MINERAÇÃO DE OPINIÕES COMPARATIVAS EM PORTUGUÊS

DANIEL PIMENTEL KANSAON

MINERAÇÃO DE OPINIÕES COMPARATIVAS EM PORTUGUÊS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Fabrício Benevenuto de Souza

Belo Horizonte

Fevereiro de 2021

Kansaon, Daniel Pimentel

D1234p Mineração de Opiniões Comparativas em Português / Daniel Pimentel Kansaon. — Belo Horizonte, 2021
xviii, 84 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Fabrício Benevenuto de Souza

1. Computação — Teses. 2. Redes — Teses.
I. Orientador. II. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Agradecimentos

Primeiramente, agradeço a Deus, pois foi Ele que me guiou durante todo esse processo, estando presente comigo em todos os momentos, permitindo a realização deste trabalho.

Depois, à minha família, que sempre me apoiou, dando o suporte necessário e incentivando na persistência dos objetivos, o que tem possibilitado o meu crescimento pessoal e profissional.

Aos amigos e colegas de laboratório, que compartilharam esse momento comigo, tornando todo o processo mais leve.

Agradeço também ao orientador por todo conhecimento compartilhado, paciência e apoio durante todo o processo, desde a escolha do tema, nos planejamentos e decisões pontuais do trabalho.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento do trabalho e a todos que contribuíram no desenvolvimento do trabalho.

Obrigado a todos!

Resumo

A constante expansão do comércio eletrônico, recentemente impulsionada pela pandemia de COVID-19, tem levado a um grande aumento no número de compras online, feitas por clientes cada vez mais exigentes, que buscam por comentários e revisões na Web para auxiliar na tomada de decisão sobre a compra de produtos. Nessas revisões, parte das opiniões encontradas são comparações, que contrastam aspectos expressando preferência a um objeto em relação a outros, o que permite, por exemplo, que empresas entendam como clientes comparam seus produtos aos de seus concorrentes. Essas informações muitas vezes são negligenciadas pelas técnicas tradicionais de análise de sentimentos, que quase sempre capturam apenas sentimentos positivos ou negativos associados a aspectos de produtos. Apesar de recentes esforços voltados para a língua inglesa, quase nenhum estudo foi feito para o desenvolvimento de soluções apropriadas que permitam a análise de comparações na língua portuguesa.

Este trabalho apresenta um dos primeiros estudos sobre opiniões comparativas na língua portuguesa. De maneira geral, o trabalho contém duas principais contribuições. Primeiramente, foi proposta uma abordagem hierárquica para a detecção de comparações, que consiste em uma etapa binária inicial, que subdivide as opiniões regulares das comparativas, para posteriormente categorizar as comparativas nos cinco grupos detalhados de opiniões: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. Os resultados obtidos se mostram promissores, alcançando 87% de Macro-F1 e 0,94 de AUC para a etapa binária, e 61% de Macro-F1 para a categorização em múltiplas classes. Por fim, na segunda contribuição, foi proposto um algoritmo para detecção da entidade expressa como preferida em sentenças comparativas, alcançando valores de 94% de Macro-F1 para as Superlativas e aproximadamente 84% para as Gradativas com Predileção.

Palavras-chave: Mineração de Opinião, Análise de Sentimentos, Opinião Comparativa.

Abstract

The constant expansion of e-commerce, recently boosted due to the coronavirus pandemic, has led to a huge increase in online shopping, made by increasingly demanding customers, who seek comments and reviews on the Web to assist in decision making regarding the purchase of products. In these reviews, part of the opinions found are comparisons, which contrast aspects expressing a preference for an object over others, allowing, for example, companies to know how customers compare their products to their competitors. However, this information is neglected by traditional sentiment analysis techniques and it is not applicable for comparisons, since they do not directly express a positive or negative sentiment. In this context, despite efforts in the English language, almost no studies have been done to develop appropriate solutions that allow the analysis of comparisons in the Portuguese language.

This work presents one of the first studies on comparative opinion in Portuguese. In general, this work contains two main contributions. First, a hierarchical approach for detecting comparisons was proposed, which consists of an initial binary step, which subdivides the regular opinions of the comparatives, to further categorize the comparatives into the five groups of opinions: (1) Non-Comparative; (2) Non-Equal Gradable; (3) Equative, (4) Superlative; and (5) Non-Gradable. The results obtained are promising, reaching 87% of Macro-F1 and 0.94 of AUC for the binary step, and 61% of Macro-F1 for classification in multiple classes. Finally, in the second contribution, an algorithm was proposed to detect the entity expressed as preferred in comparative sentences, reaching 94% of Macro-F1 for Superlative and almost 84% for Non-Equal Gradable opinions.

Keywords: Opinion Mining, Sentiment Analysis, Comparative Opinions.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O número de compras no mercado online vem aumentando e estima-se que cerca de 25% da população mundial utilizará esse mercado para compras nos próximos anos [Law, 2019]. A tendência é que esse número aumente ainda mais com a pandemia de COVID-19 [OECD, 2020]. A principal vantagem do comércio eletrônico é a capacidade de alcançar um grande número de pessoas em diferentes lugares, independentemente da distância e do tempo [Nasti et al., 2020]. Toda essa interação online em compras, vendas e avaliações gera uma grande quantidade de informações, que são utilizadas por clientes cada vez mais exigentes para tomada de decisão através de revisões em fóruns, blogs, Redes Sociais Online, entre outros.

As opiniões contidas em avaliações de produtos podem ser divididas em dois grupos [Liu, 2012]: (i) opiniões regulares, que são opiniões diretas ou indiretas sobre uma entidade, criticando ou ressaltando pontos positivos de diferentes aspectos da mesma, e (ii) opiniões comparativas, que contrastam aspectos de determinado produto aos mesmos aspectos de seus concorrentes. Enquanto as opiniões regulares expressam um sentimento acerca de uma marca ou produto, as comparativas apresentam uma maneira comum de avaliação que geralmente indica um contraste ou similaridade entre diferentes produtos. Essa capacidade de fazer comparações expressando ordem e preferência é um componente básico da cognição humana [Sapir, 1944], que se reflete na linguagem natural através de sentenças comparativas, uma maneira direta e eficiente de contrastar objetos exibindo predileção.

Uma grande parte dos trabalhos na literatura se concentram na aplicação de técnicas de análise de sentimentos para classificação de sentenças em positivas, negativas e neutras, utilizando em sua maioria abordagens léxicas [Taboada et al., 2011; Ribeiro et al., 2016; Araújo et al., 2016; Melo et al., 2019] e supervisionadas [Bespalov et al., 2011; Wang et al., 2016; de O. Carosia et al., 2019; Mehta et al., 2020]. Entretanto,

para a mineração das opiniões comparativas, as técnicas tradicionais de análise de sentimentos não são suficientes. Por exemplo, na sentença comparativa: "O celular X é *melhor* do que o Y", a obtenção da polaridade se mostra insuficiente para uma análise mais profunda que tenha como objetivo a extração de informações adicionais, como quais produtos são comparados e até mesmo qual objeto é indicado como preferido.

Nesse contexto, um dos esforços fundamentais para analisar e extrair informações úteis das comparações é a criação de um mecanismo para detectar quais sentenças dentre um conjunto de revisões podem ser classificadas como comparativas, distinguindo as sentenças comparativas das sentenças não comparativas. Essa tarefa é primordial e a mais importante, pois distinguindo corretamente as sentenças, é possível, então, direcionar esforços na aplicação de técnicas apropriadas para cada tipo de opinião, sendo essencial para soluções de recomendação de produtos, geração eficaz de plano de marketing e gerenciamento de reputação de empresas.

Dada a importância e aplicabilidade da tarefa de se identificar expressões de comparações, vários esforços propõem técnicas para resolver o problema [Jindal & Liu, 2006a,b; Bakshi et al., 2016]. Em comum, tais técnicas são dependentes da língua e voltadas para o idioma inglês. Apesar de existirem esforços voltados para outros idiomas, como árabe [El-Halees, 2012; Eldefrawi et al., 2019], chinês [Huang et al., 2008], vietnamita [Bach et al., 2015] e coreano [Yang & Ko, 2009, 2011], não há nenhum esforço que busca construir um sistema de detecção de expressões comparativas na língua portuguesa. O idioma português está entre os 10 mais falados no mundo [Souza et al., 2017] e, em especial, o Brasil, o maior país de língua portuguesa do mundo, representa um vasto mercado para comércio eletrônico que demanda soluções específicas para esse contexto. Este trabalho visa preencher essa lacuna, apresentando uma técnica para a detecção de sentenças comparativas na língua portuguesa.

Especificamente, o nosso trabalho apresenta a criação de um léxico[1] com palavras e expressões comparativas escritas em português, que é utilizado para encontrar opiniões comparativas, construindo dois conjuntos de dados de contextos diferentes, (1) sites de revisões/avaliações; e (2) Redes Sociais Online, que são posteriormente rotulados e utilizados para validação da abordagem supervisionada proposta para a classificação de sentenças comparativas.

Além disso, o nosso trabalho propõe-se uma estratégia automática baseada em algoritmos de aprendizado de máquina para a classificação de sentenças comparativas, categorizando-as em cinco classes: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. Os resultados apresentados são

[1]Léxico: é um conjunto de palavras existentes de uma determinada língua.

promissores, alcançando uma acurácia de 87% na tarefa de detecção das sentenças comparativas. Isso indica que, apesar dos desafios inerentes à língua portuguesa e similaridades existentes entre os tipos opinativos, é possível, dentre um conjunto de revisões, encontrar a maioria das comparações de maneira satisfatória, abrindo caminho para análises mais detalhadas acerca de comparações e predileções.

Finalmente, a partir da detecção das comparações, o trabalho apresenta um algoritmo para a análise das preferências expressas em uma opinião comparativa, sendo possível detectar a entidade indicada como a preferida em uma comparação.

1.1 Objetivos

Para o estudo das opiniões comparativas, é importante inicialmente compreender os diferentes tipos de comparações existentes, são eles: (1) Gradativa com Predileção; (2) Equitativa; (3) Superlativa; e (4) Não Gradativa [Liu, 2012]. Essas opiniões comparativas normalmente são encontradas próximas às opiniões regulares, separadas apenas por sentenças intercaladas em uma revisão.

Nesse contexto, um dos esforços fundamentais é a identificação, dentre um conjunto de revisões, de quais sentenças são comparativas e em quais categorias de opinião podem ser classificadas. Sendo assim, esta etapa inicial do trabalho tem por objetivo a distinção dos tipos específicos de opiniões, detalhados abaixo:

- Dado um conjunto de opiniões, tem por objetivo classificá-las nos dois tipos fundamentais de opiniões, são eles: opiniões regulares e opiniões comparativas.

- Em seguida, visto que as opiniões já foram distintas em regulares e comparativas, tem por objetivo analisar as opiniões comparativas em um nível menor de granularidade, categorizando essas comparações nos cinco diferentes tipos de opiniões, são eles: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa.

Após as opiniões serem categorizadas nos tipos específicos, é possível, então, direcionar esforços na aplicação de análises apropriadas para cada tipo de opinião. Especificamente para as opiniões comparativas, que é o foco deste trabalho, a classificação tradicional de sentimentos, que tem por objetivo classificar uma sentença em positiva, negativa ou neutra, não é aplicável [Liu, 2012]. Assim, o trabalho apresenta uma abordagem apropriada para análise das preferências expressas em uma opinião comparativa.

- As opiniões comparativas, principalmente as Gradativas com Predileção e Superlativas, comparam objetos expressando preferência a um deles. Essa informação é uma das mais importantes a serem extraídas, sendo o foco desta etapa, que é detectar qual a entidade preferida expressa em uma sentença comparativa.

De maneira geral, o objetivo do nosso trabalho é apresentar uma abordagem para estudo das opiniões comparativas na língua portuguesa, que consiste desde a detecção das opiniões comparativas até a análise minuciosa dessas estruturas para a extração de informações relevantes, úteis para a interpretação da opinião e do sentimento associado a cada uma das entidades ali presentes.

1.2 Contribuições

Esta Seção apresenta as quatro principais contribuições feitas para o estudo das opiniões comparativas na língua portuguesa:

- A construção de um léxico com palavras e expressões que são frequentemente utilizadas para se fazer comparações na língua portuguesa.

- A construção de duas bases de dados com sentenças comparativas em português, obtidas de dois importantes contextos online, que são: (1) sites de revisões/avaliações; e (2) Redes Sociais Online. Para o primeiro contexto, foi construído uma base de dados com 2.754 sentenças rotuladas. Já para o segundo, 2.053 sentenças foram rotuladas.

- Uma abordagem hierárquica é proposta para detecção de opiniões comparativas, onde inicialmente aplica-se a classificação binária, dividindo as sentenças nos dois tipos fundamentais de opiniões, as regulares e as comparativas. Em seguida, as opiniões classificadas como comparativas são categorizadas nos cinco tipos específicos, que são: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. O objetivo dessa estratégia é justamente separar as opiniões em grupos específicos, o que permite a aplicação de análises apropriadas e mais detalhadas para cada classe. Essa abordagem apresenta uma acurácia e Macro-F1 próximos a 87% na classificação binária, indicando que, apesar das peculiaridades da língua portuguesa, as comparações fazem uso de expressões que possibilitam diferenciá-las das opiniões regulares. Já para a classificação em múltiplas classes, percebe-se um desafio um pouco maior devido à semelhança existente entre os tipos comparativos, alcançando

67% de acurácia e 61% de Macro-F1. No entanto, a abordagem é capaz de agrupar as opiniões Gradativas com Predileção, Equitativas e Superlativas com taxas de revocação de até 85%, um resultado satisfatório para as comparações que realmente são relevantes para a análise de preferência proposta posteriormente.

- Por fim, é proposto um algoritmo para a análise das opiniões comparativas, utilizando características textuais de uma sentença para determinar a entidade preferida em uma comparação. Nesse contexto, a preferência é entendida como o ato de escolher um objeto em detrimento a outros, ou seja, é a indicação de superioridade de um objeto, por exemplo "O celular X é muito *melhor* do que o celular Y", onde o primeiro produto mencionado é apontado como preferido. O algoritmo proposto para detecção de preferência se mostra promissor, alcançando valores de acurácia de aproximadamente 94% de Macro-F1 para as Superlativas e aproximadamente 84% de Macro-F1 para as Gradativas com Predileção.

Em complemento, partes deste trabalho foi publicação em 2020 no Simpósio Brasileiro de Sistemas Multimídia e Web (**WebMedia**) [Kansaon et al., 2020].

1.3 Organização do Trabalho

O restante do trabalho está organizado conforme detalhado a seguir.

- **Capítulo 2 - Fundamentação Teórica.** Este capítulo apresenta os conceitos relacionados à análise de sentimentos no estudo de opiniões, que são descritas e detalhadas, apresentando as técnicas e trabalhos relacionados ao estudo das opiniões comparativas, que é o foco deste trabalho.

- **Capítulo 3 - Metodologia.** Este capítulo apresenta a metodologia utilizada para estudo das opiniões comparativas em português, que consiste na criação de um léxico, utilizado para a construção de duas bases de dados com sentenças comparativas que foram pré-processadas e rotuladas. Por fim, são analisados os possíveis impactos e limitações dessa abordagem.

- **Capítulo 4 - Classificação das Sentenças.** Neste capítulo, é apresentado a abordagem supervisionada utilizada para a classificação das opiniões, avaliando os resultados obtidos pelos algoritmos de aprendizado de máquina e discutindo os desafios existentes na língua portuguesa.

- **Capítulo 5 - Detecção de Preferência.** Este capítulo apresenta uma análise detalhada das opiniões comparativas por meio de um algoritmo proposto para detecção da entidade preferida em uma sentença.
- **Capítulo 6 - Conclusão e Trabalhos Futuros.** Por fim, é apresentada a conclusão deste trabalho, ressaltando as principais contribuições e caminhos para os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O trabalho apresenta a análise de sentimentos como uma ferramenta para o estudo de opiniões, mais especificamente as opiniões comparativas. Visto isso, este capítulo contextualiza o cenário atual da área de análise de sentimentos apresentando conceitos e definições relacionados aos tipos de opiniões existentes, bem como estratégias de classificação de texto e trabalhos relacionados que apresentam esforços no estudo das opiniões comparativas.

2.1 Análise de Sentimentos

A análise de sentimentos, também conhecida como mineração de opinião, é uma área de estudo que tem por objetivo extrair de forma automática opiniões, sentimentos e emoções expressas em um texto [Liu, 2012; Tsytsarau & Palpanas, 2012]. As pesquisas voltadas para análises de sentimentos e opiniões ganharam grande força por volta dos anos 2000 [Morinaga et al., 2002; Pang et al., 2002]. Porém, o termo análise de sentimentos só passou a ser utilizado a partir do trabalho desenvolvido por Nasukawa & Yi [2003], que apresentou uma abordagem para extração de sentimentos ligados a assuntos mais específicos, focando primeiramente em detectar quais sentimentos estão sendo expressos nas sentenças, para em seguida, verificar suas polaridades (positiva ou negativa). Já o termo análise de opinião, apareceu em um período próximo [Dave et al., 2003], nesse trabalho foram identificados propriedades e atributos para o desenvolvimento de um método utilizando técnicas de recuperação de informação com o objetivo de distinguir automaticamente revisões positivas e negativas.

Desde então, a análise de opiniões e sentimentos se tornou uma área de estudo muito atrativa [Liu, 2012]. Por ter uma aplicação comercial muito útil em processos de avaliação de opiniões, compreensão de sentimento de clientes e até mesmo geren-

ciamento de reputações, a análise de sentimentos se tornou uma grande motivação de pesquisa. Além disso, com o surgimento das mídias sociais e da Web, tem-se um grande volume de dados opinativos, que possibilitam o avanço de pesquisas e o desenvolvimento de técnicas para análise de sentimentos.

Dentro da área, existem vários outros termos que se compõem em pequenas diferentes tarefas que fazem parte da análise de sentimentos, como a mineração de opinião, extração de opinião, mineração de sentimentos, análise de subjetividade, análise de emoções, mineração de revisões [Liu, 2012]. Esses problemas de pesquisa ainda podem ser divididos em três níveis de granularidade [Liu, 2012], ou seja, quanto menor o nível de granularidade mais específica a análise, são eles:

- **Documento**: Neste nível, a classificação ocorre através da análise se as opiniões contidas em um documento expressam um sentimento positivo ou negativo. Considerando a revisão de um produto, o nível documento fornece uma visão geral das opiniões ali presentes, assumindo que o documento possui opiniões acerca de uma única característica de um produto.

- **Sentença**: Já este nível tem um objetivo mais específico, de determinar se uma determinada sentença contida em documento expressa um sentimento positivo, negativo ou até mesmo neutro. As análises são feitas em cada sentença individualmente, assumindo que cada frase possua um sentimento a ser classificado.

- **Aspecto ou Entidade**: Apesar dos níveis anteriores identificarem o sentimento expresso, eles não detalham exatamente qual o sentimento relacionado a cada aspecto mencionado. Considerando a sentença: "Este celular tem uma excelente câmera, porém, sua bateria deixa muito a desejar", percebe-se a existência de diferentes sentimentos para cada aspecto. O aspecto câmera é avaliado positivamente, já para a bateria, é expresso um sentimento negativo. Logo, este nível visa identificar o sentimento expresso para cada aspecto de uma entidade.

Para o estudo de opiniões neste trabalho, estamos interessados no estudo a nível de sentenças, uma vez que um texto pode ser composto por diversas frases opinativas. Logo, para o estudo específico deste trabalho, tratar cada sentença individualmente se torna o mais apropriado.

2.2 Definições e Terminologias

Na análise de sentimentos, uma opinião é considerada como uma composição de alguns elementos, como o alvo da opinião, seus aspectos e sentimentos atrelados a eles, que são

expressos por um formador de opinião em um determinado momento [Liu, 2012]. De maneira geral, essas opiniões podem ser classificadas entre opiniões regulares e opiniões comparativas [Jindal & Liu, 2006a,b].

1. **Opinião Regular**: A opinião regular expressa um sentimento a um aspecto de uma determinada entidade e pode ser subdividida em dois tipos:

   a) **Diretas**: Essas opiniões expressam um sentimento direto a uma entidade ou um aspecto, por exemplo, "A bateria desse celular é excelente".

   b) **Indiretas**: Expressam uma opinião sobre entidades de maneira indireta, geralmente manifestada pela consequência ou efeito de algumas entidades. Na sentença "Após dirigir este novo carro por um longo período, passei a ter dores lombares", percebe-se que a crítica feita ao carro e o sentimento negativo relacionado ao assento é demonstrado de maneira indireta, fazendo com que essas opiniões sejam mais complexas para análises.

2. **Opinião Comparativa**: As opiniões comparativas contrastam dois ou mais objetos expressando uma relação de semelhança ou diferença entre eles, por exemplo, "O celular x é bem *melhor* que o celular Y".

Os sentimentos e emoções contidos em opiniões, na maioria das vezes, são expressos através de visões subjetivas inerentes à pessoa que expressa a determinada opinião. Nesse contexto, uma opinião regular ou comparativa também pode ser classificada como subjetiva quando há uma visão individual e particular de uma pessoa a um determinado objeto, por exemplo "Achei a torta de morango muito saborosa". Por outro lado, quando o sentimento é indicado de maneira implícita, as opiniões podem ser classificadas como objetivas ou implícitas, pois são declarações feitas através de fatos expressos, por exemplo "Ontem comprei um celular, mas hoje já tive que levar para a manutenção".

No trabalho, iremos abordar o estudo das opiniões comparativas a nível de sentença, isso porque uma opinião pode ser formada por um conjunto de sentenças que, por sua vez, podem ser categorizadas em tipos diferentes de classes de opinião, sendo mais apropriado o estudo específico de cada uma das sentenças. A Seção 2.2.1 detalha os tipos de opiniões comparativas existentes.

2.2.1 Opiniões comparativas

As opiniões comparativas podem ser classificadas em dois principais grupos [Liu, 2012]: (1) comparações gradativas, que expressam relação de ordem entre as entidades compa-

radas na sentença, podendo ser de semelhança ou de superioridade; e (2) comparações não gradativas, que comparam objetos sem indicar ordem entre eles. Dentro desses grupos, encontram-se quatro categorias mais específicas que separam os tipos diferentes de opiniões comparativas [Liu, 2012; Jindal & Liu, 2006a,b]. As três primeiras fazem parte das comparações gradativas, já a última das não gradativas.

- **Gradativa com Predileção:** Contém ao menos duas entidades expressando predileção e ordem de uma em relação à outra, por exemplo, "O carro X é *melhor* que o carro Y". Eldefrawi et al. [2019] propõem uma técnica não supervisionada que utiliza a estrutura linguística dessas comparações na língua árabe para identificar a entidade preferida. Da mesma forma, Gupta et al. [2017] utilizam opiniões comparativas com predileção em um sistema que objetiva identificar e extrair entidades da literatura biomédica.

- **Equitativa:** Existem duas entidades na qual a relação entre elas é de igualdade baseada em algum aspecto, por exemplo, "A câmera do smartphone X é *igual* ao Y". Esse tipo de comparação também é considerado por Gupta et al. [2017] no estudo de texto biomédico, mas não é usado por Eldefrawi et al. [2019], porque esse tipo comparativo não foi encontrado no conjunto de dados em árabe. Além disso, Ramırez & Sánchez [2016] propõem uma abordagem de análise de sentimentos para identificar o gênero dos usuários através de suas opiniões no Twitter. No estudo, a quantidade de comparações Equitativas encontradas é pequena, 3,70% das sentenças postadas por homens e 1,28% por mulheres.

- **Superlativa:** Uma entidade possui relações do tipo maior ou menor que um grupo de outras, por exemplo, "Este é o *melhor* laptop do mundo". Esse tipo de sentença é considerado por Eldefrawi et al. [2019] no estudo de opiniões comparativas em árabe e por Ramırez & Sánchez [2016] na investigação de quais sentenças foram feitas por homens ou mulheres. No entanto, Gupta et al. [2017] não considera sentenças Superlativas, porque as entidades comparadas em textos biomédicos raramente são mencionadas em uma única sentença.

- **Não Gradativa:** Compara duas ou mais entidades, mas não expressa ordem nem predileção por nenhuma, por exemplo, "O design do laptop X possui alguns recursos diferentes do laptop Y". Além de não expressarem ordem entre os objetos, essas comparações são mais difíceis de serem detectadas, pois são mais sutis e não possuem um padrão claro [Liu, 2012].

De maneira geral, os estudos que lidam com as opiniões comparativas concentram seus esforços nas comparações que expressam relação de preferência entre os objetos, como as Gradativas com Predileção e Superlativas, dando menos importância para as Não Gradativas, uma vez que não apresentam relação entre os objetos [Liu, 2012]. No entanto, apesar das peculiaridades existentes nas diferentes classes de opinião, o trabalho aqui apresentado estuda todos esses tipos comparativos, propondo uma estratégia hierárquica para a detecção de sentenças comparativas, classificando-as nos quatro diferentes tipos comparativos.

2.3 Classificação de Texto

Dentro da classificação de texto, a detecção de sentimentos é um dos problemas abordados, que diferentemente das estratégias de classificação de documentos em tópicos, tem por objetivo classificar textos em classes de sentimentos positivas, negativas ou até mesmo neutras [Serrano-Guerrero et al., 2015; Melo et al., 2019; de O. Carosia et al., 2019; Mehta et al., 2020]. Ainda dentro dessa área, alguns outros estudos buscam a classificação de texto em classes de opiniões, regulares e comparativas [Jindal & Liu, 2006a,b; Park & Blake, 2012; Bach et al., 2015], que é a tarefa principal abordada neste trabalho. Em todos esses casos, apesar de serem tarefas com foco em tipos diferentes de classes, todas elas compartilham da utilização de técnicas semelhantes, uma vez que fazem parte da classificação de texto.

As abordagens de aprendizado de máquina, principalmente as estratégias supervisionadas, são frequentemente utilizadas como uma solução para problemas de classificação de textos [Tsytsarau & Palpanas, 2012], onde um determinado algoritmo (i.e., modelo) aprende padrões através de um conjunto de treinamento e, em seguida, ocorre a etapa de generalização, onde o modelo treinado é utilizado para a classificação dos dados ainda não vistos. Nesse contexto, pela simplicidade de implementação e capacidade de predição, nota-se a frequente utilização de algoritmos supervisionados como o Support Vector Machine (SVM) [Hartmann et al., 2019; Kowsari et al., 2019] e Naive Bayes (NB) [Sun et al., 2017; Liu, 2012] para problemas de classificação de texto. Além disso, o Logistic Regression (LR) [Kocoń et al., 2019; Prabhat & Khullar, 2017] e o Random Forest (RF) [Al Amrani et al., 2018; Selvi et al., 2017] também têm se mostrado promissores.

- **Support Vector Machine (SVM)**: É um algoritmo de aprendizado de máquina que introduz o conceito de margem, onde as classes predeterminadas são separadas em um hiperplano através de uma margem encontrada pela minimi-

zação dos erros [Sebastiani, 2002]. Normalmente, o SVM funciona bem para problemas de classificação de texto, isso se dá devido à capacidade do algoritmo em lidar com alta dimensionalidade e também porque a maioria dos problemas de classificação de texto é linearmente separável [Joachims, 1998].

- **Naive Bayes (NB)**: É um algoritmo probabilístico de classificação baseado no teorema de Bayes, que assume a independência entre as *features*[1], permitindo uma alta performance em grandes volumes de dados [McCallum et al., 1998]. Além disso, existem outras versões desse modelo, como o Complement Naive Bayes, Bernoulli Naive Bayes e Multinomial Naive Bayes. Os dois últimos apresentam distribuições específicas para as *features* [Rennie et al., 2003], o que pode ser interessante para a aplicação em contextos específicos.

- **Logistic Regression (LR)**: A regressão logística é uma técnica estatística utilizada no aprendizado de máquina por um algoritmo supervisionado de classificação, que através da análise dos coeficientes e dos valores fornecidos como entrada, determina a classe correta [Kleinbaum et al., 2002].

- **Random Forest (RF)**: O modelo utiliza o conceito de *enssemble*, que é a combinação do resultado de um conjunto de modelos de aprendizado de máquina, no caso do Random Forest, de árvores de decisões, para definir o resultado final da classificação, que normalmente é determinado pela média do resultado de cada modelo [Breiman, 2001].

No aprendizado de máquina, especialmente nas estratégias supervisionadas para classificação de texto, diferentes combinações de *features* podem ser utilizadas para treinamento dos modelos. Nesse contexto, as características presentes nos textos são extraídas para a utilização de um modelo, dentre elas, algumas se destacam:

- **Term Frequency – Inverse Document Frequency (TF-IDF)**: Pode ser entendida como uma medida que representa a importância de um termo de um documento em um corpus [Rajaraman & Ullman, 2011]. *Term Frequency* (TF) se refere à frequência de uma palavra em um documento, enquanto *Inverse Document Frequency* (IDF) se refere à frequência que o termo aparece em todos os documentos. Dessa maneira, palavras que se repetem frequentemente nos documentos, como preposições, artigos, conjunções, recebem um valor inferior se

[1] *features*: em aprendizado de máquina, as *features* são propriedades ou características dos dados analisados. Essas *features* são utilizadas pelos algoritmos de aprendizado de máquina para encontrar padrões que permitam a distinção das classes predefinidas.

comparadas às palavras que são mais relevantes para recuperação de informação. Na classificação de texto, é comum calcular essa métrica para as palavras existentes em uma sentença, porém, também pode ser obtido um valor para combinações de termos, como unigrama, bigrama, trigrama, etc.

- **N-gramas**: É uma sequência consecutiva de palavras com um tamanho n [Liu, 2007], podendo ser unigrama, bigrama, trigrama, etc. Considerando a sentença "este laptop é o *melhor* de todos", é possível formar 5 trigramas, "este laptop é", "laptop é o", "é o melhor", "o melhor de", "melhor de todos". Além da combinação de palavras, é possível também formar n-gramas de caracteres seguindo a mesma lógica. Dessa maneira, métricas como TF-IDF podem ser calculadas para obter a relevância dessas combinações de termos, possibilitando novas combinações de *features* para classificação de texto.

- **Part of Speech (PoS) Tagging**: Também conhecida como marcação gramatical, (PoS) Tagging é a análise morfológica de um texto onde cada palavra recebe uma marcação referente à sua classe gramatical, por exemplo: substantivo, verbo, pronome, preposição, entre outras. No texto: "este carro é bonito", percebe-se as seguintes classes gramaticais: "o [artigo] carro [substantivo] é [verbo] bonito [adjetivo]".

- **Dependency Parsing**: No processamento de linguagem natural (NLP), a análise de dependência se refere à análise da estrutura gramatical de uma sentença através de uma árvore de dependências, que indica as relações sintáticas entre as palavras existentes no texto [Liu, 2015].

- **Polaridade**: É um valor referente à orientação de sentimento associado a um determinado termo [Taboada et al., 2011]. Normalmente a polaridade é quantificada como positiva (+1) ou negativa (-1) e pode ser obtida para uma palavra ou até mesmo para uma sentença.

Existem também as abordagens não supervisionadas, que diferente das supervisionadas, não necessitam de um conjunto de dados para treinamento e construção de um modelo, o que simplifica a sua aplicação. Os métodos baseados em abordagens léxicas são exemplos comuns dessas estratégias não supervisionadas, pois possuem um dicionário de sentimentos de palavras [Taboada et al., 2011; Ribeiro et al., 2016], que normalmente é utilizado para classificação de sentimentos em textos sem a necessidade de treinamento de um modelo [Serrano-Guerrero et al., 2015].

As abordagens léxicas, além de possuírem uma aplicação prática, podem ser utilizadas de maneira combinada com estratégias supervisionadas. Neste trabalho, por exemplo, um léxico com palavras comparativas é construído inicialmente para mineração de opiniões visando encontrar por comparações e, em seguida, a estratégia é aprimorada com a utilização de uma abordagem supervisionada para classificação das opiniões nas diversas classes opinativas.

2.4 Trabalhos Relacionados ao Estudo de Opiniões

No estudo de opiniões, diferentes técnicas podem ser utilizadas para a extração de informações, e a escolha da técnica apropriada depende das características dos dados analisados e do tipo de informação que se deseja extrair. Nesse contexto, muitos estudos se concentram no desenvolvimento de técnicas de análise de sentimentos para a classificação dos sentimentos e/ou emoções expressos em um texto [Alaei et al., 2019; Eldefrawi et al., 2019; Melo et al., 2019; Ribeiro et al., 2016; Araújo et al., 2016], o que é muito útil quando se deseja entender qual o sentimento relacionado a um determinado evento, produto ou até mesmo um assunto.

A sumarização de texto é outra técnica que pode ser aplicada para o estudo de opiniões, fornecendo uma visão concisa de um grande conjunto de revisões. Através das entidades e aspectos mencionados em uma revisão, a sumarização pode organizar os textos de diferentes maneiras, contrastando as opiniões ou até mesmo separando em positivas e negativas [Lerman & McDonald, 2009; Kawahara et al., 2010; Paul et al., 2010; Ren & de Rijke, 2015; Ibeke et al., 2017; Kim & Zhai, 2009].

Já a clusterização, pode ser utilizada com o objetivo de agrupar as revisões que possuem características similares, tal como as revisões que se referem a um determinado aspecto de um produto [Kunneman et al., 2018; Chen & Wang, 2013; Tata & Di Eugenio, 2010; Wang et al., 2010].

Visto isso, o foco do nosso trabalho está no estudo das opiniões comparativas na língua portuguesa, especificamente nas sentenças que comparam diferentes objetos expressando semelhanças e preferências. Nesse contexto, as técnicas tradicionais de análise de sentimentos, principalmente a detecção de sentimentos, não são suficientes para o estudo das opiniões comparativas, uma vez que as comparações não possuem um sentimento direto que possa ser classificado como positivo ou negativo. Considerando a sentença "o celular x é *melhor* que o celular y", percebe-se a existência de outras questões mais importantes para serem respondidas, como quais objetos são comparados e qual entidade é a preferida.

O número de estudos que lidam com a mineração de sentenças comparativas é pequeno e é ainda menor quando se trata das sentenças em português, que é a segunda língua mais usada no Twitter e está entre as dez mais faladas no mundo [Souza et al., 2017]. Souza et al. [2017] realizaram uma revisão sistemática sobre mineração de texto em português e observaram que a maioria dos estudos anteriores se concentram na classificação de texto tradicional, o que reforça a lacuna existente em estudos de sentenças comparativas.

Para o estudo de comparações, a tarefa fundamental é a distinção, dentre um conjunto de revisões, das opiniões comparativas das regulares, pois a partir dela as técnicas apropriadas podem ser direcionadas para o estudo de cada classe de opinião. Nesse contexto, abordagens léxicas vêm sendo utilizadas por várias soluções para a detecção de sentenças comparativas em diferentes idiomas, como inglês [Jindal & Liu, 2006a,b; Park & Blake, 2012], árabe [El-Halees, 2012; Eldefrawi et al., 2019], chinês [Huang et al., 2008; Liu et al., 2013], vietnamita [Bach et al., 2015] e coreano [Yang & Ko, 2009, 2011]. Apesar das possíveis limitações das abordagens léxicas, o estudo feito por [Jindal & Liu, 2006a] mostra que grande parte das sentenças comparativas fazem uso de um conjunto mapeável de palavras para expressar comparações entre objetos, o que viabiliza a utilização de um léxico para encontrar comparações, sendo possível capturar grande parte de todos os tipos comparativos existentes [Jindal & Liu, 2006a,b].

Por outro lado, alguns estudos direcionam esforços no desenvolvimento de técnicas para a análise específicas das comparações, como a detecção da entidade preferida em uma comparação [Ganapathibhotla & Liu, 2008; Ding et al., 2008, 2009; Eldefrawi et al., 2019]. A tarefa de determinar se uma sentença possui um sentimento positivo ou negativo sem indicar qual entidade esse sentimento está associado é pouco informativa. Assim, a detecção da entidade preferia é uma das principais análises que possibilita a interpretação de uma comparação, entendendo qual entidade é indicada como superior ou inferior. Por meio das características textuais, principalmente da palavra utilizada para expressar comparação, é possível determinar qual sentimento está associado a cada entidade, viabilizando, então, a detecção da entidade preferida ou superior em uma determinada sentença.

De maneira geral, os estudos que abordam comparações esbarram em soluções específicas que se restringem a determinados idiomas, como a criação de léxicos e tratamento de peculiaridades existentes na língua [Jindal & Liu, 2006a,b; Yang & Ko, 2009; Eldefrawi et al., 2019]. Além disso, grande parte desses estudos se concentram nas tarefas fundamentais de classificação de comparações [Jindal & Liu, 2006a; Yang & Ko, 2011; Bach et al., 2015; El-Halees, 2012; Huang et al., 2008; Yang & Ko, 2009] e análise

das preferências [Ganapathibhotla & Liu, 2008; Ding et al., 2008, 2009; Eldefrawi et al., 2019].

A principal diferença entre esses estudos está justamente nas peculiaridades existentes no idioma tratado. Contudo, percebe-se também que os objetivos e tipos comparativos tratados variam dependendo do contexto e do foco apresentado em cada trabalho. Alguns estudos se concentram em tarefas de classificação binária [Jindal & Liu, 2006a,b; Park & Blake, 2012; El-Halees, 2012; Huang et al., 2008; Yang & Ko, 2009]. Já outros se restringem apenas a alguns tipos comparativos, como Superlativos e Gradativos com Predileção [Eldefrawi et al., 2019; Ding et al., 2009; Bach et al., 2015]. Mais a fundo, surgem estudos com soluções para a análise e extração de informações em sentenças comparativas [Eldefrawi et al., 2019; Chen et al., 2017; Gao et al., 2018]. Em suma, todas essas soluções são complementares, e as diferenças existentes entre as abordagens utilizadas passam pelas decisões e foco dado em cada um dos trabalhos.

Nesse contexto, apesar de existirem esforços em vários idiomas, como inglês [Jindal & Liu, 2006a,b; Park & Blake, 2012], chinês [Huang et al., 2008; Liu et al., 2013], entre outros [Eldefrawi et al., 2019; Bach et al., 2015; Yang & Ko, 2009, 2011], o português tem sido pouco estudado e negligenciado pelas técnicas tradicionais. O que evidencia a necessidade da análise da viabilidade de abordagens apropriadas para o estudo de comparações na língua portuguesa, bem como a construção de técnicas que possibilitem o tratamento e extração de informações úteis sobre preferências.

Lacuna de pesquisa.

Nosso esforço é complementar aos trabalhos anteriores, por ser o primeiro estudo a explorar a análise comparativa na língua portuguesa. Particularmente, nosso trabalho cria um léxico de palavras e expressões frequentemente utilizadas para se fazer comparações na língua portuguesa, de maneira semelhante ao que outros estudos realizaram para o inglês [Jindal & Liu, 2006a,b]. Mais importante, nosso estudo apresenta uma estratégia para ampliação desse léxico inicial para diferentes contextos, agregando verbos, advérbios e expressões regulares.

Em complemento, uma estratégia supervisionada é proposta para classificação automática de cada classe de opinião comparativa. Além disso, é proposto uma abordagem para análise das opiniões que expressam relação de superioridade ou inferioridade entre os objetos comparados, determinando através dos aspectos mencionados e da palavra utilizada para comparação qual a entidade preferida na sentença.

Capítulo 3

Metodologia

Neste Capítulo, como mostrado na Figura 3.1, apresenta-se a metodologia experimental adotada para o estudo das opiniões comparativas, que envolve a construção de um léxico na língua portuguesa (Seção 3.1), que é avaliado por meio da análise de suas principais características e limitações (Seção 3.2), para, em seguida, ser utilizado na construção das bases de dados comparativas (Seção 3.3). Por fim, são apresentadas as etapas de pré-processamento e rotulação dos dados (Seção 3.4), que serão utilizados nos próximos capítulos para o treinamento de abordagens supervisionadas na classificação das opiniões (Capítulo 4) e pelo algoritmo de detecção de preferências (Capítulo 5).

3.1 Construção do Léxico com Palavras Comparativas

Apesar da comparação ser uma figura de linguagem comumente usada para indicar preferências e semelhanças entre elementos [Liu, 2012], grande parte das revisões encontradas na Web e em fóruns de discussão são compostas por opiniões regulares, ou seja, não possuem comparação. Portanto, se faz necessário a utilização de uma estraté-

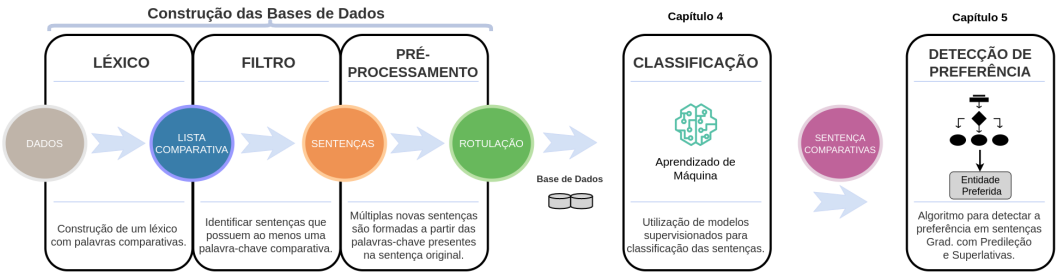


Figura 3.1: **Metodologia proposta para estudo das opiniões comparativas.**

gia para encontrar dentre todas as revisões existentes, apenas as que são comparativas, pois caso as revisões fossem obtidas indiscriminadamente, muitas opiniões não comparativas seriam encontradas, o que poderia tornar custoso o processo de mineração.

Ao observar as sentenças comparativas, nota-se a existência de um grupo restrito de palavras e expressões que é constantemente utilizado para fazer comparações, motivando a criação de um léxico em português. Esse conjunto de palavras é capaz de cobrir a maioria das comparações feitas em português. Assim, a partir da análise de frases comparativas em português e de léxicos da língua inglesa [Jindal & Liu, 2006a], um novo léxico com palavras frequentemente usadas em comparações na língua portuguesa foi construído, contendo verbos (e.g., *ganha*, *supera* e *ultrapassa*), advérbios (e.g., *mais* e *menos*), adjetivos (e.g., *melhor*, *pior* e *semelhante*) e expressões comuns na língua portuguesa (e.g., *tão bom quanto*, *fica para trás*), totalizando 59 palavras-chave comparativas.

No português, algumas comparações podem ser classificadas como não-oracionais, ou seja, o trecho comparativo da sentença não possui verbo [Rodrigues, 2002]. Nesses casos, a parte comparativa é conectada à primeira parte da sentença através do uso de conjunções comparativas (e.g., *como*, *que*, *etc*). Na sentença: "você corre *mais* **do que ele**", a parte comparativa (do que ele) não possui nenhum verbo ou palavra comparativa, mas é conectada por meio da locução conjuncional *do que*, indicando que é o aspecto *correr* que está sendo comparado.

Entretanto, o fato da parte comparativa não possuir um verbo não impede que a abordagem baseada em palavras-chave capture essas comparações. Geralmente, essas conjunções comparativas são antecedidas por palavras como: *mais*, *menos*, *maior*, *menor*, *melhor*, entre outras [da Rocha Lima, 2017], que estão inseridas no conjunto de palavras. Observando o exemplo "você corre *mais* **do que ele**", mesmo que a parte comparativa não possua um verbo, a comparação pode ser encontrada através da palavra-chave *mais*, que é referente à expressão comparativa "corre mais".

Embora o léxico contenha grande parte das palavras comparativas, os diferentes ambientes online possuem peculiaridades que podem não ser englobadas pelo conjunto inicial de palavras, motivando a ampliação desse léxico. Nosso trabalho explora dois importantes contextos opinativos existentes no ambiente online, que são: (1) sites de revisões/avaliações; e (2) Redes Sociais Online. Assim, o léxico é expandido através da inclusão de novas palavras comparativas encontradas nesses contextos.

Para o contexto de avaliações online, escolheu-se o Buscapé[1], um dos principais sites brasileiros utilizado para busca de produtos e pesquisa de preços em lojas online.

[1]Site do Buscapé: `https://www.buscape.com.br/`. Acesso em: 01 de fevereiro de 2021.

Tabela 3.1: 5 palavras-chave comparativas mais frequentes.

Dados	**Palavras-Chave**				
Buscapé	mais	como	recomendo	melhor	comprei
Twitter	mais	como	queria	melhor	parece

Tabela 3.2: 5 palavras-chave comparativas mais precisas.

Dados	**Palavras-Chave**				
Buscapé	incomparável	líder	idêntico	assemelha	supera
Twitter	incomparável	similar	idênticos	assemelha	preferível

Em tal site também é possível fazer revisões acerca de produtos que foram comprados, e as opiniões nessas plataformas se restringem apenas a produtos e marcas. Para a ampliação do léxico com novas palavras e expressões comparativas, foram lidas manualmente as revisões dos produtos mais comentados. Ao todo, foram encontradas 107 novas palavras-chave comparativas.

Ademais, o Twitter é a plataforma escolhida para o estudo de comparações em Redes Sociais Online por ser uma das principais redes opinativas, constituída em sua grande parte por textos que expressam opiniões sobre marcas, produtos e também comentários a respeito de variados assuntos, e está entre as seis Redes Sociais Online mais usadas no Brasil[2]. Para otimização do processo de construção do léxico, foram selecionadas as 35 marcas mais valiosas do Brasil e do mundo[3]. Para cada marca selecionada, foi adicionado um concorrente, obtendo um total de 70 marcas. Em seguida, filtrando por *tweets* que contenham ao menos uma marca, procurou-se manualmente por novas palavras comparativas, encontrando 10 novas palavras-chave comparativas que não foram encontradas no Buscapé, formando assim um léxico único com 176 palavras (ver Tabela A.1 no Apêndice A).

Embora as palavras-chave possam ser usadas em comparações, nem sempre são utilizadas com esse único objetivo. A Tabela 3.1 apresenta as 5 palavras-chave comparativas mais encontradas em cada contexto. Tais palavras possuem um uso prático muito amplo que nem sempre está relacionado à comparação, e.g. "não quero *mais* o produto". Isso justifica a grande ocorrência dessas palavras, apesar de nem sempre serem comparativas. Na Tabela 3.1, as palavras-chave: *mais*, *como* e *melhor* são comuns em ambos contextos, porém, palavras como: *recomendo* e *comprei* são mais frequentes

[2]Ranking de uso das Redes Sociais Online: `https://wearesocial.com/blog/2020/04/digital-around-the-world-in-april-2020`. Acesso em: 01 de fevereiro de 2021.

[3]Site BrandZ: `https://brandz.com/Global`. Acesso em: 01 de fevereiro de 2021.

em revisões online, o que nos indica uma diferença existente entre os contextos. Para as avaliações online, é frequente o uso de expressões referentes à decisão de compra e preferências por um produto, porém o mesmo não ocorre no Twitter, que possui opiniões sobre variados assuntos, como esportes, eventos, pessoas, entre outros.

Por fim, existem as palavras-chave que são pouco frequentes, mas precisas, ou seja, quando uma sentença possui tal palavra é provável a presença de comparação. A Tabela 3.2 apresenta tais palavras, que são recorrentes nos diferentes contextos. É possível encontrar exemplos como: *incomparável, preferível, líder e supera*, que são palavras utilizadas para indicar preferência por algum objeto, e *similar, assemelha e idêntico*, que retratam aspectos de similaridades entre objetos.

3.2 Avaliação da Abordagem Léxica

A utilização de palavras e expressões para encontrar por opiniões comparativas é uma das várias aplicabilidades do léxico construído. Porém, essa estratégia possui limitações que, por sua vez, podem deixar de fora algumas frases comparativas. Sendo assim, nesta Seção é realizado um estudo sobre a viabilidade do léxico em relação a sua capacidade de encontrar por comparações através das palavras-chave, onde calcula-se métricas como precisão e revocação.

A decisão de utilizar a estratégia de palavras-chave, que consiste na busca de sentenças que possuem ao menos uma das palavras presentes no léxico, passa pela avaliação se esse conjunto de palavras é suficiente para capturar a expressiva maioria das comparações, ou seja, é necessário quantificar a capacidade do léxico em encontrar comparações em um cenário prático, viabilizando a sua utilização.

Para a avaliação, a revocação se mostra uma das métricas úteis para esse tipo de análise, pois indica a fração de sentenças comparativas recuperadas dentre todas as comparações existentes em um determinado contexto. Considere, por exemplo, que existam dez comparações dentre todas as revisões de um determinado laptop e, após a aplicação da estratégia de palavras-chave, encontra-se apenas oito dentre as dez comparações existentes. Nesse caso, pode-se afirmar que a abordagem apresentou uma taxa de revocação igual a 80% (i.e., 8 dividido por 10).

Em complemento, existe a métrica de precisão, que diferente da revocação, indica a fração de sentenças que são, de fato, comparativas dentre todas as sentenças indicadas como comparativas. Considere, por exemplo, que existam 100 sentenças avaliando um determinado produto, sendo que apenas dez são comparativas. Assim, ao aplicar a estratégia de palavras-chave, trinta sentenças foram apontadas como comparativas,

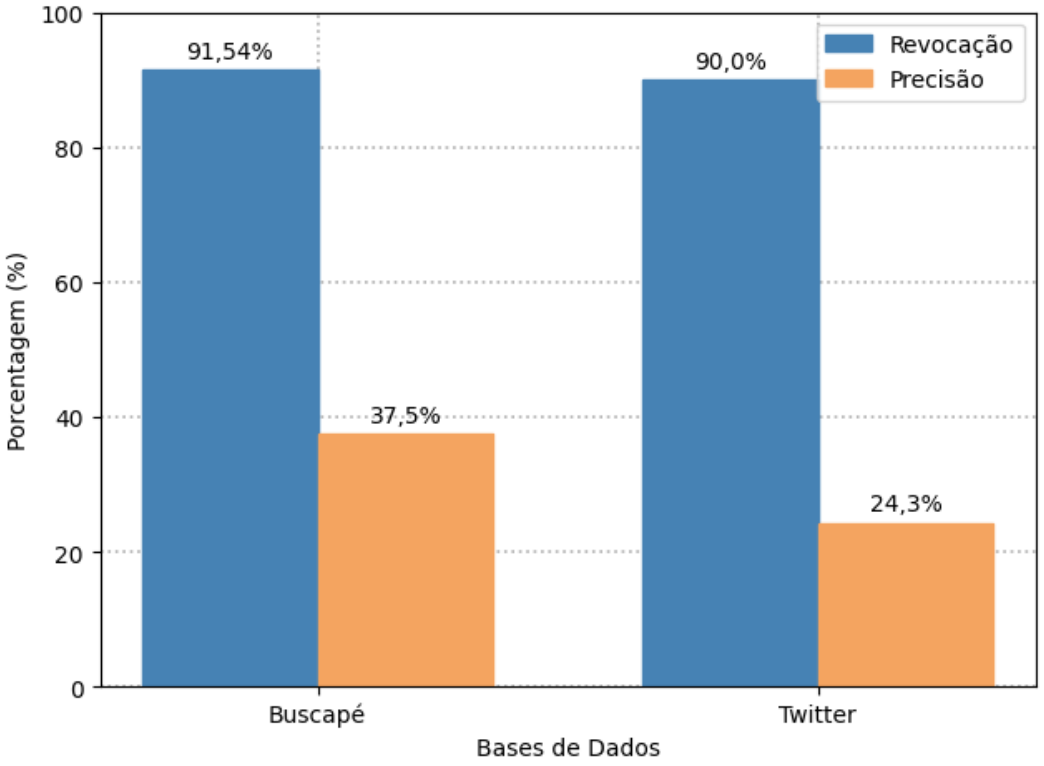


Figura 3.2: **Métricas de Revocação e Precisão obtidas na validação da abordagem léxica.**

porém apenas nove desse grupo de trinta sentenças são realmente comparativas. Logo, pode-se dizer que a estratégia apresentou uma precisão de 30% (i.e., 9 dividido por 30).

3.2.1 Experimento

Para a avaliação dessas métricas, é necessário que as sentenças do contexto analisado sejam rotuladas em sua totalidade. No entanto, devido ao grande volume de dados existente nos dois contextos abordados no trabalho, foi necessário a construção de uma amostra, rotulando todas as sentenças referentes a um dentre os vários produtos avaliados. Vale ressaltar que as sentenças que fizeram parte da etapa de expansão do léxico não estão presentes nessa amostra utilizada.

Inicialmente, foram rotuladas todas as sentenças das revisões feitas para o produto *Smartphone LG Nexus 4* na base de dados do Buscapé[4], totalizando 312 sentenças, sendo que apenas 71 são comparativas.

Já para o Twitter, por não possuírem uma especificação do produto avaliado,

[4]Amostra do Buscapé: A amostra foi construída a partir da mesma base de dados utilizada para encontrar por comparações na Seção 3.3.

buscou-se por *tweets* em português que continham a palavra *Iphone* e que foram postados no Brasil na data 10/02/2019. Ao todo, 309 sentenças foram rotuladas, sendo que apenas 20 dessas sentenças são comparativas.

Com os dois conjuntos de dados preparados, pode-se iniciar a avaliação das métricas de precisão e revocação. Logo, aplicando a estratégia de buscar apenas as sentenças que contêm ao menos uma das palavras-chave existentes no léxico, observa-se uma taxa de revocação de 91,5% para o Buscapé e 90% para o Twitter, o que significa que essa abordagem, de fato, é capaz de capturar a grande parte das comparações feitas, já que essas comparações fazem uso das palavras existentes no léxico para expressar comparação. Por outro lado, avaliando a métrica de precisão, percebemos um valor de 37,5% para o conjunto de sentenças do Buscapé e 24,3% para o Twitter, como mostrado na Figura 3.2.

3.2.2 Análise dos resultados e avaliação do léxico

Analisando de maneira conjunta os valores de precisão e revocação, é possível concluir que a abordagem léxica pode ser muito útil na tarefa de encontrar comparações, sendo favorável às estratégias de mineração que tem como foco detectar comparações em um amplo conjunto de revisões.

No entanto, observando a métrica de precisão, percebe-se que a utilização exclusiva do léxico pode ser insuficiente em alguns cenários, visto que a abordagem apresenta um número considerável de falsos positivos, pois nem todas sentenças capturadas são, de fato, comparativas.

Contudo, isso não denota um problema, uma vez que a abordagem léxica pode ser aprimorada através de soluções de aprendizado de máquina, como é detalhado no Capítulo 4. Além disso, a abordagem léxica tem ampla aplicabilidade nesse problema de mineração de texto, pois minimiza o esforço em processos de rotulação e construção de bases de dados, que posteriormente podem ser utilizadas para treinamento de modelos de aprendizado de máquina.

Observando os casos de exceção, percebe-se que o fato de uma comparação não ser construída através da utilização de palavras-chave dificulta a captura proposta por meio do léxico, uma vez que não existe um elemento que explicitamente indique o contraste feito entre objetos, como "Você raramente verá um *lag* nesse produto (muito frequente nos aparelhos da marca X)". Nesse exemplo, observa-se uma crítica feita aos aparelhos da marca X através de uma comparação. Já no exemplo: "O Nexus 4 deveria estar no intervalo de preço de 700-1300, já o Nexus 5 entre 1500-2000", é apontado uma diferença de preço entre os produtos, indicando que o Nexus 4 é mais atrativo

levando em consideração apenas o aspecto preço. Nesse caso, percebe-se uma oração coordenada assindética, em que as orações são conectadas por meio da vírgula sem a utilização de uma conjunção coordenativa.

De maneira geral, essas comparações são complexas de serem identificadas, já que não possuem uma palavra direta que permita identificar a comparação. Entretanto, esses casos são minoria dentre as comparações existentes, o que não inviabiliza a utilização da abordagem para encontrar comparações.

Mesmo com essas peculiaridades, a abordagem se mostrou promissora e capaz de encontrar cerca de 90% das sentenças comparativas. Já em relação aos casos de exceção, apesar de serem minorias, percebe-se a prevalência de comparações do tipo Não Gradativas, que dentre todos os tipos comparativos existentes, são os menos suscetíveis às análises posteriores, já que não expressam relação entre os objetos comparados [Jindal & Liu, 2006b], além de não possuírem um padrão, tornando o processo de classificação muito mais complexo [Liu, 2012]. Justamente devido a isso, alguns trabalhos direcionam mais esforços no estudo dos demais tipos comparativos [Jindal & Liu, 2006b; Ganapathibhotla & Liu, 2008; Bach et al., 2015; Yang & Ko, 2009], que expressam algum tipo relação entre os objetos, seja de igualdade ou de contraposição.

Por fim, apesar das limitações, os resultados evidenciam que a estratégia é promissora e capaz de encontrar a ampla maioria de comparações, sendo aliada a processos de mineração que visam filtrar e selecionar apenas as sentenças com mais probabilidade de serem comparações, simplificando a construção e rotulação de conjuntos de dados, detalhado na Seção 3.3.

3.3 Estratégia para Encontrar Opiniões Comparativas

Após a construção do léxico, o conjunto de palavras comparativas é utilizado para a descoberta de comparações realizadas no Buscapé e Twitter.

Para o Buscapé, utilizou-se um grande conjunto de revisões em português coletadas em setembro de 2013 [Hartmann et al., 2014]. Esse conjunto de dados ainda contém uma linguagem atual com revisões acerca de 230 produtos diferentes, considerando eletrônicos, carros, cosméticos, entre outros, que foram obtidas por meio de um coletor Web, totalizando 85.910 revisões. Filtrando apenas as revisões que possuem ao menos uma palavra-chave ou expressão comparativa, encontrou-se 48.311 revisões. Essas opiniões encontradas não são apenas mais formais que o Twitter, mas também

são extensas e mais complexas, e conjectura-se que isso esteja relacionado ao propósito de criação de cada uma das plataformas.

Já no Twitter, foram coletadas publicações escritas no idioma português postadas em um dia escolhido aleatoriamente (10/01/2018), totalizando 759.111 *tweets*, considerando opiniões comparativas e não comparativas. Em seguida, o léxico com palavras comparativas foi utilizado para filtrar todos os *tweets* que possuem ao menos uma palavra-chave, obtendo 130.459 *tweets*.

3.4 Pré-Processamento e Construção das Bases de Dados

O estudo de opiniões comparativas a nível de sentença requer inicialmente a extração das sentenças presentes nas revisões obtidas através da estratégia de palavras-chave. Observando as sentenças comparativas, percebe-se algumas características importantes que precisam ser consideradas para a tarefa de mineração.

1. ***A entidade comparada não é especificada na sentença.*** Existem revisões que não citam explicitamente a entidade comparada, e especulamos algumas razões para tal: (1) Antes de fazer uma revisão online, seleciona-se qual produto será avaliado. Portanto, o produto não é mencionado explicitamente no texto, pois se trata de uma informação já fornecida. No exemplo: "é *melhor* que o Celular X", não é informado qual produto é melhor que o Celular X. Em outros casos, utiliza-se um pronome demonstrativo ou até mesmo expressões como "o produto", para se referir ao produto avaliado; e (2) Na língua portuguesa, as orações podem ter o sujeito oculto, ou seja, o sujeito não está presente na sentença. No exemplo: "O *melhor* carro de todos", não é detalhado qual carro é o melhor.

   Solução proposta: Mesmo que os objetos comparados não sejam especificados na sentença, em ambos os casos, buscando por detalhes sobre o produto ou mesmo em sentenças anteriores, a entidade comparada pode ser inferida a partir do contexto em que a comparação foi feita.

2. ***Múltiplas comparações.*** Há sentenças com múltiplas comparações, por exemplo, "A TV é incrível, vale a pena comprá-la, o preço é *superior* ao da TV X, mas *supera* todas as outras em qualidade". A sentença possui duas comparações distintas: (1) o preço é *superior* ao da TV X; e (2) *supera* todas as outras em qualidade. Nesses casos, é necessário identificar não somente se a sentença é ou não comparativa, mas também suas diferentes partes comparativas.

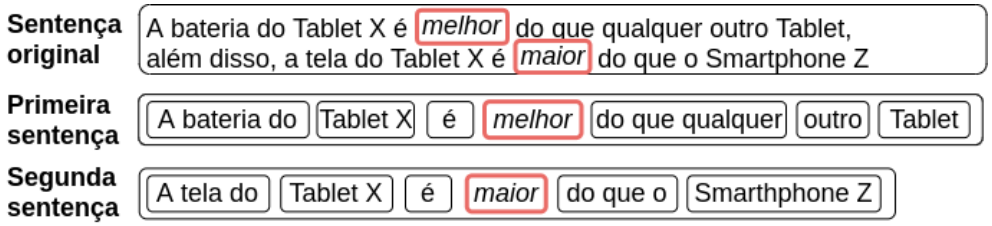


Figura 3.3: **Estratégia utilizada para extrair múltiplas comparações das sentenças.**

Solução proposta: Em vez de considerar uma sentença como uma estrutura única, é importante enxergá-la como uma estrutura que possui inúmeras partes, que podem ser ou não comparativas. Assim, uma sentença pode receber múltiplos rótulos referentes às partes comparativas encontradas. Considerando o exemplo acima, encontra-se duas partes comparativas, que são detectadas através das palavras-chave *superior* e *supera*. Na primeira parte, a palavra-chave *superior* é usada para fazer uma comparação Gradativa com Predileção. Já na segunda parte, a palavra-chave *supera* indica a existência de um superlativo.

Uma estratégia para lidar com as múltiplas comparações é a divisão da sentença em partes. Para cada uma das palavras-chave comparativas na sentença, obtém-se um intervalo de três outras palavras localizadas antes e depois da respectiva palavra-chave, de maneira a garantir que uma sentença possua uma única comparação mantendo o sentido original da frase, formando assim novas sentenças menores. No exemplo: "A bateria do Tablet X é *melhor* do que qualquer outro Tablet, além disso, a tela do Tablet X é *maior* do que o Smartphone Z", existem duas palavras-chave comparativas, *melhor* e *maior*. Aplicando a estratégia de divisão, duas novas sentenças são obtidas conforme a Figura 3.3: (1) "A bateria do Tablet X é *melhor* do que qualquer outro Tablet" e (2) "a tela do Tablet X é *maior* do que o Smartphone Z". Assim, uma sentença complexa com várias comparações é dividida em sentenças mais simples com apenas uma comparação.

Após obter todas as sentenças de cada revisão e processá-las extraindo as múltiplas comparações, criou-se duas bases de dados[5]. Rotular todas as sentenças exige um grande esforço devido à quantidade de dados disponíveis. Dessa forma, para as sentenças obtidas, uma amostra foi criada mantendo a distribuição original. As sentenças da amostra foram rotuladas manualmente por um grupo de três pessoas voluntárias que indicaram se a sentença é comparativa ou não.

[5]Acesso às Bases de Dados: `http://doi.org/10.5281/zenodo.4124410`

Tabela 3.3: Sentenças rotuladas em cada base de dados.

Sentenças	**Buscapé**	**Twitter**	**Total**
Comparativas	1.282	918	2.200
Não Comparativas	1.472	1.135	2.607
Total	2.754	2.053	4.807

No processo de rotulação, caso no mínimo dois do grupo concordem, a opinião é aceita como o rótulo final. Para os casos de divergência, o grupo discute as opiniões chegando a um consenso. O coeficiente Fleiss Kappa [Cohen, 1960] foi calculado e obteve-se uma concordância entre os três rotuladores de 88,09% (±0,007) para o Buscapé e 87,73% (±0,009) para o Twitter.

Além dessa rotulação, também foram incluídas informações sobre o tipo comparativo (Seção 2.2.1) de cada comparação encontrada as entidades e aspectos mencionados em cada uma delas. No total, 2.754 sentenças foram rotuladas no Buscapé, nas quais foram encontradas 1.282 comparações e 1.472 não comparações. Para o Twitter, 2.053 sentenças foram rotuladas, sendo 918 comparações e 1.135 não comparações. A Tabela 3.3 detalha a quantidade de comparações rotuladas em cada base de dados.

De maneira complementar a estratégia de palavras-chave utilizada para construção do conjunto de dados, abordagens supervisionadas são propostas visando aprimorar a acurácia na detecção de sentenças comparativas, apresentadas no Capítulo 4. A base de dados construída é importante para as etapas de treinamento e avaliação dos algoritmos propostos nos Capítulos 4 e 5.

Capítulo 4

Classificação das Sentenças

Esta seção apresenta a abordagem hierárquica para classificação automática de sentenças comparativas, que foi dividida em duas etapas, conforme apresentado na Figura 4.1. Após utilizar o filtro léxico para encontrar as prováveis comparações, é apresentado uma abordagem para classificação das sentenças, separando as comparativas das não comparativas (Seção 4.1). Em seguida, para as sentenças comparativas detectadas no primeiro passo, é aplicado uma estratégia de classificação, possibilitando categorizar os resultados em cinco grupos, que representam cada tipo de opinião (Seção 4.2).

4.1 Identificação das Sentenças Comparativas

Um dos passos fundamentais para a mineração de opiniões é separar as comparativas das não comparativas. A divisão das opiniões é a tarefa mais prática e importante, pois viabiliza a aplicação de técnicas de classificação e de extração de informações mais detalhadas sobre as comparações.

Após o processamento das sentenças obtidas, o desempenho de quatro classificadores de aprendizado de máquina que utilizam abordagens diferentes foi analisado para classificação das sentenças: Multinomial Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) e Random Forest (RF). Os algoritmos foram aplicados com a combinação de três *features* textuais: (1) TF-IDF de palavras, (2) TF-IDF de bigrama de palavras, (3) TF-IDF de trigrama de palavras. Para a análise de desempenho dos classificadores foram utilizadas métricas comumente usadas em tarefas de aprendizado de máquina e recuperação da informação [Baeza-Yates et al., 1999]. A Tabela 4.1 apresenta a média dos resultados obtidos para os experimentos conduzidos, que foram replicados 35 vezes para permitir o cálculo e reporte do intervalo de confiança (95%), com a aplicação de uma validação cruzada de 5 partições.

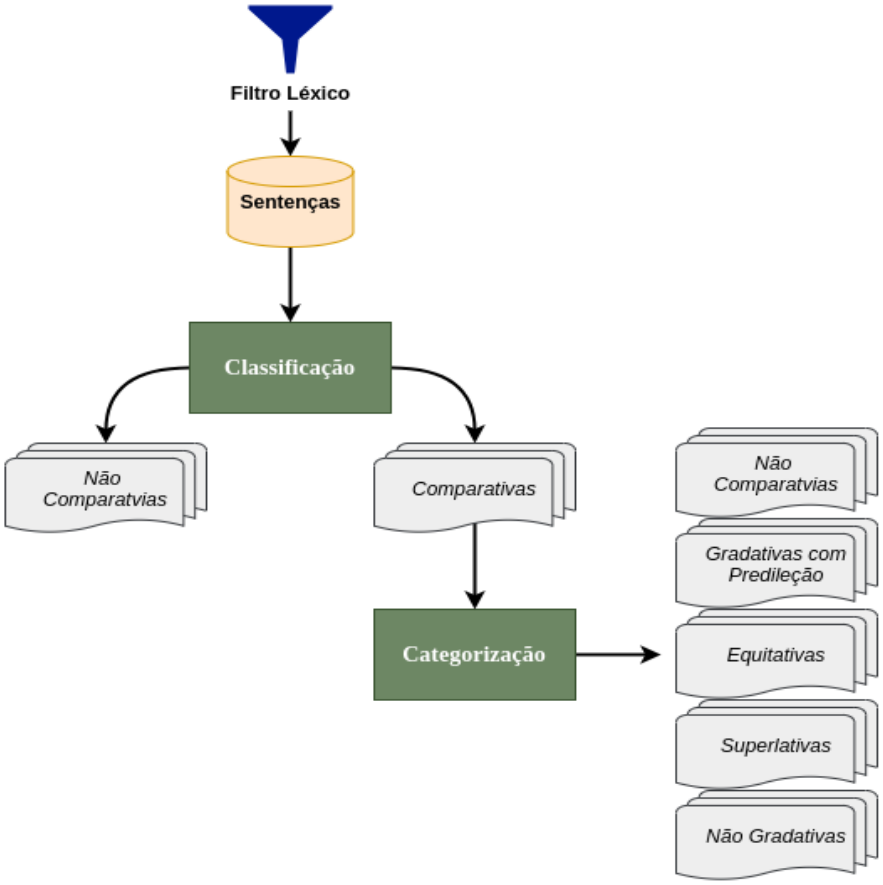


Figura 4.1: **Abordagem hierárquica para classificação.**

Para ambos conjuntos de dados, o algoritmo probabilístico NB apresentou os melhores resultados em termos de acurácia (ACC.) (i.e., Buscapé=87,3%; Twitter=86,2%). É válido ressaltar que o SVM (i.e., Buscapé=87,2%; Twitter=85,1%) apresentou valores estatisticamente similares aos do NB, considerando as métricas de acurácia e Macro-F1. Porém, pode-se afirmar que o modelo probabilístico é superior, pois apresenta uma revocação superior para a classe comparativa.

Além da acurácia, que indica a proporção de sentenças preditas corretamente, a revocação é uma métrica importante na abordagem proposta para classificação de sentenças, pois indica a frequência do modelo em detectar os exemplos de cada classe. Na abordagem hierárquica, o classificador inicial encontra as sentenças comparativas para posteriormente separá-las nos diversos tipos. Assim, é fundamental otimizar a métrica de revocação para as sentenças comparativas.

No conjunto de dados do Buscapé, apesar do NB apresentar um valor de acurácia similar ao do SVM, o primeiro possui uma taxa de revocação de 90,3% para a classe comparativa, bem acima dos demais modelos, que possuem valores próximos a 80%. O

Tabela 4.1: Precisão (Prec.), Revocação (Rev.) e F1-Score (F1) para o Buscapé e Twitter com 95% de confiança.

	Buscapé							
	Não Comparativa			**Comparativa**			**Média**	
	Prec.	Rev.	F1	Prec.	Rev.	F1	**ACC.**	**Macro F1**
RF	0,741 ± 0,006	0,801 ± 0,008	0,77 ± 0,006	0,749 ± 0,008	0,679 ± 0,01	0,712 ± 0,007	0,744 ± 0,006	0,741 ± 0,006
LR	0,861 ± 0,006	0,863 ± 0,007	0,862 ± 0,005	0,843 ± 0,007	0,839 ± 0,008	0,841 ± 0,006	0,852 ± 0,005	0,851 ± 0,006
SVM	0,869 ± 0,005	**0,895** ± 0,006	**0,882** ± 0,004	**0,875** ± 0,007	0,845 ± 0,006	**0,86** ± 0,005	**0,872** ± 0,005	**0,871** ± 0,005
NB	**0,909** ± 0,005	0,847 ± 0,006	**0,877** ± 0,005	0,838 ± 0,006	**0,903** ± 0,006	**0,869** ± 0,005	**0,873** ± 0,004	**0,873** ± 0,004
	Twitter							
	Não Comparativa			**Comparativa**			**Média**	
	Prec.	Rev.	F1	Prec.	Rev.	F1	**ACC.**	**Macro F1**
RF	0,741 ± 0,007	0,84 ± 0,011	0,787 ± 0,008	0,764 ± 0,013	0,637 ± 0,011	0,695 ± 0,01	0,749 ± 0,008	0,741 ± 0,009
LR	0,831 ± 0,006	0,874 ± 0,007	0,851 ± 0,005	0,833 ± 0,008	0,779 ± 0,01	0,805 ± 0,007	0,831 ± 0,006	0,828 ± 0,006
SVM	0,834 ± 0,007	**0,912** ± 0,005	**0,871** ± 0,005	**0,878** ± 0,007	0,775 ± 0,011	0,823 ± 0,007	**0,851** ± 0,006	0,847 ± 0,006
NB	**0,894** ± 0,007	0,851 ± 0,008	**0,872** ± 0,005	0,827 ± 0,007	**0,874** ± 0,009	**0,85** ± 0,006	**0,862** ± 0,005	**0,861** ± 0,005

mesmo ocorre na base de dados do Twitter, na qual o NB apresenta a melhor taxa de revocação, com 87,4%.

Embora as bases de dados possuam comparações feitas em diferentes contextos, tal característica não influenciou os resultados obtidos, não apresentando uma diferença significativa entre as bases de dados. A estratégia utilizada de palavras-chave juntamente com a abordagem de divisão, que separa uma sentença em múltiplas sentenças, formam novas pequenas sentenças que possuem apenas a parte opinativa. Assim, as características textuais não influenciam tanto nos resultados da classificação binária, mesmo existindo diferenças nas sentenças de cada base de dados.

A Tabela 4.2 exibe a matriz de confusão obtida com o NB, que apresentou os melhores resultados para a distinção das duas classes. O modelo consegue detectar corretamente 90,3% das sentenças comparativas existentes no Buscapé e 87,4% no Twitter, o que ressalta a boa capacidade do modelo na cobertura das comparações.

Tabela 4.2: Frequência de classificação para cada classe com o Multinomial Naive Bayes (NB).

Buscapé			
		Label Predito	
		Não Comparativa	Comparativa
Label Real	Não Comparativa	**84,7%**	15,3%
	Comparativa	9,7%	**90,3%**
Twitter			
		Label Predito	
		Não Comparativa	Comparativa
Label Real	Não Comparativa	**85,1%**	14,9%
	Comparativa	12,6%	**87,4%**

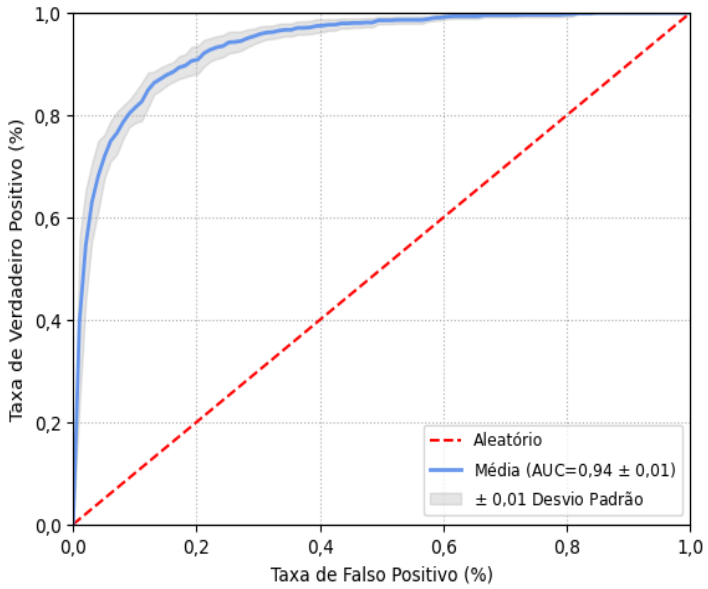


Figura 4.2: **Buscapé: Curva ROC para o Multinomial Naive Bayes (NB).**

Por fim, os melhores resultados foram obtidos através do NB, com AUC de 0,94 para ambas bases de dados. Observando a curva ROC do algoritmo na Figura 4.3, nota-se a possibilidade de escolher um limiar de classificação para detectar corretamente quase 90% de todas as comparações com apenas 10% de erro de classificação (taxa de falso positivo $\approx$ 0,1). Isso pode ser interessante para as abordagens que focam em detectar sentenças com maior chance de serem comparativas.

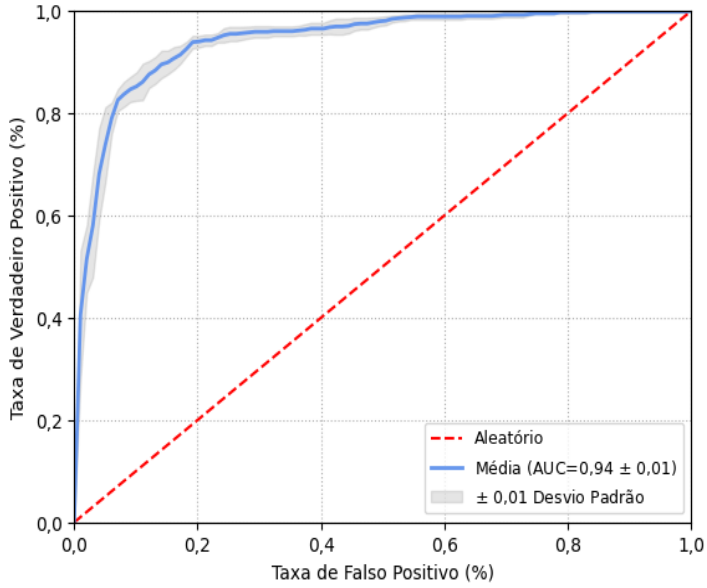


Figura 4.3: **Twitter: Curva ROC para o Multinomial Naive Bayes (NB).**

Tabela 4.3: Detalhamento das sentenças classificadas como comparativas através do Multinomial Naive Bayes (NB).

Sentenças	**Buscapé**	**Twitter**	**Total**
Gradativa com Predileção	502	279	781
Equitativa	255	172	427
Superlativa	290	270	560
Não Gradativa	115	81	196
Total Comparativas	1.162	802	1.964
Total Não Comparativas	234	170	404

4.2 Categorização das Comparações em Grupos

Após separação binária das sentenças, iniciou-se a classificação dos resultados (i.e., hierárquica) com objetivo de categorizar as sentenças previamente classificadas como comparativas em cinco grupos, ou seja, (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. O grupo de sentenças não comparativas se faz presente devido aos eventuais falsos positivos oriundos da classificação binária. Essa categorização dos resultados tem um papel importante para fornecer mais detalhes acerca das comparações encontradas, facilitando a visualização das informações e possibilitando análises mais sistemáticas.

As sentenças classificadas como comparativas na etapa anterior com o Multinomial Naive Bayes (NB), que apresentou os melhores resultados, foram adicionadas a um novo conjunto de dados. A Tabela 4.3 apresenta a quantidade de sentenças obti-

Tabela 4.4: Precisão (Prec.), Revocação (Rev.) e F1-Score (F1) para a classificação em múltiplas classes para as bases de dados do Buscapé e Twitter com 95% de confiança com o Multinomial Naive Bayes (NB).

	Buscapé																
	Não Comparativa			**Grad. com Predileção**			**Equitativa**			**Superlativa**			**Não Gradativa**			**Média**	
	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	**ACC.**	**Macro F1**
NB>RF	0,284 ± 0,019	0,28 ± 0,019	0,28 ± 0,017	0,575 ± 0,013	0,694 ± 0,017	0,628 ± 0,011	0,653 ± 0,021	0,579 ± 0,023	0,611 ± 0,019	0,694 ± 0,016	0,63 ± 0,017	0,659 ± 0,014	0,696 ± 0,042	0,379 ± 0,031	0,483 ± 0,029	0,564 ± 0,01	0,532 ± 0,011
NB>LR	**0,412** ± 0,017	**0,343** ± 0,022	**0,372** ± 0,018	**0,711** ± 0,011	0,751 ± 0,013	0,73 ± 0,008	**0,713** ± 0,017	0,743 ± 0,023	**0,727** ± 0,018	**0,766** ± 0,017	0,798 ± 0,018	**0,78** ± 0,012	0,527 ± 0,03	0,457 ± 0,035	0,485 ± 0,03	**0,667** ± 0,008	**0,619** ± 0,01
NB>SVM	0,396 ± 0,02	0,3 ± 0,018	0,34 ± 0,017	0,682 ± 0,01	**0,817** ± 0,011	**0,743** ± 0,009	0,705 ± 0,017	0,672 ± 0,02	0,687 ± 0,015	0,745 ± 0,018	0,742 ± 0,023	0,741 ± 0,016	**0,611** ± 0,034	0,432 ± 0,038	**0,499** ± 0,032	**0,657** ± 0,008	**0,602** ± 0,01
NB>NB	0,401 ± 0,027	0,244 ± 0,02	0,302 ± 0,022	**0,765** ± 0,012	0,654 ± 0,015	0,704 ± 0,012	0,674 ± 0,014	**0,789** ± 0,021	0,726 ± 0,013	0,699 ± 0,012	**0,876** ± 0,017	0,777 ± 0,01	0,369 ± 0,02	**0,509** ± 0,035	0,424 ± 0,023	0,644 ± 0,008	0,587 ± 0,009
	Twitter																
	Não Comparativa			**Grad. com Predileção**			**Equitativa**			**Superlativa**			**Não Gradativa**			**Média**	
	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	Prec.	Rev.	F1	**ACC.**	**Macro F1**
NB>RF	0,25 ± 0,019	0,362 ± 0,034	0,293 ± 0,023	0,627 ± 0,019	0,63 ± 0,029	0,625 ± 0,018	0,728 ± 0,026	0,671 ± 0,026	0,694 ± 0,018	0,676 ± 0,016	0,598 ± 0,026	0,633 ± 0,019	0,728 ± 0,05	0,385 ± 0,037	0,496 ± 0,038	0,561 ± 0,011	0,548 ± 0,011
NB>LR	0,351 ± 0,025	0,264 ± 0,024	0,298 ± 0,022	0,678 ± 0,019	0,801 ± 0,021	0,732 ± 0,013	**0,773** ± 0,025	0,811 ± 0,024	**0,788** ± 0,017	0,725 ± 0,016	0,733 ± 0,023	0,727 ± 0,016	0,703 ± 0,041	**0,471** ± 0,033	**0,556** ± 0,031	**0,662** ± 0,011	**0,62** ± 0,012
NB>SVM	**0,4** ± 0,023	**0,379** ± 0,025	**0,386** ± 0,02	0,72 ± 0,014	**0,856** ± 0,015	**0,781** ± 0,01	0,672 ± 0,026	0,789 ± 0,023	0,723 ± 0,019	**0,776** ± 0,015	0,669 ± 0,02	0,717 ± 0,014	**0,759** ± 0,049	0,357 ± 0,038	0,477 ± 0,039	**0,667** ± 0,009	**0,616** ± 0,012
NB>NB	0,345 ± 0,033	0,181 ± 0,021	0,235 ± 0,025	**0,728** ± 0,017	0,719 ± 0,018	0,723 ± 0,016	0,726 ± 0,021	**0,848** ± 0,02	0,78 ± 0,016	0,675 ± 0,014	**0,811** ± 0,013	**0,735** ± 0,009	0,49 ± 0,038	**0,471** ± 0,039	0,474 ± 0,033	0,653 ± 0,01	0,589 ± 0,013

das para cada tipo comparativo. O novo conjunto de dados possui cerca de 88% das sentenças comparativas rotuladas inicialmente. Além das comparações, foram trazidas algumas poucas sentenças não comparativas identificadas como falsas positivas na etapa anterior, 170 para o Twitter e 234 para o Buscapé. Portanto, além de classificar as sentenças nos quatro tipos comparativos existentes, é necessário também distinguir esse grupo pequeno de falso positivo.

Para a classificação, utilizou-se os quatro algoritmos de aprendizado de máquina explorados anteriormente, com três conjuntos de *features* textuais: (1) TF-IDF de palavras, (2) TF-IDF de bigrama de palavras, (3) TF-IDF de trigrama de palavras. Após aplicar os algoritmos para separação das sentenças em cinco classes, através de 35 replicações realizadas por meio da validação cruzada com 5 partições, verificou-se que LR e o SVM apresentam resultados similares de acurácia e Macro F1 em ambas bases de dados. No entanto, o LR com uma acurácia de 66,7% (±0,008) e Macro F1 de 61,9% (±0,01) para o Buscapé, e o SVM acurácia de 66,7% (±0,009) e Macro F1 de 61,6% (±0,012) para o Twitter, se mostram superiores na distinção da classe não comparativa dos demais grupos comparativos, como mostra a Tabela 4.4.

Apesar da acurácia ser uma métrica importante, a frequência de classificação de

Tabela 4.5: Frequência de classificação para o Buscapé (LR - ACC = 66,7%±0,008) e Twitter (SVM - ACC = 66,7%±0,09).

Buscapé						
		Label Predito				
		Não Comparativa	Grad. com Pred.	Equitativa	Superlativa	Não Gradativa
Label Real	Não Comparativa	**34,3%**	29,7%	13,1%	14,2%	8,7%
	Gradativa com Predileção	9,8%	**75,1%**	5,4%	5,7%	4,0%
	Equitativa	11,3%	9,9%	**74,3%**	2,3%	2,2%
	Superlativa	5,9%	11,6%	2,3%	**79,8%**	0,5%
	Não Gradativa	17,3%	22,7%	10,6%	3,7%	**45,7%**
Twitter						
		Label Predito				
		Não Comparativa	Grad. com Pred.	Equitativa	Superlativa	Não Gradativa
Label Real	Não Comparativa	**37,9%**	21%	15,9%	20,7%	4,5%
	Gradativa com Predileção	5,6%	**85,6%**	3,7%	4,9%	0,3%
	Equitativa	11,0%	7,8%	**78,9%**	2,2%	0,1%
	Superlativa	13,8%	16,5%	2,5%	**66,9%**	0,3%
	Não Gradativa	33,0%	0,7%	30,2%	0,5%	**35,6%**

cada classe (i.e., revocação) é essencial para a tarefa de categorização das comparações em diferentes classes. A Tabela 4.5 apresenta o resultado da classificação das sentenças, onde os valores indicam a frequência de classificação de cada classe.

Observando os resultados, percebe-se que apesar de existirem diferentes tipos de comparações, o principal desafio ainda continua sendo a distinção das sentenças comparativas das não comparativas. Essas sentenças identificadas como falsas positivas na primeira etapa possuem muitas semelhanças com as sentenças comparativas, sendo complexas para classificação, pois muitas vezes não possuem um padrão, o que dificulta a distinção desses tipos. Entretanto, a baixa precisão para o grupo de sentenças não comparativas não é um problema, pois cerca de 85% das sentenças já foram separadas na classificação binária, além de serem minorias no novo conjunto de dados.

Em relação aos dois contextos analisados, apesar das diferenças existentes, os resultados mostram que não existe uma diferença significativa na tarefa de classificação das sentenças nas bases de dados.

Já se tratando dos tipos Superlativos e Equitativos, percebe-se que tais comparações fazem uso de expressões próprias que permitem distingui-las mais facilmente das demais, como é visto na Tabela 4.5, com valores de 79,8% e 66,9% para os Superlativos e 74,3% e 78,9% para as sentenças Equitativas do Buscapé e Twitter. O oposto ocorre com as comparações Não Gradativas, que são bem mais complexas por não possuírem um padrão claro. Tais sentenças podem ser frequentemente confundidas com as Gradativas com Predileção ou até mesmo com as sentenças não comparativas.

Para a classificação das comparações, a estrutura sintática das sentenças é uma característica muito importante. Em alguns casos, devido à posição da palavra-chave comparativa ou até mesmo algum complemento utilizado, as sentenças com estruturas sintáticas semelhantes podem ser classificadas em tipos comparativos diferentes. Na sentença: "Smartphone X é o *melhor*", nota-se um superlativo. No entanto, adicionando um único complemento (do que Y), a sentença passa a ser classificada como Gradativa com Predileção, mesmo as duas frases possuindo estruturas semelhantes e utilizando a mesma palavra-chave comparativa (e.g., "Smartphone X é *melhor* do que o Y").

Apesar desses desafios, as comparações normalmente possuem algumas preposições e advérbios próximos à palavra-chave comparativa, o que viabiliza a detecção das comparações através dos bigramas e trigramas de palavras. No exemplo acima, a expressão *do que* juntamente com a palavra-chave *melhor* forma um trigrama: "*melhor* do que", que na maioria das vezes é utilizado para comparar dois objetos expressando predileção, o que possibilita diferenciar a frase de um superlativo.

Capítulo 5

Detecção de Preferência

A capacidade de fazer comparações expressando ordem entre objetos é um componente básico da cognição humana [Sapir, 1944] e uma maneira prática de avaliar um objeto contrastando suas características com outras entidades. Em uma comparação, o sentimento expresso a cada uma das entidades é uma das informações mais relevantes a serem analisadas, pois permite a identificação da entidade preferida em uma sentença. No trabalho, a preferência é entendida como o ato de escolher um objeto em detrimento a outros ou também pela ação de indicar superioridade de um objeto em relação a outros. Assim, este Capítulo apresenta o algoritmo proposto que utiliza elementos e características das opiniões para determinar a entidade preferida em uma sentença comparativa.

As opiniões comparativas são utilizadas para contrastar objetos, logo, a tarefa de detecção de preferência surge como uma maneira de encontrar, dentre os objetos comparados, qual é o indicado como preferido. No entanto, o fato de uma entidade ser preferida em uma sentença não significa que essa mesma entidade será mantida como preferida quando comparada com outras entidades em outras sentenças. Por isso a importância da análise de cada comparação individualmente, considerando as características ali presentes.

Exemplificando, ao analisar a sentença "O celular X é *melhor* do que o Y", percebe-se a existência de dois objetos, celular X e celular Y, sendo que o primeiro é apontado como preferido. Já na sentença "A bateria do celular X é quase *melhor* do que a do laptop Y", apesar de pressupor que o aspecto do primeiro produto também é avaliado positivamente, pode-se inferir que a bateria do celular X não é a melhor, ou seja, ainda é inferior ao laptop Y, apontado como o preferido.

Uma opinião comparativa é composta por um grupo de elementos que precisam ser identificados e analisados em conjunto para que seja possível a extração das informações

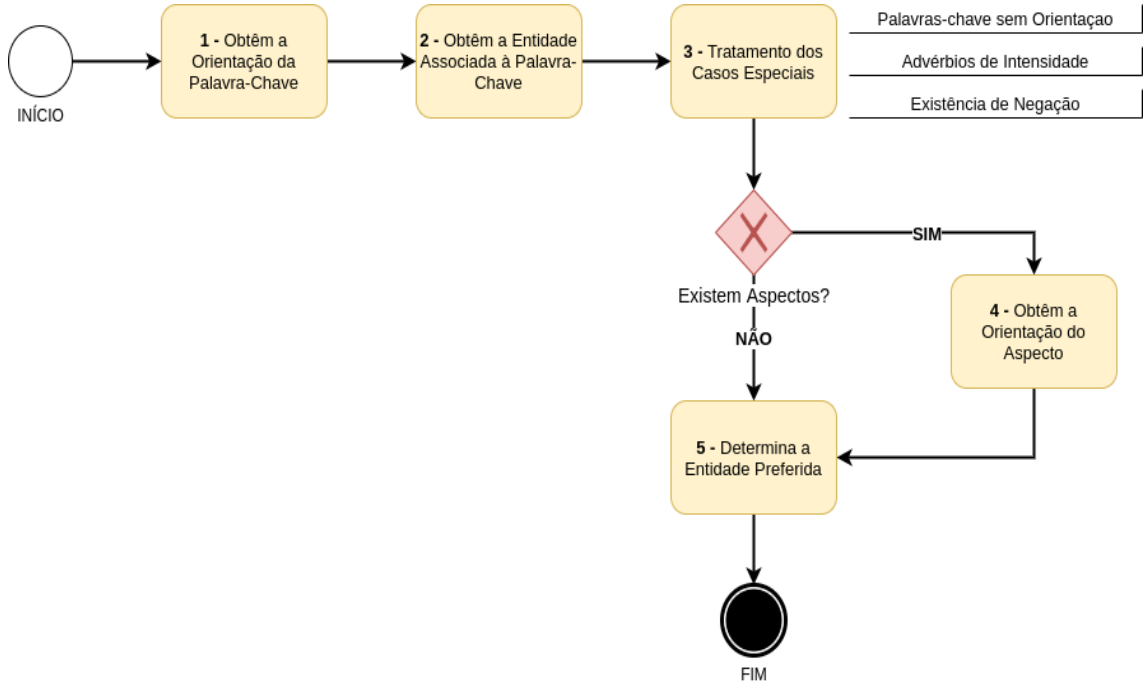


Figura 5.1: **Etapas do algoritmo para determinar a preferência.**

relevantes. Uma comparação pode ser representada por uma tupla que contém seis elementos, como (E_1, E_2, A, PE, h, t) [Liu, 2012], onde E_1 e E_2 representam as entidades que são comparadas na sentença, sendo que E_1 é mencionada antes da E_2. Já o elemento A representa o aspecto associado às entidades comparadas. Por fim, PE representa a entidade preferida, que é indicada por um formador de opinião h em um determinado tempo t.

A obtenção das preferências parte do princípio que a comparação expressa uma relação de superioridade ou inferioridade entre os objetos contrastados, visto que não existe preferência em relações de igualdade. Por esse motivo, o algoritmo não é aplicável nas comparações Equitativas e Não Gradativas. No entanto, as comparações que expressam preferências são as mais frequentes, representando cerca de 70% dentre todas as opiniões comparativas existentes nas bases de dados. Sendo assim, o algoritmo proposto se concentra apenas na detecção de preferência das opiniões Gradativas com Predileção e Superlativas.

Visto isso, o algoritmo proposto na Figura 5.1 é composto por um conjunto de etapas, que inicialmente consiste na detecção da orientação da palavra-chave utilizada para se fazer comparação (Seção 5.1) e, em seguida, encontra-se a entidade associada a essa palavra (Seção 5.2), sendo possível determinar a preferência (Seção 5.3). Já na Seção 5.4 e Seção 5.5, são tratados os casos especiais que necessitam da análise de contexto para determinar a preferência. Por fim, o algoritmo é avaliado na Seção 5.8,

onde observa-se um valor de acurácia próximo a 83% na detecção da entidade preferida nas opiniões Gradativas com Predileção e, para as opiniões Superlativas, observa-se um resultado de acurácia próximo a 95%.

5.1 Obtendo a Orientação das Palavras-Chave Comparativas

As opiniões comparativas são constituídas por uma palavra-chave comparativa que estabelece uma relação entre as entidades comparadas. Por exemplo, "A TV X é *melhor* do que a TV Y", onde o adjetivo *melhor* traz sentido à frase, conectando as entidades TV X e TV Y e expressando preferência à primeira.

Uma das maneiras práticas e eficazes de interpretar uma opinião comparativa é justamente analisar a palavra-chave, pois ela contém propriedades que contribuem para entender a relação entre as entidades comparadas na sentença. Nas análises tradicionais, a polaridade de uma palavra é uma propriedade frequentemente utilizada em problemas de classificação de sentimento [Araújo et al., 2016; Melo et al., 2019], onde um léxico que possui um mapeamento do sentimento de palavras é aplicado para determinar o sentimento final de uma frase, podendo ser positivo, negativo ou neutro [Taboada et al., 2011]. No entanto, para as opiniões comparativas, o sentimento atrelado à palavra-chave não é a propriedade mais relevante para se determinar a preferência, isso porque a palavra-chave em uma comparação é utilizada com o intuito de expressar ordem e graus entre os objetos, e a polaridade é uma informação que não capta esse tipo de relação.

Analisando as palavras-chave comparativas, observa-se que grande parte não possui polaridade (e.g., *menor*, *maior*, *mais*, *superior*, etc), sendo assinaladas como neutras. No entanto, tais palavras possuem uma orientação, que expressa uma ideia de gradação entre as entidades mencionadas na sentença, podendo ser de superioridade ou inferioridade de uma entidade em relação a outra. No exemplo da sentença "A bateria do Celular X é bem *maior* do que o Celular Y", a palavra-chave *maior* não possui uma polaridade positiva ou negativa, mas indica superioridade da bateria do Celular X em relação ao Celular Y. Logo, percebe-se que a detecção de preferência não está interessada no sentimento, mas no tipo de relação (i.e., superioridade ou inferioridade) existente entre as entidades comparadas.

Tendo em vista isso, nota-se a importância da obtenção de informações sobre a orientação das palavras-chave comparativas existentes no léxico construído e apresentado no Capítulo 3. Assim, para cada uma das palavras foi incluído um valor referente

à orientação, onde um valor positivo (+1) significa que a respectiva palavra é utilizada para indicar superioridade de uma entidade em relação à outra, enquanto um valor negativo (-1) indica inferioridade (ver Tabela A.1 no Apêndice A). Dessa maneira, as palavras maior, superior, acima, etc, recebem um valor positivo (+1), enquanto outras palavras, como menor, inferior, abaixo, etc, recebem um valor negativo (-1).

Já para os casos de exceção, onde não é possível determinar a orientação, as palavras-chave recebem um valor igual a zero, ou seja, considerando apenas a palavra-chave individualmente não se pode inferir superioridade ou inferioridade. No entanto, essa orientação pode ser encontrada através da análise do contexto, sendo possível determinar um valor de maneira automática por meio de palavras próximas à palavra-chave na sentença. Esses casos são detalhados na Seção 5.4.

5.2 Detecção da Entidade Associada à Palavra-Chave Comparativa

Além de obter a orientação da palavra-chave, é muito importante também identificar a entidade relacionada a ela, uma vez que isso permite estabelecer a relação de graus existente entre as entidades mencionadas, possibilitando determinar a entidade superior. Na sentença "O Iphone é *melhor* do que o Galaxy", a palavra-chave *melhor* indica superioridade, contudo, somente é possível determinar a preferência após identificar que o Iphone é a entidade associada diretamente com à palavra-chave comparativa.

Em uma opinião comparativa, uma entidade pode ser um produto, serviço, marca, organização, pessoa, evento, situação ou tópico [Liu, 2012]. Essas entidades possuem propriedades e aspectos que eventualmente podem ser mencionados nas comparações, principalmente quando se deseja fornecer um nível maior de detalhes, tal como "A qualidade da tela do Laptop X é *melhor* que o Laptop Y". No entanto, as entidades podem ser contrastadas diretamente sem a necessidade de explicitar o aspecto, por exemplo: "O refrigerador X é bem *superior* ao refrigerador Y que estava em promoção".

Nesse contexto, a análise das dependências sintáticas de uma sentença é uma das estratégias para encontrar a relação existente entre as palavras, permitindo descobrir quais termos se relacionam diretamente com a palavra-chave.

Na Figura 5.2, tem-se a sentença "O Iphone é *melhor* do que o Galaxy", onde a palavra-chave *melhor*, que é um adjetivo, possui uma relação de dependência do tipo *nsubj*[1] para a entidade Iphone, o que significa que Iphone é o objeto relacionado à

[1]Relação do tipo (*nsubj*): `https://universaldependencies.org/u/dep/nsubj.html`. Acesso em: 01 de fevereiro de 2021.

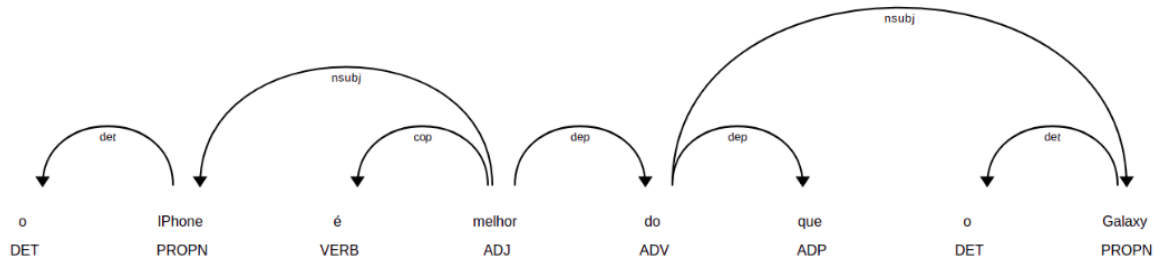


Figura 5.2: **Dependências sintáticas de uma sentença comparativa.**

palavra-chave *melhor* e é também a entidade indicada como preferida.

Para a encontrar a entidade associada à palavra-chave de maneira automática, propomos uma estratégia que consiste na utilização de um analisador de dependência[2], que é uma ferramenta que recebe uma determinada sentença como parâmetro e retorna a relação sintática e as dependências entre as palavras, como mostrado na Figura 5.2.

Com a relação sintática entre as palavras, a estratégia proposta percorre a lista de dependências a partir da palavra-chave comparativa, priorizando relações de dependências do tipo *nsubj* e *obj*[3], que são encontradas quando existe a ligação de uma palavra ao sujeito da oração ou a um objeto, que muitas das vezes é uma das entidades comparadas.

No exemplo da Figura 5.2, existem dois objetos sendo comparados, Iphone e Galaxy. Logo, percorrendo a lista de dependências, o primeiro objeto comparado encontrado (i.e., Iphone) é escolhido como a entidade associada à palavra-chave.

Essa estratégia é aplicável em grande parte das comparações, uma vez que as relações sintáticas das palavras podem ser encontradas com a análise das dependências das sentenças, o que permite identificar a entidade associada à palavra-chave.

5.2.1 Comparações com entidade oculta

As opiniões comparativas contrastam duas ou mais entidades em uma mesma sentença, no entanto, em algumas situações, uma das entidades pode ser omitida, não sendo mencionada de maneira explícita na comparação. Normalmente, isso ocorre por dois motivos principais.

O primeiro deles é quando a entidade referida já foi mencionada em alguma sentença anterior. Nesses casos, o objeto comparado é indicado através do uso pronomes

[2]Analisador de dependência: Foi utilizado o analisador de dependências da biblioteca spaCy v2.2.4, que pode ser encontrado: `https://spacy.io/usage/linguistic-features#morphology`. Acesso em: 01 de fevereiro de 2021.

[3]Relação do tipo (*obj*): `https://universaldependencies.org/u/dep/obj.html`. Acesso em: 01 de fevereiro de 2021.

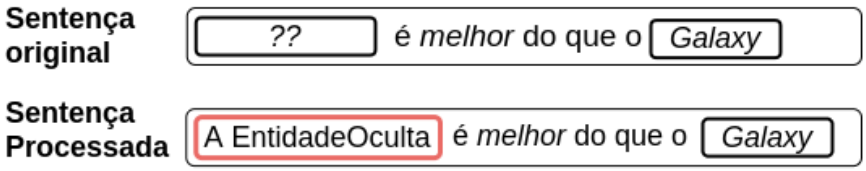


Figura 5.3: **Processamento das comparações com entidade oculta.**

demonstrativos ou de expressões como "o produto", como é o exemplo da sentença "O produto é *melhor* do que o Galaxy" ou até mesmo da sentença "ele é *melhor* do que o Galaxy". Nesses exemplos, apenas considerando as informações fornecidas na sentença não é possível determinar o exato nome do objeto contraposto com o Galaxy, sendo necessário a análise de contexto.

O segundo motivo é quando a sentença comparativa possui o sujeito oculto, que é uma situação comum em orações na língua portuguesa. Esse fenômeno indica que, apesar da existência do sujeito, ele não é mencionado de maneira explícita na frase, por exemplo "é *melhor* do que o Galaxy".

Em ambos os casos, nota-se a possibilidade de encontrar a entidade oculta através do contexto, isso porque ao se fazer uma revisão, seleciona-se o produto que será avaliado, o que permite inferir qual a entidade oculta.

Sendo assim, para garantir que a análise das dependências sintáticas funcione de maneira correta, permitindo identificar a entidade associada à palavra-chave, é necessário que os objetos comparados estejam explicitados na sentença. Dessa forma, para os casos em que a entidade está oculta, inclui-se automaticamente no início da sentença uma palavra que indique a entidade omitida, possibilitando percorrer a árvore de dependências para encontrar a entidade associada à palavra-chave. Essa abordagem inclui uma palavra (e.g., EntidadeOculta) que substitui a respectiva entidade omitida.

Na Figura 5.3 é apresentado um exemplo de uma sentença referente à revisão de um Iphone, em que a entidade avaliada foi omitida, "é *melhor* do que o Galaxy". Após a inclusão da palavra que representa a entidade oculta, obtêm-se a sentença "A *EntidadeOculta* é *melhor* do que o Galaxy". Dessa maneira, a sentença passa a ter uma estrutura sintática que possibilita a correta identificação da entidade associada à palavra-chave comparativa.

Finalmente, para os casos em que se tem pronomes demonstrativos, não é necessário executar esse procedimento, uma vez que o pronome já indica a existência de uma entidade.

5.3 Determinando a Preferência

Após o mapeamento da orientação e da entidade associada à palavra-chave, pode-se iniciar a análise para determinar a entidade preferida em uma opinião comparativa. De início, foi proposto um algoritmo que utiliza essas informações para determinar a preferência. Esse algoritmo implementa uma lógica condicional, em que caso a palavra-chave comparativa utilizada na sentença indique superioridade, automaticamente a entidade associada a ela é apontada como a preferida, caso contrário, a entidade contraposta na comparação é indicada como a preferida, como apresentado no Algoritmo 1.

Em relação aos parâmetros, o Algoritmo 1 recebe a sentença comparativa (*texto*), a palavra-chave utilizada para comparação (*palavra-chave*) e as entidades que são comparadas na sentença (*ent1* e *ent2*). Em seguida, nas duas primeiras linhas são executados dois procedimentos, um para a obtenção da orientação da palavra-chave através da consulta ao léxico (linha 2) e outro para detectar a entidade associada à palavra-chave comparativa (linha 3). Logo após, uma operação condicional é realizada verificando se a orientação da palavra-chave é positiva (linha 4), ou seja, verifica se indica superioridade. Caso verdadeiro, a entidade associada a ela é retornada como preferida, caso contrário, a entidade contraposta é apontada como a preferida.

Algoritmo 1 Algoritmo que determina a entidade preferida.

```
1:  procedure getPreference(texto, palavra-chave, ent1, ent2)
2:      orientacaoPalavra = getWordOrientation(palavra-chave)
3:      entRelacionada = getRelatedEntity(texto, palavra-chave, ent1, ent2)
4:      if orientacaoPalavra > 0 then
5:          entidadePreferida = entRelacionada
6:      else
7:          entidadePreferida = if entRelacionada == ent1 then ent2 else ent1
8:      end if
9:      return entidadePreferida
10: end procedure
```

- **getWordOrientation()** - Obtém a orientação da palavra-chave comparativa (i.e., superioridade ou inferioridade).
- **getRelatedEntity()** - Obtém a entidade relacionada à palavra-chave comparativa.

A estratégia para a detecção da entidade preferida é prática e aplicável na grande maioria das sentenças, pois com a obtenção dos elementos fundamentais da opinião e com o mapeamento da relação entre as entidades é possível determinar a preferência de maneira direta. No entanto, em alguns casos especiais, a estratégia precisa lidar com situações como quando há a existência de negação ou até mesmo quando a utilização

de advérbios de intensidade modifica o sentido da frase. Esses casos especiais são apresentados e detalhados na Seção 5.4.

5.4 Casos Especiais

Esta Seção apresenta as características encontradas nas opiniões comparativas que precisam de tratamentos especiais, pois fogem do controle do fluxo de execução normal apresentado pelo Algoritmo 1. Apesar da lógica permanecer a mesma, o Algoritmo 1 é aprimorado com a inclusão de etapas de processamento que lidam com essas peculiaridades.

5.4.1 Palavras comparativas sem orientação

Apesar de grande parte das palavras-chave comparativas carregar a informação de orientação, em alguns casos não é possível determinar essa propriedade sem a análise de contexto, como feito na Seção 5.1. Visto isso, esta Seção apresenta uma estratégia que através da análise das palavras próximas na sentença determina a orientação final das palavras-chave neutras.

Algumas expressões como *chega aos pés*, *mesmo nível*, *como* e *tem a mesma*, apesar de não possuírem uma orientação positiva ou negativa, normalmente são antecedidas por negação, formando expressões que passam a expressar inferioridade às entidades associadas a elas, tais como **não** *chega aos pés*, **não** está no *mesmo nível*, **nada** *como* e **não** *tem a mesma*. Na sentença "O Laptop X não *chega aos pés* ao Laptop Y", nota-se que a palavra-chave *chega aos pés* expressa uma relação de igualdade quando analisada de maneira isolada. Porém, ao observar a palavra-chave juntamente com a negação, percebe-se que a expressão não *chega aos pés* indica uma relação de inferioridade do Laptop X em relação ao Laptop Y.

Dessa maneira, para lidar com essas palavras que indicam igualdade, a abordagem proposta inicia verificando a existência de negação nos termos que antecedem a palavra-chave observada e, caso exista negação, atribui-se o valor de orientação referente à palavra-chave.

Por outro lado, existem alguns casos em que a busca por negação não é suficiente para determinar orientação, como é o caso das palavras *comparado*, *diferença*, *relação* e *tão bom como*. Uma alternativa para lidar com essa situação é a análise do contexto e do sentimento envolvido nas palavras próximas, uma vez que as sentenças comparativas formadas por essas palavras-chave possuem termos próximos que indicam o sentimento relacionado às entidades, permitindo identificar a orientação e, consequentemente, a

entidade preferida. Na sentença "O Celular X é bom se *comparado* ao Celular Y", apesar da palavra-chave *comparado* ser neutra, a palavra *bom* representa um sentimento positivo, indicando que o Celular X é superior ao Celular Y. O mesmo ocorre na sentença "Iphone é bom em *relação* ao Galaxy", onde novamente a palavra *bom* indica que a primeira entidade mencionada Iphone é a preferida na sentença.

Para tratar esse caso em específico, é proposto uma estratégia que consiste em encontrar o termo mais próximo da palavra-chave utilizada na comparação, seguindo a ordem de prioridade (adjetivos, advérbios e verbos)[4], ou seja, caso não existam adjetivos, é escolhido um advérbio, porém caso não seja encontrado um advérbio, o verbo mais próximo é escolhido. Em seguida, o sentimento associado a esse termo encontrado é obtido através da combinação de três léxicos em português, o SentiStrength [Thelwall & Buckley, 2013], LIWC [Balage Filho et al., 2013] e o Onto.PT [Gonçalo Oliveira & Gomes, 2014]. Nesse caso, o valor final de sentimento pode representar também o tipo de orientação, onde um valor positivo (+1) indica superioridade e um valor negativo (-1) indica inferioridade da entidade associada a esse termo. Sendo assim, a palavra encontrada passa a ser considerada a nova palavra-chave da sentença, o que permite determinar a entidade preferida através das etapas estratégia descritas na Seção 5.3.

Além das etapas detalhadas na Seção 5.1, esses procedimentos para detectar a orientação estão encapsulados internamente no método *getWordOrientation()*, utilizado para obter a orientação da palavra-chave, como apresentado no Algoritmo 1.

5.4.2 Advérbios de intensidade

Os advérbios de intensidade são palavras modificadoras utilizadas quando se deseja intensificar o efeito de adjetivos, verbos e até mesmo de advérbios, formando uma tupla (advérbio de intensidade + adjetivo) [da Rocha Lima, 2017]. No exemplo da frase "O produto X é **muito** *melhor* do que o produto Y", o advérbio de intensidade **muito** é introduzido, intensificando o efeito do adjetivo *melhor* na sentença, reforçando que o produto X é **muito** superior ao produto Y. Esses advérbios de intensidade são frequentemente encontrados em sentenças comparativas, visto que as comparações têm por função contrastar objetos, sendo comum a utilização desses advérbios para dar ênfase em aspectos e características de um objeto em relação a outro, pois são uma maneira prática de expressar predileção e superioridade a uma entidade.

[4]Part of Speech (PoS) Tagging: Para identificar a classes gramatical das palavras, utilizou-se a biblioteca *nlpnet* com um modelo treinado em um corpus em português, disponível no endereço: `http://nilc.icmc.usp.br/nlpnet/models.html#pos-portuguese`. Acesso em: 01 de fevereiro de 2021.

No entanto, a utilização de um advérbio próximo a uma palavra-chave pode alterar o significado e até mesmo a maneira como a sentença comparativa deve ser interpretada, impactando na estratégia utilizada para determinar a entidade preferida. Na sentença "O Produto X é **quase** *melhor* do que o Produto Y", apesar da palavra-chave comparativa *melhor* indicar superioridade da entidade associada à palavra-chave (i.e., Produto X), percebe-se que a utilização do advérbio quase, que forma a expressão (**quase** *melhor*), faz com que a entidade associada à palavra *melhor* seja a inferior, indicando que o Produto Y é o preferido. Sendo assim, caso haja um advérbio de intensidade, além de considerar a palavra-chave comparativa, é necessário analisar a expressão completa (advérbio de intensidade + palavra-chave comparativa).

Os advérbios de intensidade que impactam a análise das comparações foram separados em dois grupos, nomeados de advérbios de incremento e advérbios de decremento (Ver Tabelas B.1 e B.1 no Apêndice B). Os advérbios chamados de incremento são aqueles que intensificam o significado de uma palavra, por exemplo: mais, bastante, bem, tão, entre outros. Já os advérbios de decremento são opostos a esses, por exemplo: menos, quase, menor, entre outros.

Após o mapeamento dos advérbios, é proposto uma estratégia para interpretar as expressões (advérbio de intensidade + palavra-chave comparativa), composta por quatro regras:

1. Advérbio incremento + Palavra-chave (superioridade **(+1)**) = **Superioridade**
2. Advérbio incremento + Palavra-chave (inferioridade **(-1)**) = **Inferioridade**
3. Advérbio decremento + Palavra-chave (superioridade **(+1)**) = **Inferioridade**
4. Advérbio decremento + Palavra-chave (inferioridade **(-1)**) = **Superioridade**

A estratégia proposta é similar a regra de sinais da matemática, em que um advérbio de decremento (-1) seguido de uma palavra-chave de superioridade (+1) denota uma expressão de inferioridade para entidade associada à palavra-chave. Por outro lado, um advérbio de decremento (-1) seguido de uma palavra-chave de inferioridade (-1) indica uma expressão de superioridade para a entidade associada a essa palavra-chave, como na sentença "Smartphone X é **quase** *pior* do que Smartphone Y", em que a palavra-chave *pior* representa inferioridade (-1) e é precedida pelo advérbio de decremento quase (-1). Multiplicando a orientação desses termos, é obtido um valor positivo (quase (-1) * pior (-1) = +1), que indica superioridade à entidade Smartphone X.

5.4.3 Negação

A existência de negação em uma sentença é outra situação que precisa ser tratada, uma vez que as negações alteram a maneira que a sentença deve ser interpretada, pois modificam as afirmações feitas, invertendo a polaridade e o significado de uma frase. As palavras não, nada, nunca, nem e nenhum, são alguns exemplos de negações empregadas. No exemplo da frase "O produto X **não** é *superior* do que o produto Y", a palavra-chave comparativa *superior* indica superioridade do produto X, porém a negação modifica a afirmação, fazendo com que o produto Y seja o preferido.

Nesse contexto, a inversão da orientação é uma das técnicas adotadas para tratar a existência de negação. Sendo assim, a estratégia aqui proposta inverte a orientação original da palavra-chave comparativa caso exista negação associada a ela. No exemplo da frase "O produto X **não** é *superior* ao produto Y", a palavra-chave comparativa *superior* indica superioridade, porém essa orientação é invertida devido à existência da negação. Assim, executando a estratégia de detecção de preferência após a alteração da orientação, é possível determinar que o produto Y é o preferido.

Algoritmo 2 Algoritmo que determina a entidade preferida considerando os casos especiais.

```
1:  procedure getPreference(texto, palavra-chave, ent1, ent2)
2:      orientacaoPalavra = getWordOrientation(palavra-chave)
3:      entRelacionada = getRelatedEntity(texto, palavra-chave, ent1, ent2)
4:      if palavra-chave contains expressão de decremento then
5:          orientacaoPalavra = orientacaoPalavra * (-1)
6:      else if palavra-chave contains expressão de incremento then
7:          orientacaoPalavra = orientacaoPalavra * (1)
8:      end if
9:      if palavra-chave contains expressão negativa then
10:         orientacaoPalavra = orientacaoPalavra * (-1)
11:     end if
12:     if orientacaoPalavra > 0 then
13:         entidadePreferida = entRelacionada
14:     else
15:         entidadePreferida = if entRelacionada == ent1 then ent2 else ent1
16:     end if
17:     return entidadePreferida
18: end procedure
```

- **getWordOrientation()** - Obtém a orientação da palavra-chave comparativa (i.e., superioridade ou inferioridade).
- **getRelatedEntity()** - Obtém a entidade relacionada à palavra-chave comparativa.

5.4.4 Adicionando os casos especiais no algoritmo

Após todos os tratamentos apresentados, o Algoritmo 1 proposto para detecção de preferência é aprimorado com a inclusão das estratégias para lidar com casos especiais, como mostra o Algoritmo 2.

Nas duas primeiras linhas, obtém-se a orientação e a entidade associada à palavra-chave. Em seguida, é realizada uma operação condicional (linha 4) que verifica a existência de advérbios de decremento ou incremento. Caso existam, a orientação original da palavra é invertida, conforme a regra proposta na Seção 5.4.2. Posteriormente, caso exista uma expressão de negação relacionada à palavra-chave comparativa (linha 9), a orientação é invertida. Por fim, caso a orientação final seja positiva, a entidade associada a ela é retornada como a preferida, caso contrário, a entidade contraposta é apontada como preferida.

5.5 Comparações com Aspectos

Um dos principais desafios na detecção de preferência é identificar a entidade preferida em uma sentença cuja comparação está associada a um aspecto. As comparações entre objetos podem ser feitas de maneira direta ou de maneira mais detalhada, onde se fornece mais informações, normalmente através de aspectos que justificam a preferência ao objeto mencionado.

Nas sentenças comparativas sem aspectos, após identificar a orientação da palavra-chave utilizada na comparação, obtém-se qual entidade a palavra está associada e, em seguida, define-se a entidade preferida. Já nas comparações com aspectos, a dinâmica é um pouco diferente, uma vez que a orientação da palavra-chave comparativa sozinha não é suficiente para determinar a preferência, sendo necessário informações sobre o aspecto mencionado.

As entidades são compostas por um conjunto de aspectos, que são características inerentes a um objeto, tais como tamanho, preço, peso, design, entre outros. Esses aspectos são mencionados nas comparações quando se deseja detalhar a preferência, especificando quais características de um objeto se sobressaem a outros, como é o exemplo da sentença "A câmera do Smartphone X é *superior* ao do Smartphone Y" e da sentença "A bateria do Smartphone X tem *mais* durabilidade do que a do Smartphone Y". Em ambos exemplos, especifica-se o aspecto sobressalente, o que significa que, na perspectiva daquele aspecto mencionado, o objeto inicialmente referido é o superior. No entanto, essa preferência pode ser alterada ao avaliar diferentes características.

Apesar de informativas, as opiniões comparativas que possuem aspectos normal-

mente precisam ser interpretadas analisando o contexto da entidade avaliada, visto que os aspectos são características específicas a um objeto, não sendo possível determinar a preferência apenas com a palavra-chave comparativa. Observando o exemplo da sentença "A bateria do Smartphone X é *maior* que o Smartphone Y", percebe-se que a combinação (bateria, maior) remete a uma característica positiva, fazendo com que o Smartphone X seja o superior. No entanto, analisando a sentença "O aplicativo A tem um *maior* tempo de execução do que o aplicativo B", percebe-se que (tempo de execução, maior) refere a algo negativo. Logo, a mesma palavra-chave *maior* pode ter um sentido completamente diferente dependendo do aspecto mencionado, pois enquanto uma maior carga de bateria é algo positivo, um maior tempo de execução é uma característica negativa. Por isso, ao mencionar aspectos, a palavra-chave comparativa precisa ser analisada juntamente com o aspecto citado para determinar a preferência.

Sendo assim, esta Seção apresenta a estratégia utilizada para lidar com as comparações que possuem aspectos. A estratégia é dividida em três etapas, que inicialmente lida com os casos das comparações formadas por palavras determinantes (Seção 5.5.1) e com as comparações que possuem aspectos comuns a outros objetos (Seção 5.5.2). Por fim, considera-se os casos em que é necessário analisar o contexto para determinar a preferência (Seção 5.5.3 e Seção 5.5.4).

5.5.1 Palavras-chave comparativas determinantes

Existem algumas palavras-chave comparativas utilizadas para expressar comparação que não sofrem variação pelo contexto, ou seja, independentemente do aspecto comparado, essas palavras mantêm a mesma orientação (i.e., superioridade ou inferioridade), o que permite determinar a preferência sem a análise do aspecto. No exemplo da sentença "o preço do produto X é *melhor* do que o produto Y", o aspecto preço poderia ser substituído por qualquer outro aspecto, e ainda assim o produto X se manteria como o preferido devido ao uso da palavra *melhor*. O mesmo também ocorre na sentença "considerando a qualidade da imagem, *prefiro* o produto X ao invés do Y".

As palavras melhor, pior, prefiro, preferível, recomendo, entre outras, são determinantes para identificar a preferência em uma comparação. Logo, para esses casos, o aspecto comparado pode ser ignorado na análise, visto que não tem influência para a identificação da entidade preferida, importando apenas a orientação da palavra-chave.

Dessa maneira, caso uma sentença comparativa seja composta por alguma dessas palavras-chave comparativas determinantes e existam aspectos mencionados no texto, o fluxo padrão do algoritmo para determinar a preferência em comparações sem aspectos pode ser aplicado.

5.5.2 Aspectos comuns

Alguns aspectos como preço, custo-benefício, peso, entre outros, são características comuns compartilhadas entre a grande maioria dos produtos. Esses aspectos normalmente possuem a mesma orientação independentemente do contexto e do produto analisado, como é o caso do aspecto preço, em que não importa qual seja o produto, um preço menor é sempre preferível. O mesmo se aplica no aspecto ruído, onde um menor ruído é sempre desejável. Já para outros aspectos como custo-benefício a interpretação é modificada, pois um menor custo-benefício representa uma avaliação negativa.

Para análise da comparação com aspectos é introduzido o conceito de orientação de um aspecto, que é semelhante ao conceito de orientação de uma palavra-chave comparativa. Um aspecto recebe uma orientação positiva (+1) caso uma maior intensidade do aspecto seja vista como algo desejável, como é o exemplo do aspecto custo-benefício, em que a associação das palavras mais, maior, superior, etc, indica uma avaliação positiva, pois quanto maior o custo-benefício de um produto, melhor a avaliação. Já uma orientação negativa (-1) indica o oposto, ou seja, uma menor intensidade do aspecto é vista como positiva. Analisando os aspectos preço e ruído, percebe-se que independentemente do produto, quanto menor o preço ou quanto menor o ruído de um objeto, melhor. Logo, esses aspectos recebem uma orientação negativa (-1).

Dessa maneira, foram mapeados os principais aspectos comuns que possuem orientações constantes independentemente do produto, tais como preço, consumo de energia, ruído e tempo de execução, assinalando se a orientação é positiva (+1) ou negativa (-1).

Visto isso, para detectar a preferência quando se menciona um aspecto é necessário analisar a expressão formada pela combinação do aspecto e da palavra-chave (aspecto, palavra-chave). A estratégia proposta para identificar a orientação da expressão é composta por quatro regras:

1. Aspecto **(+1)** + Palavra-chave (superioridade **(+1)**) = **Superioridade**
2. Aspecto **(+1)** + Palavra-chave (inferioridade **(-1)**) = **Inferioridade**
3. Aspecto **(-1)** + Palavra-chave (superioridade **(+1)**) = **Inferioridade**
4. Aspecto **(-1)** + Palavra-chave (inferioridade **(-1)**) = **Superioridade**

Analisando as regras, percebe-se que a orientação da expressão (aspecto, palavra-chave) é obtida através da multiplicação da orientação do aspecto e da palavra-chave comparativa, o que permite identificar se a expressão indica ou não preferência.

Nesse contexto, um aspecto com orientação negativa (-1) (e.g., preço, ruído, consumo de energia, etc) seguido de uma palavra-chave comparativa que indica superioridade resulta em uma expressão que designa inferioridade à entidade associada à palavra-chave (regra 3). Por exemplo, "o preço (-1) do produto X é *maior* (+1) do que o produto Y". Por outro lado, um aspecto negativo seguido de uma palavra-chave que indica inferioridade resulta em uma expressão de superioridade (regra 4). Por exemplo, "o ruído (-1) do produto X é *menor* (-1) do que o produto Y".

5.5.3 Aspectos dependentes de contexto

Alguns aspectos, apesar de serem compartilhados entre diferentes produtos, podem ter sentidos completamente diferentes, dependendo do objeto mencionado. Um exemplo é o aspecto tamanho, onde uma maior dimensão pode ser algo positivo ou indesejável, dependendo do contexto avaliado. Considerando uma televisão, uma maior dimensão da tela pode ser uma característica positiva, no entanto, levando em conta o mesmo aspecto em um produto com finalidades portáteis e práticas, um maior comprimento pode ser visto como um problema e algo a ser evitado. O mesmo pode ser aplicado para o aspecto de temperatura, onde normalmente uma temperatura elevada em dispositivos eletrônicos é totalmente indesejável. No entanto, considerando um aquecedor, a temperatura alta e o aquecimento completo de um ambiente é sua principal finalidade.

Um dos principais desafios da detecção de preferência é tratar as sentenças que mencionam aspectos que são dependentes do contexto, pois para analisá-las é necessário observar o contexto e as finalidades específicas do objeto mencionado para que seja possível determinar se a avaliação feita ao produto é negativa ou positiva, indicando superioridade de um objeto em detrimento ao outro.

Nesse contexto, a grande parte das estratégias adotadas para estudo desses casos decorrem da análise específica do produto avaliado e de palavras com sentimentos próximas, para, então, determinar se a expressão relacionada ao objeto indica preferência ou não.

Apesar dessas peculiaridades, pensando em um cenário prático, os produtos possuem quantidades limitadas de aspectos, que podem ser facilmente enumerados assinalando suas orientações (e.g., superioridade ou inferioridade). Considerando um contexto comercial onde se deseja analisar comparações acerca de produtos específicos, as companhias conhecem a fundo seus produtos, em que as informações sobre os aspectos podem ser facilmente adquiridas, o que permite a análise dessas comparações de maneira automática.

Visto isso, como solução para análise das preferências através de contexto, para

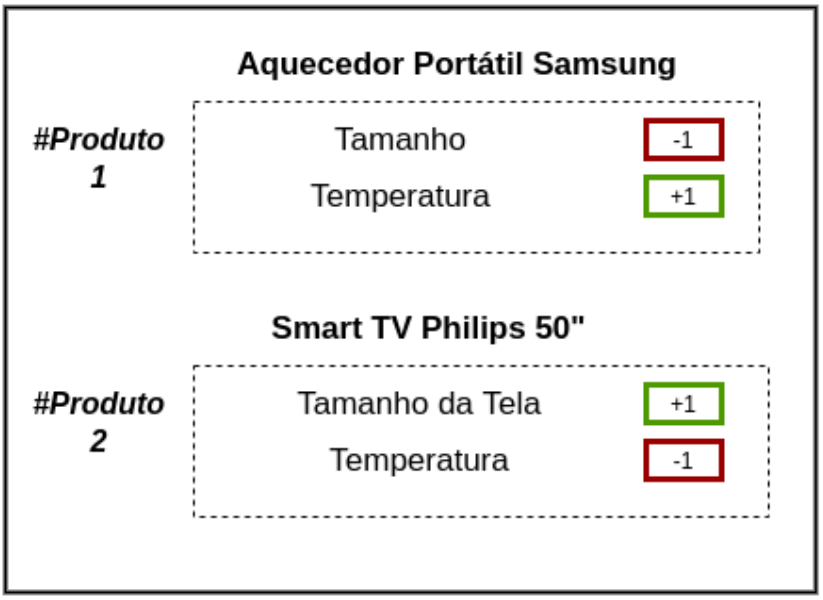


Figura 5.4: **Lista com o mapeamento da orientação dos aspectos que dependem do contexto.**

cada um dos produtos analisados, é criado uma lista onde são mapeados os aspectos cuja orientação depende da análise do contexto, destacando a orientação de cada um deles para cada um dos produtos, formando uma estrutura similar a um dicionário, conforme mostra a Figura 5.4. Os aspectos de cada um dos produtos recebem uma orientação, que pode ser positiva (+1), quando uma maior intensidade do aspecto é vista como algo desejável, caso contrário é assinalado um valor negativo (-1).

Sendo assim, após a obtenção da orientação de cada um desses aspectos, é possível estabelecer a preferência através da utilização da regra de multiplicação da orientação do aspecto e da palavra-chave. Dessa maneira, um aspecto com orientação negativa seguido de uma palavra-chave negativa indica uma expressão de superioridade à entidade associada, por exemplo "A temperatura (-1) da Smart TV X é *menor* (-1) do que o Smart TV Y".

1. Aspecto **(+1)** + Palavra-chave (superioridade **(+1)**) = **Superioridade**
2. Aspecto **(+1)** + Palavra-chave (inferioridade **(-1)**) = **Inferioridade**
3. Aspecto **(-1)** + Palavra-chave (superioridade **(+1)**) = **Inferioridade**
4. Aspecto **(-1)** + Palavra-chave (inferioridade **(-1)**) = **Superioridade**

5.5.4 Comparações com as palavras-chave mais e menos

As comparações que envolvem as palavras-chave comparativas *mais* e *menos* normalmente são acompanhadas de aspectos, o que aponta para a necessidade da avaliação de contexto para interpretá-las. Além disso, as palavras-chave *mais* e *menos* são utilizadas nas comparações como advérbios de intensidade e normalmente são sucedidas por algum adjetivo.

Nesses casos, não é possível determinar a preferência em uma sentença apenas analisando a palavra-chave e o aspecto, sendo necessário também observar o termo que sucede a palavra-chave comparativa. No exemplo "O design do produto X é *mais* bonito que o produto Y", considerando apenas a expressão (design, mais), percebe-se que não é suficiente para determinar a preferência, sendo imprescindível incluir o adjetivo mencionado (design, mais bonito), pois somente com a expressão *mais bonito* é possível identificar o produto X como preferido na sentença.

No entanto, considerando a mesma frase modificando apenas o adjetivo mencionado "O design do produto X é *mais* feio que o produto Y", ocorre um efeito contrário, onde o produto Y passa a ser o preferido, pois a expressão *mais feio* denota um sentimento negativo e de inferioridade para a entidade associada a ela.

Sendo assim, de maneira similar à análise das comparações com aspectos, para lidar com as comparações que utilizam palavras comparativas *mais* e *menos*, é necessário observar o contexto do aspecto e da expressão comparativa formada pela combinação da palavra-chave mais/menos e do termo que sucede, formando a expressão (mais/menos + palavra sucedida). Assim, para cada produto, é necessário mapear a orientação dos aspectos juntamente com a expressão (mais/menos + palavra sucedida), adicionando em uma lista similar à feita na Seção 5.5.3.

Em um cenário prático, onde companhias conhecem sobre seus produtos, as orientações das expressões com as palavras *mais* e *menos* podem ser facilmente identificadas. Com isso, é possível determinar a preferência indicada em uma sentença através da multiplicação da orientação, como o exemplo da sentença "O design (+1) relógio X é *mais* elegante do que o relógio Y", onde a expressão (design, mais elegante) tem uma orientação positiva (+1), fazendo com o que o relógio X seja o preferido.

Algoritmo 3 Algoritmo final que determina a entidade preferida nas sentenças Gradativas com Predileção.

```
1: procedure getPreference(texto, palavra-chave, aspecto, ent1, ent2)
2:     orientacaoPalavra = getWordOrientation(palavra-chave)
3:     entRelacionada = getRelatedEntity(texto, palavra-chave, ent1, ent2)
4:     if palavra-chave contains expressão de decremento then
5:         orientacaoPalavra = orientacaoPalavra * (-1)
6:     else if palavra-chave contains expressão de incremento then
7:         orientacaoPalavra = orientacaoPalavra * (1)
8:     end if
9:     if palavra-chave contains expressão negativa then
10:        orientacaoPalavra = orientacaoPalavra * (-1)
11:    end if
12:    if aspecto is not nulo then
13:        orientacaoAspecto = getFeatureOrientation(aspecto)
14:        orientacaoPalavra = orientacaoPalavra * orientacaoAspecto
15:    end if
16:    if orientacaoPalavra > 0 then
17:        entidadePreferida = entRelacionada
18:    else
19:        entidadePreferida = if entRelacionada == ent1 then ent2 else ent1
20:    end if
21:    return entidadePreferida
22: end procedure
```

- **getWordOrientation()** - Obtém a orientação da palavra-chave comparativa (i.e., superioridade ou inferioridade).
- **getRelatedEntity()** - Obtém a entidade relacionada à palavra-chave comparativa.
- **getFeatureOrientation()** - Obtém a orientação do aspecto de um determinado produto, executando os procedimentos conforme descrito na Seção 5.5.

5.6 Algoritmo Final para as Sentenças Gradativas com Predileção

Com a análise de cada um dos elementos contidos em uma opinião comparativa, juntamente com o tratamento dos casos especiais e também da análise do contexto quando existem aspectos mencionados, é possível implementar um algoritmo que permite a identificação da entidade preferida na grande parte das sentenças comparativas.

O Algoritmo 3 sumariza todas as etapas e tratamentos propostos nas seções anteriores para identificação da entidade preferida em uma sentença comparativa. O algoritmo recebe cinco parâmetros, que são: a sentença comparativa (*texto*), a palavra-

chave comparativa utilizada na sentença (*palavra-chave*), o aspecto (*aspecto*) e a primeira e segunda entidade (*ent1* e *ent2*) mencionada na comparação.

Inicialmente, as informações sobre a orientação da palavra-chave comparativa e a entidade associada à palavra-chave comparativa (linha 2 e 3) são obtidas. Com essas informações, é executada uma operação condicional (linha 4) que verifica a existência de advérbios de intensidade que podem modificar a orientação padrão da palavra-chave comparativa. De maneira similar, caso existam expressões de negações relacionadas à palavra-chave (linha 9), a orientação da palavra-chave é invertida.

Em seguida, através do método *getFeatureOrientation()*, caso existam aspectos na comparação, é obtido a orientação do aspecto mencionado (linha 12), que é multiplicado pela orientação da palavra-chave comparativa.

Por fim, caso a orientação (*orientacaoPalavra*) seja positiva, a entidade associada a ela é retornada como a preferida, caso contrário, a entidade contraposta é indicada como preferida.

5.7 Algoritmo Final para as Sentenças Superlativas

A principal diferença entre as comparações Gradativas com Predileção e as Superlativas é o número de entidades mencionadas na comparação. Enquanto o primeiro grupo conta com duas ou mais entidades, as Superlativas comparam um objeto a um grupo de outros objetos, expressando superioridade ou inferioridade entre eles.

Sendo assim, para detectar a entidade preferida em Superlativos é necessário realizar alguns ajustes no Algoritmo 3. Este algoritmo inicialmente proposto para as Gradativas com Predileção recebe como parâmetro as duas entidades mencionadas na comparação, o que não faz sentido para as Superlativas, uma vez que elas não possuem duas entidades explícitas.

Para lidar com isso, ao invés de encontrar qual a entidade preferida, o foco dado para as Superlativas está em detectar se a entidade mencionada é avaliada positivamente, ou seja, se a entidade mencionada é superior ao grupo de objetos contrapostos, o que permite identificar a entidade mencionada como a preferida. Como as Superlativas em grande parte mencionam uma única entidade, não é necessário executar a etapa de obtenção da entidade associada à palavra-chave como é feito para as Gradativas com Predileção. Dessa maneira, caso a entidade mencionada não seja a preferida, automaticamente significa que o grupo de produtos é superior à entidade mencionada.

Na sentença "O celular X é *superior* a todos os outros", pode-se considerar o celular X como a entidade mencionada e "todos os outros" como o grupo de objetos

Algoritmo 4 Algoritmo que indica se a entidade mencionada em uma sentença superlativa é a preferida ou não.

```
1: procedure isSuperlativeEntityPreferred(texto, palavra-chave, aspecto)
2:     orientacaoPalavra = getWordOrientation(palavra-chave)
3:     if palavra-chave contains expressão negativa then
4:         orientacaoPalavra = orientacaoPalavra * (-1)
5:     end if
6:     if palavra-chave contains expressão de decremento then
7:         orientacaoPalavra = orientacaoPalavra * (-1)
8:     else if palavra-chave contains expressão de incremento then
9:         orientacaoPalavra = orientacaoPalavra * (1)
10:    end if
11:    if aspecto is not nulo then
12:        orientacaoAspecto = getFeatureOrientation(aspecto)
13:        orientacaoPalavra = orientacaoPalavra * orientacaoAspecto
14:    end if
15:    if orientacaoPalavra > 0 then
16:        Return True
17:    else
18:        Return False
19:    end if
20: end procedure
```

- **getWordOrientation()** - Obtém a orientação da palavra-chave comparativa (i.e., superioridade ou inferioridade).
- **getFeatureOrientation()** - Obtém a orientação do aspecto de um determinado produto, executando os procedimentos conforme descrito na Seção 5.5.

contrapostos, sendo que o celular X é indicado como o preferido.

O Algoritmo 4 detalha os passos utilizados para encontrar preferência em comparações Superlativas. Na assinatura da função, o algoritmo recebe os parâmetros: sentença comparativa (*texto*), palavra-chave comparativa utilizada (*palavra-chave*), o aspecto do produto mencionado (*aspecto*). O algoritmo inicia obtendo a orientação da palavra-chave comparativa (linha 2). Em seguida, é verificado a existência de negação (linha 3), de advérbios modificadores que alteram a orientação padrão da palavra-chave comparativa (linha 9) e de aspectos (linha 11). Por fim, caso a orientação seja positiva, significa que a entidade mencionada no texto é a preferida, caso contrário, o grupo de entidades contrapostas é superior.

Tabela 5.1: Resultado do algoritmo de detecção de preferência para as sentenças Gradativas com Predileção na base do Buscapé e Twitter.

Buscapé			
Preferência	Precisão	Revocação	F1-Score
Primeira Entidade Preferida	0,799	0,846	0,822
Segunda Entidade Preferida	0,873	0,832	0,852
Macro-F1	**0,837**		
Acurácia	**0,838**		
Twitter			
Preferência	Precisão	Revocação	F1-Score
Primeira Entidade Preferida	0,802	0,873	0,836
Segunda Entidade Preferida	0,877	0,808	0,841
Macro-F1	**0,839**		
Acurácia	**0,839**		

5.8 Resultados

Esta Seção apresenta a avaliação do Algoritmo 3 e Algoritmo 4, propostos para detecção da entidade preferida das opiniões Gradativas com Predileção e Superlativas. A principal diferença entre eles, é que o procedimento realizado para detectar a preferência nas sentenças Superlativas não necessita da execução do *dependency parsing*, uma vez que não existem duas entidades mencionadas na sentença. Por essas pequenas peculiaridades no tratamento de cada tipo de opinião, os algoritmos são analisados e avaliados separadamente.

Para a avaliação, foram empregadas métricas utilizadas na avaliação de modelos de aprendizado de máquina em problemas de classificação [Baeza-Yates et al., 1999]. Além disso, foram utilizadas as sentenças rotuladas nas bases de dados do Twitter e Buscapé, construídas na Seção 3.4. Ao todo, existem 910 sentenças assinaladas como Gradativas com Predileção e 602 sentenças como Superlativas. Essas sentenças, além de possuírem a marcação do respectivo tipo opinativo, possuem informações adicionais sobre as entidades mencionadas na sentença, a entidade expressa como preferida e os aspectos envolvidos na comparação.

Inicialmente, o algoritmo proposto para detecção da entidade preferida nas Gradativas com Predileção (Algoritmo 3) foi testado em cada uma das bases de dados, alcançando uma acurácia de 83,8% na base de dados do Buscapé e 83,9% na base do Twitter, conforme mostra a Tabela 5.1. Os resultados são sumarizados em duas classes. A primeira, é quando a primeira entidade que aparece na sentença é indicada como preferida, já a segunda classe é quando a segunda entidade mencionada é apontada

Tabela 5.2: Resultado do algoritmo de detecção de preferência para as sentenças Superlativas na base do Buscapé e Twitter.

Buscapé			
Preferência	Precisão	Revocação	F1-Score
Entidade Mencionada Preferida	0,993	0,979	0,986
Entidade Mencionada Não Preferida	0,778	0,913	0,840
Macro-F1	**0,913**		
Acurácia	**0,974**		
Twitter			
Preferência	Precisão	Revocação	F1-Score
Entidade Mencionada Preferida	0,979	0,974	0,976
Entidade Mencionada Não Preferida	0,895	0,911	0,903
Macro-F1	**0,940**		
Acurácia	**0,962**		

como preferida na comparação.

Observando os resultados em ambas bases de dados, não percebe-se uma diferença significativa entre as sentenças analisadas em cada contexto. Apesar das redes sociais serem conhecidas por uma linguagem mais informal e, se tratando de comparações, existir uma maior variedade de objetos comparados que vão além do contexto de produtos e serviços comuns em sites de revisões, todas essas diferenças passam despercebidas ao observar apenas os resultados obtidos.

Uma das razões é que, como estamos trabalhando a nível de sentença, os fatores de contexto que poderiam gerar uma disparidade nos resultados são atenuados devido aos tratamentos propostos. Como é o caso da entidade oculta, que sem o tratamento que adiciona a entidade faltante na sentença, poderia ocasionar inconsistências ao determinar a entidade associada à palavra-chave.

Já analisando as sentenças Superlativas, o Algoritmo 4 foi avaliado nas duas bases de dado de maneira similar às Gradativas com Predileção. Uma vez que essas sentenças não comparam duas entidades explicitamente como as Gradativas com Predileção, as classes utilizadas para avaliação da abordagem se referem quando: (1) a entidade mencionada na sentença é a preferida; e (2) a entidade mencionada é inferior ao grupo de objetos contrapostos, ou seja, a entidade mencionada no texto não é a superior ou preferida na comparação.

Após executar o algoritmo na base de dados, percebe-se um resultado de acurácia superior se comparado com o resultado obtido para as opiniões Gradativas com Predileção. Para o Buscapé, obteve-se um valor de acurácia de 97,4%, já para o Twitter obteve-se um valor de 96,2%, conforme mostra a Tabela 5.2.

Tabela 5.3: Razões que levam a detecção incorreta da entidade preferida nas Gradativas com Predileções.

Razões para o erro na classificação *(Gradativa com Predileção)*	**Buscapé**	**Twitter**
Entidade Oculta	2,09%	0%
Erro do Dependency Parse	11,65%	11,94%
Orientação da palavra-chave mal estimada	1,22%	0,59%
Preferência modificada pelo contexto	0,87%	2,09%
Negação	0,34%	1,5%
Erro Total (%)	**16,17%**	**16,12%**

Tabela 5.4: Razões que levam a detecção incorreta da entidade preferida nas Superlativas.

Razões para o erro na classificação *(Superlativa)*	**Buscapé**	**Twitter**
Orientação da palavra-chave mal estimada	1,28%	2,41%
Negação	0,64%	0%
Preferência modificada pelo contexto	0,64%	1,38%
Erro Total (%)	**2,56%**	**3,79%**

Essa disparidade nos resultados acontece pela característica de cada opinião e pelas número de etapas necessárias para identificar a preferência em cada caso. Os resultados são melhor entendidos ao observar as razões que levam o algoritmo identificar incorretamente a a preferência, detalhadas na Tabela 5.3 e 5.4.

5.8.1 Analisando as razões para a incorreta detecção da preferência

Para compreender os motivos dessa diferença nos resultados observados, é importante analisar as características das opiniões Gradativas com Predileção e Superlativas.

Analisando a Tabela 5.3, é possível observar as razões que levam aos erros de classificação na detecção da preferência nas opiniões Gradativas com Predileção. Dos quase 16% de erro, pouco mais de 11% é devido ao *dependency parsing*, que quando não captura corretamente as dependências da sentença, acaba não sendo possível encontrar a entidade associada à palavra-chave e, consequentemente, a entidade preferida.

Apesar do *dependency parsing* ser uma ferramenta muito útil e, de maneira geral, ter uma boa taxa de acerto, ainda existem muitos desafios a serem enfrentados quando se trata da língua portuguesa. A existência de vícios de linguagem, falta de acentuação, ausência de vírgulas, expressões informais não conhecidas, entre outras características gramaticais, são exemplos de situações que podem impactar consideravelmente no desempenho ao analisar uma sentença.

Mesmo com essas questões, pode-se dizer que o *dependency parsing* alcança re-

sultados satisfatórios para o português. Os resultados mostram uma assertividade de aproximadamente 90% para as sentenças Gradativas com Predileção.

Analisando a Tabela 5.4, que apresenta as razões para os erros de classificação das opiniões Superlativas, percebe-se que a ausência da etapa de *dependency parsing*, resulta em um erro agregado bem inferior ($\approx 3\%$), uma vez que a principal tarefa, que é detectar a entidade associada à palavra-chave, não é necessária.

A ausência da etapa de detecção da entidade relacionada à palavra-chave, é um dos principais motivos que justificam uma maior acurácia para as opiniões Superlativas. Tendo em vista que essas sentenças possuem apenas uma entidade explicitamente mencionada, pode-se assumir que a palavra-chave comparativa se refere a essa entidade mencionada. Isso permite determinar a preferência sem a necessidade de analisar as dependências (*dependency parsing*), o que minimiza o erro agregado e simplifica o processo de determinar a preferência das opiniões Superlativas.

Além do *dependency parsing*, foram encontradas outras razões para os erros de classificação, são elas: quando o tratamento da entidade oculta não é suficiente, a existência de expressões em contextos específicos que modificam a maneira de determinar a preferência, erros para identificar uma negação e, por fim, situações onde a palavra-chave não possui orientação e, mesmo analisando palavras próximas, não é possível estimar corretamente a orientação.

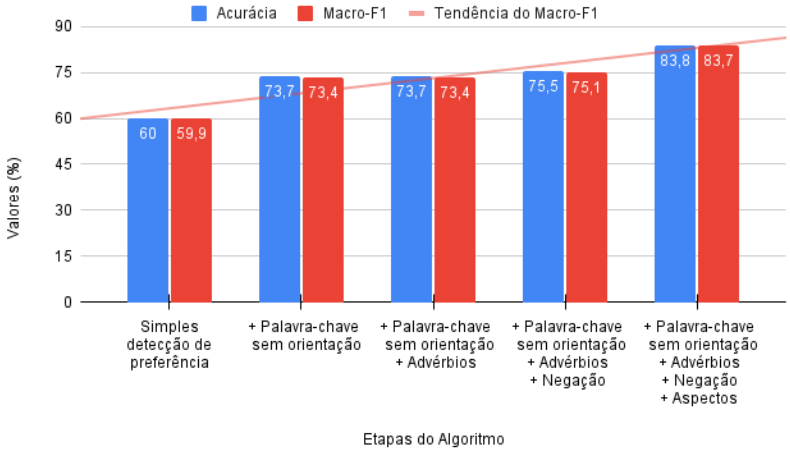
No entanto, essas situações são menos frequentes e, de maneira geral, os algoritmos propostos conseguem cobrir a ampla maioria das comparações capturando corretamente a entidade preferida das sentenças.

5.8.2 Avaliando a importância de cada etapa do algoritmo

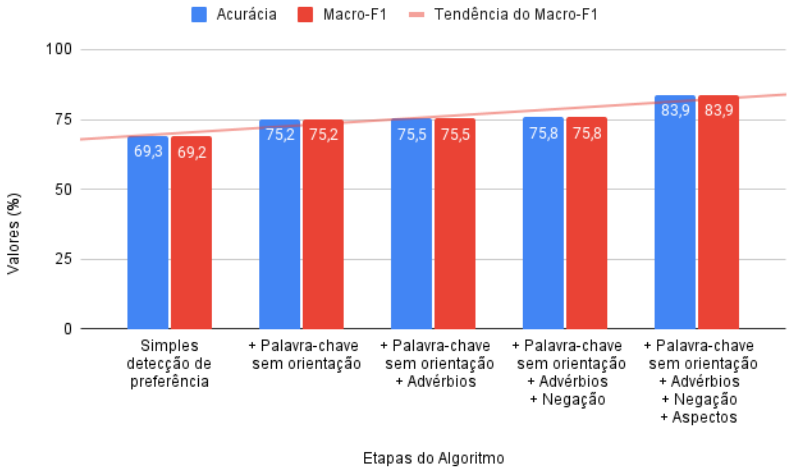
Após discutir os resultados de maneira agregada, é importante analisar individualmente cada etapa do algoritmo, para entender o impacto de cada passo no processo de determinar a preferência.

Nas Figuras 5.6(a) e 5.6(b), encontra-se as métricas de acurácia e Macro-F1 para cada uma das cinco etapas do algoritmo. A primeira barra do gráfico representa o procedimento simples de detectar a preferência, que basicamente encontra a entidade associada à palavra-chave e determina a preferência sem nenhum tratamento.

Em seguida, temos a inclusão dos tratamentos para os três casos especiais, são eles: (1) quando a palavra-chave não possui orientação; (2) advérbios de intensidade; e (3) negação. Por fim, a última barra representa o algoritmo completo, com todas as etapas anteriores mais a inclusão do tratamento de aspectos.

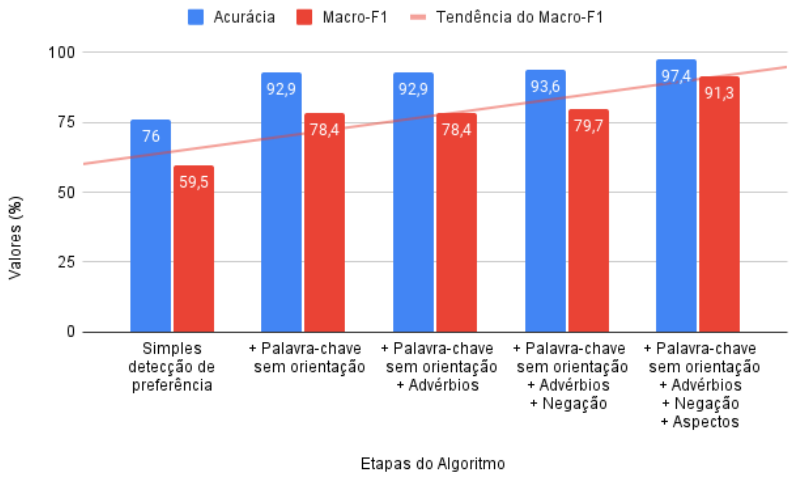


(a) Buscapé

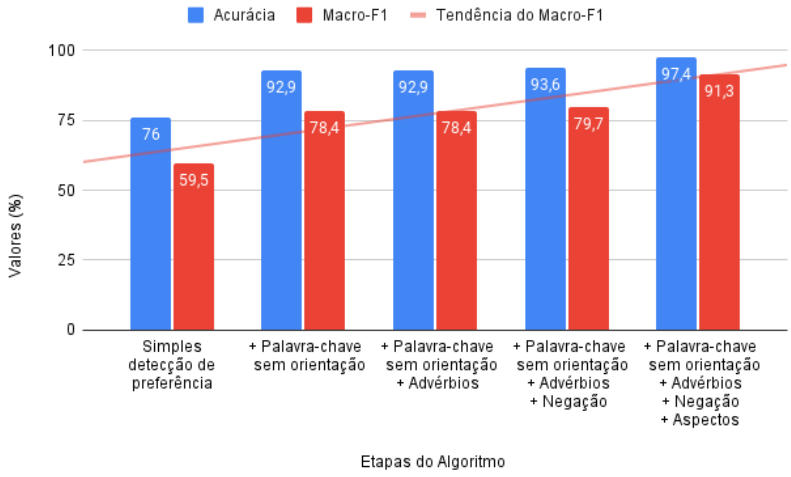


(b) Twitter

Figura 5.5: **Gradativa com Predileção - Avaliação de cada etapa do algoritmo proposto.**



(a) Buscapé



(b) Twitter

Figura 5.6: **Superlativa - Avaliação de cada etapa do algoritmo proposto.**

Analisando cada uma das etapas representadas nos gráficos, percebe-se que o tratamento para as palavras-chave sem orientação (segunda barra) e o tratamento de aspectos (última barra) são os mais importantes, uma vez que são os principais responsáveis pelo aumento considerável da assertividade do algoritmo. Essas duas etapas juntas são responsáveis por um aumento entre 14% a 20% na acurácia do algoritmo de detecção de preferência.

Por outro lado, as etapas de tratamento de advérbios de intensidade e negação tem um impacto pouco significativo na precisão do algoritmo. Uma das razões é que, apesar de serem tratamentos importantes, a base de dados não possui muitas sentenças que necessitam desses tratamentos. Logo, esses procedimentos acabam não gerando uma

diferença significativa no resultado final do algoritmo.

De maneira geral, pode-se concluir que a detecção de preferência sem nenhum tratamento não é tão eficaz, porém, com a inclusão das etapas apresentadas, o algoritmo se torna capaz de cobrir a grande maioria das comparações de maneira satisfatória, tanto as Gradativas com Predileção, quanto para as Superlativas.

Mesmo com a necessidade de tratar os casos de exceção, representada por cada uma das etapas, a estratégia de considerar a orientação da palavra-chave como uma maneira de expressar preferência e, ao mesmo tempo, incluir os tratamentos, se mostra uma maneira eficaz de interpretar as preferências.

Observando os resultados analisados, cada uma das etapas colabora com o aumento da cobertura e precisão do algoritmo, que se mostra capaz de encontrar a preferência mesmo em contextos diferentes, como é o caso do Buscapé e Twitter.

Por fim, os resultados obtidos se mostram promissores, e a abordagem proposta para detecção da entidade preferida se mostra capaz de identificar grande parte das preferências expressas nas comparações.

Capítulo 6

Conclusão e Trabalhos Futuros

A utilização dos mercados online para compras vem aumentando, e cada vez mais clientes utilizam desses meios para buscar informações que auxiliem na tomada de decisão de comprar ou não um determinado produto. Normalmente essas informações são obtidas em sites de revisões e avaliações de produtos, mas também podem ser encontradas em fóruns e Redes Sociais Online. As opiniões nesses contextos podem ser classificadas em dois principais tipos. As opiniões regulares, que são aquelas que expressam um sentimento direto ou indireto acerca de um determinado produto e as opiniões comparativas, que contrastam dois ou mais objetos, expressando uma relação de semelhança ou diferença.

As opiniões comparativas são de grande valor para companhias que desejam entender as preferências de seus clientes. No entanto, as técnicas tradicionais de análise de sentimentos não são suficientes para o estudo das comparações, sendo necessário a utilização de técnicas apropriadas. Nesse contexto, propomos uma estratégia para o estudo das opiniões comparativas no português, onde inicialmente modelos de classificação são aplicados para distinguir as comparativas das regulares, para que posteriormente seja feita a análise e a extração de informações sobre as preferências ali expressas.

No estudo de comparações, uma das tarefas fundamentais é a distinção dos tipos comparativos e regulares, pois a partir dela é possível a aplicação de técnicas apropriadas para lidar com cada opinião. Sendo assim, neste trabalho apresentamos uma metodologia experimental para o estudo das opiniões comparativas em português, onde é proposto uma estratégia supervisionada de classificação hierárquica para a categorização dos tipos específicos de opiniões, são eles: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa.

Analisando a estrutura das sentenças comparativas, percebe-se a existência de um conjunto restrito de palavras que são frequentemente utilizadas para expressar opiniões

comparativas, como *melhor*, *pior*, *superior*, *inferior*, entre outras, que indicam a prevalência de um objeto em relação a outros. Visto isso, propomos uma abordagem léxica para a mineração de opiniões, onde um léxico com 176 palavras-chave comparativas foi construído e utilizado para encontrar sentenças comparativas.

Nota-se, que a utilização de um léxico para a busca de sentenças que possuem ao menos uma das palavras-chave é uma estratégia capaz de capturar a grande maioria das comparações existentes em um determinado contexto. Testes realizados mostraram uma taxa de revocação superior a 90%, o que indica que a estratégia de filtrar apenas por sentenças que possuem ao menos uma das palavras-chave é eficiente para encontrar comparações, favorecendo a construção de dois conjuntos de dados em dois importantes contextos, que são: (1) sites de revisões/avaliações; e (2) Redes Sociais Online, formando duas bases de dados com 2.754 e 2.053 sentenças rotuladas.

Inicialmente, a estratégia de classificação hierárquica proposta no trabalho foi dividida em duas etapas, onde é realizada uma classificação binária com o objetivo de separar as sentenças nos dois tipos fundamentais, comparativos e regulares. Nessa etapa, quatro modelos de aprendizado de máquina foram avaliados, onde o Multinomial Naive Bayes (NB) apresentou os melhores resultados, com uma acurácia de 86,4% na base de dados do Buscapé e 85,4% no Twitter. Esses resultados apontam que, apesar dos desafios inerentes à língua portuguesa, as comparações fazem uso de expressões que normalmente permitem distingui-las das opiniões regulares. Além disso, apesar das peculiaridades existentes em cada um dos contextos analisados, não se encontrou uma diferença significativa nos resultados.

Em seguida, essas sentenças classificadas como comparativas na etapa anterior são categorizadas nos cinco tipos opinativos, são eles: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. O principal objetivo dessa estratégia é agrupar as opiniões nos seus respectivos grupos, o que permite a análise mais detalhada de cada opinião. Nessa etapa, os quatro algoritmos de aprendizado de máquina foram avaliados e, diferente da classificação binária, o Logistic Regression (LR) e o Support Vector Machine (SVM) apresentaram os melhores resultados, com 66,7% de acurácia e 61,9% de Macro-F1 no Buscapé, 66,7% de acurácia e 61,6% de Macro-F1 no Twitter. Contrastando com os resultados obtidos na classificação binária, percebe-se um declínio da métrica de acurácia, o que é justificável devido ao aumento no número de classes e também devido à similaridade existente entre os tipos de comparações, o que aumenta a complexidade na classificação das sentenças.

Porém, considerando a classificação em múltiplas classes como uma estratégia de agrupamento, percebe-se que é possível agrupar as comparações Gradativas com Predileção, Equitativas e Superlativas com taxas de revocação de até 85%, o que indica que

a abordagem é capaz de capturar as principais comparações que realmente importam para a análise de preferências.

Por outro lado, as Não Gradativas apresentam um resultado inferior, já que não possuem um padrão claro que permita uma melhor identificação. No entanto, isso não é um problema, visto que dentre as opiniões comparativas, elas são as que menos favorecem uma análise mais detalhada, pois não indicam relação de graus entre os objetos comparados, inviabilizando a aplicação da estratégia de detecção de preferência.

A preferência em uma comparação é entendida como o ato de escolher um objeto em detrimento a outros e é uma das principais informações a serem extraídas, por exemplo "O produto X é *melhor* que o produto Y", onde a primeira entidade mencionada é apontada como superior ao ser contrastada com o produto Y. Dentre as opiniões, as Gradativas com Predileção e Superlativas são as comparações que tendem a receber uma maior atenção, pois as Equitativas e Não Gradativas não possuem relações de preferência, não fazendo sentido a aplicação dessas técnicas de detecção de preferência.

Sendo assim, propomos um primeiro algoritmo para a detecção de preferência em sentenças comparativas no português, onde dado uma sentença, é retornado qual a entidade é indicada como preferida. O algoritmo inicialmente recebe os elementos contidos na sentença, como as entidades, os aspectos e a palavra-chave comparativa mencionada. Logo, a partir das dependências das palavras na sentença e da orientação da palavra-chave mencionada no texto, é possível, então, determinar qual a entidade preferida. Analisando os resultados, a estratégia se mostra promissora, atingindo uma acurácia próxima a 84% para as opiniões Gradativas com Predileção e aproximadamente 97% de acurácia para os Superlativos.

Esses resultados se mostram promissores, detectando a maior parte das preferências expressas nas comparações, que é uma das informações mais relevantes para serem obtidas, tendo uma aplicação prática no gerenciamento de reputação e em estratégias que visam entender a perspectiva de clientes na busca de novas oportunidades.

De maneira geral, os resultados mostram que a abordagem hierárquica para classificação das comparações e a análise da entidade preferida são promissoras, possibilitando análises mais sistemáticas e detalhadas para cada tipo opinativo. Em suma, o nosso trabalho apresenta o primeiro esforço para o estudo das opiniões comparativas na língua portuguesa, criando um arcabouço que viabiliza a detecção de comparações e análise das preferências ali expressas.

Nesse contexto, percebe-se que a falta de dados rotulados é um fator que ainda inibe o desenvolvimento de soluções práticas para análise de comparações na língua portuguesa, que ainda é muito pouco explorada. Sendo assim, este estudo propôs a criação de um léxico com palavras comparativas, viabilizando a mineração de opiniões

na construção de conjuntos de dados. Nesse sentido, um primeiro esforço foi feito na língua portuguesa para a construção de duas bases de dados e para a criação de um mecanismo que possibilitasse detectar automaticamente comparações, extraindo delas informações relevantes sobre preferências.

Ao todo, nosso trabalho apresenta quatro principais contribuições: (1) uma abordagem hierárquica de aprendizado de máquina para detecção de opiniões comparativas; (2) um algoritmo para detecção da entidade preferida de uma comparação; (3) a construção de duas bases de dados com aproximadamente cinco mil sentenças rotuladas; e (4) um léxico com palavras e expressões comparativas utilizadas na língua portuguesa.

Em relação aos trabalhos futuros, planeja-se inicialmente desenvolver uma ferramenta que implemente as técnicas desenvolvidas neste trabalho. Isso possibilitará que pesquisadores e empresas apliquem de maneira prática as estratégias de mineração de opinião para detecção de opiniões comparativas e análise das preferências. Nesse contexto ainda existem alguns desafios a serem enfrentados. Um deles é ampliação do estudo para outros idiomas, que normalmente exigem soluções específicas para tratamento das peculiaridades existentes, visto que a comparação é uma figura de linguagem que possui aspectos que nem sempre são compartilhados entre os diferentes idiomas.

Além disso, alguns outros contextos podem ser melhor analisados, como fóruns de discussão, onde encontram-se cascatas de discussões, em que uma opinião comparativa pode ser descoberta a partir de respostas feitas a um comentário, muitas vezes sendo necessário a análise de todo o fluxo para determinar a existência da comparação.

As técnicas para processamento das comparações ainda podem ser aprimoradas para lidar com ironia e sarcasmo, que são características complexas que demandam soluções específicas. Outra característica é a hipérbole, que é uma figura de linguagem que expressa exagero, utilizada quando se deseja intensificar o aspecto de uma determinada entidade. Normalmente, é realizada através da dramaticidade, como é o caso da sentença "Esse celular *parece* um caminhão", em que ocorre a aproximação de objetos de universos diferentes, um fenômeno conhecido como símile híbrida, que indica a existência de comparação e metáfora em uma mesma oração. Nesses casos, a sentença não deve ser interpretada de maneira literal, pois um celular e um caminhão são objetos totalmente diferentes, apesar de possuírem alguns aspectos em comum. A expressão "parece um caminhão" citada anteriormente deseja dar ênfase a algum aspecto negativo do celular, podendo ser o peso, o tamanho ou até mesmo o design do aparelho.

Finalmente, a construção e ampliação de conjuntos de dados é uma das tarefas fundamentais para continuação do estudo de opiniões, pois possibilita a aplicação de novas técnicas, que na grande maioria das vezes exigem um conjunto maior de dados para treinamento. Assim, técnicas podem ser aprimoradas e novas análises podem ser

realizadas buscando novos padrões e *features* semânticas que auxiliem na classificação de opiniões que possuem traços maiores de similaridade, como é o caso das Não Gradativas. Por fim, a combinação de métodos pode ser utilizada de maneira alternativa ou até mesmo complementar as técnicas já existentes.

Referências Bibliográficas

Al Amrani, Y.; Lazaar, M. & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127:511--520.

Alaei, A. R.; Becken, S. & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research*, 58(2):175--191.

Araújo, M.; Diniz, J. P.; Bastos, L.; Soares, E.; Júnior, M.; Ferreira, M.; Ribeiro, F. & Benevenuto, F. (2016). ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Proceedings of the International AAAI Conference on Web-Blogs and Social Media (ICWSM)*, volume 10, p. 758—759.

Bach, N. X.; Van, P. D.; Tai, N. D. & Phuong, T. M. (2015). Mining vietnamese comparative sentences for sentiment analysis. In *Proceedings of the 7th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 162--167.

Baeza-Yates, R.; Ribeiro-Neto, B. et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bakshi, R. K.; Kaur, N.; Kaur, R. & Kaur, G. (2016). Opinion mining and sentiment analysis. In *Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 452--455.

Balage Filho, P.; Pardo, T. A. S. & Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pp. 215—219.

Bespalov, D.; Bai, B.; Qi, Y. & Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM International Conference on Information and knowledge management (CIKM)*, pp. 375--382.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Chen, L. & Wang, F. (2013). Preference-based clustering reviews for augmenting e-commerce recommendation. *Knowledge-Based Systems*, 50:44--59.

Chen, T.; Xu, R.; He, Y. & Wang, X. (2017). Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221--230.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37--46.

da Rocha Lima, C. H. (2017). *Gramática normativa da língua portuguesa*. José Olympio, 53 edição. ISBN 9788503010221.

Dave, K.; Lawrence, S. & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web (WWW)*, pp. 519--528.

de O. Carosia, A. E.; Coelho, G. P. & Silva, A. E. d. (2019). The influence of tweets and news on the brazilian stock market through sentiment analysis. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web (WebMedia)*, pp. 385--392.

Ding, X.; Liu, B. & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2nd International Conference on Web Search and Data Mining (WSDM)*, pp. 231--240.

Ding, X.; Liu, B. & Zhang, L. (2009). Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1125--1134.

El-Halees, A. M. (2012). Opinion mining from arabic comparative sentences. In *Proceedings of the 13th International Arab Conference on Information Technology (ACIT)*, pp. 265--271.

Eldefrawi, M. M.; Elzanfaly, D. S.; Farhan, M. S. & Eldin, A. S. (2019). Sentiment analysis of arabic comparative opinions. *SN Applied Sciences*, 1(5):411.

Ganapathibhotla, M. & Liu, B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pp. 241--248.

Gao, S.; Tang, O.; Wang, H. & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, 71:19--32.

Gonçalo Oliveira, H. & Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation Journal*, 48(2):373--393.

Gupta, S.; Mahmood, A. A.; Ross, K.; Wu, C. & Vijay-Shanker, K. (2017). Identifying comparative structures in biomedical text. In *Proceedings of the 16th Biomedical Natural Language Processing Workshop (BioNLP)*, pp. 206--215.

Hartmann, J.; Huppertz, J.; Schamp, C. & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20--38.

Hartmann, N.; Avanço, L.; Balage Filho, P. P.; Duran, M. S.; Nunes, M. D. G. V.; Pardo, T. A. S.; Aluísio, S. M. et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pp. 3865--3871.

Huang, X.; Wan, X.; Yang, J. & Xiao, J. (2008). Learning to identify comparative sentences in chinese text. In *Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pp. 187--198.

Ibeke, E.; Lin, C.; Wyner, A. & Barawi, M. H. (2017). Extracting and understanding contrastive opinion through topic relevant sentences. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 395--400.

Jindal, N. & Liu, B. (2006a). Identifying comparative sentences in text documents. In *Proceedings of the 29th International Conference on Special Interest Group on Information Retrieval (SIGIR)*, pp. 244--251.

Jindal, N. & Liu, B. (2006b). Mining comparative sentences and relations. In *Proceedings of the 21st Conference on Artificial Intelligence (AAAI)*, p. 9.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pp. 137--142.

Kansaon, D.; Brandão, M. A.; Reis, J. C.; Barbosa, M.; Matos, B. & Benevenuto, F. (2020). Mining portuguese comparative sentences in online reviews. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pp. 333--340.

Kawahara, D.; Inui, K. & Kurohashi, S. (2010). Identifying contradictory and contrastive relations between statements to outline web information on a given topic. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pp. 534--542.

Kim, H. D. & Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM International Conference on Information and knowledge management (CIKM)*, pp. 385--394.

Kleinbaum, D. G.; Dietz, K.; Gail, M.; Klein, M. & Klein, M. (2002). *Logistic regression. A Self-Learning Text.* Springer, 3 edição.

Kocoń, J.; Zaśko-Zielińska, M. & Miłkowski, P. (2019). Multi-level analysis and recognition of the text sentiment on the example of consumer opinions. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 559--567.

Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L. & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

Kunneman, F.; Wubben, S.; van den Bosch, A. & Krahmer, E. (2018). Aspect-based summarization of pros and cons in unstructured product reviews. In *Proceedings of the of the 32nd International Conference on Computational Linguistics (COLING)*, pp. 2219--2229.

Law, T. J. (2019). 19 Powerful ECommerce Statistics That Will Guide Your Strategy in 2020. `https://www.oberlo.com/blog/ecommerce-statistics`. Acesso em: 28 de janeiro de 2021.

Lerman, K. & McDonald, R. (2009). Contrastive summarization: an experiment with consumer reviews. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 113--116.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data.* Springer Science & Business Media.

Liu, B. (2012). *Sentiment analysis and opinion mining*, volume 5. Morgan & Claypool Publishers.

Liu, B. (2015). Sentiment analysis: mining sentiments, opinions, and emotions.

Liu, Q.; Huang, H.; Zhang, C.; Chen, Z. & Chen, J. (2013). Chinese comparative sentence identification based on the combination of rules and statistics. In *Proceedings of the 9th International Conference on Advanced Data Mining and Applications (ADMA)*, pp. 300--310.

McCallum, A.; Nigam, K. et al. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning For Text Categorization (AAAI)*, volume 752, pp. 41--48.

Mehta, R. P.; Sanghvi, M. A.; Shah, D. K. & Singh, A. (2020). Sentiment analysis of tweets using supervised learning algorithms. In *Proceedings of the 1st International Conference on Sustainable Technologies for Computational Intelligence (ICTSCI)*, pp. 323--338.

Melo, P. F.; Dalip, D. H.; Junior, M. M.; Gonçalves, M. A. & Benevenuto, F. (2019). 10sent: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *Journal of the Association for Information Science and Technology*, 70(3):242--255.

Morinaga, S.; Yamanishi, K.; Tateishi, K. & Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 341--349.

Nasti, S. J.; Asger, M. & Butt, M. A. (2020). Automatic extraction of product information from multiple e-commerce web sites. In *Proceedings of the 2nd International Conference on Robotics and Intelligent Control (ICRIC)*, pp. 739--747.

Nasukawa, T. & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. pp. 70–77.

OECD (2020). E-commerce in the times of COVID-19. `https://read.oecd-ilibrary.org/view/?ref=137_137212-t0fjgnerdb&title=E-commerce-in-the-time-of-COVID-19`. Acesso em: 28 de janeiro de 2021.

Pang, B.; Lee, L. & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79--86.

Park, D. H. & Blake, C. (2012). Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (DSSD)*, pp. 1--9.

Paul, M. J.; Zhai, C. & Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 66--76.

Prabhat, A. & Khullar, V. (2017). Sentiment classification on big data using naïve bayes and logistic regression. In *Proceedings of the 7th International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1--5.

Rajaraman, A. & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Ramırez, M. & Sánchez, O. (2016). Ye shall know them by their verbs: How gender express their opinion in twitter. *Advances in Computational Linguistics*, p. 23.

Ren, Z. & de Rijke, M. (2015). Summarizing contrastive themes via hierarchical nonparametric processes. In *Proceedings of the 38th International Conference on Special Interest Group on Information Retrieval (SIGIR)*, pp. 93--102.

Rennie, J. D.; Shih, L.; Teevan, J. & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 616--623.

Ribeiro, F. N.; Araújo, M.; Gonçalves, P.; Gonçalves, M. A. & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1--29. ISSN 2193-1127.

Rodrigues, V. V. (2002). As construções comparativas em língua portuguesa. *Revista do GELNE*, 4(1):1--6.

Sapir, E. (1944). Grading, a study in semantics. *Philosophy of science*, 11(2):93--116.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1--47.

Selvi, S. T.; Karthikeyan, P.; Vincent, A.; Abinaya, V.; Neeraja, G. & Deepika, R. (2017). Text categorization using rocchio algorithm and random forest algorithm. In *Proceedings of the 8th International Conference on Advanced Computing (ICoAC)*, pp. 7--12.

Serrano-Guerrero, J.; Olivas, J. A.; Romero, F. P. & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18--38.

Souza, E.; Costa, D.; Castro, D. W.; Vitório, D.; Teles, I.; Almeida, R.; Alves, T.; Oliveira, A. L. & Gusmão, C. (2017). Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49--75.

Sun, S.; Luo, C. & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10--25.

Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267--307.

Tata, S. & Di Eugenio, B. (2010). Generating fine-grained reviews of songs from album reviews. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1376--1385.

Thelwall, M. & Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608--1617.

Tsytsarau, M. & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478--514.

Wang, H.; Lu, Y. & Zhai, C. (2010). Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 783--792.

Wang, Y.; Huang, M.; Zhu, X. & Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 606--615.

Yang, S. & Ko, Y. (2009). Extracting comparative sentences from korean text documents using comparative lexical patterns and machine learning techniques. In *Proceedings of the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 153--156.

Yang, S. & Ko, Y. (2011). Extracting comparative entities and predicates from texts using comparative type classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language technologies (ACL-HLT)*, pp. 1636--1644.

Apêndice A

Léxico com Palavras Comparativas

A Tabela A.1 contém todas as palavras comparativas presentes no léxico proposto e utilizado durante no trabalho. A coluna orientação, se refere à orientação da palavra em questão, sendo positiva (+1) se a palavra indica uma relação de superioridade ao objeto associado a ela, negativa (-1) se indicar inferioridade e, por fim, neutra (0) se a palavra não indicar superioridade nem inferioridade.

Tabela A.1: Orientações das palavras do Léxico comparativo construído.

Índice	**Palavra Comparativa**	**Orientação**
1	mais	1
2	maior	1
3	maiores	1
4	menos	-1
5	menor	-1
6	menores	-1
7	melhor	1
8	melhores	1
9	pior	-1
10	piores	-1
11	acima	1
12	superior	1
13	superiores	1
14	inferior	-1
15	inferiores	-1
16	abaixo	-1
		Continua na próxima página...

Tabela A.1 – Palavras e Orientações do léxico Comparativo Construído

Índice	**Palavra Comparativa**	**Orientação**
17	igual	0
18	iguais	0
19	igualmente	0
20	igualar	0
21	iguala	0
22	igualou	0
23	igualando	0
24	identico/idêntico	0
25	identica/idêntica	0
26	idênticos/identicos	0
27	identicas/idênticas	0
28	mesmo nivel/mesmo nível	0
29	diferente	0
30	diferencia	0
31	diferença	0
32	parece	0
33	parecido	0
34	parecida	0
35	parecidos	0
36	parecidas	0
37	parecendo	0
38	parecia	0
39	equivalente	0
40	similar	0
41	semelhante	0
42	assemelha	0
43	lembra	0
44	tipo	0
45	aproxima	0
46	aproximando	0
47	próximo/proximo	0
48	perto	0
49	passou	1

Continua na próxima página...

Tabela A.1 – Palavras e Orientações do léxico Comparativo Construído

Índice	**Palavra Comparativa**	**Orientação**
50	passar	1
51	passa	1
52	ultrapassou	1
53	ultrapassa	1
54	ultrapassando	1
55	disputa	0
56	disputam	0
57	disputando	0
58	enfrentou	0
59	compete	0
60	competir	0
61	competindo	0
62	ganha	1
63	ganhou	1
64	ganhando	1
65	perder	-1
66	perde	-1
67	perdeu	-1
68	perdendo	-1
69	número 1/numero 1	1
70	número um/numero um	1
71	prefiro	1
72	preferível/preferivel	1
73	prefere	1
74	preferiu	1
75	invés/inves	1
76	domina	1
77	dominar	1
78	dominado	1
79	dominando	1
80	dominou	1
81	bater	1
82	bate	1

Continua na próxima página...

Tabela A.1 – Palavras e Orientações do léxico Comparativo Construído

Índice	**Palavra Comparativa**	**Orientação**
83	bateu	1
84	batendo	1
85	superar	1
86	supera	1
87	superou	1
88	superando	1
89	vence	1
90	venceu	1
91	vencendo	1
92	topo	1
93	vantagem	1
94	vs	0
95	versus	0
96	dobro	1
97	único/unico	1
98	trocar	-1
99	troquei	-1
100	trocando	-1
101	troco	-1
102	alternativa	1
103	opção	1
104	optar	1
105	optando	1
106	optei	1
107	opte	1
108	mudar	1
109	mudando	1
110	em vez de	1
111	que nem	0
112	tão x como	0
113	tão x quanto	0
114	recomendo	1
115	recomendaria	1

Continua na próxima página...

Tabela A.1 – Palavras e Orientações do léxico Comparativo Construído

Índice	**Palavra Comparativa**	**Orientação**
116	frente	1
117	atrás;atras	-1
118	dúvida;duvida	0
119	ouro	1
120	vezes	0
121	comparar	0
122	comparação/comparacao	0
123	comparações/comparacoes	0
124	comparando	0
125	compara	0
126	comparei	0
127	comparado	0
128	comparada	0
129	comparável/comparavel	0
130	preferi	1
131	preferia	1
132	preferem	1
133	prefiram	1
134	preferimos	1
135	preferindo	1
136	escolher	1
137	escolho	1
138	escolhido	1
139	escolha	1
140	escolhi	1
141	escolheria	1
142	concorrer	0
143	concorrendo	0
144	concorrente	0
145	concorrentes	0
146	rival	0
147	cogitei	1
148	cogitando	1

Continua na próxima página...

Tabela A.1 – Palavras e Orientações do léxico Comparativo Construído

Índice	**Palavra Comparativa**	**Orientação**
149	uso	1
150	comprei	1
151	comprando	1
152	comprem	1
153	como	0
154	quero	1
155	queria	1
156	relação	0
157	principal	1
158	principais	1
159	líder/lider	1
160	lidera	1
161	sem precedentes	1
162	tem tudo o que o	0
163	tem tudo que a	0
164	tem o mesmo	0
165	tem a mesma	0
166	possui o mesmo	0
167	possui a mesma	0
168	com o mesmo	0
169	com a mesma	0
170	fica para trás/fica para tras	-1
171	fora de série/fora de serie	1
172	chega aos pés/chega aos pes	0
173	incomparável/incomparavel	1
174	inigualável/inigualavel	1
175	jamais visto	1
176	nunca vi	1

Apêndice B

Modificadores de Intensidade

As Tabelas contêm os principais advérbios de intensidade que modificam a orientação original das palavras-chave comparativas. Eles são separados em advérbios de incremento, como mostrado na Tabela B.1, e advérbios de decremento, apresentado na Tabela B.2. Os advérbios de incremento são aqueles que intensificam o significado de uma palavra-chave comparativa, enquanto os advérbios de decremento são opostos aos advérbios de incremento.

Tabela B.1: Advérbios de incremento utilizados como modificadores.

Índice	**Palavra**	**Orientação**
1	mais	incremento
2	muito	incremento
3	superior	incremento
4	tanto	incremento
5	bocado	incremento
6	ligeiramente	incremento
7	totalmente	incremento
8	minimamente	incremento
9	bastante	incremento
10	demais	incremento
11	tão	incremento
12	bem	incremento

Tabela B.2: Advérbios de decremento utilizados como modificadores.

Índice	**Palavra**	**Orientação**
1	menos	decremento
2	menor	decremento
3	inferior	decremento
4	quase	decremento