Mining Portuguese Comparative Sentences in Online Reviews

Daniel Kansaon
UFMG, Brasil
daniel.kansaon@dcc.ufmg.br

Michele A. Brandão
IFMG, Brasil
michele.brandao@ifmg.edu.br

Julio C. S. Reis
UFMG/FUMEC, Brasil
julio.reis@dcc.ufmg.br

Matheus Barbosa
UFMG, Brasil
matheusbarbosa@dcc.ufmg.br

Breno Matos
UFMG, Brasil
brenomatos@dcc.ufmg.br

Fabrício Benevenuto
UFMG, Brasil
fabricio@dcc.ufmg.br

ABSTRACT

The constant expansion of e-commerce, recently boosted due to the coronavirus pandemic, has led to a huge increase in online shopping. More and more, customers demand online reviews of products and comments on the Web to make decisions about buying a product over another. In this context, sentiment analysis techniques constitute the traditional way to summarize user's opinions that criticizes or highlights the positive aspects of a product. Sentiment analysis of reviews usually relies on extracting positive and negative aspects of products, neglecting comparative opinions. Such opinions do not directly express a positive or negative view but contrast aspects of products from different competitors. In this paper, we present the first effort towards detecting comparative sentences in Portuguese. Identifying comparative sentences is a key task for companies to know how users are comparing a product with their competitors and is essential for developing sentiment summarization applications for the end user. In addition, we present a supervised approach to automatically detect Portuguese comparative sentences, classifying them into five distinct groups: (1) Non-Comparative, (2) Non-Equal Gradable, (3) Equative, (4) Superlative e (5) Non-Gradable. To that end, this paper provides three main contributions: (1) a Portuguese lexicon list with words used to make comparisons; (2) two new Portuguese datasets with comparative sentences; and (3) a hierarchical approach for detecting multiple comparisons and classify the sentences in different groups by using state-of-art classification algorithms, reaching an accuracy of 87%.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Online shopping*.

KEYWORDS

opinion mining, comparative opinion, sentiment analysis

WebMedia '20, November 30-December 4, 2020, São Luís, Brazil

ACM ISBN 978-1-4503-8196-3/20/11…$15.00
https://doi.org/10.1145/3428658.3431081

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[1]. A tendência é que esse número aumente ainda mais com a pandemia de COVID-19[2]. A principal vantagem do comércio eletrônico é a capacidade de alcançar um grande número de pessoas em diferentes lugares, independente da distância e do tempo [20]. 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 [18]: (i) opiniões regulares, que são opiniões diretas sobre o produto, criticando ou ressaltando pontos positivos de diferentes aspectos do mesmo, 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. A capacidade de fazer comparações expressando ordem e preferência é um componente básico da cognição humana [23], 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 grande maioria abordagens léxicas [5, 26] e supervisionadas [6, 9, 19, 27]. 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

[1] 19 Powerful E-commerce Statistics That Will Guide Your Strategy in 2020: https://www.oberlo.com/blog/ecommerce-statistics. Acesso em: 25 de outubro de 2020.
[2] As pandemic pushes on, online sales grow 76% in June: https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail. Acesso em: 25 de outubro de 2020.

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 [4, 16, 17]. Em comum, tais técnicas são dependentes da língua e voltadas para o idioma inglês. Apesar de existirem esforços voltados para outros idiomas, como árabe [10, 11], chinês [15], vietnamita [2] e coreano [28], não há nenhum esforço que busca construir um sistema de detecção de expressões comparativas para o português. O idioma português está entre os 10 mais falados no mundo [25] 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 em português.

Especificamente, este trabalho apresenta a criação de um léxico[3] com palavras e expressões comparativas escritas em português, que é utilizado para encontrar opiniões comparativas, construindo então 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, neste 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 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, este trabalho contém três principais contribuições: (1) a criação de um léxico com palavras e expressões frequentemente usadas em comparações em português; (2) construção de duas bases de dados com sentenças comparativas; e (3) uma abordagem hierárquica para detecção de múltiplas comparações, classificando as sentenças em diferentes grupos.

2 REFERENCIAL TEÓRICO

Esta seção apresenta trabalhos relacionados que exploram o contexto de opiniões comparativas (Seção 2.1) e, em seguida, define alguns conceitos abordados neste trabalho (Seção 2.2).

2.1 Trabalhos Relacionados

Muitos estudos vêm desenvolvendo técnicas de análise de sentimentos [1, 11, 12, 24] e, em geral, existem três níveis de granularidade para a classificação de sentimentos e/ou emoções: (1) em nível de documentos, (2) sentenças e (3) baseadas em aspectos e entidades [18]. O foco deste trabalho está na segunda direção, ou seja, identificar dentre as revisões online as sentenças que expressam opiniões comparativas acerca de diferentes objetos. Especificamente, estamos interessados nas sentenças que comparam diferentes objetos e que podem expressar semelhanças e preferências entre eles [11, 24].

O número de estudos que lidam com a mineração de sentenças comparativas ainda é pequeno, e esse número é ainda menor para sentenças em português, que é a segunda língua mais usada no Twitter e está entre as dez mais faladas no mundo [25]. Souza *et al.* [25] 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. Nesse contexto, abordagens léxicas vêm sendo utilizadas por várias soluções em diferentes idiomas [10, 16, 17, 28]. Apesar das possíveis limitações das abordagens léxicas, o estudo feito por Jindal e Liu [16] mostra que grande parte das sentenças comparativas são caracterizadas pelo uso de um conjunto restrito 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 [16, 17].

De forma geral, este trabalho cria um léxico de palavras e expressões comparativas em português. Neste quesito, nosso esforço é complementar aos estudos anteriores que exploram outros idiomas [16, 17]. No entanto, este trabalho se difere dos demais pois apresenta uma estratégia para ampliar esse léxico inicial para diferentes contextos, agregando verbos, advérbios e expressões regulares, que em conjunto com estratégias supervisionadas, podem ser aplicadas para detecção de comparações, classificando as sentenças em comparativas e não comparativas.

2.2 Conceitos

As comparações podem ser separadas em dois principais grupos [18]: (1) comparações gradativas, que expressam relação de ordem entre as entidades comparadas na sentença, podendo ser de semelhança ou de superioridade; e (2) comparações não gradativas, que comparam objetos sem expressar ordem entre eles. Liu [18] define quatro categorias para organizar as opiniões comparativas, 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.* [11] 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.* [13] 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.* [13] no estudo de texto biomédico, mas não é usado por Eldefrawi *et al.* [11], porque esse tipo comparativo não foi encontrado no conjunto de dados em árabe. Além disso, Ramırez e Sánchez [21] propõem uma abordagem de análise de sentimentos para identificar o gênero dos usuários através de suas opiniões no

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

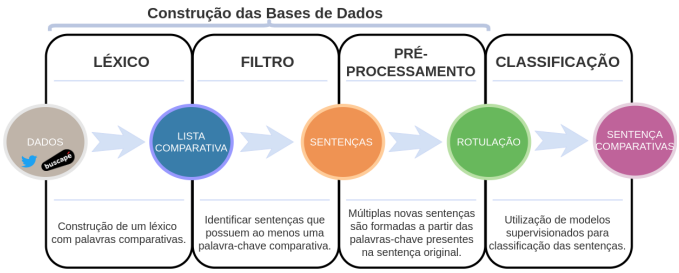


Figura 1: Etapas utilizadas na metodologia.

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.* [11] no estudo de opiniões comparativas em árabe e por Ramırez e Sánchez [21] na investigação de quais sentenças foram feitas por homens ou mulheres. No entanto, Gupta *et al.* [13] 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 dessas comparações não expressarem ordem entre os objetos, as comparações não gradativas geralmente são mais difíceis de serem detectadas, pois são mais sutis e não possuem um padrão claro [18].

Apesar das peculiaridades existentes nas classes comparativas, o trabalho explora todos esses tipos comparativos, apresentando uma estratégia para detectar sentenças comparativas, classificando-as em diferentes grupos.

3 METODOLOGIA

Esta seção apresenta detalhes relacionados à metodologia experimental adotada neste trabalho, que envolve a construção de um léxico em português (Seção 3.1), utilizado para construção das bases de dados comparativas (Seção 3.2). Além disso, é apresentado o processo de tratamento desses dados (Seção 3.3) para a etapa de classificação, conforme mostrado na Figura 1.

3.1 Construção do Léxico

Apesar da comparação ser uma figura de linguagem comumente usada para indicar preferências e semelhanças entre elementos [18], 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 estratégia 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. 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 [16], 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 em português (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 [22]. 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 [8], 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 as principais palavras comparativas, os diferentes ambientes online possuem peculiaridades que podem não ser englobadas pelo conjunto inicial de palavras, motivando a ampliação desse léxico. Este 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[4], um dos principais sites brasileiros utilizado para busca de produtos e pesquisa de preços em lojas online. 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[5]. Para otimização do processo de construção do léxico, foram selecionadas as 35 marcas mais valiosas do Brasil e do mundo[6]. 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

[4]Site do Buscapé: https://www.buscape.com.br/
[5]Ranking de uso das Redes Sociais Online: https://wearesocial.com/. Acesso em: 25 de outubro de 2020.
[6]Site BrandZ: https://brandz.com/Global

Tabela 1: 5 palavras-chave comparativas mais frequentes.

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

Tabela 2: 5 palavras-chave comparativas mais precisas.

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

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[7].

Embora as palavras-chave possam ser usadas em comparações, nem sempre são utilizadas com esse único objetivo. A Tabela 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 1, as palavras-chave: *mais*, *como* e *melhor* são comuns em ambos contextos, porém, palavras como: *recomendo* e *comprei* são mais frequentes 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 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 Processo de Obtenção das Comparações

Após a construção do léxico, esse conjunto de palavras comparativas foi utilizado para a descoberta de comparações no conjunto de revisões do Buscapé e Twitter.

Para o Buscapé, utilizou-se um grande conjunto de revisões em português coletadas em setembro de 2013 [14]. 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. Através da lista comparativa do Buscapé, encontrou-se 48.311 revisões que possuem ao menos uma palavra-chave ou expressão comparativa. As 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, a lista comparativa formada com palavras do Twitter foi utilizada para filtrar todos os *tweets* que possuem ao menos uma palavra-chave, obtendo então 130.459 *tweets*.

[7] Acesso ao léxico: http://doi.org/10.5281/zenodo.4124410

3.3 Pré-processamento e Construção das Bases de Dados

O estudo de opiniões comparativas no nível da 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 especificado 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.
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

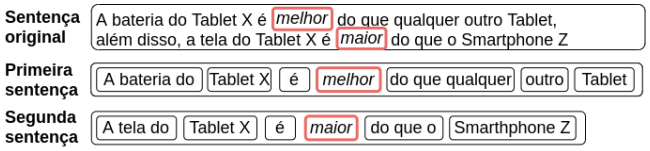


Figura 2: Estratégia utilizada para extrair múltiplas comparações das sentença.

Tabela 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

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 2: (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[8]. 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 foram rotuladas manualmente por um grupo de três pessoas voluntárias que indicaram se a sentença é comparativa ou não. 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 [7] foi calculado e obteve-se uma concordância entre os três rotuladores de 83,46% para o Buscapé e 83,21% para o Twitter. Além dessa rotulação, informações adicionais foram identificadas acerca de qual tipo comparativo e quais objetos são comparados nas sentenças. 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 detalha a quantidade de comparações rotuladas em cada base de dados.

Para melhorar a identificação de comparações, a abordagem que utiliza palavras-chave é aprimorada com o uso de abordagens complementares, apresentadas na Seção 4.

4 CLASSIFICAÇÃO DAS SENTENÇAS

Esta seção apresenta a abordagem supervisionada para classificação automática de sentenças comparativas, que foi dividida em duas etapas, conforme apresentado na Figura 3. Após utilizar o filtro léxico para encontrar as prováveis comparações, é apresentado uma abordagem para classificação das sentenças, separando as comparativas

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

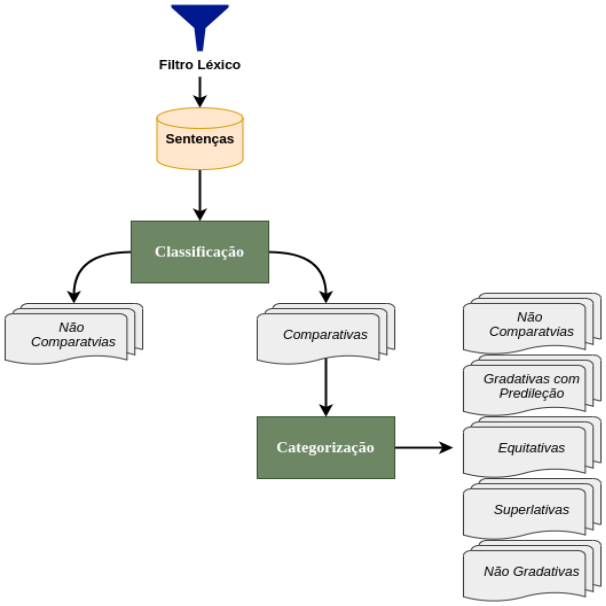


Figura 3: Abordagem hierárquica para classificação.

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[9] 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 [3]. A Tabela 4 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.

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, considerando as métricas de acurácia e Macro-F1.

[9]Tf–idf: é uma métrica que indica a importância de uma palavra em um corpus.

Tabela 4: Precisão (Prec.), Revocação (Rec.) e F1-Score (F1) com 95% de confiança para o Buscapé e Twitter.

	Buscapé							
	Não Comparativa			Comparativa			Média	
	Prec.	Rec.	F1	Prec.	Rec.	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.	Rec.	F1	Prec.	Rec.	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

Tabela 5: 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%**

Tabela 6: Detalhamento das sentenças classificadas como comparativas através do Multinomial Naive Bayes.

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

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 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.

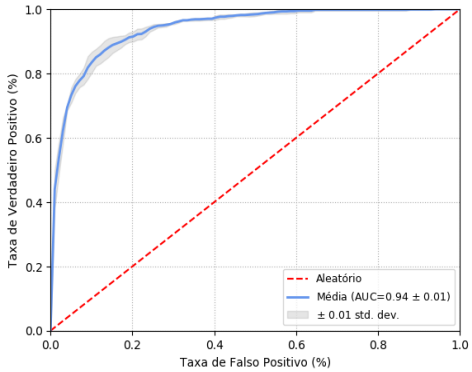
A Tabela 5 exibe a matriz de confusão obtida com o NB, que apresentou os melhores resultados para a distinção das duas classes. Nota-se que o modelo consegue detectar corretamente 90,3% das sentenças comparativas existentes no Buscapé e 87,4% para as comparações do Twitter, o que ressalta a boa capacidade do modelo na cobertura das comparações.

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, 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.

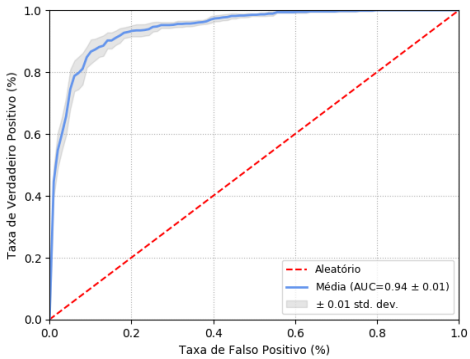
4.2 Classificação em Múltiplas Classes

Após separação das sentenças, iniciou-se a classificação dos resultados (i.e. hierárquica) com objetivo de classificar 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. A 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 6 apresenta a quantidade de sentenças obtidas 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



(a) Buscapé (AUC=0.94 ± 0.01)



(b) Twitter (AUC=0.94 ± 0.01)

Figura 4: Curva ROC para o Multinomial Naive Bayes.

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.

Apesar da acurácia ser uma métrica importante, a frequência de classificação de cada classe (i.e. revocação) é essencial para a tarefa de categorização das comparações em diferentes classes. A Tabela 7 apresenta o resultado da classificação das sentenças, onde os valores indicam a frequência de classificação de cada classe. 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.

Apesar das diferenças nos contextos analisados, os resultados mostram que não existe uma diferença significativa na tarefa de classificação das sentenças nas bases de dados. Já se tratando dos superlativos e das comparações equitativas, percebe-se que tais comparações fazem uso de expressões próprias que permitem distinguilas mais facilmente das demais, como é visto na Tabela 7, 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 mais complexas e geralmente não possuem um padrão claro. Tais sentenças podem ser frequentemente confundidas com comparações 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, sentenças com estruturas sintáticas semelhantes podem ser de tipos comparativos diferentes. Por exemplo, na sentença: "Smartphone X é o *melhor*", nota-se um superlativo. No entanto, adicionando um único complemento (do que Y), a comparação passa a ser classificada como gradativa com predileção para o Smartphone X ("Smartphone X é *melhor* do que Smartphone Y"), mesmo as frases tendo estruturas semelhantes e utilizando a mesma palavra-chave comparativa. 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, que possibilita diferenciar a frase de um superlativo.

5 CONCLUSÃO

Este trabalho apresentou um novo estudo para mineração de opiniões em revisões online na língua portuguesa. Uma das tarefas fundamentais na mineração de opinião é separar as opiniões regulares das comparativas. Assim, uma abordagem léxica baseada em palavras-chave comparativas foi proposta como um dos primeiros passos para encontrar sentenças comparativas dentre um amplo conjunto de revisões online, criando bases de dados de dois importantes contextos: (1) sites de revisões/avaliações; e (2) Redes Sociais Online. De forma complementar à abordagem léxica, aplicou-se uma estratégia supervisionada por meio do algoritmo Multinomial Naive Bayes (NB), separando as sentenças comparativas das sentenças não comparativas com uma acurácia de 87%. Em seguida, as sentenças detectadas como comparativas foram classificadas em cinco grupos: (1) Não Comparativa; (2) Gradativa com Predileção; (3) Equitativa; (4) Superlativa; e (5) Não Gradativa. Os resultados mostram que

Tabela 7: 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%**

a abordagem hierárquica é promissora na detecção de sentenças comparativas e possibilita análises mais sistemáticas sobre opiniões e preferências. Em suma, este trabalho apresentou três principais contribuições: (1) a criação de um léxico com palavras e expressões usadas em comparações em português; (2) a construção de duas bases de dados com sentenças comparativas; e (3) uma abordagem hierárquica para detecção de múltiplas comparações.

Para futuros trabalhos, planeja-se desenvolver uma ferramenta que implemente as técnicas desenvolvidas neste trabalho. Isso possibilitará que pesquisadores e empresas apliquem as estratégias de mineração de opinião para detecção de opiniões comparativas.

6 AGRADECIMENTOS

Este trabalho foi financiado pelo MPMG, projeto Capacidades Analíticas, pelo CNPq, CAPES, Fapemig e pelo programa de bolsas da Big Data.

REFERENCES

[1] Ali Reza Alaei, Susanne Becken, and Bela Stantic. 2019. Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research* 58, 2 (2019), 175–191.
[2] Ngo Xuan Bach, Pham Duc Van, Nguyen Dinh Tai, and Tu Minh Phuong. 2015. Mining Vietnamese comparative sentences for sentiment analysis. In *Proc. of the KSE*. 162–167.
[3] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.
[4] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In *Proc. of the INDIACom*. 452–455.
[5] Fabrício Benevenuto, Filipe Ribeiro, and Matheus Araújo. 2015. Métodos para análise de sentimentos em mídias sociais. In *Proc. of the Webmedia*.
[6] Dmitriy Bespalov, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. 2011. Sentiment classification based on supervised latent n-gram analysis. In *Proc. of the CIKM*. 375–382.
[7] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[8] Carlos Henrique da Rocha Lima. 2017. *Gramática normativa da língua portuguesa* (53 ed.). José Olympio.
[9] Arthur E de O. Carosia, Guilherme P Coelho, and Ana EA da Silva. 2019. The influence of tweets and news on the brazilian stock market through sentiment analysis. In *Proc. of the WebMedia*. 385–392.
[10] Alaa M El-Halees. 2012. Opinion mining from Arabic comparative sentences. In *Proc. of the ACIT*. 265–271.
[11] Mai M Eldefrawi, Doaa S Elzanfaly, Marwa S Farhan, and Ahmed S Eldin. 2019. Sentiment analysis of Arabic comparative opinions. *SN Applied Sciences* 1, 5 (2019), 411.
[12] Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proc. of the Coling*. 241–248.
[13] Samir Gupta, ASM Ashique Mahmood, Karen Ross, Cathy Wu, and K Vijay-Shanker. 2017. Identifying comparative structures in biomedical text. In *Proc. of the BioNLP*. 206–215.
[14] Nathan Hartmann, Lucas Avanço, Pedro Paulo Balage Filho, Magali Sanches Duran, Maria Das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo, Sandra M Aluísio, et al. 2014. A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words.. In *Proc. of the LREC*. 3865–3871.
[15] Xiaojiang Huang, Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2008. Learning to identify comparative sentences in Chinese text. In *Proc. of the PRICAI*. 187–198.
[16] Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proc. of the SIGIR*. 244–251.
[17] Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. In *Proc. of the AAAI*. 9.
[18] Bing Liu. 2012. *Sentiment analysis and opinion mining*. Vol. 5. Morgan & Claypool Publishers. 1–167 pages.
[19] Raj P Mehta, Meet A Sanghvi, Darshin K Shah, and Artika Singh. 2020. Sentiment analysis of tweets using supervised learning algorithms. In *Proc of the ICTSCI*. 323–338.
[20] Samiah Jan Nasti, M Asger, and Muheet Ahmad Butt. 2020. Automatic Extraction of Product Information from Multiple e-Commerce Web Sites. In *Prof. of the ICRIC*. 739–747.
[21] Madai Ramırez and Octavio Sánchez. 2016. Ye shall know them by their verbs: How gender express their opinion in Twitter. *Advances in Computational Linguistics* (2016), 23.
[22] Violeta Virginia Rodrigues. 2002. As construções comparativas em língua portuguesa. *Revista do GELNE* 4, 1 (2002), 1–6.
[23] Edward Sapir. 1944. Grading, a study in semantics. *Philosophy of science* 11, 2 (1944), 93–116.
[24] Jesus Serrano-Guerrero, Jose A Olivas, Francisco P Romero, and Enrique Herrera-Viedma. 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences* 311 (2015), 18–38.
[25] Ellen Souza, Danilo Costa, Dayvid W Castro, Douglas Vitório, Ingryd Teles, Rafaela Almeida, Tiago Alves, Adriano LI Oliveira, and Cristine Gusmão. 2017. Characterising text mining: a systematic mapping review of the Portuguese language. *IET Software* 12, 2 (2017), 49–75.
[26] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.
[27] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proc. of the EMNLP*. 606–615.
[28] Seon Yang and Youngjoong Ko. 2009. Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques. In *Proc. of the AACL-IJCNLP*. 153–156.