estabelecidas para que futuramente sejam aprofundados, tendo fornecido um importante conhecimento nesta área de investigação.

Área de Aplicação: Aplicação de Análise de Sentimentos em comentários do TripAdvisor

Palavras-Chave: Análise de Sentimento, TripAdvisor, Python, Processamento de Linguagem Natural, Machine Learning

Índice

Capítulo I.....	1
1. Introdução	1
1.1. Contexto.....	2
1.2. Motivação.....	3
1.2.1. Aprendizagem	3
1.2.2. Relevância Prática e Profissional.....	3
1.2.3. Desafios e Oportunidades.....	4
1.3. Objetivos	5
1.3.1. Compreender a Metodologia de Análise de Sentimentos	5
1.3.2. Explorar os Desafios na Análise de Sentimentos	5
1.4. Estrutura do documento	6
Capítulo II.....	7
2. Análise de sentimentos	7
2.1. O que é a análise de sentimentos?.....	7
2.2. Por que a análise de sentimentos é importante?	7
2.3. Como funciona a análise de sentimentos?.....	8
2.3.1. Pré-processamento	8
2.3.2. Análise das palavras-chave.....	8
2.4. Quais são as abordagens para a análise de sentimentos?	9
2.4.1. Abordagem Baseada em Regras	9
2.4.2. <i>Machine Learning</i>	10
2.4.2.1. Métodos de <i>Machine Learning</i>	11
2.4.3. Híbrida	12
2.5. Quais são os diversos tipos de análise de sentimento?	12
2.6. Quais são os desafios na análise de sentimentos?	13
2.7. Alguns Trabalhos Realizados.....	15
Capítulo III.....	16

3.	O Processo de Análise de Sentimento	16
Capítulo IV		18
4.	Caso de Estudo – Análise de Sentimentos em Comentários do TripAdvisor.....	18
4.1.	Recolha dos Dados.....	18
4.2.	Processamento dos Dados	19
4.3.	Deteção de Sentimento	22
4.4.	Apresentação de resultados.....	25
4.5.	Machine Learning	26
Capítulo V.....		29
5.	Conclusões.....	29
6.	Referências	30

Índice de Figuras

Figura 1- Abordagens da Análise de Sentimento.....	14
Figura 2 - Processo da Análise de Sentimento	17
Figura 3 - Resultados da Extração de Comentários.....	18
Figura 4 - Função Regex_Remove_Punctuation.....	20
Figura 5 - Função Normalize.....	20
Figura 6 - Tratamento de Emojis	21
Figura 7 - Função _punctuation_emphasis.....	21
Figura 8 - Tokenização e Normalização de Palavras	22
Figura 9 - Função score_valence	23
Figura 10 - Primeiros Resultados do Vader.....	24
Figura 11 - Função fix_sentence.....	24
Figura 12 - Comparação de Resultados no Vader.....	25
Figura 13 - Polaridade dos Comentários Extraídos	26
Figura 14 - Resultados da Máquina de Vetores de Suporte (SVM).....	27
Figura 15 - Resultados do Naive Bayes Multinomial.....	27
Figura 16 - Resultados da Regressão Logística.....	28
Figura 17 - Comparação entre algoritmos de Machine Learning.....	28

Capítulo I

1. Introdução

Este relatório apresenta uma análise sobre a aplicação de técnicas de Processamento de Linguagem Natural (PLN) na análise de sentimentos dos comentários do TripAdvisor na língua portuguesa. Ao longo das próximas seções, exploraremos o contexto do PLN, as motivações por trás deste estudo e os objetivos que pretendemos alcançar, bem como o processo por detrás desta análise de sentimentos.

No contexto do PLN, enfrentamos desafios significativos devido às ambiguidades e complexidades da linguagem humana, especialmente num idioma tão diverso como é o português. Desde a sua origem nas décadas de 1940 nos Estados Unidos, o campo do PLN evoluiu consideravelmente, impulsionado pela crescente procura de soluções que permitam às máquinas compreender e processar a linguagem humana de forma mais eficaz.

As motivações para este estudo são diversas e abrangentes. Primeiramente, o projeto representa uma oportunidade valiosa de aprendizagem e aplicação de conhecimentos numa área de ampla evolução. Ao enfrentar desafios reais de análise de sentimentos nos comentários do TripAdvisor, estamos não apenas a expandir os nossos conhecimentos em PLN, Programação e Análise de Dados, mas também a contribuir para o avanço das capacidades de análise de dados na língua portuguesa.

Além disso, reconhecemos a relevância prática e profissional deste estudo. A capacidade de analisar sentimentos de forma automatizada em grandes volumes de dados é uma ferramenta fundamental para empresas e organizações, influenciando diretamente as estratégias de negócios e a experiência do cliente. Ao adquirir competências nessa área, os participantes deste projeto estão também a preparar-se para enfrentar desafios futuros e aproveitar oportunidades no mercado de trabalho.

Os objetivos deste relatório são apresentar os resultados da análise de sentimentos realizada nos comentários do TripAdvisor na língua portuguesa, identificar padrões e tendências nos dados analisados e propor recomendações com base nas conclusões alcançadas. Compreender a metodologia utilizada e os desafios enfrentados permitirá

que avancemos no desenvolvimento de soluções mais eficazes para a análise de sentimentos em textos portugueses.

1.1. Contexto

O Processamento de Linguagem Natural (PLN) é uma área da Ciência da Computação dedicada a capacitar máquinas para entender e processar a linguagem humana.

Inicialmente, procurava-se a criação de sistemas de tradução automatizada, o PLN evoluiu consideravelmente. No entanto, os desafios decorrentes das ambiguidades e complexidades da linguagem humana têm sido uma constante ao longo do tempo.

Apesar dos desafios, o campo da PLN tem mostrado um crescimento significativo, impulsionado pela necessidade crescente de soluções para variedade de tarefas relacionadas ao processamento da linguagem. O surgimento de bibliotecas e ferramentas específicas para diferentes idiomas tem proporcionado avanços consideráveis nesse campo.

No entanto, quando se trata da língua portuguesa, que apresenta uma diversidade linguística significativa, a disponibilidade de recursos de PLN ainda é limitada. Este documento propõe-se a explorar e descrever os desafios enfrentados na aplicação da análise de sentimentos nos comentários do TripAdvisor em língua portuguesa e as soluções desenvolvidas para enfrentá-los.

1.2. Motivação

Nas próximas subsecções, apresentaremos algumas motivações que justifiquem o caso de estudo enfrentado pelo grupo de trabalho.

1.2.1. Aprendizagem

Este projeto representa uma oportunidade valiosa de adquirir e aplicar conhecimentos numa área em ampla expansão. Ao enfrentar desafios reais como analisar sentimentos nos comentários do TripAdvisor, o grupo de trabalho enfrentou uma oportunidade de expandir e consolidar os conhecimentos em Processamento de Linguagem Natural (PLN), Programação e Análise de Dados.

Ao longo do processo de aprendizagem numa área tão complexa, foi-nos oferecida a possibilidade de explorar diversas abordagens, testar hipóteses até encontrar uma solução adequada, o que contribuiu imenso para um processo de aprendizagem dinâmico e abrangente devido à natureza multidisciplinar do projeto que combina conceitos da língua portuguesa, Ciência da Computação e Análise de Dados.

Portanto, a aprendizagem é uma motivação essencial para o projeto, por oferecer ao grupo de trabalho uma oportunidade única de adquirir conhecimento numa área dinâmica e multidisciplinar.

1.2.2. Relevância Prática e Profissional

Este projeto assume uma elevada relevância prática e profissional devido à sua importância crucial na análise de sentimentos para empresas e organizações. Ao adquirirem conhecimento nesta área, os membros do grupo estão a preparar-se para aproveitar oportunidades futuras nas suas carreiras profissionais. Esta iniciativa oferece uma oportunidade valiosa para o desenvolvimento de conhecimentos relevantes, altamente valorizadas no mercado de trabalho.

A capacidade de analisar sentimentos de forma automatizada em grandes volumes de dados é uma ferramenta essencial para empresas e organizações. Esta capacidade permite obter informações valiosas sobre as opiniões dos clientes, fundamentais para tomar decisões estratégicas e aprimorar a experiência do cliente.

Ao participarem neste projeto, os membros do grupo estão a recolher bases importantes, além de obter uma compreensão aprofundada da área de Processamento de Linguagem

Natural. Estes fundamentos adquiridos poderão ser extremamente úteis nas suas futuras carreiras, permitindo-lhes contribuir de forma significativa para o sucesso das organizações em que trabalharem.

1.2.3. Desafios e Oportunidades

Este projeto envolve a exploração de desafios e oportunidades únicos apresentados pela análise de sentimentos em comentários do TripAdvisor. Um dos principais desafios reside na diversidade linguística dos comentários, que pode complicar a análise devido às variações na linguagem utilizada pelos diferentes utilizadores do TripAdvisor, tornando a interpretação dos sentimentos uma tarefa complexa e desafiadora.

No entanto, é importante destacar que esses desafios não são apenas obstáculos, mas também oportunidades para explorar soluções. Ao enfrentar a diversidade linguística dos comentários, os participantes do projeto têm a oportunidade de experimentar e avaliar diferentes ferramentas e abordagens na tentativa de encontrar a melhor solução para o problema. Isso pode envolver o uso de algoritmos de processamento de linguagem natural adaptados para lidar com diferentes idiomas, procurando o que melhor se adapta à língua portuguesa.

Em suma, os desafios encontrados na análise de sentimentos nos comentários do TripAdvisor representam oportunidades valiosas para o desenvolvimento de competências e a busca por soluções que melhorem a precisão e eficácia da análise sendo que essa experiência enriquece o conhecimento dos elementos participantes no projeto.

1.3. Objetivos

Nas subsecções seguintes, vamos apresentar os objetivos que o grupo de trabalho para este projeto.

1.3.1. Compreender a Metodologia de Análise de Sentimentos

Compreender a metodologia de análise de sentimentos nos comentários do TripAdvisor envolve entender os métodos e técnicas utilizadas para identificar e classificar as emoções que são expressas em textos. Geralmente, isso inclui o uso de algoritmos de processamento de linguagem natural, que ajudam a extrair as informações relevantes que estão presentes nos comentários e as técnicas de aprendizado de máquina, que aprimoram a precisão da análise. No contexto do TripAdvisor, é importante considerar a diversidade linguística dos comentários, o que vai exigir técnicas adicionais de pré-processamento de texto para melhorar a precisão da análise. Esta compreensão é fundamental para avaliar a eficácia da análise de sentimento e identificar possíveis melhorias.

1.3.2. Explorar os Desafios na Análise de Sentimentos

Para avançar no projeto de análise de sentimentos nos comentários do TripAdvisor, é essencial identificar e compreender os desafios enfrentados durante o processo. Estes desafios incluem lidar com a diversidade linguística presente nos comentários em língua portuguesa, o que pode complicar a interpretação devido às variações na linguagem utilizada pelos seus utilizadores. Além disso, a ambiguidade associada à língua portuguesa pode dificultar a identificação precisa das emoções expressas nos comentários. Ao explorarmos esses desafios de forma aprofundada, estaremos preparados para desenvolver estratégias eficazes e soluções que melhorem a precisão e eficácia da análise de sentimentos no contexto do TripAdvisor.

1.4. Estrutura do documento

Este documento é composto por 5 capítulos principais, estruturados da seguinte maneira:

Capítulo 1. Introdução: Introduz o problema que nos propomos abordar e descreve o contexto e motivação que definem este projeto.

Capítulo 2. Análise de Sentimentos: Abrange os conceitos centrais da PNL que são relevantes para melhor se compreender a metodologia seguida.

Capítulo 3. Processo de Análise de Sentimentos: Descreve um processo (e as suas várias etapas) de como se desenvolve uma aplicação de análise de sentimentos.

Capítulo 4. Caso de Estudo: Descreve o trabalho realizado no desenvolvimento da aplicação de análise de sentimentos a comentários do TripAdvisor e cada uma das suas etapas desenvolvidas, bem como as ferramentas utilizadas.

Capítulo 5. Conclusão: Conclui este documento e apresenta os resultados adquiridos pelo grupo de trabalho.

Capítulo II

2. Análise de sentimentos

Nesta secção, vamos fornecer uma visão geral acerca da análise de sentimentos, como por exemplo, por que razão a análise de sentimentos é importante.

2.1. O que é a análise de sentimentos?

A análise de sentimentos é o processo de avaliar as emoções presente em textos, determinando se é positivo, negativo ou neutro. Por meio de ferramentas especializadas, empresas podem automatizar essa análise em grandes volumes de dados textuais, como e-mails, chats de atendimento ao cliente e avaliações em redes sociais. Isso permite compreender a opinião dos clientes e adaptar estratégias de negócio de forma eficaz.

2.2. Por que a análise de sentimentos é importante?

A análise de sentimentos, também conhecida como mineração de opiniões, é uma ferramenta crucial para ajudar as empresas a aprimorar os seus produtos e serviços. Aqui estão alguns benefícios dessa prática:

- Fornecer informações objetivas: Ferramentas de análise de sentimentos ajudam as empresas a evitar tendências pessoais na avaliação de comentários, garantindo resultados consistentes e imparciais sobre as opiniões dos clientes.
- Criar melhores produtos e serviços: Com base no feedback genuíno e específico dos clientes, a análise de sentimentos fornece uma orientação valiosa para melhorar produtos e serviços, focando em áreas identificadas como negativas.
- Análise em alta escala: Ferramentas baseadas na nuvem permitem que as empresas processem grandes volumes de dados não estruturados, como e-mails e comentários sobre produtos, revelando as emoções presentes nesses textos de forma acessível.
- Resultados em Tempo Real: Em um mercado dinâmico, é essencial que as empresas possam responder rapidamente às tendências. Com a análise de sentimentos, os profissionais de marketing podem entender os sentimentos dos clientes sobre a marca, produtos e serviços em tempo real, permitindo ações imediatas com base nessas análises.

2.3. Como funciona a análise de sentimentos?

A análise de sentimentos é uma aplicação de tecnologias de Processamento de Linguagem Natural (PLN) que visa ensinar máquinas a compreender texto de maneira semelhante aos humanos. Este processo geralmente passa por várias fases antes de gerar resultados significativos.

2.3.1. Pré-processamento

Durante a fase de pré-processamento, a análise de sentimentos identifica palavras-chave que possam destacar a mensagem principal contida no texto. Isso envolve:

- Tokenização: Dividir uma frase em elementos individuais (*tokens*), permitindo que cada palavra seja tratada separadamente para processamento.
- Lematização: Converter cada palavra para sua forma básica, por exemplo, a palavra "sou" será convertida para "ser", garantindo que diferentes formas da mesma palavra sejam tratadas da mesma maneira.
- Remoção de stopwords: Eliminar palavras que não contribuem para o significado da frase, como artigos, preposições e conjunções.

Essas etapas de pré-processamento ajudam a simplificar o texto e a prepará-lo para análise de sentimentos mais detalhada.

2.3.2. Análise das palavras-chave

As tecnologias de Processamento de Linguagem Natural (PLN) analisam as palavras-chave que foram extraídas e atribuem a cada uma pontuação de sentimento. Essa pontuação é uma escala que mede o sentimento da palavra no sistema de análise de sentimentos, oferecendo uma visão relativa da emoção expressa no texto para fins analíticos. Por exemplo, num contexto de análise a comentários de clientes sobre um produto, a palavra 'ótimo' deverá ter uma pontuação mais alta quando comparado com a palavra 'mau' que deverá ter uma pontuação mais baixa. Estas pontuações proporcionam uma visão relativa da emoção expressa no texto para efeitos de análise.

2.4. Quais são as abordagens para a análise de sentimentos?

Existem 3 abordagens principais usadas por software de análise de sentimentos que serão apresentadas a seguir, tal como os seus prós e contras.

2.4.1. Abordagem Baseada em Regras

Nesta abordagem, o software identifica, classifica e atribui pontuações a palavras-chave específicas com base em léxicos predeterminados. Léxicos são compilações de palavras que representam a polaridade de cada termo num contexto específico. Para determinar se uma frase é positiva, negativa ou neutra, o software procura as palavras contidas no léxico e calcula uma pontuação de sentimento. A pontuação final é então comparada com limites pré-definidos para determinar a carga emocional geral da frase.

Prós e Contras:

- **Prós:** Um sistema de análise de sentimentos baseado em regras é de fácil configuração e compreensão. Ele permite uma abordagem transparente na atribuição de sentimentos às palavras-chave.
- **Contras:** No entanto, é difícil escalá-lo, pois exigiria a expansão contínua dos léxicos para melhorar os resultados. Além disso, essa abordagem pode ser menos precisa ao lidar com frases influenciadas por diferentes culturas, uma vez que os léxicos podem não abranger todas as nuances de linguagem.

2.4.2. *Machine Learning*

Nesta abordagem, são utilizadas técnicas de Machine Learning e algoritmos de classificação de sentimentos, como redes neurais e aprendizado profundo, para ensinar o software a identificar os sentimentos emocionais em um texto. Este processo envolve a criação de um modelo de análise de sentimentos e o treino contínuo desse modelo com dados conhecidos, a fim de permitir que ele faça previsões precisas sobre dados desconhecidos.

Treino

Durante o treino, é utilizado um conjunto de dados de análise de sentimentos contendo diversos exemplos. O software de *Machine Learning* utiliza esses conjuntos de dados como entrada e é treinado para produzir resultados coerentes. Ao ser exposto a uma grande variedade de exemplos, o software aprende a reconhecer como diferentes combinações de palavras que afetam a pontuação final do sentimento.

Prós e Contras:

- **Prós:** A análise de sentimentos por meio de Machine Learning é vantajosa porque processa uma ampla variedade de informações textuais com precisão. Com um treino adequado, o software pode prever com precisão os sentimentos expressos nas mensagens.
- **Contras:** No entanto, um modelo de Machine Learning treinado é específico para uma determinada área, o que significa que um software treinado para análise de sentimentos em marketing, por exemplo, pode não ser adequado para identificar sentimentos em redes sociais sem ajustes no treino.

Alguns dos algoritmos de classificação normalmente utilizados na análise de sentimentos incluem:

Regressão Linear: Um algoritmo estatístico que descreve um valor com base num conjunto de características. Na análise de sentimentos, pode ser usado para prever a polaridade do sentimento com base em características específicas do texto.

Naive Bayes: Um algoritmo que utiliza o teorema de Bayes para categorizar palavras num bloco de texto. Ele assume independência entre as características, o que o torna eficiente e rápido para tarefas de classificação, como análise de sentimentos.

Máquinas de Vetores de Suporte (SVM): Um algoritmo de classificação rápido e eficiente usado para resolver problemas de classificação de dois grupos. SVMs são frequentemente utilizados na análise de sentimentos devido à sua capacidade de lidar com dados complexos e não-lineares.

Deep Learning: Também conhecido como rede neural artificial, o *Deep Learning* é uma técnica avançada de *Machine Learning* que utiliza múltiplos algoritmos para imitar o funcionamento do cérebro humano. Na análise de sentimentos, redes neurais profundas podem aprender representações complexas dos dados textuais, captando nuances e padrões subtis de sentimentos.

2.4.2.1. Métodos de *Machine Learning*

Na secção que se segue, iremos explorar as diversas abordagens do Machine Learning: Supervisionado e Não Supervisionado.

No método supervisionado, os algoritmos são treinados utilizando conjuntos de dados já rotulados. Em cada exemplo, o algoritmo é apresentado com dados acompanhados de rótulos, e espera-se que seja capaz de aprender a partir desses dados para depois se produzir respostas para dados não rotulados. Durante o processo de treino, o algoritmo ajusta os seus parâmetros para possibilitar a realização de previsões sobre novos dados. Este método é frequentemente utilizado em tarefas como previsão de vendas ou classificação de e-mails como spam ou não spam.

Por outro lado, na abordagem não supervisionada, os algoritmos são treinados sem a utilização de dados rotulados. O objetivo é explorar a estrutura subjacente dos dados, identificando padrões e estruturas intrínsecas. Os algoritmos não supervisionados podem ser particularmente úteis para descobrir informações valiosas em grandes conjuntos de dados.

Estas abordagens possuem vantagens distintas e são aplicáveis em contextos diversos, dependendo das características dos dados e dos objetivos da análise.

2.4.3. Híbrida

A abordagem híbrida para análise de sentimentos combina recursos baseados em regras e em *Machine Learning* para otimizar a precisão e a velocidade do processo. Ao integrar elementos das duas técnicas, essa abordagem procura obter os benefícios de ambas, aproveitando a compreensão profunda das regras e a capacidade de aprendizagem e adaptação do *Machine Learning*.

Embora seja altamente precisa, essa abordagem requer mais recursos, como tempo e capacidade técnica, em comparação com as abordagens baseadas apenas em regras ou *Machine Learning*. No entanto, os resultados podem ser mais robustos e adaptáveis a diferentes contextos e tipos de dados, o que pode ser especialmente vantajoso em cenários complexos e em constante mudança.

2.5. Quais são os diversos tipos de análise de sentimento?

Existem vários tipos de análise de sentimento que as organizações utilizam, dependendo das suas necessidades específicas. Os três tipos mais comuns são a análise de sentimento baseada em emoções, a análise de sentimento detalhada e a análise de sentimento baseada em aspetos (ABSA). Todos eles dependem da capacidade do software de avaliar a polaridade, ou seja, o sentimento geral transmitido pelo texto.

A polaridade pode ser descrita de forma geral como positiva, negativa ou neutra, mas o texto pode ser ainda mais categorizado em subgrupos, como "extremamente positivo" ou "extremamente negativo". Alguns modelos de análise de sentimento podem identificar emoções mais específicas. A polaridade de um texto é a métrica normalmente utilizada para medir a emoção contida num determinado texto e é expressa pelo software como uma classificação numérica, por exemplo, numa escala de 1 a 100.

Os três tipos de análise de sentimentos mais utilizados são:

- Detalhada (em níveis): A análise de sentimento detalhada ou em níveis agrupa o texto em diversas emoções e no nível de emoção que está a ser expresso. A emoção é então classificada numa escala de 0 a 100, semelhante à forma como os sites de consumidores utilizam classificações por estrelas para medir a satisfação dos clientes.
- Baseada em Aspeto (ABSA): A análise de sentimento baseada em aspeto restringe o que está a ser examinado num texto a um único aspeto de um

produto, serviço ou experiência do cliente que uma empresa deseja analisar. Por exemplo, uma app de viagens pode utilizar o ABSA para avaliar a eficácia de um *chatbot* de atendimento ao cliente. O ABSA pode ajudar empresas a entender melhor como seus produtos estão a ser recebidos pelos clientes.

- Deteção Emocional: A análise de sentimento de deteção emocional procura entender o estado psicológico do indivíduo por trás de um texto, incluindo a sua estrutura de espírito quando estava a escrever as suas intenções. É mais complexa do que a análise detalhada ou o ABSA e normalmente é utilizada para capturar uma compreensão mais profunda da motivação ou do estado emocional de uma pessoa. Em vez de utilizar polaridades, como a deteção emocional positiva, negativa ou neutra, é capaz de identificar emoções específicas num texto, como frustração, indiferença, inquietação e choque.

2.6. Quais são os desafios na análise de sentimentos?

Os desafios na análise de sentimentos persistem, apesar dos avanços nas tecnologias de processamento de linguagem natural (PLN). Máquinas ainda enfrentam dificuldades em entender nuances da comunicação humana, como:

- Sarcasmo: Identificar corretamente ironia e sarcasmo num texto é um desafio para os algoritmos de análise de sentimentos. O tom de voz e expressão facial, que são indicadores importantes na comunicação humana, não estão presentes no texto, tornando difícil para as máquinas entender o verdadeiro significado.
- Negação: O uso de palavras negativas para transmitir uma inversão de significado na frase pode confundir os algoritmos de análise de sentimentos, especialmente se a negação ocorrer em frases separadas, como em "Achei que o carro fosse barato. Mas não foi".
- Falta de Contexto: O contexto é fundamental para compreender a emoção expressa num texto. A falta de contexto pode levar as ferramentas de análise de sentimentos a cometerem erros na interpretação das emoções transmitidas.
- Expressões Idiomáticas: O uso de expressões idiomáticas pode confundir as ferramentas de análise de sentimentos e os algoritmos de *Machine Learning*, pois nem sempre seguem um significado literal.

- Multipolaridade: Quando uma frase contém mais de um sentimento, torna-se desafiador para o software interpretar a emoção da frase, pois os diferentes sentimentos podem se sobrepor ou contradizer uns aos outros.

Estes desafios destacam a complexidade de análise de sentimentos e a necessidade contínua de se desenvolver as tecnologias para lidar com eles.

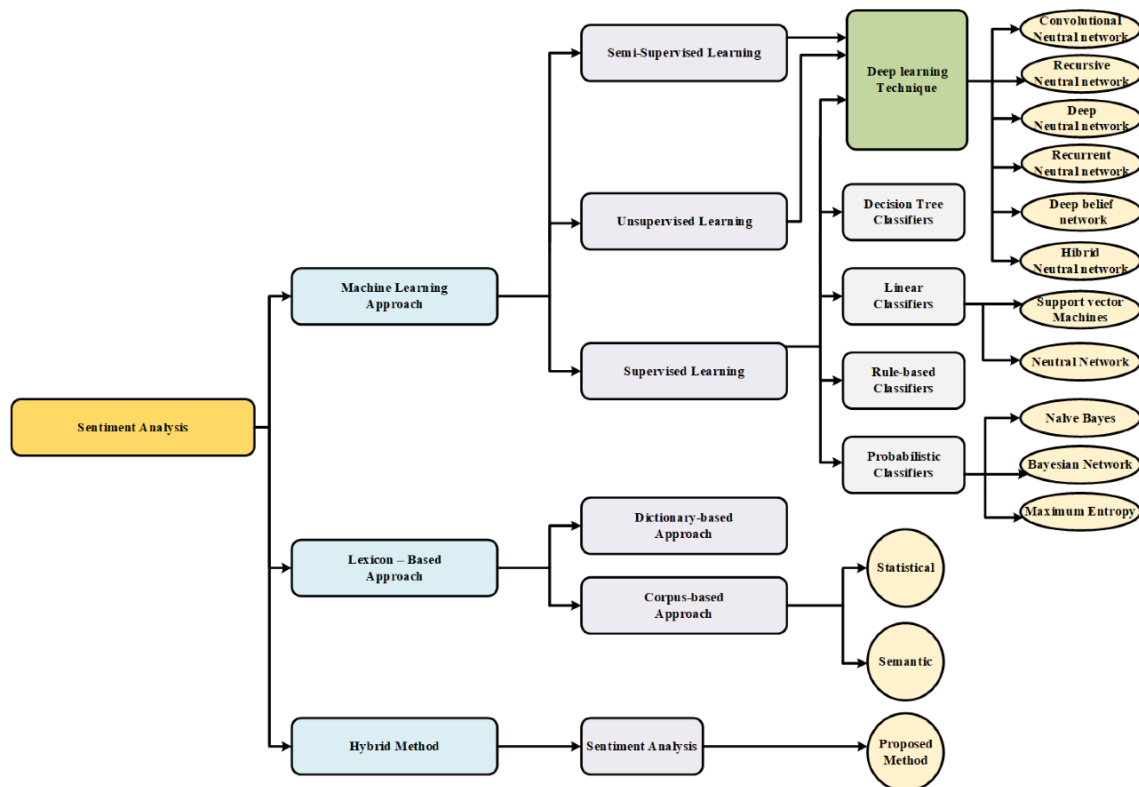


Figura 1- Abordagens da Análise de Sentimento

A figura 1, mostra as várias abordagens existentes para realizar análise de sentimento

2.7. Alguns Trabalhos Realizados

Nesta subsecção pretende-se dar alguns exemplos de trabalhos realizados na área de Processamento de Linguagem Natural, tal como são os exemplos.

No trabalho de Dosciatti e Ferreira (2013) [11], os autores apresentam uma abordagem que utiliza o SVM para identificar emoções em textos escritos em português do Brasil. O trabalho usa um texto composto de notícias retiradas de um jornal online. Cada texto é previamente rotulado com as seis emoções básicas: medo, surpresa, raiva, desgosto, tristeza e alegria. Os textos previamente rotulados são submetidos a um classificador SVM em configuração multi-classe, obtendo uma taxa de precisão de 61%.

Neethu e Rajasree (2013) [12] realizaram a classificação de postagens do Twitter em duas categorias, positivas e negativas, utilizando diversas técnicas de Machine Learning. Os desafios enfrentados pelos classificadores ao lidar com textos do Twitter são notáveis, devido à natureza dessas publicações serem geralmente curtas, repletas de gírias e erros ortográficos. Para lidar com isso, os autores optaram por realizar um pré-processamento dos *tweets*, corrigindo os erros ortográficos e substituindo as gírias e abreviações pelos seus significados correspondentes. Na classificação, foram empregues técnicas de Machine Learning como Naïve Bayes, SVM e Máxima Entropia.

Outra aplicação importante de análise de sentimentos em textos do Twitter é descrita no trabalho de Pak e Paroubek (2010) [13]. Eles destacam que o *microblogging* se tornou uma ferramenta de comunicação extremamente popular na internet, onde milhões de utilizadores expressam opiniões sobre diversos aspetos da vida quotidiana. Os textos dessas redes sociais são uma valiosa fonte de dados para análise de sentimentos. Pak e Paroubek (2010) automatizaram a construção de um texto para gerar um classificador de sentimentos, capaz de identificar sentimentos positivos, negativos e neutros em textos de *microblogging* em inglês originados no Twitter. Eles desenvolveram dois classificadores, um baseado em Naïve Bayes e outro em SVM, com o Naïve Bayes a apresentar melhores resultados. Uma diferença significativa desse trabalho é que o texto usado para validar os classificadores foi construído automaticamente, em contraste com anotações manuais como em outros estudos.

Por fim, temos o exemplo de Freitas e Vieira (2015) [14] propuseram um classificador semântico de polaridade (positiva ou negativa) para analisar comentários em português do TripAdvisor, uma página online de avaliações de conteúdos relacionados ao turismo.

Eles investigaram diferentes recursos léxicos em português e o impacto de diversas regras linguísticas na atribuição de sentimentos. O conjunto de dados utilizado por eles consiste em 194 comentários, publicados entre março de 2010 e maio de 2014. A atribuição dos sentimentos aos comentários foi realizada manualmente por dois anotadores, um linguista e um cientista da computação, ambos nativos em português.

Capítulo III

3. O Processo de Análise de Sentimento

O Processo de Análise de Sentimento é essencial para lidar com a grande quantidade de informações disponíveis na *web* nos dias de hoje. Para realizar essa análise de forma eficaz, são utilizadas técnicas capazes de lidar com grandes volumes de dados. O Processo de Análise de Sentimento geralmente envolve cinco etapas principais:

1. Recolha dos Dados: Esta etapa é crucial no processo de análise de sentimentos. Diversas fontes são exploradas para obter dados relevantes, incluindo redes sociais, sites de notícias, fóruns online, e-mails e aplicativos de mensagens. É essencial garantir a qualidade e a relevância dos dados recolhidos para uma análise precisa e significativa.
2. Preparação dos Dados: Os dados recolhidos passam por uma fase de preparação, onde são limpos, organizados e estruturados para facilitar a análise. Processos como remoção de informações irrelevantes, correção de erros ortográficos, tratamento de dados ausentes e técnicas de pré-processamento, como tokenização, lematização e *stemming*, são aplicadas para normalizar os dados textuais.
3. Deteção de Sentimento: Nesta fase, as técnicas de processamento de linguagem natural são aplicadas para identificar e extrair o sentimento presente nos dados textuais. Estas técnicas incluem análise de polaridade baseada em léxicos ou modelos de aprendizagem de máquina. A análise leva em consideração as palavras-chave, gramática, contexto e outras características linguísticas para determinar a polaridade do sentimento (positivo, negativo ou neutro).
4. Classificação de Sentimento: Com base na deteção de sentimento, os dados são classificados em categorias ou rótulos de acordo com a sua polaridade. Exemplos de categorias incluem "positivo", "negativo", "neutro", "irritado",

"feliz", entre outros, dependendo do domínio da análise. Isso permite agrupar os dados em diferentes classes de sentimentos para análise mais detalhada.

5. Apresentação dos Resultados: Os resultados da análise de sentimento são apresentados de forma clara e concisa para facilitar a compreensão e a tomada de decisões. Isso geralmente é feito por meio de relatórios ou gráficos que destacam as principais conclusões e informações obtidas a partir dos dados analisados. Por exemplo, gráficos de barras, tabelas, estatísticas ou outras representações visuais são usadas para ilustrar os *insights* e tendências identificadas durante a análise, influenciando decisões estratégicas em áreas como marketing, atendimento ao cliente, desenvolvimento de produtos, entre outros.

Na figura 2, temos uma representação do processo de análise de sentimento.

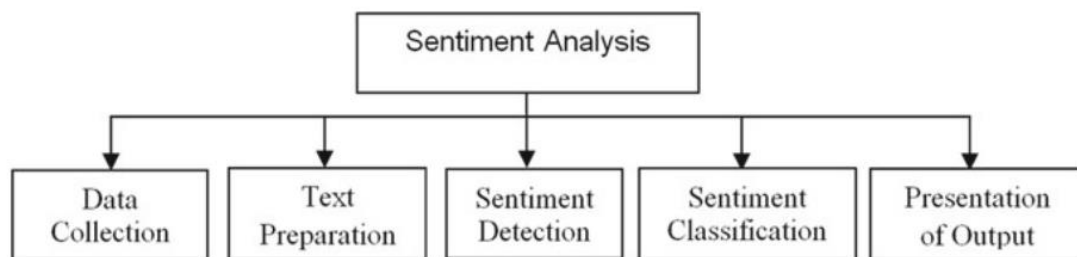


Figura 2 - Processo da Análise de Sentimento

Cada etapa deste processo desempenha um papel crucial na obtenção de resultados precisos e significativos. Seguir rigorosamente esses passos não apenas assegura a qualidade da análise automatizada de sentimentos, mas também permite que se aproveite ao máximo o poder dessa abordagem.

Capítulo IV

4. Caso de Estudo – Análise de Sentimentos em Comentários do TripAdvisor

Esta secção apresenta um caso de estudo focado na análise de sentimentos em comentários retirados da plataforma TripAdvisor. O processo realizado pelo grupo de trabalho será descrito desde a recolha dos dados e pré-processamento até à aplicação das técnicas de Machine Learning para classificar os comentários. Os resultados obtidos serão exibidos, fornecendo informações sobre a opinião dos clientes em relação a determinado estabelecimento turístico.

4.1. Recolha dos Dados

No âmbito do projeto de análise de sentimentos proposto, a primeira etapa envolveu a recolha de dados provenientes do site TripAdvisor, como foi sugerido pelo Coordenador do Projeto. O TripAdvisor, uma plataforma com uma extensa base de avaliações e opiniões de pessoas, oferece um conjunto de dados grande e interessante para realizar a análise de sentimentos.

Os membros do grupo realizaram a extração manual dos dados, optando por extrair o comentário, o *username* do autor do comentário e o *rating* atribuído pelo mesmo. Essa abordagem manual garantiu a precisão na extração e a seleção cuidadosa desses elementos que foram considerados importantes pelos membros do grupo.

Os resultados desta extração manual foram registados num ficheiro Excel com o formato indicado na figura seguinte.

Comentário	Username	Rating
Juro que não entendo como acham esta francesinha como uma das melhores... muito adocicado, independentemente dos gostos, uma francesinha não deve ter muito adocicado, carne com gordura... estas alterações modernas, vão estragando a francesinha original. Já fui a várias cidades e aos vinhos onde são "badaalados" estão sandes.	Escape55974470877	2
Atendimento excelente! Apesar de terem a esplanada em obras e não poder entrar no restaurante com a minha cadela arranjaram uma mesa na rua e o atendimento não deixou de ser excelente! 0* RECOMENDO!	Guide48979331606	5
A francesinha é excelente o muito diversão e a carne muito boa. Os empregados muito prestáveis e atentos. O tempo de espera poderá ser elevado, caso não se chegue bastante cedo.	MargandaF275	4
Há algo bastante desagradável: um dos funcionários grita constantemente com os outros, quer para chamar para limpar mesas quer para trazer contas de clientes, dando a sensação de ser chefe. Já não vivemos numa época de se gritar com funcionários, muito menos na restauração onde é constante a queixa de falta de pessoal. Fica a dica, para melhorar não só o ambiente com os clientes mas também com funcionários.	brunandipenera	5
Para mim é uma das melhores francesinhas do hotel! Boas carnes, bom molho e excelente sabor! Atendimento rápido. Espaço acolhedor. Recomendo experimentar!	Ana Paula V	5
Oportunidade! É sempre excelente a melhor das melhores! Hingem deveria montar um experimental! Aconselho a todos.	rtas5000261	5
Luca destaca! A francesinha é excelente e o atendimento super rápido, respeitamos sempre que possível! 5 Estrelas. Durante anos com francesinhas em todo o país, mas nada se compara a francesinha que comi aqui com a minha namorada. Fizemos 600 km para experimentar e certamente vamos repetir. Staff atencioso e simpático, e muita atenção ao detalhe.	JoanaSantos89	5
A melhor Francesinha do Mundo, o melhor molho de Francesinha do Mundo e combinação com uma ótima batata frita e uma Bush! Só existe um ponto de melhoria que é abrir uma Taberna no Porto. Dizem que é melhor francesinha de Portugal, mas está longe disso. É bastante boa sem dúvida, mas muito longe disso. Como atendimento e ótimas batatas. Recomendo.	Paradise52913355965	4
Com o objetivo de conhecer a famosa francesinha da Taberna Beiga, fomos aqui almoçar no dia 14 de fevereiro de 2023, dia de São Valentim, uma terça-feira. Antes da hora de abertura, às 12:00 horas, já havia uma fila com cerca de dez metros, mas assim que abriu o restaurante, a entrada decorreu com toda a normalidade e com lugares suficientes no interior. Não tínhamos feito reserva.	igbasso	5
A funcionária que nos levou ao nosso lugar foi correta, despaçada e simpática, sem stress. Já estão preparados para estes momentos.	JoanaSantos89	5
A comida é nos apresentada no papel que vai ficar por baixo do prato. O interior é bonito e com bons acessos, mesmo a pessoas em cadeira de rodas, conta com várias salas de diferentes tamanhos, todas no mesmo piso da entrada/saída, e nota-se a publicidade a cervejas estrangeiras, que também comercializam. Pedimos uma francesinha (10.90€) + meia francesinha (8.90€) + uma dose de batatas fritas (1.90€) que se nota não serem pré-fritas, que foram entregues em curto espaço de tempo. Pedimos também dois ice-tea (1.90€ cada).	Rui_F74	5
Todas as francesinhas trazem um estratagem em cima e está, apesar de visualmente parecer igual a todas outras, é de facto diferente, a começar pelo molho, não picante e ligeiramente doce, passando pela carne de vaca muito tenra. Achei a minha francesinha um pouco cara, 2.10€ mais barata que a interior.	GordinhoAvaliador	1
Valeu a pena ter vindo aqui experimentar a francesinha tradicional desta casa, havendo outras com lombo de porco, legumes, lombo de boi, e peru ou frango, com grejos diferentes.	marianab405	4
Estacionamento gratuito nas imediações e relativamente fácil de conseguir.	Leandro_Santos30	5
Depois de tanto ouvir falar da melhor francesinha do País, fui experimentar. O molho é doce, o que não é para toda a gente, e sabe demasiado a queijo. A carne (pele) cheia de ligamentos e gordura. Os empregados foram simpáticos. Parabéns pelo seu restaurante e pela co-cina!		
Provamos um prato tipo português chamado francesinha. Estava delicioso. Aqueles que apreciam sanduiches e carne é uma grande experiência.		
Francesinha de enorme qualidade. Com um molho muito bom, que não é demasiado picante o que é o que pretendia.		
A variedade de cervejas é astronómica e é difícil escolher.		
O staff também é muito atencioso e o espaço casa muito bem com o tema do restaurante.		

Figura 3 - Resultados da Extração de Comentários

O documento contém cerca de 136 comentários acerca da Taberna Belga em Braga, número considerado bastante satisfatório para a tarefa pretendida pelos elementos do grupo.

4.2. Processamento dos Dados

Para realizar esta tarefa de análise de sentimento, o grupo de trabalho teve de dedicar algum tempo a pesquisas e à leitura de artigos fornecidos pelo coordenador deste projeto [1,2,4,5,7,9], a fim de encontrar as melhores ferramentas e métodos para processar os dados e detetar sentimentos. Após estas pesquisas, o grupo de trabalho descobriu uma ferramenta chamada Vader (Valence Aware Dictionary and Sentiment Reasoner) [15], já preparada para análise de textos em português. Esta ferramenta é baseada em regras e especificamente pensada para analisar textos de redes sociais. O Vader é um modelo pré-treinado que atribui uma pontuação de sentimento a um determinado texto.

O Vader utiliza um ou mais léxicos para determinar o sentimento de texto, neste caso, para analisar o sentimento de vários comentários. Ele atribui um valor a cada palavra para determinar a sua positividade ou negatividade, numa escala que varia de -4 a +4, onde -4 representa o sentimento mais negativo e +4 o mais positivo. O Vader também leva em conta a intensidade do sentimento associada à pontuação, por exemplo, todas as letras maiúsculas podem indicar um sentimento mais intenso.

Este Vader utiliza quatro léxicos distintos para realizar a tarefa de análise de sentimento. Esses léxicos são:

- 1) Léxico de negação: Contém algumas palavras negativas da língua portuguesa, que são utilizadas para identificar negações em texto. Quando uma palavra negativa é identificada, ela influencia a pontuação de sentimento do texto.
- 2) Léxico de emojis: Contém uma lista de emojis acompanhada das suas descrições textuais. Isso permite que emojis sejam tratados como palavras durante a análise de sentimentos.
- 3) Léxico de boosters: Contém palavras que podem aumentar ou diminuir o sentimento associado a uma palavra. Cada palavra no léxico de boosters é acompanhada da sigla 'INCR' ou 'DECR', indicando se ela aumenta ou diminui o sentimento.

- 4) Léxico principal: Contém as palavras em português juntamente com as polaridades atribuídas. Além da polaridade, cada palavra no léxico principal foi avaliada por 10 pessoas, que atribuíram valores de -4 a +4. A partir dessas avaliações, foi calculado um desvio padrão que é utilizado posteriormente no cálculo da pontuação final.

O código do Vader [21], escrito em Python, contém algumas funções de pré-processamento que foram feitas para prepararem os comentários iniciais extraídos. Estas etapas de pré-processamento incluem:

- 1) Remoção de pontuação: Utilizando a função '*Regex_Remove_Punctuation*', a pontuação é eliminada dos comentários. Isso é fundamental para garantir que apenas palavras e *emojis* sejam considerados na análise de sentimentos, sem a interferência de caracteres de pontuação.

```
# Remoção de Pontuação
REGEX_REMOVE_PUNCTUATION = re.compile('%s' % re.escape('!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'))
```

Figura 4 - Função *Regex_Remove_Punctuation*

- 2) Normalização: A função *normalize* é responsável por normalizar as pontuações de sentimento para que estejam dentro de um intervalo específico, geralmente entre -1 e +1. Essa padronização facilita a comparação entre diferentes comentários.

```
'''A função normalize tem como objetivo normalizar um valor de pontuação de sentimento para que esteja dentro de um intervalo específico, geralmente entre -1 e 1'''
def normalize(score, alpha=15): #o score é dado no léxico (parâmetro a seguir à palavra). varia de -4 a 4
    norm_score = score / math.sqrt((score * score) + alpha)
    if norm_score < -1.0:
        return -1.0
    elif norm_score > 1.0:
        return 1.0
    else:
        return norm_score
```

Figura 5 - Função *Normalize*

- 3) Tratamento de emojis: *Emojis* são convertidos nas suas descrições textuais. Isso é feito para assegurar que os *emojis* sejam tratados como palavras durante o processamento do texto, permitindo a sua correta análise em relação ao sentimento que expressam.

```
''' converte emojis para as suas descrições textuais---> no lexico está "😊 rosto sorridente" '''
text_token_list = text.split()
text_no_emoji_lst = []
for token in text_token_list:
    if token in self.emojis:
        # obtem a descrição do emoji
        description = self.emojis[token]
        text_no_emoji_lst.append(description)
    else:
        text_no_emoji_lst.append(token)
text = " ".join(x for x in text_no_emoji_lst)

sentitext = SentiText(text)
```

Figura 6 - Tratamento de Emojis

- 4) Verificação de ênfase de pontuação: A função ‘*_punctuation_emphasis*’ adiciona ênfase ao texto com base na presença de pontos de exclamação e interrogação. Esses sinais de pontuação podem indicar intensidade de sentimento e influenciar a polaridade do texto.

```
def _punctuation_emphasis(self, text):
    '''Adiciona ênfase a partir de pontos de exclamação e interrogação.'''
    ep_amplifier = self._amplify_ep(text) # Amplifica devido a pontos de exclamação
    qm_amplifier = self._amplify_qm(text) # Amplifica devido a pontos de interrogação
    punct_emph_amplifier = ep_amplifier + qm_amplifier # Soma as ênfases

    return punct_emph_amplifier
```

Figura 7 - Função _punctuation_emphasis

- 5) Tokenização e normalização de palavras: O texto é dividido em palavras e *emojis*, com a remoção da pontuação e a normalização das palavras para minúsculas. Esse processo de tokenização e normalização é crucial para garantir que todas as palavras sejam tratadas de forma consistente durante a análise de sentimentos.

```
def __init__(self, text):
    if not isinstance(text, str):
        text = str(text).encode('utf-8')
    self.text = text
    self.words_and_emojis = self._words_and_emojis()

    self.is_cap_diff = allcap_differential(self.words_and_emojis)

'''#constrói um dicionário que associa palavras e pontuações às suas formas combinadas'''
def _words_plus_punc(self):

    no_punc_text = REGEX_REMOVE_PUNCTUATION.sub('', self.text) #Remove a pontuação do texto de entrada

    #Divide o texto sem pontuação em uma lista de palavras
    words_only = no_punc_text.split()

    #Remove palavras únicas e mantém apenas aquelas que têm mais de uma letra
    words_only = set(w for w in words_only if len(w) > 1)

    # cria um dicionário onde as chaves são strings formadas pela concatenação
    # de um caractere de pontuação com uma palavra, e os valores são o segundo elemento de cada par
    punc_before = {''.join(p): p[1] for p in product(PUNC_LIST, words_only)} #{'', 'hello': 'hello', ', world': 'world', '.hello': 'hello'}
    # product(p[0], p[1]) -> p[1] corresponde a words_only

    punc_after = {''.join(p): p[0] for p in product(words_only, PUNC_LIST)} #{'hello', ':', 'hello', 'hello.': 'hello', 'hello!': 'hello'}
    #product(p[0], p[1]) -> p[0] corresponde a words_only

    #Combina os dicionários punc_before e punc_after, criando um dicionário final
    words_punc_dict = punc_before
    words_punc_dict.update(punc_after)

    return words_punc_dict
```

Figura 8 - Tokenização e Normalização de Palavras

4.3. Detecção de Sentimento

O cálculo da pontuação final acontece através de um processo em que cada palavra ou emoji do texto recebe um valor de polaridade com base nos léxicos. Além disso, são consideradas as avaliações fornecidas por 10 pessoas diferentes para cada palavra, numa escala de -4 a +4. Essas avaliações são usadas para calcular um desvio padrão, que é uma medida de dispersão das avaliações.

O desvio padrão é importante porque reflete a variabilidade das avaliações para a mesma palavra. Se as avaliações estiverem muito dispersas, isso indica que a palavra pode ter um impacto emocional bastante diferente em diferentes pessoas. O cálculo final leva em conta tanto a polaridade quanto o desvio padrão. Palavras com um desvio padrão elevado podem ter o seu impacto final reduzido, pois representam uma incerteza grande sobre o seu efeito emocional.

Este processo de cálculo visa capturar como diferentes palavras afetam o sentimento geral do texto, levando em consideração tanto a polaridade das palavras como a variabilidade das percepções das pessoas sobre essas palavras.

```
def score_valence(self, sentiments, text):  
    '''Calcula a valência do sentimento com base nos scores de sentimentos.'''  
    if sentiments:  
        sum_s = float(sum(sentiments)) # Soma os scores de sentimentos  
        # Adiciona ênfase da pontuação no texto  
        punct_emph_amplifier = self._punctuation_emphasis(text)  
        # Ajusta a valência com base na ênfase da pontuação  
        if sum_s > 0:  
            sum_s += punct_emph_amplifier  
        elif sum_s < 0:  
            sum_s -= punct_emph_amplifier  
  
        # Normaliza o valor da valência (linha 78)  
        compound = normalize(sum_s)  
        # Separa os scores de sentimentos em positivos, negativos e neutros  
        pos_sum, neg_sum, neu_count = self._sift_sentiment_scores(sentiments)  
  
        # Adiciona a ênfase da pontuação aos scores de sentimentos positivos ou negativos  
        if pos_sum > math.fabs(neg_sum):  
            pos_sum += punct_emph_amplifier  
        elif pos_sum < math.fabs(neg_sum):  
            neg_sum -= punct_emph_amplifier  
  
        # Calcula as porcentagens de sentimentos positivos, negativos e neutros  
        total = pos_sum + math.fabs(neg_sum) + neu_count  
        pos = math.fabs(pos_sum / total)  
        neg = math.fabs(neg_sum / total)  
        neu = math.fabs(neu_count / total)  
  
        # Caso não haja scores de sentimentos, retorna valores neutros  
    else:  
        compound = 0.0  
        pos = 0.0  
        neg = 0.0  
        neu = 0.0  
  
    # Retorna um dicionário com os valores de valência (arredonda)  
    sentiment_dict = {  
        'neg': round(neg, 3),  
        'neu': round(neu, 3),  
        'pos': round(pos, 3),  
        'compound': round(compound, 4)  
    }  
  
    return sentiment_dict
```

Figura 9 - Função score_valence

Após examinar o código do Vader, o grupo de trabalho considerou que estavam prontos para calcular a polaridade dos comentários extraídos, resultando nos dados ilustrados na Figura 10.

	A	B
49	Restaurante ótimo, tivemos um bocado a espera, mas valeu apenas. As moelas estavam ok	0,8762
50	Bom local para comer francesinha. Estava maravilhosa, bom atendimento, rapidez, e simp	0,9231
51	Francesinha excelente, com boa carne. Muito boa lista de cervejas europeias. Boa organiz	0,9042
52	Para mim, o sítio para comer francesinha de eleição recomendo a sua visita, atendimento	0,6808
53	É um local ótimo! A francesinha é maravilhosa! O atendimento excelente! Uma variedade	0,9515
54	Muito tentam imitar, muito queriam estar ao nível da qualidade desta francesinha, mas atr	0,9097
55	Mande vir a francesinha, lamento mas não gostei nada 😞 o molhe era tipo sopa a saber	-0,0258
56	Sou do Porto e para ser franco não gostei da francesinha, molho nada a ver e a carne com	0,2263
57	Uma referência na Francesinha! Quando passo por Braga a francesinha da Taberna Belga f	0,8257
58	Bom restaurante e bem situado. Por vezes tem uma espera demorada mas de pois de sent	0,8176
59	Apesar da comida ser boa, o tempo de espera no exterior ao frio, com pessoas de idade, e	-0,296
60	Fui ontem á noite dia 27.08.2021 comer á Taberna Belga com 3 amigos. Infelizmente onten	-0,8996
61	A taberna belga devia ter vergonha de apresentar este prato. Não é preciso muita imaginaç	-0,9136
62	Apesar de continuar a ter grandes enchentes, o Taberna Belga conseguiu melhorar o que já	0,9861
63	De facto a francesinha é uma das melhores que já comi e com carne de excelente qualidad	0,9485
64	Visitámos o espaço pela fama (foi-nos amplamente recomendado). Chegámos ao local e d	0,9442
65	Francesinha incrível, batata top e cerveja ainda melhor 😊🍷 atendimento espetacular. C	0,8906
66	A melhor francesinha que já comemos. Definitivamente um local de vista obrigatória em B	0,9042
67	Sem dúvida alguma a melhor francesinha que já provei. Sou muito fã deste local, pela com	0,743
68	Boa francesinha e boas cervejas Os preços são acessíveis Não tinha expectativa nenhuma n	0,6696
69	Amei! Espetacular, ótimo serviço, ótima comida, ótimos funcionários. Estava tudo bom a c	0,5696
70	Melhor francesinha do Norte. Apesar de esperar um pouco, não desilude nada. Vale muito	-0,128
71	Simplesmente é o meu local favorito para francesinhas. Tem algo no molho que não sei ex	0,6428
72	Fomos para experimentar a francesinha que muitos elogiam, de fato achamos muito sabor	0,2942
73	Francesinha muito boa, pães feitos na hora, entradas muito boas! Recomendo vivamente	0,8268
74	Melhor francesinha que já provámos! Já tínhamos ouvido falar e realmente vale mesmo a	-0,4493
75	Fui almoçar lá com família. Havia muita gente à espera e a nossa reserva estava marcada	-0,1154
76	Carne de má qualidade, molho extremamente gorduroso e enjoativo. Das últimas vezes qu	-0,6808
77	Após visitarmos vários sítios em Braga o dia todo. bateu aquela fome. Seguimos a sugestão	0,6705

Figura 10 - Primeiros Resultados do Vader

A seguir, os colaboradores do projeto analisaram manualmente o conjunto de comentários extraídos e discordaram de algumas das avaliações feitas, concentrando-se em corrigir algumas polaridades. Depois de refletirem sobre o assunto, decidiram criar uma função para remover os acentos das palavras (Figura 11), pois observaram que o léxico não continha nenhum tipo de acentuação.

```

replace_dict = {'á': 'a', 'é': 'e', 'í': 'i', 'ó': 'o', 'ú': 'u', 'ã': 'a', 'õ': 'o', 'ç': 'c',
               'à': 'a', 'è': 'e', 'ì': 'i', 'ò': 'o', 'ù': 'u', 'ê': 'e', 'î': 'i', 'ô': 'o', 'û': 'u',
               'ä': 'a'}

def fix_sentence(sentence):
    final_sentence = ""
    for word in sentence:
        final_word = ""
        for letter in word:
            if letter.lower() in replace_dict:
                final_word += replace_dict[letter.lower()]
            else:
                final_word += letter
        final_sentence += final_word
    return final_sentence

```

Figura 11 - Função fix_sentence

Depois de criada a função, os comentários foram novamente submetidos ao Vader com essa remoção de acentos, e foram observados alguns resultados diferentes, como ilustrado na Figura XXX.

	A	B		A	B
9	Restaurante ótimo, tivemos um bocado a espera, mas valeu a pena. As moelas estavam ok	0,8762	49	Restaurante otimo, tivemos um bocado a espera, mas valeu a pena. As moelas estavam ok	0,8762
10	Bom local para comer francesinha. Estava maravilhosa, bom atendimento, rapidez, e simp.	0,9231	50	Bom local para comer francesinha. Estava maravilhosa, bom atendimento, rapidez, e simp.	0,9231
11	Francesinha excelente, com boa carne. Muito boa lista de cervejas europeias. Boa organiz	0,9042	51	Francesinha excelente, com boa carne. Muito boa lista de cervejas europeias. Boa organiz	0,9042
12	Para mim, o sitio para comer francesinha de eleição recomendo a sua visita, atendimento	0,6808	52	Para mim, o sitio para comer francesinha de eleicao recomendo a sua visita, atendimento	0,7085
13	É um local ótimo! A francesinha é maravilhosa! O atendimento excelente! Uma variedade	0,9513	53	e um local otimo! A francesinha e maravilhosa! O atendimento excelente! Uma variedade	0,9513
14	Muito tentam imitar, muito queriam estar ao nivel da qualidade desta francesinha, mas atr	0,9097	54	Muito tentam imitar, muito queriam estar ao nivel da qualidade desta francesinha, mas atr	0,9097
15	Mandeí vir a francesinha, lamento mas não gostei nada 😞 o molhe era tipo sopa a saber	-0,0258	55	Mandeí vir a francesinha, lamento mas nao gostei nada 😞 o molhe era tipo sopa a saber	-0,0258
16	Sou do Porto e para ser franco não gostei da francesinha, molho nada a ver e a carne com	0,2263	56	Sou do Porto e para ser franco nao gostei da francesinha, molho nada a ver e a carne com	0,2263
17	Uma referência na Francesinha! Quando passo por Braga a francesinha da Taberna Belga f	0,8257	57	Uma referencia na Francesinha! Quando passo por Braga a francesinha da Taberna Belga f	0,8257
18	Bom restaurante e bem situado. Por vezes tem uma espera demorada mas de pois de sent	0,8178	58	Bom restaurante e bem situado. Por vezes tem uma espera demorada mas de pois de sent	0,8178
19	Apesar da comida serboa, o tempo de espera no exterior ao frio, com pessoas de idade, e	-0,296	59	Apesar da comida serboa, o tempo de espera no exterior ao frio, com pessoas de idade, e	-0,296
20	Fui ontem á noite dia 27.08.2021 comer á Taberna Belga com 3 amigos. Infelizmente onte	-0,8996	60	Fui ontem a noite dia 27.08.2021 comer a Taberna Belga com 3 amigos. Infelizmente onte	-0,9056
21	A taberna belga devia ter vergonha de apresentar este prato. Não é preciso muita imaginac	-0,9136	61	A taberna belga devia ter vergonha de apresentar este prato. Nao e preciso muita imaginac	-0,9136
22	Apesar de continuar a ter grandes enchentes, o Taberna Belga conseguiu melhorar o que já	0,9861	62	Apesar de continuar a ter grandes enchentes, o Taberna Belga conseguiu melhorar o que ja	0,9861
23	De facto a francesinha é uma das melhores que já comi e com carne de excelente qualidac	0,9489	63	De facto a francesinha e uma das melhores que ja comi e com carne de excelente qualidac	0,9489
24	Visitámos o espaço pela fama (foi-nos amplamente recomendado). Chegámos ao local e d	0,9442	64	Visitamos o espaco pela fama (foi-nos amplamente recomendado). Chegamos ao local e d	0,9442
25	Francesinha incrível, batata top e cerveja ainda melhor 😊 atendimento espetacular. C	0,8906	65	Francesinha incrível, batata top e cerveja ainda melhor 😊 atendimento espetacular. C	0,8906
26	A melhor francesinha que já comemos. Definitivamente um local de vista obrigatória em B	0,9042	66	A melhor francesinha que ja comemos. Definitivamente um local de vista obrigatoria em B	0,9042
27	Sem dúvida alguma a melhor francesinha que já provei. Sou muito fã deste local, pela com	0,743	67	Sem duvida alguma a melhor francesinha que ja provei. Sou muito fa deste local, pela com	0,743
28	Boa francesinha e boas cervejas Os preços são acessíveis Não tinha expectativa nenhuma	0,6698	68	Boa francesinha e boas cervejas Os precos sao acessiveis Nao tinha expectativa nenhuma	0,7124
29	Amei! Espetacular, ótimo serviço, ótima comida, ótimos funcionários. Estava tudo bom a c	0,5696	69	Amei! Espetacular, otimo servico, otima comida, otimos funcionarios. Estava tudo bom a c	0,5696
30	Melhor francesinha do Norte. Apesar de esperar um pouco, não desilude nada. Vale muito	-0,128	70	Melhor francesinha do Norte. Apesar de esperar um pouco, nao desilude nada. Vale muito	0,4224
31	Simplemente é o meu local favorito para francesinhas. Tem algo no molho que não sei ex	0,6428	71	Simplemente e o meu local favorito para francesinhas. Tem algo no molho que nao sei ex	0,6428
32	Fomos para experimentar a francesinha que muitos elogiam, de fato achamos muito saboi	0,2942	72	Fomos para experimentar a francesinha que muitos elogiam, de fato achamos muito saboi	0,2942
33	Francesinha muito boa, pães feitos na hora, entradas muito boas! Recomendo vivamente	0,8268	73	Francesinha muito boa, paes feitos na hora, entradas muito boas! Recomendo vivamente	0,8268
34	Melhor francesinha que já provámos! Já tínhamos ouvido falar e realmente vale mesmo a	-0,4493	74	Melhor francesinha que ja provamos! Ja tinhamos ouvido falar e realmente vale mesmo a	-0,4493
35	Fui almoçar lá com família. Havia muita gente à espera e a nossa reserva estava marcada	-0,1154	75	Fui almoçar la com familia. Havia muita gente a espera e a nossa reserva estava marcada	-0,7754
36	Carne de má qualidade, molho extremamente gorduroso e enjoativo. Das últimas vezes qu	-0,6808	76	Carne de ma qualidade, molho extremamente gorduroso e enjoativo. Das ultimas vezes qu	-0,6808
37	Apos visitarmos varios sitios em Braga o dia todo, bateu aquela fome. Seguimos a sugestao	0,6702	77	Apos visitarmos varios sitios em Braga o dia todo, bateu aquela fome. Seguimos a sugestao	0,6702

Figura 12 - Comparação de Resultados no Vader

4.4. Apresentação de resultados

Para apresentar os resultados, o grupo de trabalho estabeleceu os limites para definir se um comentário é positivo, negativo ou neutro. Chegaram a um consenso de que comentários com um valor superior a +0.35 seriam considerados positivos, enquanto aqueles com um valor inferior a -0.35 seriam classificados como negativos. Todos os valores restantes seriam considerados neutros. Em seguida, procederam à criação de um novo arquivo Excel que refletisse adequadamente a polaridade dos comentários.

Concluída a deteção de sentimento e dos resultados foi criado um gráfico para representar a distribuição dos comentários.

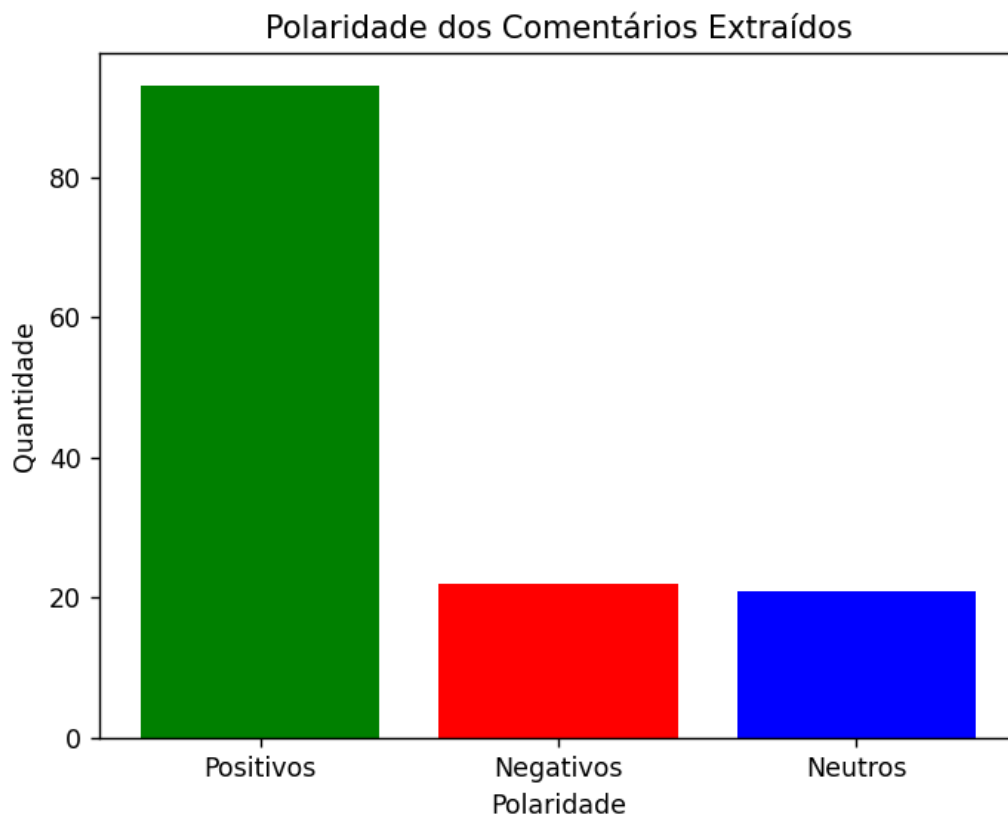


Figura 13 - Polaridade dos Comentários Extraídos

Podemos concluir a partir do gráfico anterior que os comentários positivos predominam sobre os restantes, sugerindo uma satisfação geral dos clientes com o restaurante em questão, dos quais os comentários foram extraídos. No entanto, é importante notar que o número de comentários neutros pode ser mais elevado devido a possíveis limitações do léxico utilizado na análise de sentimentos.

4.5. Machine Learning

O grupo de trabalho testou a eficácia de três algoritmos de Machine Learning. Pretendeu-se comparar os resultados que cada algoritmo produziu para determinar qual destes algoritmos tem uma maior precisão na análise no conjunto de dados usado.

O conjunto de dados foi dividido em duas partes: um conjunto de teste e um conjunto de treino. Este procedimento é habitualmente usado em Machine Learning quando se pretende avaliar o desempenho dos algoritmos, pois permite que o algoritmo seja treinado conhecendo uma parte dos dados, geralmente 80% dos dados. Depois, o

algoritmo é desafiado a fazer previsões precisas nos restantes 20% dos dados através do conhecimento obtido nos dados conhecidos.

Inicialmente, usou-se o algoritmo SVM [16] e registou-se os posteriores resultados:

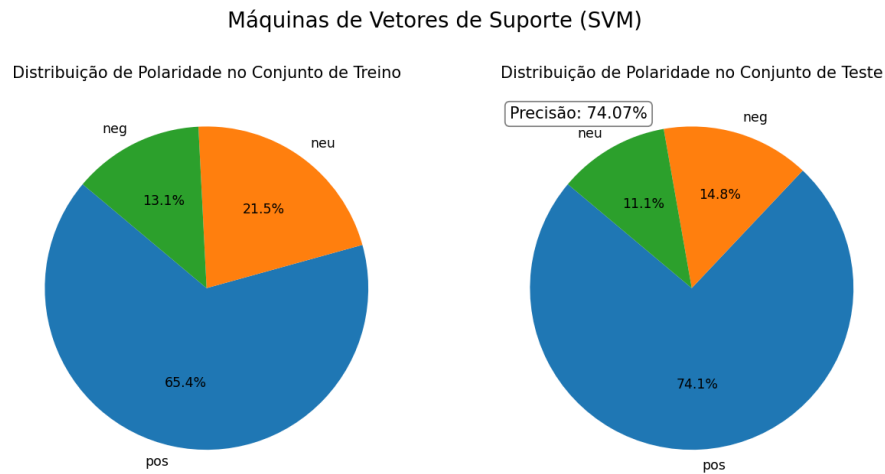


Figura 14 - Resultados da Máquina de Vetores de Suporte (SVM)

Logo de seguida, o grupo utilizou o algoritmo de Bayes [19,20] que se baseia no algoritmo probabilístico de Bayes e obteve os resultados da figura 15:

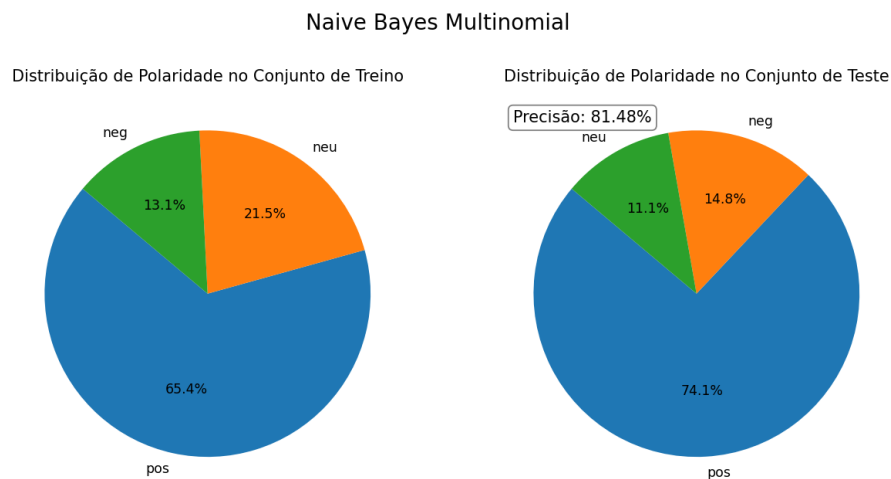


Figura 15 - Resultados do Naive Bayes Multinomial

Por último, aplicou-se o algoritmo estatístico de Regressão Linear [10] cujo objetivo é prever os resultados com base numa relação linear entre variáveis e registou-se os seguintes resultados:

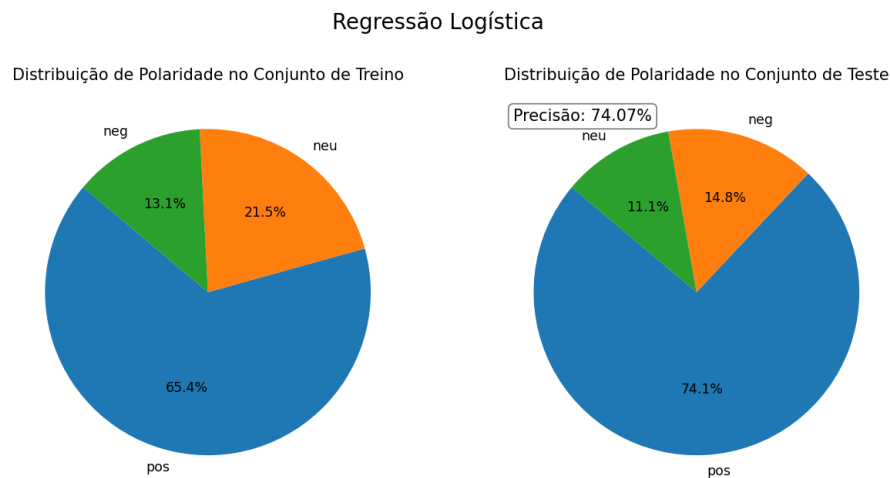


Figura 16 - Resultados da Regressão Logística

Depois de ambos os três algoritmos terem produzido os seus resultados, procedeu-se à comparação destes através da seguinte tabela:

Algoritmo	Precisão
Máquinas de Vetores de Suporte (SVM)	74.07%
Naive Bayes Multinomial	81.48%
Regressão Logística	74.07%

Figura 17 - Comparação entre algoritmos de Machine Learning

Através da análise da tabela, podemos concluir que a precisão do SVM e da regressão logística é similar, ao passo que o Naive Bayes Multinomial apresenta uma vantagem de cerca de 5 pontos percentuais. Dessa forma, é evidente que o Naive Bayes Multinomial se destaca como o algoritmo mais adequado para o conjunto de dados em questão. Este resultado pode ser atribuído possivelmente à sua abordagem na consideração de cada característica de forma independente, o que o torna menos vulnerável à qualidade do léxico. Além disso, a simplicidade inerente ao referido algoritmo pode contribuir para evitar o *overfitting*, um fenómeno comum em *Machine Learning* que ocorre quando um modelo se ajusta excessivamente aos dados de treino. Assim, a superioridade do Naive Bayes Multinomial sugere que é uma escolha sólida para este conjunto de dados específico

Capítulo V

5. Conclusões

Neste estudo, foram apresentados conceitos relativos à Análise de Sentimentos e às fases necessárias para a realização dessa tarefa. Posteriormente, fizemos a análise de sentimentos num conjunto de comentários previamente definido pelo grupo de trabalho, comparando algoritmos de Machine Learning para determinar qual seria o mais adequado para os nossos dados.

Identificamos que a riqueza do léxico é crucial para se obter resultados precisos. Neste contexto, é relevante mencionar que o grupo de trabalho dispunha de um léxico bastante rico[17], específico para o domínio da restauração, que poderia ter sido utilizado para enriquecer a análise. No entanto, não foi possível adaptar esse léxico para o formato necessário para o uso no Vader, a ferramenta utilizada na Análise de Sentimento.

Além disso, reconhecemos a importância de melhorar a sensibilidade dos algoritmos de análise de sentimentos às particularidades da língua portuguesa. Isso inclui melhorar a identificação e interpretação de conectivos e casos especiais da língua, para minimizar a influência de palavras neutras nos resultados.

Outro ponto crítico, diz respeito ao pré-processamento dos textos. Embora essencial para preparar os dados para análise, o pré-processamento padrão pode não ser adequado para a língua portuguesa, que apresenta diversas nuances. Uma abordagem mais sofisticada, levaria em consideração aspetos tais como a colocação das vírgulas e outras características específicas da língua.

Em resumo, para melhorar os resultados apresentados neste estudo, seria necessário um esforço para enriquecer o léxico, adaptá-lo às ferramentas existentes, melhorar os algoritmos para levar em conta as nuances da língua e aperfeiçoar o pré-processamento dos textos. Essas melhorias são essenciais para garantir análises mais precisas e úteis, contribuindo assim para melhores resultados na análise de sentimentos.

6. Referências

1. Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences*, 13(7), 4550 (BELO)
2. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). *A survey on sentiment analysis methods, applications, and challenges*. Springer Nature.
3. Goel, S., & Uzuner, O. (2016). *Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports*. Wiley Online Library.
<https://doi.org/10.1002/isaf.1392>
4. Aljumah, A., Altuwijri, A., Alsuhailani, T., Selmi, A., & Alruhaily, N. (2021). *Identification of Fraud Apps Using Sentiment Analysis Techniques*. International Journal of Interactive Mobile Technologies, 15(23), 178-185.
<https://doi.org/10.3991/ijim.v15i23.27361>
5. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). *A comprehensive survey on sentiment analysis: Approaches, challenges and trends*. Knowledge-Based Systems, 226, 107134. <https://doi.org/10.1016/j.knosys.2021.107134>
6. Sousa, R. C. C. de. (2016). *Identificando sentimentos de texto em português com o SentiWordNet traduzido* (Trabalho de Conclusão de Curso). Universidade Federal do Ceará, Campus de Quixadá, Curso de Ciência da Computação, Quixadá. Orientação: Prof. Dr. Paulo de Tarso Guerra Oliveira.
7. Dieksona, Z. A., Prakosoa, M. R. B., Putra, M. S. Q., Syaputra, M. S. A. F., Achmad, S., & Sutoyo, R. (2023). Sentiment analysis for customer review: Case study of Traveloka. *Procedia Computer Science*, 216, 682–690.
8. Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Georgia Institute of Technology, Atlanta, GA 30032. Email: cjhutto@gatech.edu, gilbert@cc.gatech.edu.
9. Pearl, L., & Steyvers, M. (Ano). *Identifying emotions, intentions, and attitudes in text using a game with a purpose*. Department of Cognitive Sciences, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA 92697, USA. Emails: lpearl@uci.edu, msteyver@uci.edu.
10. Aliman, G. B., Nivera, T. F. S., Olazo, J. C. A., Ramos, D. J. P., Sanchez, C. D. B., Amado, T. M., Arago, N. M., Jorda Jr., R. L., Virrey, G. C., & Valenzuela, I. C. (2022). *Sentiment analysis using logistic regression*. Journal of Computational Innovations and Engineering Applications, JULY 2022, 35-40.

11. DOSCIATTI, M. M.; FERREIRA, E. C. L. P. C. Identificando emoções em textos em português do brasil usando máquina de vetores de suporte em solução multiclasse. ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil, 2013.
12. NEETHU, M.; RAJASREE, R. Sentiment analysis in twitter using machine learning techniques. In: IEEE. Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. [S.l.], 2013. p. 1–5.
13. PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. [S.l.: s.n.], 2010. v. 10, p. 1320–1326.
14. FREITAS, L. A. de. Feature-level sentiment analysis applied to brazilian portuguese reviews. Tese (Doutorado) — Pontificia Universidade Católica do Rio Grande do Sul, 2015.
15. Geetha, L. (2023, 28 de fevereiro). Vader: A Comprehensive Guide to Sentiment Analysis in Python. *Medium*. Recuperado de <https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e>
16. Zafar, A. (2023, 22 de dezembro). SVM Implementation in Python from Scratch- Step by Step Guide. *Machine Learning Tutorials*. Recuperado de <https://www.mltut.com/svm-implementation-in-python-from-scratch/#:~:text=SVM%20in%20Python%201%201.%20Import%20the%20Libraries-,training%20set.%20...%205%205.%20Perform%20Feature%20Scaling>
17. Melo, T. (2021). *Building a Restaurant-Specific Sentiment Lexicon via Probability Theory*. In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia), p. 129-132. Recuperado de <https://tiagodemelo.info/datasets.html>
18. Hutto, C. J., & Gilbert, E. E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.
19. DataCamp. (2023). *Naive Bayes Classification Tutorial using Scikit-learn*. *DataCamp*. Recuperado de <https://www.datacamp.com/tutorial/naive-bayes-scikit-learn>
20. Brownlee, J. (2019, 25 de outubro). Naive Bayes Classifier From Scratch in Python. *Machine Learning Mastery*. Recuperado de <https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>

21. Almeida, R. J. A. (2018). *LeIA - Léxico para Inferência Adaptada*. GitHub repository. Recuperado de <https://github.com/rafjaa/LeIA>