UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Sumarização contrastiva de opinião

Raphael Rocha da Silva

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R672s

Rocha da Silva, Raphael
Sumarização contrastiva de opinião / Raphael Rocha da Silva; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2020.
151 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Processamento de Linguagem Natural. 2. Sumarização de texto. 3. Mineração de opinião. I. Salgueiro Pardo, Thiago Alexandre, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Raphael Rocha da Silva

Sumarização contrastiva de opinião

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
Janeiro de 2020

Raphael Rocha da Silva

Contrastive opinion summarization

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
January 2020

AGRADECIMENTOS

- Ao **Thiago Pardo**, professor que me orientou e contribuiu com valiosas ideias.

- Aos **professores membros da banca**, por sua atenta avaliação do projeto.

- Ao **Otávio Sousa**, que ajudou na construção do conjunto de dados usado neste projeto.

- A **Marco Sobrevilla**, **Roney Santos**, **Henrico Brum**, **Guilherme Hiromoto**, **Márcio Lima**, **Ana Caroline** e **Sidney Leal** que atuaram como voluntários na tarefa de avaliação de resumos deste projeto.

- Aos **colegas do NILC**, pela convivência fraterna e produtivas discussões;

- À minha **família** e **amigos**, pelo constante apoio e incentivo.

- À **Wavy**, empresa onde trabalho desde maio de 2019, que me deu total liberdade para participar das atividades do mestrado.

- Ao **ICMC**[1] e à **Universidade de São Paulo**, que possibilitaram a execução deste projeto.

- À **CAPES**[2], que forneceu suporte financeiro a este projeto entre março de 2017 e agosto de 2017.

- À **FAPESP**[3], que forneceu suporte financeiro a este projeto entre setembro de 2017 e fevereiro de 2019 (processo 17/12236-0).

[1] Instituto de Ciências Matemáticas e de Computação
[2] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
[3] Fundação de Amparo à Pesquisa do Estado de São Paulo

"Cada pessoa fala uma língua diferente."
(Cleon Lopes)

"PLN é sempre divertido!"
(Thiago Pardo)

RESUMO

SILVA, R. R. **Sumarização contrastiva de opinião**. 2020. 151 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Esta dissertação apresenta métodos que permitem comparar entidades por meio da geração de um resumo que realce diferenças entre elas a partir do processamento automático de textos opinativos. Métodos de sumarização contrastiva de opinião foram descritos e avaliados. Três métodos foram trazidos da literatura e um método foi criado. Os métodos foram testados em textos opinativos pré-anotados sobre eletrônicos de uso pessoal extraídos da Web.

Embora existam alguns métodos publicados anteriormente, não houve um estudo que os compare: os métodos foram testados em conjuntos de dados diferentes e avaliados com métricas diferentes. Partindo da hipótese que os métodos gerarão sumários com características diferentes para um mesmo conjunto de dados, este trabalho busca suprir essa lacuna montando um conjunto de dados diversificado e definindo métricas padronizadas para testar características desejáveis dos sumários gerados por cada método.

A importância da sumarização automática se dá porque ela permite o desenvolvimento de ferramentas que ajudam usuários a absorver melhor as informações de um conjunto de texto, especialmente se esse conjunto for muito grande, como ocorre com dados coletados em massa da Internet. A sumarização contrastiva de opinião toca uma parte mais específica do problema: o caso em que um usuário deseja comparar duas entidades a partir de um grande volume de textos opinativos.

Esta pesquisa permite identificar quanto os resumos gerados por diferentes métodos são úteis para os usuários; conjectura-se que eles são mais eficazes do que resumos de opinião simples na tarefa de ajudar as pessoas a entender diferenças entre duas entidades. Isso pode ser benéfico para uma pessoa que quer comprar um produto e está em dúvida entre duas marcas ou dois modelos. Também pode ser útil para um fabricante entender como seus produtos se posicionam segundo a opinião popular em relação a seus concorrentes.

Espera-se que esta pesquisa traga contribuições tanto no âmbito acadêmico quanto no contexto prático. Do ponto de vista prático, ela tem o potencial de permitir o desenvolvimento de ferramentas as quais empresas e usuários demandam. Na academia, ela se unirá às iniciativas recentes de pesquisa em Processamento de Linguagem Natural e Mineração de Opinião que têm ganhado destaque no Brasil, dando continuidade a seus trabalhos e somando a elas ideias novas que poderão ser futuramente utilizadas por outros pesquisadores.

Palavras-chave: Processamento de Linguagem Natural, Sumarização Contrastiva de Opinião.

ABSTRACT

SILVA, R. R. **Contrastive opinion summarization**. 2020. 151 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

This theses presents automatic techniques for comparing opinions by generating summaries that highlight differences and similarities between two entities given a set of opinionated text. We describe and evaluate different methods for comparative opinion summarization. Three methods are brought from previous work and one is created. The input for tests consists of reviews about consumer electronic products written in Portuguese and extracted from the Web.

Although there are some previously published methods, there was no study comparing them: the methods were tested on different datasets and evaluated with different metrics. Assuming that the methods will generate summaries with different characteristics for the same dataset, this paper fills this gap by building a diverse dataset and defining standardized metrics to test desirable characteristics of summaries generated by each method.

Automatic summarization is important because it allows the development of tools that help users to better absorb information from a set of texts. This is especially useful if the set is too large, such as batch data collected from the Internet. Comparative opinion summarization reaches a more specific part of the problem: the case where a user wants to compare two entities based on a large volume of text that contains other people's opinions.

This research leads to a survey on how useful summaries generated by different methods are. We hypothesize that they are more effective than single-entity opinion summaries to help people understand differences between two entities. This can be beneficial for a person who wants to buy a product and is in doubt between two brands or two models. It can also be useful for a manufacturer to understand how their products rank in relation to their competitors according to popular opinion.

We expect this research brings contributions both in the academic context and in the practical context. From the practical point of view, it has the potential to enable the development of tools that companies and users demand. In the academy, it will join recent research initiatives in Natural Language Processing and Opinion Mining that have gained prominence in Brazil; this project will proceed their work and bring new ideas that may be used in the future by other researchers.

Keywords: Natural Language Processing, Contrastive Opinion Summarization.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização e Motivação

O termo **opinião** é provido pelo dicionário Michaelis[1] com as seguintes acepções:

1. Modo de pensar, de julgar, de ver: *'Segundo a sua opinião, qualquer escândalo doméstico ficava muito mal a um negociante de certa ordem.'*
2. Ponto de vista ou posição tomada sobre assunto em particular (social, político, religioso etc.); teoria, tese: *'Sr. Lambertosa é então de opinião que o casamento convém às enfermidades nervosas?'*
3. Parecer emitido sobre determinado assunto em que muito se refletiu e deliberou; partido, voto.
4. Juízo de valor que se faz sobre alguém ou alguma coisa; conceito: *'Qual é a opinião que você tem de mim?'*
5. Consenso partilhado por um grupo de pessoas sobre um ou mais assuntos; julgamento coletivo: *'A opinião nacional estava dividida. Criara-se um impasse perigoso que talvez só pudesse ser resolvido por meios violentos.'*
6. Ideia ou hipótese sem fundamento; aquilo que se presume sem certeza; presunção: *'Deixemos de lado as opiniões e passemos à demonstração dos fatos.'*
7. (Filosofia) Liberdade de ter e adotar, como verdadeiras, preferências e convicções religiosas e políticas, a despeito de estarem sujeitas a dúvidas e questionamentos sobre sua validade e seus pressupostos.
8. Capricho voluntarioso; birra, teimosia: *'Sua opinião não lhe deixa dar ouvidos à outra versão da história.'*

[1] Versão online, consulta em 30 de janeiro de 2019.

9. (Coloquial) Sentimento pretensioso sobre si mesmo; presunção, pretensão, vaidade: *'Ele é cheio de opinião e vive a gabar-se.'*

O mesmo dicionário informa que **subjetividade** é o antônimo de **objetividade**, e é o 'caráter ou qualidade de subjetivo', ou seja, aquilo 'que exprime ou manifesta apenas as ideias ou preferências da própria pessoa; individual, particular, pessoal'; em uma visão filosófica, é aquilo 'que é característico da realidade como ela é percebida por um indivíduo, em oposição ao que ela de fato é, independente da mente que a percebe'.

A área de Análise de Sentimento (também conhecida como Mineração de Opinião) atua na análise e extração de conteúdo relevante dentro de conjuntos de dados que contêm textos subjetivos (LIU, 2012). Por causa da interatividade e da publicação de conteúdo independente que a Web tem proporcionado e popularizado, a Análise de Sentimento tem-se tornado cada vez mais útil para o melhor aproveitamento de textos subjetivos que são publicados na Internet.

Como as pessoas têm acesso a grandes conjuntos de dados de vários tipos através da Rede Mundial, é válido pensar em novas maneiras de explorar esses dados para filtrar as informações mais relevantes. Idealmente, usuários que precisam de uma informação pontual poderiam obtê-la simplesmente digitando algumas palavras-chave ou fazendo perguntas em linguagem natural (escrita ou falada) em vez de ter que visitar várias páginas da Web e minerar eles mesmos as informações importantes até resolverem a questão que têm.

Muitas tarefas podem ilustrar o uso de ferramentas automáticas que ajudam pessoas a obter informações em tempo hábil. A busca na Web é uma delas. Outrora limitadas a encontrar documentos contendo palavras-chave específicas, agora as ferramentas de busca são capazes de responder perguntas de maneira muito natural. As perguntas podem ser feitas como se fossem direcionadas a um humano: *'como vai estar o tempo no fim de semana?'*, *'quantos reais são oitocentos dólares?'*. Alguns serviços de busca, em vez de mostrar uma lista de páginas que pode ajudar usuários a encontrar o que eles precisam, respondem essas perguntas com sentenças curtas que as respondam diretamente (Figura 1.1). Isso é feito com técnicas que permitem isolar, dentro do conjunto de dados a que a ferramenta tem acesso, a informação que é realmente requisitada na situação dada.

Quando se tem um conjunto muito grande de dados, a **sumarização** é uma ação que pode ser executada para viabilizar a absorção de informação por humanos. Seu objetivo é identificar as partes mais relevantes de um conjunto de dados para apresentá-lo de maneira resumida para um usuário (MANI, 2001a).

A **sumarização de opinião** é voltada para um tipo específico de dado: textos que contêm opiniões sobre um assunto. É fácil coletar textos assim: basta consultar redes sociais ou qualquer outra página virtual que aceite publicações de usuários, e em alguns minutos podem ser extraídos milhares de comentários opinativos. O problema então é como extrair desses dados

Figura 1.1 – Serviço de busca do Google respondendo a uma pergunta de maneira sucinta.

Fonte: Captura de tela do buscador Google

informações úteis. Devido à quantidade de dados, pode ser humanamente custoso separar as partes relevantes e, principalmente, interpretar as informações.

A sumarização de opinião é uma atividade que começou a ser investigada no começo da década de 2000 e foi ganhando cada vez mais importância com o acentuado crescimento da participação de pessoas em fóruns da Web e em redes sociais virtuais que ocorreu nessa década (LIU, 2012, pp. 5;8;10), que foi acompanhado por um aumento natural na quantidade de textos opinativos publicados.

Este trabalho será voltado para a **sumarização contrastiva de opinião**, que tem como objetivo a comparação entre duas ou mais entidades a partir de textos opinativos escritos sobre elas (LERMAN; MCDONALD, 2009). Por exemplo, uma pessoa que deseja comprar um celular pode estar em dúvida entre dois modelos; para essa pessoa, pode ser útil uma forma de compará-los a partir das opiniões de pessoas que já compraram um dos dois modelos. Este trabalho estuda formas de fazer essa comparação de modo automático, o que pode ajudar o usuário por permitir que o conjunto de textos usado para análise seja maior do que ele conseguiria analisar manualmente.

As técnicas investigadas neste trabalho podem permitir o desenvolvimento de algo semelhante ao mostrado na Figura 1.1 para o caso específico em que uma pessoa queira comparar duas entidades com base em avaliações escritas sobre elas: em vez de ter que ler textos avaliativos de várias páginas, deseja-se que essa pessoa obtenha um quadro sucinto trazendo essas informações.

Estudos mostram que as pessoas consideram importante ler avaliações sobre produtos antes de efetuar uma compra. Com o crescimento da quantidade de pessoas publicando textos na rede mundial a fim de oferecer informações em relação às próprias experiências, muitos passaram a fazer uso frequente dessas informações. Uma pesquisa da TNS Research, noticiada por portais

como G1[2] e UOL[3], revelou que já em 2010 aproximadamente 90% dos brasileiros costumavam consultar a Internet para ajudá-los a decidir se devem ou não efetuar uma compra, dos quais 76% procuravam informações em blogues e fóruns, tipos de páginas que frequentemente contêm textos subjetivos. Dos entrevistados, 63% disseram que já escreveram e publicaram suas experiências com produtos e serviços.

O papel da sumarização automática de opinião é receber esses dados obtidos da Web e transformá-los em informações concisas que sejam úteis para um determinado fim, geralmente auxiliando usuários a adquirir apenas as informações mais relevantes de um conjunto de dados, que é uma atividade difícil de ser executada manualmente quando o conjunto é muito grande e seus elementos não estão bem organizados.

Para exemplificar um cenário de uso, suponha que um certo governo publique repentinamente alguma medida polêmica, e um pesquisador deseje saber o impacto desse anúncio na população. Uma boa maneira de ouvir a voz popular é por meio de redes sociais virtuais, que permitem a qualquer pessoa com acesso à Internet publicar textos de forma instantânea. O pesquisador pode então querer usar a rede social Twitter para tentar entender o que as pessoas estão falando sobre o governo nas últimas horas. Ele pode conseguir ler as postagens mais recentes e mais populares pesquisando por palavras-chave relacionadas ao tema, mas dificilmente conseguiria atingir um volume de leitura que o permitisse fazer um levantamento preciso da porcentagem de pessoas que aprovam a decisão do governo. Além de ser um conjunto grande de texto, muitos dos resultados que a busca retornar serão inúteis para a finalidade que esse pesquisador deseja (certamente haverá muitos tuítes com a palavra 'governo' que não contenham opinião alguma; vários podem ser apenas manchetes de notícias). Se ele tiver uma ferramenta capaz de separar automaticamente as opiniões positivas das negativas, ele pode executar esse trabalho muito mais facilmente.

Um exemplo real trazido por Hoque e Carenini (2016) fala do caso do problema de deformação observado no modelo de telefone celular iPhone 6 Plus[4]. Logo após o lançamento do celular, ele foi alvo de vários comentários de usuários reclamando que o telefone entortava depois de alguns dias de uso. O incidente causou tumulto em páginas como Macrumors, blogue especializado que publica notícias relacionadas à Apple (fabricante do produto) e permite que leitores publiquem comentários. Em poucos dias, surgiram dezenas de conversações envolvendo milhares de comentários em várias linhas de assuntos: o que os usuários haviam reportado sobre o problema do entortamento, o que a Apple disse para defender o produto, quais foram as reações das companhias concorrentes, etc. Nessa situação, pode-se enxergar três tipos de pessoas desejando explorar essas conversas. Primeiro, um potencial comprador que tem a

[2] Disponível em g1.globo.com/tecnologia/noticia/2010/04/mais-de-90-pesquisam-na-web-antes-de-comprar-diz-pesquisa.html, acesso em 24/2/2018.

[3] Disponível em economia.uol.com.br/ultimas-noticias/infomoney/2010/04/19/internet-92-dos-usuarios-pesquisam-produtos-ou-comparam-precos-pela-web.jhtm, acesso em 24/2/2018.

[4] Caso ocorrido em 2014 com o produto da Apple Inc (noticiado em tecnoblog.net/166179/iphone-6-plus-entorta/, acesso em 25/2/2018).

intenção de adquirir o produto e quer saber se o problema do entortamento é mesmo grave; segundo, um jornalista que quer publicar uma matéria sobre o que as pessoas estão dizendo a respeito do problema; terceiro, um analista da Apple pode querer usar os comentários para entender como o problema está afetando os consumidores e o que deve ser feito para amenizá-lo. Em todos os três casos, observam Hoque e Carenini, dada a grande quantidade de comentários, pode ser extremamente difícil e demorado para alguém explorar e analisar toda a informação útil contida no conjunto. Os pesquisadores observam também que a maioria das páginas virtuais oferece uma interface que torna a descoberta de informação um tanto escassa, porque se limita a oferecer acesso sequencial de comentários, algumas vezes com uma (geralmente básica) ferramenta que permite procurar texto por palavras-chave.

Dadas as motivações elucidadas anteriormente, o principal interesse do uso da Análise de Sentimento neste trabalho é coletar um grande conjunto de texto escrito por várias pessoas diferentes (talvez também seja interessante que os textos tenham sido escritos em vários períodos de tempo diferentes) e processá-los a fim de distinguir sentenças que implicam sentimentos positivos de sentenças que implicam sentimentos negativos. Não haverá interesse em lidar com as etapas iniciais da análise de sentimento: os conjuntos de dados usados nos testes já terão sido previmante processados.

1.2 Definição do Problema

Para compreender o problema estudado neste projeto, serão necessários conceitos básicos de sumarização textual e de análise de sentimento, enunciados a seguir.

Sumarização de texto refere-se ao ato de resumir um ou mais textos de forma que se obtenha um texto menor contendo as informações mais relevantes do conjunto de origem (HAHN; MANI, 2000). Também existe a opção de se gerar não um texto menor, mas uma imagem (por exemplo, um gráfico contendo informações quantitativas do conjunto fonte) (GAMBHIR; GUPTA, 2017). Em ambos os formatos, a saída obtida é chamada de '**sumário**' ou '**resumo**'[5] (RIBALDO; PARDO; RINO, 2011). O Quadro 1.1 mostra um exemplo de resumo automático de textos jornalísticos, obtido de um sistema baseado em grafos implementado por Ribaldo et al. (2012).

A **Análise de Sentimento** (também chamada de **Mineração de Opinião**) é a área de pesquisa que se interessa por extrair automaticamente informações relevantes de textos que contêm opinião (LIU, 2012, pp. 7). Um **texto opinativo** é aquele que contém ideias subjetivas

[5] Na língua portuguesa cotidiana, a palavra 'sumário' costuma-se referir à lista de divisões e seções de um documento estruturado (também chamado de 'índice' e 'tabela de conteúdo'), enquanto 'resumo' é um texto contendo as informações mais relevantes de uma fonte (ABNT, 2011). Mas na área de Sumarização Automática, os dois termos têm sido usados sem distinção e são ainda mais abrangentes, podendo nomear outros elementos com formas diferentes desses dois (todavia com funções similares às deles).

Quadro 1.1 – Exemplo de resumo de notícias.

Primeiro texto	Segundo texto
A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos. O revezamento terminará em 8 de agosto, primeiro dia das Olimpíadas de Pequim.	Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim. Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos. O Brasil não faz parte do trajeto da tocha olímpica. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril. Aos 16 anos, Jade conquistou três medalhas no Pan: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no solo. Ao todo, a chama olímpica percorrerá 20 países antes de chegar a Pequim para a abertura da competição, no dia 8 de agosto.
Resumo	
A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.	

Fonte: Resumo de Ribaldo et al. (2012, p. 265-266), textos do córpus CSTNews (CARDOSO et al., 2011)

(críticas, resenhas, expressão de sentimentos) em relação a uma entidade. Uma **entidade** pode ser um produto, um indivíduo, uma empresa, um acontecimento, uma obra de arte, entre outros. Uma **opinião** pode então ser entendida como um sentimento expresso por alguém a respeito de uma entidade. Exemplos de textos que têm essas características são as resenhas de filmes e de produtos. O Quadro 1.2 mostra um trecho de um texto opinativo escrito por um crítico de cinema. A Figura 1.2 mostra comentários extraídos de uma loja virtual.

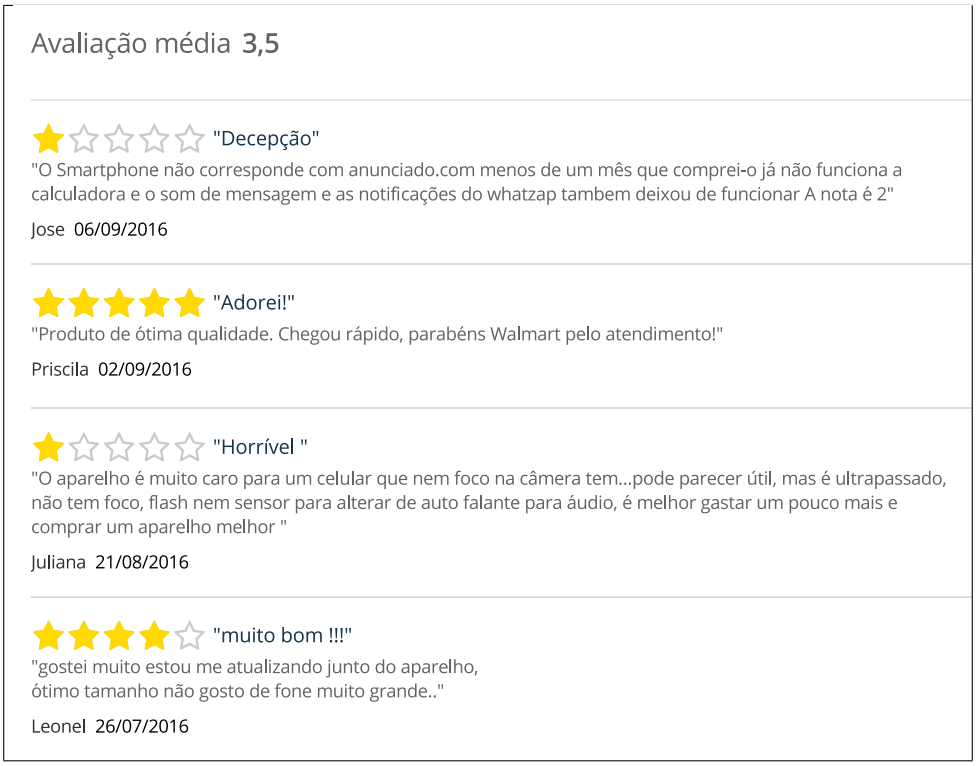
Quadro 1.2 – Trecho de crítica do filme Procurando Nemo (2003).

Como já se tornou padrão nas produções da Pixar, Procurando Nemo mergulha o espectador (com o perdão do trocadilho) em um universo multicolorido e repleto de nuances: habitado por uma infinidade de criaturas engraçadinhas (e outras nem tanto), o oceano visto aqui possui suas próprias versões de sinais de trânsito, rodovias e até mesmo de grupos de apoio como os Alcoólatras Anônimos. Preenchendo cada centímetro da tela com detalhes que revelam um preciosismo admirável, a equipe de animadores confere atenção particular à elaboração das expressões faciais de seus personagens, transformando-os em figuras carismáticas e divertidíssimas - reparem, por exemplo, as mudanças no rosto da tartaruga Crush enquanto esta conversa com Marlin e observe, também, o olhar de orgulho que surge no rosto de Nemo quando este ouve uma determinada notícia sobre seu pai. Em momentos como estes, a expressividade dos personagens é tamanha que os diálogos tornam-se até mesmo dispensáveis.

Fonte: Texto de Pablo Villaça para a página Cinema em Cena

A área da Análise de Sentimento tem reunido pesquisadores de Processamento de Linguagem Natural, Mineração de Dados, Análise de Redes Sociais e Recuperação de Informação (AGGARWAL; ZHAI, 2012). A importância de pesquisa nesse tema deve-se fortemente ao fato de opiniões estarem presentes em praticamente todas as atividades humanas e terem um grande poder de influenciar comportamento (LIU, 2012, p. 8): empresas se interessam em saber as opiniões de seus consumidores sobre seus produtos e serviços, consumidores buscam opiniões de outros consumidores antes de fazer uma compra, eleitores ouvem o ponto de vista dos candidatos antes de decidir o voto. Ações como essas, que sempre fizeram parte da vida em

Figura 1.2 – Comentários opinativos de compradores do Walmart sobre o aparelho de telefone móvel Samsung Galaxy J1 Mini SM-J105B/DL.



Fonte: Captura de tela da página virtual do Walmart

sociedade, têm nos últimos anos estado ainda mais presentes por causa da popularização da Internet, que permite que milhares de opiniões sejam semeadas e colhidas facilmente. Esse é outro motivo que alavanca a importância da Mineração de Opinião: como se dispõe de um grande volume de dados, é válido investir em maneiras automáticas de interpretá-los (LIU, 2012, p. 9).

Por causa de sua natureza essencialmente subjetiva, geralmente não é suficiente conhecer a opinião de uma única pessoa, porque pessoas diferentes têm pontos de vista diferentes sobre uma mesma coisa. Quando se tem uma grande quantidade de texto subjetivo com opiniões escritas por pessoas diferentes, pode ser desejável fazer um resumo das informações para ajudar um leitor a entender o conteúdo contido no conjunto (LIU, 2012, p. 102). A **sumarização de opinião** é o tipo de sumarização que lida especificamente com textos opinativos.

O Quadro 1.3 mostra um exemplo de sumário de opinião obtido por López Condori (2015). As sentenças do texto original são separadas por aspecto (característica do produto que avaliam); o sumário contabiliza as sentenças positivas e negativas de cada aspecto e mostra as principais sentenças.

Os sumarizadores de opinião tradicionais são bastante úteis porque oferecem uma forma de se extrair as informações mais relevantes em textos opinativos considerando a entidade sobre a qual eles falam e suas características. Porém, podem não funcionar para se comparar duas

Quadro 1.3 – Exemplo de sumário de opinião (tradicional). O método usado para obter este sumário é uma adaptação de Hu e Liu (2004a).

```
Aspecto: Samsung Smart TV
   Sentenças Positivas: 16
   - Ótima televisão , ou melhor , central multimídia fiquei
   impressionado pela quantidade de recursos (Wi-Fi , Aplicativos
   e All Share) como tenho um celular Samsung (Omnia W) com
   Windows 7.5 tive uma interação completa.
   Sentenças Negativas: 11
   - Custo-benefício não compensa.
Aspecto: Preço
   Sentenças Positivas: 0
   Sentenças Negativas: 2
   - O que não gostei: Valor muito alto
Aspecto: Durabilidade
   Sentenças Positivas: 0
   Sentenças Negativas: 2
   - A TV é maravilhosa, porém queimou com menos de dois anos de
   uso e o conserto na autorizada fica o preço de uma nova.
Aspecto: Câmera
   Sentenças Positivas: 1
   - A câmera embutida na tv com sensor de movimento e presença
   funciona bem e impressiona quem não conhece.
   Sentenças Negativas: 1
   - O que não gostei: A câmera poderia ter movimentação
   horizontal de visão e não só apenas vertical.
Aspecto: Qualidade da Imagem
   Sentenças Positivas: 1
   - Excelante aparelho de TV com ótima qualidade de imagem e
   recursos.
   Sentenças Negativas: 0
```

Fonte: López Condori (2015, p. 74)

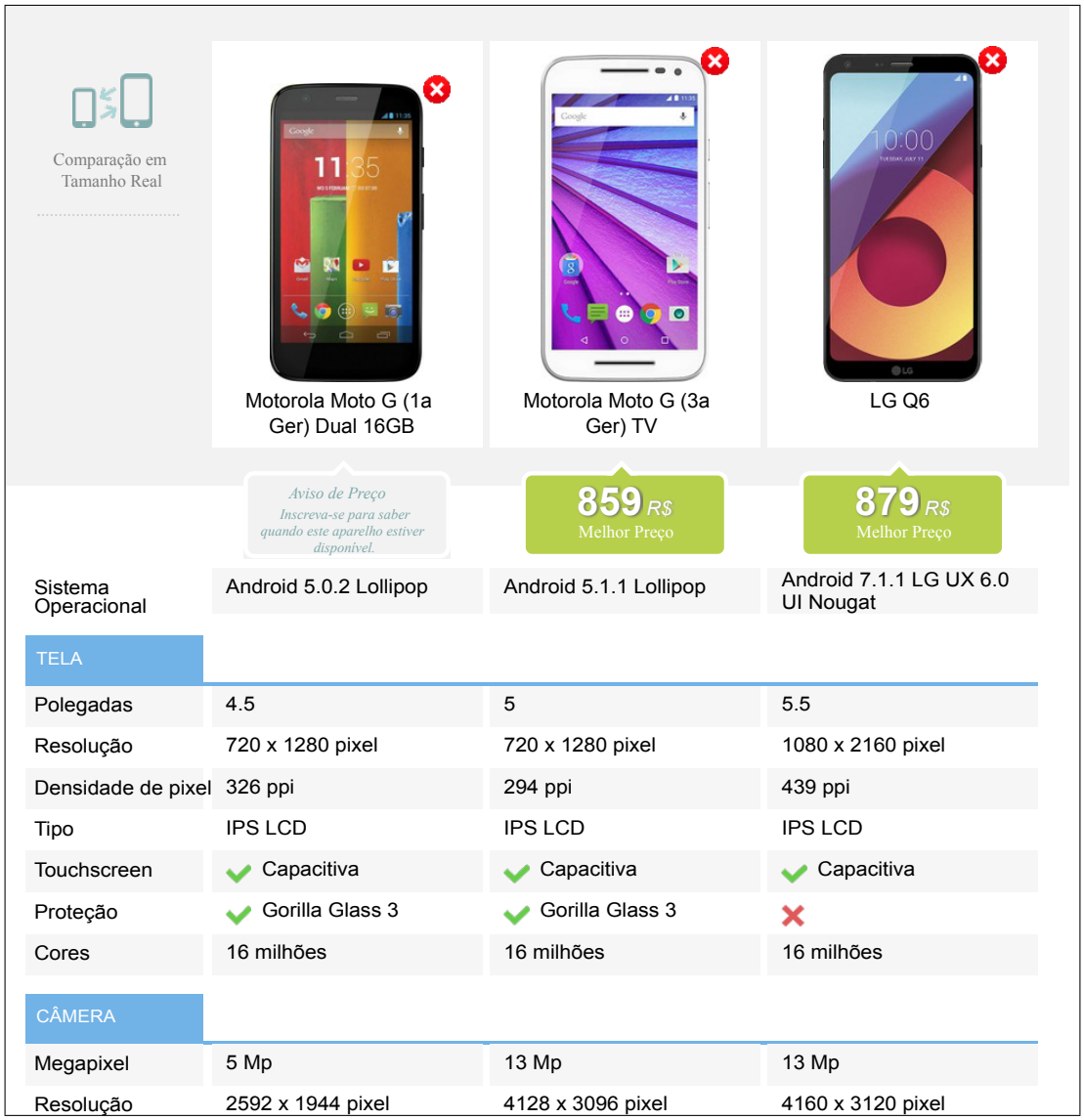
entidades: se for feito separadamente um resumo para cada entidade, esses resumos são gerados de forma independente, donde não há garantia que os dois resumos sejam comparáveis; é possível que eles falem de características disjuntas (não comparáveis) de cada entidade, e que não apresentem as melhores informações possíveis para se fazer uma comparação de maneira adequada.

Para comparar dois produtos, existem sistemas comerciais em português que fazem sumários contrastivos de dados objetivos. Um exemplo de sumário contrastivo (que não é de opinião) é disponibilizado pela página Tudo Celular[6]. Na página, é possível selecionar dois ou mais aparelhos de celular e o sistema mostra uma tabela comparando os dados técnicos deles (Figura 1.3). Neste caso, o sumário é feito não a partir de um conjunto de textos em linguagem natural, mas a partir de um conjunto de dados estruturados; então, para gerar a tabela, basta que ele copie as informações de um banco de dados, sem que seja necessário haver um processamento linguístico e tampouco análise de opinião.

Um sumário contrastivo de dados estruturados como o da Figura 1.3 é muito útil porque oferece uma maneira prática e principalmente bem organizada de se conhecer as diferenças e similaridades entre aparelhos. Mas ele se limita a considerar os dados informados pelos próprios fabricantes. Ocorre que a experiência de uso não depende apenas desse tipo de

[6] Disponível em tudocelular.com/compare/3281-3277-3631.html, acesso em 2/2/2018.

Figura 1.3 – Captura de tela do serviço de comparação da página Tudo Celular, mostrando uma parte da tabela gerada.



Comparação em Tamanho Real

	Motorola Moto G (1a Ger) Dual 16GB	Motorola Moto G (3a Ger) TV	LG Q6
	Aviso de Preço: Inscreva-se para saber quando este aparelho estiver disponível.	859 R$ Melhor Preço	879 R$ Melhor Preço
Sistema Operacional	Android 5.0.2 Lollipop	Android 5.1.1 Lollipop	Android 7.1.1 LG UX 6.0 UI Nougat
TELA			
Polegadas	4.5	5	5.5
Resolução	720 x 1280 pixel	720 x 1280 pixel	1080 x 2160 pixel
Densidade de pixel	326 ppi	294 ppi	439 ppi
Tipo	IPS LCD	IPS LCD	IPS LCD
Touchscreen	✓ Capacitiva	✓ Capacitiva	✓ Capacitiva
Proteção	✓ Gorilla Glass 3	✓ Gorilla Glass 3	✗
Cores	16 milhões	16 milhões	16 milhões
CÂMERA			
Megapixel	5 Mp	13 Mp	13 Mp
Resolução	2592 x 1944 pixel	4128 x 3096 pixel	4160 x 3120 pixel

Fonte: Captura de tela da página Tudo Celular[6], modificada para melhor diagramação

informação: é possível, por exemplo, que uma câmera fotografe melhor do que uma concorrente em algumas situações mesmo tendo uma resolução nominal inferior a esta; ou que haja uma incompatibilidade entre o aparelho e um aplicativo muito popular. Informações como essas têm um nível de detalhe que não seriam informados pelos fabricantes nas especificações técnicas. A melhor forma de obtê-las seria por meio da análise dos relatos de uso desses produtos.

Neste trabalho, será estudado um tipo de sumarização de opinião cujo interesse é a geração de resumos que destacam diferenças entre entidades a partir de textos opinativos escritos sobre elas. Na literatura, essa tarefa é encontrada com várias nuances, e não existe uma consolidação de nomenclatura para ela. Por exemplo:

- Liu, Hu e Cheng (2005) desenvolvem um sistema que compara quantitativamente as opiniões positivas e negativas encontradas em textos opinativos sobre duas ou mais entidades. Eles chamam a tarefa de 'análise e comparação de opiniões[7]'.
- Lerman e McDonald (2009) definem uma tarefa exatamente como a proposta deste trabalho e a alcunham 'sumarização contrastiva[8]'.
- Kim e Zhai (2009) estudam a tarefa de encontrar opiniões divergentes sobre uma mesma entidade. A tarefa é chamada no texto de 'sumarização contrastiva de opinião[9]', e os resumos formados com as opiniões encontradas são chamados de 'sumários contrastivos de opiniões contraditórias[10]'. Outros autores seguem a mesma definição e mesma nomenclatura de Kim e Zhai: Özsoy e Çakıcı (2014), Guo et al. (2015), Thonet (2017)[11].
- Liu (2012) chama de 'sumarização de visão contrastiva[12]' qualquer tarefa que envolva a sumarização de opinião por meio da identificação de pontos de vista contrastivos.
- Jin, Ji e Gu (2016) desenvolvem uma tarefa similar à proposta neste projeto, com a diferença que seus resumos trazem, além das diferenças, as similaridades entre as entidades. Além disso, cada resumo traz informações sobre somente um tópico relacionado à entidade, previamente escolhido. Eles chamam a tarefa de 'identificação de requisitos contrastivos de consumidores[13]'.

A atividade foco deste trabalho, chamada aqui de 'sumarização contrastiva de opinião' (ou simplesmente '**sumarização contrastiva**'), é a tarefa de geração de resumos que comparam diferentes textos opinativos sobre diferentes entidades de um mesmo tipo (isto é, entidades comparáveis) de modo a ressaltar as diferenças entre elas de acordo com as opiniões consultadas. Um exemplo é o caso quando se tem diferentes textos contendo avaliações críticas sobre dois modelos de telefone celular. Nesse caso, cada modelo de telefone é uma entidade, e elas todas são do mesmo tipo ('telefone celular'). A sumarização contrastiva de opinião processa esses textos a fim de encontrar as principais informações que permitem comparar os aparelhos.

O Quadro 1.4 mostra o que pode ser considerado um sumário contrastivo de opinião[14]. O quadro é um modelo de formato feito com informação fictícia, e é trazido como um modelo que pode ser adotado para a exibição das informações computadas através dos métodos estudados

[7] No texto original em inglês, 'analyzing and comparing opinions'.
[8] No texto original em inglês, 'contrastive summarization'.
[9] No texto original em inglês, 'contrastive opinion summarization'.
[10] No texto original em inglês, 'comparative summaries of contradictory opinions'.
[11] No texto original em francês, 'résumés d'opinions contrastés'.
[12] No texto original em inglês, 'contrastive view summarization'.
[13] No texto original em inglês, 'identifying comparative customer requirements'.
[14] As seções de destaques e de visão geral do resumo podem ser geradas por um sumarizador de opinião tradicional (não contrastivo), pois dizem respeito apenas à entidade correspondente, e não à concorrente. Contudo, pode ser útil exibir essas informações lado a lado pois, mesmo não sendo diretamente comparáveis, permitem ao usuário conhecer melhor cada produto, porque pode acontecer de um sumário contrastivo deixar de refletir os destaques individuais de cada produto pelo fato de ele priorizar opiniões que podem ser diretamente comparadas.

neste trabalho. O quadro separa cada entidade em um lado e destaca as opiniões positivas com um triângulo apontando para cima e negativas com um triângulo apontando para baixo. Em 'Visão geral', as áreas dos triângulos são proporcionais à respectiva porcentagem de comentários, e as áreas dos quadrados da última linha são proporcionais à respectiva quantidade total de comentários.

Quadro 1.4 – Possível formato de sumário contrastivo de opinião entre dois computadores.

Computador portátil A	Computador portátil B
DESTAQUES POSITIVOS	
A câmera é excepcional. ▲	▲ A tela é muito nítida.
O alto-falante é alto. ▲	▲ A bateria dura bastante.
DESTAQUES NEGATIVOS	
O alto-falante distorce um pouco o som. ▼	▼ É pesado demais.
PRINCIPAIS DIFERENÇAS	
As teclas são pequenas demais. ▼	▲ O teclado é muito bom.
A tela é sensível ao toque. ▲	▼ Este produto não possui tela tátil.
PRINCIPAIS SIMILARIDADES	
O painel tátil fica numa posição boa. ▲	▲ O painel tátil é bem posicionado.
VISÃO GERAL	
45% dos comentários são positivos. ▲	▲ 40% dos comentários são positivos.
22% dos comentários são negativos. ▼	▼ 18% dos comentários são negativos.
Total de comentários: 330 ■	■ Total de comentários: 180

Fonte: Original

Este trabalho estuda métodos de sumarização contrastiva de opinião conforme a definição da tarefa feita anteriormente. Também será útil estudar técnicas de tarefas ligeiramente diferentes da definição, pois as similaridades entre as tarefas podem contribuir com conhecimento útil e ideias interessantes, e, como existem poucos trabalhos publicados sobre sumarização contrastiva, pode valer a pena aproveitar alguns métodos de publicações de tarefas similares.

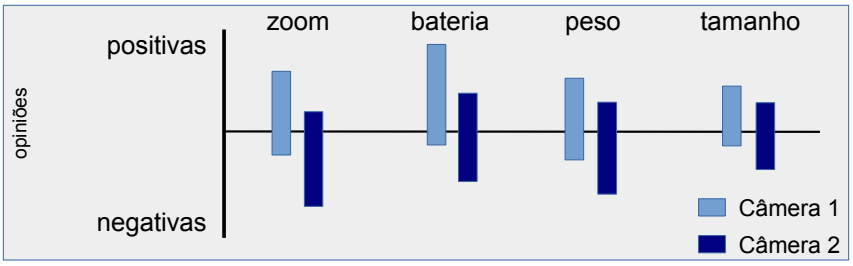
Neste trabalho, um **sumário contrastivo de opiniões** (referido também como 'sumário' ou 'resumo' quando subentendido que se trata desse tipo de sumário) é um resumo cuja função é exibir uma visão geral das opiniões coletadas de um grupo de pessoas sobre duas entidades diferentes (porém comparáveis, i.e., para os quais uma comparação faz sentido).

Entidade é um termo genérico, neste trabalho referindo-se a qualquer coisa sobre a qual faça sentido fazer um sumário de opinião: produtos, empresas, serviços, filmes, livros, destinos turísticos, etc. O trabalho usará textos opinativos sobre produtos eletrônicos para testar os métodos.

Um sumário contrastivo de opinião pode ter vários formatos, como o ilustrado no Quadro 1.4, que separa as opiniões sobre dois produtos e mostra suas principais diferenças, similaridades, e os pontos fortes e fracos de cada um. Pode-se também usar um formato gráfico, como na Figura 1.4, que mostra a comparação entre duas câmeras. Na figura, o comprimento de cada

barra é proporcional ao número de comentários que a característica do produto correspondente recebeu, e sua posição é relativa à proporção de avaliações positivas e negativas. Percebe-se que o produto 'Câmera 1' foi melhor avaliado do que 'Câmera 2' em todos os aspectos. O método e a interface da implementação também permitem comparar mais de dois produtos. Esse formato permite a comparação visual das opiniões de consumidores para dois produtos diferentes. Esse formato é usado por Liu, Hu e Cheng (2005).

Figura 1.4 – Possível formato de sumário contrastivo de opinião entre duas câmeras.



Fonte: Original, baseado em Liu, Hu e Cheng (2005, p. 342)

É válido ter como verdade que um sumário contrastivo deve destacar apenas as diferenças entre as entidades analisadas, mas alguns autores (JIN; JI; GU, 2016) destacam também as similaridades. Também é natural admitir que, para que seja de fato contrastivo, o sumário idealmente tem que falar das mesmas características para todas as entidades. Por exemplo, ao se comparar dois celulares, é pouco útil que o sumário fale apenas da tela de um deles e da bateria do outro, a menos que isso realmente represente a informação contida na origem (como no caso em que a tela de um aparelho seja o único aspecto comentado de um aparelho e a bateria seja o único do outro). Outra coisa desejável é que os sumários tragam a maior quantidade de informação possível da fonte; para isso, deve-se evitar repetições para que haja mais espaço para tópicos diversificados. Os conceitos falados neste parágrafo serão mais adiante representados por medidas chamadas comparabilidade, representatividade e diversidade.

1.3 Objetivo

O foco desta pesquisa é a investigação de métodos de sumarização contrastiva de opinião como definida na seção anterior. Para os testes, são usados textos em português, mas os métodos de sumarização contrastiva estudados não dependem de língua: o que depende da língua é a etapa inicial de identificação de opinião. Foram encontrados alguns métodos na literatura voltados para tarefas iguais ou similares à sumarização contrastiva como definida aqui. Este trabalho reúne alguns desses métodos, sintetiza a teoria desenvolvida neles e os implementa para testes práticos. São propostas variações dos métodos encontrados na literatura e também abordagens inéditas.

Uma parte do estudo é dedicada às formas de avaliação. Avaliar um sumarizador não é uma tarefa direta, pois, para decidir quão bom um sumário é, existem muitos fatores a serem levados em conta, e esses fatores são geralmente subjetivos (i.e., não há consenso sobre o que é melhor e pior). Assim, são desenvolvidas métricas de avaliação que reflitam características que se consideram importantes em um sumário contrastivo. Os resumos produzidos também são avaliados por humanos de uma maneira mais intuitiva, onde cada avaliador pode julgar quão bom um sumário é para a sua própria necessidade. Tem-se como hipótese que os métodos terão performances diferentes (ou seja, alguns deles vão se sobressair em determinados cenários) e que os métodos de sumarização contrastiva são mais adequados para comparar entidades do que os métodos de sumarização simples. Além disso, as avaliações humanas permitirão investigar a hipótese de que os resumos gerados realmente apresentam alguma utilidade.

Embora existam trabalhos publicados sobre sumarização contrastiva, não há um estudo que compare os diferentes métodos disponíveis, sendo isso uma lacuna na área. As publicações encontradas usam diferentes conjuntos de dados e diferentes métricas de avaliação, e não é possível saber como os métodos se comparam uns com os outros. Neste trabalho, todos os métodos serão testados sob os mesmos parâmetros. Foi criado um conjunto de dados adequado para os testes e foram estudadas formas de avaliar os métodos para esse conjunto de dados.

Todo o estudo é suplementado por implementações práticas porque elas são indispensáveis para haver teste e avaliação dos métodos. Uma pergunta a ser respondida ao final do trabalho é como a performance dos métodos varia em consequência de características do conjunto analisado (por exemplo, quantidade de dados, razão entre opiniões positivas e negativas, tipos de opinião). É possível que haja um método diferente que se sobressaia em cada cenário.

Espera-se que os métodos descritos no trabalho permitam no futuro o desenvolvimento de outras ferramentas por pesquisadores da área, e que o trabalho assim contribua com a pesquisa em Processamento de Linguagem Natural, somando a ela ideias novas que poderão ser futuramente utilizadas.

1.4 Estrutura do texto

O restante do texto está organizado da seguinte maneira:

- No Capítulo **2**, são mostrados conceitos elementares das áreas de pesquisa das quais este estudo faz parte.
- No Capítulo **3**, resumem-se alguns trabalhos relacionados à pesquisa que já foram publicados por outros autores.
- O Capítulo **4** descreve detalhadamente cada método e relata a implementação de cada um deles.

- O Capítulo **5** relata o trabalho prático deste projeto. Ele descreve o conjunto de dados usado para testes, descreve os procedimentos de avaliação, mostra os resultados obtidos com a avaliação de cada método e faz uma análise geral dos resultados obtidos, com comparações entre os métodos, interpretação dos resultados e observações sobre pontos que podem ser aperfeiçoados.
- O Capítulo **6** fecha o texto com um resumo sobre os principais aprendizados deste trabalho e sugestões de trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, estão sintetizados os conceitos que são fundamentais para o restante do texto. São trazidos elementos de Sumarização Automática de Texto, Análise de Sentimento, Sumarização de Opinião tradicional e Sumarização Contrastiva de Opinião.

2.1 Sumarização Automática de Texto

A sumarização automática tradicional, que lida principalmente com textos objetivos, é a que será abordada nesta seção. Esse tipo de sumarização não considera características específicas de opiniões; entretanto, muitos dos conceitos de sumarização tradicional são herdados pela sumarização de opinião.

Sumarizar um conjunto de dados textuais, chamado de **texto-fonte** ou **conjunto-fonte** (ou **conjunto de entrada**, ou simplesmente **entrada** ou **fonte**), significa transformá-lo em um conjunto de menor tamanho (o **resumo** ou **sumário**) que contenha as informações mais relevantes[1] do conjunto original. Mani (1999, p. 3) fornece uma definição[2] ampla: 'um **sumarizador** é um sistema cujo objetivo é produzir uma representação reduzida do conteúdo de sua entrada para fim de consumo humano'.

Existe uma vasta gama de tarefas na área de sumarização automática, e os trabalhos da área podem ser agrupados segundo diferentes critérios, que serão mencionados a seguir.

Quanto à sua função principal, um sumário pode ser (MANI, 2001a):

- **Indicativo** Indica os principais conteúdos do texto-fonte sem detalhá-los. Um leitor, ao ler um sumário indicativo, sabe qual tipo de conteúdo vai encontrar ao ler o

[1] A relevância de informações é um conceito subjetivo, portanto debatível e não exato, e varia com cada contexto e necessidade; Lloret e Palomar (2010) apresentam alguns critérios que podem ser usados para determinar relevância.

[2] Traduzida aqui do original em inglês.

texto-fonte, mas não recebe as informações contidas no texto-fonte por meio do sumário. Alguns autores usam uma definição mais genérica, dizendo que os sumários indicativos podem conter qualquer metainformação sobre o texto-fonte (como autor, número de páginas, idioma, etc) (KAN; KLAVANS; MCKEOWN, 2002). O sumário[3] desta monografia é um exemplo de sumário indicativo.

- **Informativo** Exibe as principais informações do texto-fonte. Pode, em alguns casos, substituir o texto-fonte (naturalmente com menos riqueza de informação devido ao seu tamanho reduzido), pois carrega sua mensagem principal. O resumo[4] desta monografia é um exemplo. Também são exemplos o Quadro 1.1 e o Quadro 1.4.
- **Avaliativo**[5] Contém pontos opinativos do autor do resumo em relação à fonte. Um sumário avaliativo pode ser tanto indicativo quanto informativo[6]. Um sumário avaliativo informativo apresenta uma descrição da entidade sendo avaliada, além de elementos subjetivos; a crítica de filme do Quadro 1.2 (página 24) pode ser considerada um sumário avaliativo informativo do filme em questão. Um sumário avaliativo indicativo pode ser uma lista de entidades feita com base em alguma opinião (por exemplo, 'Lista de novelas da década passada que vale a pena não ver de novo') ou um comentário que aponte uma característica subjetiva de uma entidade sem detalhar, como *'este livro é bom, aquele outro é ruim'*.

Quanto à quantidade de textos-fonte, a sumarização pode ser (MANI, 2001a):

- **Monodocumento**: a entrada é formada por apenas um documento.
- **Multidocumento**: a entrada é um conjunto com dois ou mais textos. Quando esse é o caso, o sistema deve levar em consideração que pode haver redundância ou informações contraditórias entre os documentos.

Quanto ao procedimento para gerar o conteúdo da saída, a sumarização pode ser:

- **Extrativa**: seleciona os trechos do texto-fonte e os copia como são para o resumo. O sumário mostrado no Quadro 1.1 (página 24) faz isso; pode-se notar que as sentenças contidas nele são exatamente iguais a sentenças do texto-fonte (MANI, 2001a).
- **Abstrativa**: seleciona trechos do texto-fonte e interpreta-os de alguma maneira para gerar um resumo que os represente de alguma forma, sem copiá-los integralmente. Ganesan (2013, p. 52) nomeiam duas formas principais de sumarização abstrativa:

[3] Na língua portuguesa corriqueira, em particular neste contexto, a palavra 'sumário' tem um significado menos genérico do que na Sumarização Automática; assim, esse sumário é intitulado simplesmente 'sumário' porque a palavra 'sumário' geralmente remete a esse tipo de sumário.

[4] Na língua portuguesa corriqueira, em particular neste contexto, a palavra 'resumo' tem um significado menos genérico do que na Sumarização Automática; assim, esse resumo é intitulado simplesmente 'resumo' porque a palavra 'resumo' geralmente remete a esse tipo de resumo.

[5] No texto original em inglês (MANI, 2001a) é chamado de '*critical evaluative abstract*'.

[6] A divisão em três categorias não é mutualmente exclusiva.

– Por **geração de linguagem natural**: após selecionar as principais informações do texto-fonte, utiliza técnicas de geração de linguagem natural para construir um novo texto a partir delas, de maneira similar a um típico resumo feito por um humano. Um exemplo é o trabalho de Genest e Lapalme (2011).

– Por **modelo** (template): possui padrões de texto previamente definidos que são preenchidos com as informações selecionadas. Alguns serviços de meteorologia usam esse modelo: eles têm trechos predefinidos onde as lacunas são preenchidas com informações coletadas de um banco de dados, a fim de resumir essas informações para facilitar a leitura e absorção delas. A Figura 2.1 mostra dois exemplos obtidos em busca no portal Weather Spark que sumarizam as condições meteorológicas típicas de duas cidades. Note que o texto para as duas cidades é exatamente o mesmo exceto pelas informações que diferem entre elas, pois ambos foram feitos usando um mesmo modelo.

Figura 2.1 – Exemplos de sumários abstrativos com modelos predefinidos.

Fonte: Captura de tela de páginas do Weather Spark.

Para avaliar um sumário, algumas medidas propostas em Mani (2001b) são:

- **Informatividade**: mede o conteúdo do sumário para avaliar quanta informação da fonte foi preservada nele;
- **Coerência**: mede quanto as informações do sumário estão em contexto adequado, evitando lacunas em sua estrutura retórica (como anáforas sem referência, por exemplo).

Embora muitas definições de sumarização impliquem a transformação de um texto em um texto menor, existem publicações que fazem essa transformação com outros tipos de dados também sob o nome de sumarização. Por exemplo, há iniciativas de pesquisa que fazem uma sumarização textual de dados não textuais, como Carberry et al. (2004) e Demir, Carberry e McCoy (2008), que transformam certos tipos de gráficos em textos explicativos sobre a

informação contida neles. Os autores dizem que esse tipo de imagem é muito comum em reportagens escritas, e que muitas vezes a informação contida no gráfico não está repetida no texto, daí a importância de se analisar esse tipo de dado. O sentido contrário é igualmente válido: uma sumarização também pode utilizar recursos gráficos para sintetizar as informações extraídas de um texto, o que pode ser muito útil para abstrair dados numéricos e para comparar entidades diferentes. Por exemplo, pode-se desejar analisar um conjunto de textos jornalísticos para saber a quantidade de homicídios noticiados por cidade em determinado período. Após contabilizar as informações, pode-se apresentá-las em um gráfico de barras.

2.2 Análise de Sentimento

Esta seção traz características da Análise de Sentimento, ainda sem considerar a sumarização de opinião, que será assunto da próxima seção.

Como define Liu (2012)[7], 'a **Análise de Sentimento** (também chamada de **Mineração de Opinião**) é a área de estudo que analisa opiniões [...] de pessoas para com entidades [...] e seus atributos'.

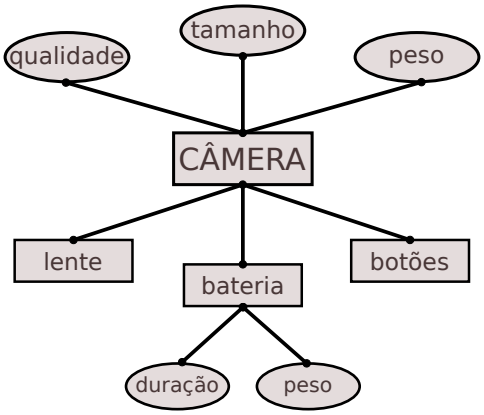
A definição inicial de Liu inclui não só opiniões como o foco de análise da área, mas também sentimentos, avaliações, apreciações[8], atitudes e emoções, mas o autor opta por não definir e distinguir cada uma dessas palavras a fim de simplificar a apresentação do texto, e passa a usar o termo genérico '**opinião**' para designar qualquer uma delas.

Entidade, ainda segundo Liu, é o objeto alvo da opinião; pode ser um produto, serviço, organização, indivíduo, questão[9], evento ou tópico. Ela pode ser definida como um par $e : (T, W)$ onde T é uma hierarquia de partes e subpartes da entidade e (com vários níveis, se necessário) e W é um conjunto de atributos de e. Uma **parte** é um elemento que compõe uma entidade ou uma de suas partes: 'teclado' é uma parte de 'computador', 'tecla enter' é uma parte de 'teclado'. Um **atributo** é uma característica de uma entidade ou uma de suas partes: 'peso' e 'tamanho' são atributos de 'computador'.

A definição de entidade feita por Liu descreve uma decomposição hierárquica da entidade baseada na relação 'é parte de'; o nó raiz da hierarquia é a entidade propriamente dita, e cada nó dos níveis abaixo da raiz representa uma parte do elemento representado por seu nó-pai. Por exemplo, uma câmera pode ter uma série de atributos, como qualidade da foto, tamanho, peso; além de ter atributos, ela tem partes, como: bateria, lente, botões. A bateria, por sua vez, também tem atributos: duração, peso, etc. A Figura 2.2 mostra um esquema de organização hierárquica.

[7] Tradução do original em inglês.
[8] No original em inglês, '*appraisal*'.
[9] No original em inglês, '*issue*'.

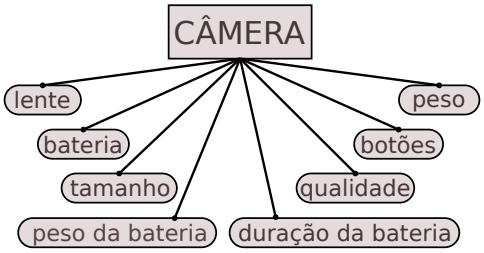
Figura 2.2 – Possível organização hierárquica (não exaustiva) da entidade 'câmera'. A entidade é a raiz da árvore, as partes estão representadas por retângulos e os atributos estão representados por elipses.



Fonte: Original

Para simplificar a representação de entidade, Liu reduz a hierarquia proposta a apenas dois níveis e passa a chamar tanto as partes quanto os atributos de **aspectos** da entidade. A Figura 2.3 ilustra como isso pode ser representado.

Figura 2.3 – Possível organização hierárquica (não exaustiva) da entidade 'câmera'. A entidade é a raiz da árvore e os demais nós são seus aspectos.



Fonte: Original

A mesma entidade pode ser representada por vários diferentes termos. Por exemplo, 'celular', 'telefone' e 'Sony Ericsson K300' podem referir-se à mesma entidade. Cada um desses termos é identificado como uma **expressão de entidade**. O mesmo ocorre com aspectos, e é definido, de forma análoga, o termo **expressão de aspecto**. Um exemplo de várias expressões de aspecto para o mesmo aspecto são 'imagem', 'foto' e 'fotografia' para a entidade 'câmera'.

Um aspecto pode ser explícito ou implícito. Um **aspecto explícito** é aquele cuja expressão é um substantivo ou sintagma nominal. Um exemplo é 'tela' em *'a tela desse celular é boa'*. Os aspectos que não aparecem como substantivos ou sintagmas nominais são chamados de **aspectos implícitos**. Por exemplo, em *'este produto é caro'*, a palavra 'caro' é um adjetivo que implicitamente representa o aspecto 'preço' (LIU, 2012).

Uma **opinião** é definida por (LIU, 2012) como uma quíntupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, onde e_i é uma entidade, a_{ij} é um aspecto de e_i, s_{ijkl} é um sentimento sobre o aspecto a_{ij}, h_k é a pessoa

detentora da opinião e t_l é o tempo em que a opinião foi expressada. Os índices subscritos enfatizam as correspondências entre os elementos da quíntupla.

O sentimento de uma opinião pode ser positivo, negativo ou neutro; essa medida se chama **orientação** ou **polaridade** da opinião (LIU, 2012). Quando o sentimento for positivo (negativo, neutro), pode-se dizer também que a opinião é positiva (negativa, neutra).

Uma opinião pode ser implícita ou explícita. Uma **opinião explícita** é uma afirmação subjetiva, por exemplo: *'este produto é bom'*, ou *'este produto é melhor que aquele'*. Uma **opinião implícita** é uma afirmação objetiva que sugere uma opinião, geralmente expressando um fato desejável ou indesejável, como: *'eu comprei esse produto semana passada e ele já estragou'* e *'a bateria deste telefone dura mais do que a daquele'* (LIU, 2012).

Uma sentença **objetiva** apresenta informação factual, e uma sentença **subjetiva** (também chamada de **opinativa** neste contexto) contém sentimentos, visões e crenças pessoais (LIU, 2012).

Os trabalhos de Análise de Sentimento tratam os conjuntos de entrada com diferentes **níveis de granularidade**. Pode-se processar um texto com interesse voltado para (LIU, 2012; FELDMAN, 2013):

- **Documento**: avalia-se todo o texto escrito pelo autor, do começo ao fim, a fim de entender se aquele texto expressa, no geral, uma opinião positiva, negativa ou neutra. Assume que o documento todo trata de uma única entidade.
- **Sentença**: distingue-se as sentenças positivas, negativas e neutras de um mesmo documento, processando cada sentença individualmente.
- **Entidade e aspecto**: considera-se que uma mesma sentença pode conter mais de uma opinião, e que cada uma dessas opiniões pode ter um alvo diferente e uma polaridade diferente. Um exemplo é *'mesmo com a organização do evento tendo sido horrível, valeu a pena ter ido ao show porque a banda teve uma excelente performance'*. Seria estranho dizer que essa sentença é positiva ou negativa; pode ser desejável identificar que 'horrível' refere-se à organização do evento e 'excelente' refere-se à banda. Desse modo, identifica-se que na sentença há duas opiniões diferentes.

Liu, Hu e Cheng (2005) distinguem três **níveis de detalhe** sob os quais um conjunto de opiniões pode ser analisado ou visualizado, onde o nível maior é o que traz mais detalhamento sobre o significado semântico da opinião:

- O nível de menor detalhe (nível 1) consiste apenas em separar as opiniões positivas das negativas.
- No nível seguinte, foca-se nos aspectos de cada produto, e as opiniões são separadas em positivas e negativas dentro de cada aspecto.

- No último nível, são estudados problemas específicos de cada aspecto, por exemplo: *'a foto fica borrada'* e *'a foto fica escura'* são reconhecidos como problemas distintos embora ambos sejam negativos e sejam sobre o mesmo aspecto.

Considerando os dois níveis de menor detalhe e o nível de granularidade da entidade e aspecto, é comum transformar cada sentença em um conjunto de duplas compostas apenas pelo aspecto e pelo sentimento sobre aquele aspecto, descartando-se todas as outras palavras. Por exemplo, a sentença *'as fotos e os vídeos têm uma boa qualidade, mas o áudio fica ruim'* pode ser representada como $\{(\text{fotos}, \text{boa qualidade}), (\text{vídeos}, \text{boa qualidade}), (\text{áudio}, \text{ruim})\}$. Pode-se abstrair ainda mais e descartar a informação léxica do sentimento, transformando seu valor semântico para um número dentro de uma escala que representa tão somente a polaridade do sentimento. Lerman, Blair-Goldensohn e McDonald (2009) fazem isso. Assim, a sentença do exemplo pode ser representada como $\{(\text{fotos}, 80), (\text{vídeos}, 80), (\text{áudio}, 10)\}$, supondo uma escala de 0 a 100 onde o valor é maior quanto mais positivo for o sentimento.

2.3 Sumarização de Opinião

A sumarização de opinião é um processo que herda características tanto da Sumarização Automática quanto da Análise de Sentimento. A sumarização de texto seleciona as informações mais importantes de uma fonte; a análise de sentimento extrai informações de textos opinativos; a **sumarização de opinião** seleciona as informações mais importantes de um conjunto de textos opinativos (LIU, 2012).

Liu (2012) diz que a sumarização de opinião pode ser vista como uma forma de sumarização multidocumento, porque ela visa agrupar informações contidas em vários textos escritos por autores diferentes. Não obstante, alguns trabalhos já foram feitos com sumarização de opinião monodocumento, onde o sistema deve resumir a opinião do autor de um único texto subjetivo; Beineke et al. (2004) executam essa tarefa sob o nome de "sumarização de sentimento", tratando-a como um problema de classificação no nível da sentença e resolvendo-a com modelos estatísticos e técnicas de otimização.

Identificam-se algumas diferenças fundamentais entre a sumarização de opinião e a sumarização tradicional (LIU, 2012, p. 102):

- A sumarização de opinião deve ser quantitativa, ou seja, contabilizar informação repetida (porque 20% dos usuários falando mal de um produto é bem diferente de 90% falando mal) enquanto a sumarização objetiva multidocumento normalmente não se preocupa em contabilizar informação repetida.
- A sumarização de opinião costuma preocupar-se com entidades e aspectos, isto é, associa a informação obtida sempre a um objeto alvo. Na sumarização objetiva, não se costuma identificar essas relações.

- Embora possam usar um formato de texto puro para representar a saída (igual aos sumarizadores mais tradicionais), muitas vezes pode ser mais adequado usar um formato estruturado (como o do Quadro 1.3, página 26), principalmente por causa da natureza quantitativa desse tipo de sumarização e pelo interesse em separar as opiniões por polaridade e por aspecto.

Apesar das diferenças entre a sumarização de opinião e a tradicional, nos primórdios da Mineração de Opinião já foram usadas técnicas puras da sumarização objetiva para se extrair os trechos mais relevantes de textos subjetivos, como Beineke et al. (2004), que usam um conjunto de dados da página Rotten Tomatoes para encontrar a melhor sentença dentro de um texto de opinião que o represente, usando Naive Bayes e modelos de regressão logística calculados sobre a posição das sentenças no texto e sobre a escolha de palavras. Atualmente, quase a totalidade dos trabalhos que lidam com textos subjetivos usa técnicas de Análise de Sentimento (RAVI; RAVI, 2015).

Muitos trabalhos de sumarização de opinião usam uma abordagem **baseada em aspectos**, o que quer dizer que elas capturam a essência da opinião por meio de dois elementos: o **alvo** da opinião (entidade ou aspecto) e o **sentimento** sobre o alvo (que é a opinião propriamente dita) (LIU, 2012). Liu (2012, p. 103) afirma que esse tipo de abordagem sempre deve ser quantitativo e informar em sua saída o número de pessoas que têm opiniões negativas e positivas sobre cada aspecto. Mas isso não é uma lei: alguns trabalhos (LERMAN; BLAIR-GOLDENSOHN; MCDONALD, 2009; JIN; JI; GU, 2016) fazem uma sumarização baseada em aspecto onde o resumo não traz informações sobre a quantidade de opiniões no conjunto-fonte, e a parte quantitativa da sumarização é somente internamente considerada (não exibida na saída), porquanto opiniões mais frequentes têm mais chance de serem selecionadas para o resumo de acordo com os métodos de sumarização elaborados.

Uma preocupação recorrente nos trabalhos de sumarização de opinião é que o sumário gerado seja representativo e diversificado (XU; MENG; CHENG, 2011; WANG; ZHU; LI, 2013; JIN; JI; GU, 2016). As medidas que medem essas características são a representatividade e a diversidade, e existem várias maneiras diferentes de calculá-las segundo diferentes trabalhos.

A **representatividade** mede o quanto o resumo reflete de maneira justa as informações contidas na fonte. Se em um conjunto de opiniões coletadas de várias pessoas ocorrer metade das sentenças falando bem de um produto e metade falando mal, é injusto que o sumário contenha apenas sentenças positivas; diz-se então que esse resumo não é representativo. No nível da entidade e aspecto, a representatividade também se relaciona com os tópicos mencionados: se 90% das pessoas fala sobre a bateria de uma câmera e apenas 5% das pessoas fala sobre a lente, é de maior valor incluir no sumário uma sentença sobre a bateria do que uma sobre a lente.

A **diversidade** é uma medida para indicar a quantidade de informação contida no sumário. É desejável que o sumário contenha a maior quantidade possível de informação em seu espaço limitado. Por isso, um sumário que fala só sobre um único tópico pode ser considerado ruim.

A diversidade e a representatividade podem ser tanto aliadas quanto concorrentes, ou seja, pode acontecer de uma ser automaticamente beneficiada quando se maximiza a outra, ou pode acontecer de uma ser prejudicada quando se maximiza a outra. Isso depende das características do conjunto-fonte, porque o conjunto-fonte, por sua vez, pode ser diversificado ou não. Um conjunto-fonte que contém aproximadamente a mesma quantidade de cada tópico, se sumarizado de forma a obter um resumo representativo, tem-se de imediato que esse resumo é diversificado, porque ser representativo significa refletir de maneira justa o conteúdo da fonte, e como a fonte é diversificada, o resumo também é. Porém, se a fonte não for diversificada (como no caso em que 90% das pessoas falam sobre a bateria de uma câmera e 5% das pessoas falam sobre a lente), um resumo representativo dela também não será. Nesse caso, um resumo mais diversificado seria aquele que contém 50% das sentenças falando sobre a bateria e 50% falando sobre a lente. Mas isso vai contra a representatividade, onde as opiniões mais frequentes na fonte têm maior valor. Nota-se então que é preciso fazer uma troca entre a representatividade e a diversidade a fim de balancear as características do resumo. A importância relativa entre as duas medidas é debatida futuramente em vários pontos deste texto.

2.4 Sumarização Contrastiva de Opinião

A tarefa foco deste projeto é a **sumarização contrastiva de opinião**, cujo objetivo é encontrar as principais informações que permitam comparar duas ou mais entidades dados conjuntos de textos opinativos sobre cada uma delas (LIU, 2012; KIM; ZHAI, 2009; LERMAN; MCDONALD, 2009; JIN; JI; GU, 2016).

O trabalho mais antigo de análise contrastiva de opinião que se conhece é Liu, Hu e Cheng (2005), que processa opiniões sobre produtos concorrentes. O objetivo principal foi fazer um sistema que permita ao usuário entender facilmente os pontos fortes e fracos de cada produto. O formato de sumário proposto é um gráfico de barras que separa os aspectos de cada produto, como o mostrado na Figura 1.4 (página 30). Como esse trabalho apenas separa as opiniões positivas das negativas de cada aspecto e as contabiliza, ele não usa técnicas de sumarização informativa, porque as opiniões são somente contabilizadas, não sendo necessário selecionar as mais relevantes.

O primeiro trabalho que usa técnicas de sumarização na análise contrastiva de opinião é Lerman e McDonald (2009), que tem como principal objetivo extrair informações de pares de entidades para gerar um sumário que realce as diferenças entre elas. A aplicação mais direta sugerida pelos autores é o domínio das avaliações de consumidores, onde um possível comprador pode

usar o resumo para ver diferenças entre opiniões sobre produtos sem ter que ler todas as avaliações de cada produto.

Lerman e McDonald (2009) definem a **sumarização contrastiva de opinião** como o problema de gerar um resumo que realce apenas as diferenças entre duas entidades recebendo como entrada textos opinativos sobre cada uma. Alguns pesquisadores (JIN; JI; GU, 2016), além de destacar as diferenças, fazem um sumário que mostra também as similaridades. Muitos trabalhos executam a tarefa sobre duas entidades (LERMAN; MCDONALD, 2009; JIN; JI; GU, 2016), mas também se pode defini-la para um número ilimitado de entidades (LIU; HU; CHENG, 2005).

Comumente, nos trabalhos de análise contrastiva de opinião, assume-se que os textos opinativos são previamente separados de modo que se saiba a qual entidade cada um deles se refere (LIU; HU; CHENG, 2005; LERMAN; MCDONALD, 2009; JIN; JI; GU, 2016). Uma exceção é Ibeke et al. (2017), que inclui no trabalho a tarefa de separar entidades diferentes mesmo que elas estejam mencionadas em um mesmo texto.

Na seção anterior, foram definidos os conceitos de representatividade e diversidade como sendo características desejáveis em um sumário de opinião. Quando se trata de sumários contrastivos de opinião, além dessas duas, uma outra característica é desejável: a **comparabilidade**, que mede quanto as sentenças contidas no sumário falam sobre assuntos similares das entidades sendo comparadas (JIN; JI; GU, 2016). Por exemplo, se se está comparando duas câmeras, é de pouco valor ter um sumário que fale somente da tela de uma delas e somente da bateria da outra, porque essas informações não permitem que haja uma comparação apropriada. Idealmente, para cada sentença sobre a bateria de uma das câmeras, deveria haver uma sentença sobre a bateria da outra.

Uma medida similar à comparabilidade é a **contrastividade**, que, além de exigir que as sentenças tratem sobre tópicos similares, exige que elas tenham polaridades opostas, isto é, apresentem opiniões discordantes (KIM; ZHAI, 2009). Neste projeto, esta medida será mais útil do que a comparabilidade, pois o foco aqui é obter informações divergentes sobre as entidades.

As três medidas de característica desejáveis – representatividade, diversidade e contrastividade – podem ser concorrentes, ou seja, pode ser que, ao se maximizar uma, outras sejam prejudicadas. Por isso, é preciso que haja um balanceamento entre as três; priorizar uma pode significar abrir mão de outra (JIN; JI; GU, 2016). Este texto discutirá isso com exemplos de trabalhos da literatura e dos resultados originais obtidos.

Alguns trabalhos descrevem a sumarização contrastiva de opinião como um problema de otimização combinatorial: dados dois conjuntos E_1 e E_2, cada um contendo textos opinativos sobre uma entidade diferente, definida uma função L que estime quão bom um resumo $R \in E_1 \cup E_2$ é, e escolhido um limite de tamanho t, o problema resume-se a resolver $\arg\max L(R)$ com $|R| \leq t$. Isso é um problema NP-difícil, donde a prática mais efetiva é usar um algoritmo

guloso para se encontrar soluções subótimas (WANG et al., 2012). Uma grande parte do estudo é dedicada a definir uma L adequada (LERMAN; MCDONALD, 2009; JIN; JI; GU, 2016). Também pode ser interessante definir mais de uma função L com cada uma medindo uma característica do resumo (Jin, Ji e Gu (2016) usam as três medidas mencionadas no parágrafo anterior).

2.4.1 Formato de sumário

Escolher adequadamente um conjunto de informação para fazer parte de um resumo é apenas uma das preocupações da sumarização contrastiva. Outra preocupação importante é a apresentação do resumo: uma vez escolhidas as informações que comporão a saída, como apresentá-las ao usuário? A apresentação pode ser textual ou gráfica. Alguns trabalhos usam gráficos de barras para exibir quantitativamente as opiniões encontradas (LIU; HU; CHENG, 2005). Dos que optam por um formato textual, é comum usar modelo bem estruturado, como uma tabela (KIM; ZHAI, 2009; JIN; JI; GU, 2016); uma exceção é Lerman e McDonald (2009), que não descrevem um formato específico para a saída, dando a entender que as sentenças de cada entidade são simplesmente concatenadas como texto simples.

O resumo gerado por Jin, Ji e Gu (2016) (mostrado na Figura 2.4) é alinhado, o que significa que as sentenças vêm aos pares (chamado no original de 'grupos'): cada par de sentença fala sobre um mesmo assunto e cada sentença do par é sobre uma das entidades de interesse. Um par pode trazer uma semelhança ou diferença entre os produtos, dependendo dos sentimentos considerados: positivo (de uma entidade) com positivo (da outra), positivo com negativo, etc. Para cada uma das quatro combinações de sentimentos, existem dois grupos, onde cada um é um par de sentenças que comparam as duas entidades.

Kim e Zhai (2009) fazem resumos contrastivos que trazem pares mostrando opiniões divergentes encontradas sobre uma mesma entidade. A Figura 2.5 mostra o formato ilustrado na publicação.

Sanchan, Bontcheva e Aker (2016) fazem um estudo explorando diferentes formatos de apresentação e quão útil cada um deles é para os usuários. Para fazer os sumários, eles usaram textos argumentativos sobre o tema de debate 'As mudanças climáticas são causadas pelo homem?' Um dos formatos apresentados no trabalho está ilustrado na Figura 2.6. Ele foi o melhor avaliado segundo os experimentos dos autores. Ele é formado por um gráfico de barras contendo os vários tópicos de argumentação; cada barra tem tamanho proporcional à quantidade de vezes que o tópico correspondente foi mencionado, e sua posição indica a quantidade de comentários que concorda ou discorda do tema. Por exemplo, para o tópico 'CO2', existem 38 comentários que o mencionam dentro de textos escritos por pessoas que advogam a favor do tema (ou seja, que defendem que as mudanças climáticas são causadas pelo homem) e 14 de pessoas que escreveram para tentar refutar o tema. Acoplado ao gráfico de barras, existe uma tabela que mostra os comentários para cada tópico.

Figura 2.4 – Formato de sumário gerado por Jin, Ji e Gu (2016).

Sentiment	Group #	Pair of sentences
Positive vs. positive	1	The battery life is really good. The battery life is pretty good.
	2	Very good battery life too. Good battery life.
Negative vs. negative	1	Battery life got worse as it was used. Battery life is not much longer than I expected.
	2	Battery life so far average. I've accidently drained the battery.
Negative vs. positive	1	Battery life got worse as it was used. Battery life is excellent as well.
	2	Battery life so far average. Battery life is good.
Positive vs. negative	1	Battery life is excellent as well. Battery life is not much longer than I expected.
	2	Battery life is good enough for the amount of processing the phone does. I've accidently drained the battery.

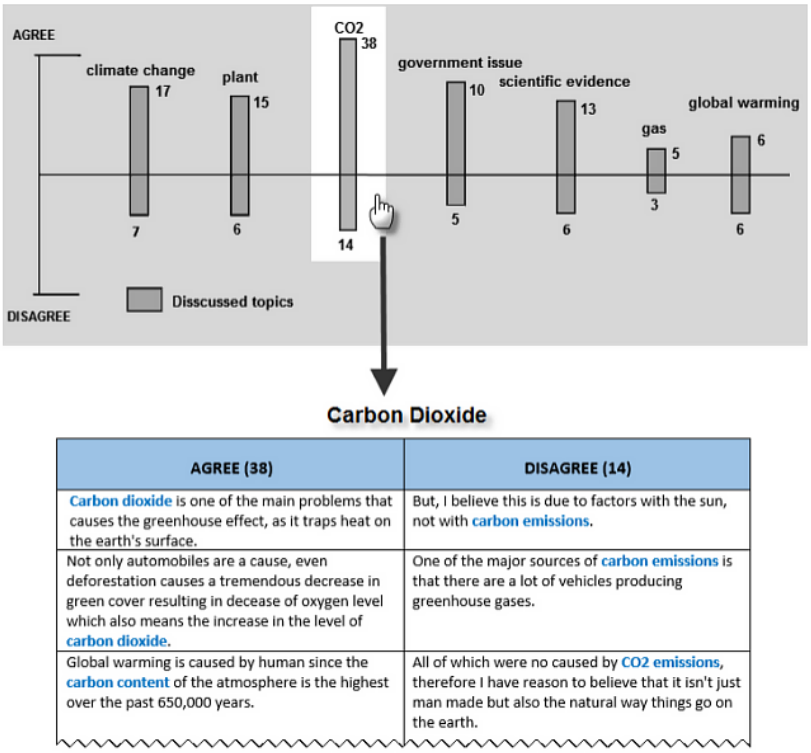
Fonte: Jin, Ji e Gu (2016)

Figura 2.5 – Formato de sumário gerado por Kim e Zhai (2009).

No	Positive	Negative
1	oh ... and file transfers are fast & easy .	you need the software to actually transfer files
2	i noticed that the micro adjustment knob and collet are well made and work well too.	the adjustment knob seemed ok, but when lowering the router, i have to practically pull it down while turning the knob.
3	the navigation is nice enough , but scrolling and searching through thousands of tracks , hundreds of albums or artists , or even dozens of genres is not conducive to save driving .	difficult navigation - i wo n't necessarily say " difficult ," but i do n't enjoy the scrollwheel to navigate .
4	i imagine if i left my player untouched (no backlight) it could play for considerably more than 12 hours at a low volume level.	there are 2 things that need fixing first is the battery life. it will run for 6 hrs without problems with medium usage of the buttons.

Fonte: Kim e Zhai (2009)

Figura 2.6 – Um dos formatos de sumário contrastivo propostos por Sanchan, Bontcheva e Aker (2016).



Fonte: Sanchan, Bontcheva e Aker (2016, p. 83)

O presente trabalho focará na seleção do conteúdo do sumário e adotará um formato simples para a apresentação desse conteúdo: um quadro dividido em dois lados onde cada lado mostra as opiniões de uma das entidades. Esse formato é exemplificado na Figura 2.7. Ele é funcionalmente equivalente ao formato chamado por Sanchan, Bontcheva e Aker (2016) de '*table summary*', exceto que em Sanchan, Bontcheva e Aker (2016) a tabela tem uma única coluna e as entidades aparecem uma embaixo da outra.

Figura 2.7 – Formato de sumário adotado neste trabalho.

Celular A	Celular B
A tela desse celular é muito boa!	A câmera não é das melhores.
A bateria não dura nada.	Demora para focar.
Gostei da câmera.	A resolução da tela é perfeita.

Fonte: Original

2.4.2 Atividades relacionadas

Não sendo exatamente o problema abordado neste projeto, a sumarização contrastiva de opinião como concebida por Kim e Zhai (2009) tem como meta formar um resumo com as

opiniões divergentes a respeito de uma mesma entidade. Os dois problemas são, todavia, muito parecidos. Apesar de ter foco final distinto da tarefa deste projeto, os procedimentos usados em ambas são muito similares, porque ambas buscam formas de se encontrar o melhor conjunto de opiniões opostas que representam um conjunto-fonte. A grande diferença é que neste presente estudo esse conjunto-fonte é separado em duas partes (uma para cada entidade) e os pares de sentenças com opiniões opostas devem conter uma sentença de cada lado dele.

Outras tarefas se assemelham à sumarização contrastiva de opinião, mas têm diferenças que devem ser observadas:

- A mineração de sentenças comparativas como feita por Jindal e Liu (2006) toma um conjunto de textos avaliativos extraído da Web e identifica as sentenças comparativas do texto para extrair suas relações de comparação. Um trabalho similar é chamado de 'análise comparativa de sentimento' por Feldman (2013) e é definido como uma subtarefa da Análise de Sentimento que lida com sentenças onde comparações aparecem explicitamente (por meio de adjetivos e advérbios contrastivos, superlativos, e algumas construções específicas: 'mais', 'menos', 'maior', 'do que', 'prefiro', 'número um', 'superior') (Liu (2012) chama esse tipo de sentença de 'opinião comparativa'). Esses dois trabalhos são sobre sentenças explicitamente comparativas. A sumarização contrastiva de opinião, por outro lado, compara entidades usando dois conjuntos de texto separados, um para cada entidade, de modo que se pode assumir que não há sentenças explicitamente comparativas, e as comparações são inferidas através das sentenças de cada entidade individualmente.
- A busca contrastiva, como desenvolvida por Sun et al. (2006), é uma tarefa de extração de informação que basicamente faz uma busca na Web que, em vez de permitir uma única consulta simultânea como nas ferramentas de busca mais comuns, permite ao usuário inserir duas consultas e retorna pares de resultados onde cada par corresponde a duas páginas da Web (uma para cada uma das consultas) que possuam alta similaridade entre si. Por exemplo, ao pesquisar 'Leonardo' e 'Rafael', o primeiro par retornado pode ser formado pela página da Wikipédia sobre Leonardo da Vinci e a página da Wikipédia sobre Rafael Sanzio, porque ambas as personalidades são pintores italianos do Alto Renascimento, e essas duas páginas, sendo ambas do mesmo portal, possuem grande similaridade quanto ao tipo de conteúdo disponibilizado. O sistema também é capaz de mostrar as diferenças entre os tópicos buscados. A diferença óbvia entre esse sistema e um de sumarização contrastivo de opinião é que ele não considera subjetividade.

CAPÍTULO

3

TRABALHOS RELACIONADOS

Este capítulo resume algumas linhas de trabalho que se considera importante ter como exemplos para o desenvolvimento desta pesquisa. O Quadro 3.1 resume as principais características dos trabalhos encontrados: Liu, Hu e Cheng (2005), Lerman e McDonald (2009), Jin, Ji e Gu (2016), Kim e Zhai (2009), Paul, Zhai e Girju (2010), Park, Lee e Song (2011), Özsoy e Çakıcı (2014), Guo et al. (2015) e Ibeke et al. (2017). As células destacadas em verde são pontos que melhor concordam com a proposta do presente trabalho (de fazer sumários que comparem dois ou mais produtos); as células em vermelho são pontos em total desacordo. Os quatro primeiros trabalhos do quadro são os que mais se alinharam a esta pesquisa; eles serão resumidos neste capítulo.

Quadro 3.1 – Lista de trabalhos relacionados encontrados.

	Trabalho		Formato do sumário	Domínio	Idioma	Nº entidades	Método	Avaliação
1	**Liu, Hu e Cheng**	**2005**	Gráfico de barras	Produtos	Inglês	n	Contagem de opiniões	✗
2	**Lerman e McDonald**	**2009**	Texto simples	Produtos	Inglês	2	**Distribuição de probabilidade**	Enquete
3	**Jin, Ji e Gu**	**2016**	Tabela pareada	Produtos	Inglês	2	**Similaridade de tópicos**	Automática
4	**Kim e Zhai**	**2009**	Tabela pareada	Produtos Aspartame	Inglês	1	**Agrupamento de sentenças**	Comparação com trabalho manual
5	Paul, Zhai e Girju	2010	Tabela pareada	Política	Inglês	1	Passeio aleatório	Comparação com trabalho manual
6	Park, Lee e Song	2011	✗ Classificação	Notícias	Coreano	1	Busca de tópicos induzida por hiperlink (HITS)	Comparação com trabalho manual
7	Özsoy e Çakici	2014	Tabela pareada	Filmes Produtos Aspartame	Turco Inglês	1	Diversificação max-sum	Comparação com trabalho manual
8	Guo et al.	2015	Tabela de tópicos	Controvérsias	Inglês	1	Análise probabilística de semântica latente (PLSA)	Comparação com trabalho manual
9	Ibeke et al.	2017	Tabela de tópicos	Software Plano de saúde	Inglês	1	Modelo de variável latente	Automática

Poucos trabalhos fazem a sumarização contrastiva como definida por este estudo (apenas Lerman e McDonald (2009) definem o problema de maneira idêntica). Então, serão mencionados também alguns trabalhos que executam tarefas similares, por dois motivos, explanados nos próximos dois parágrafos.

Tarefas similares podem ajudar a ilustrar o problema e conhecer formas alternativas de resolvê-lo, como Liu, Hu e Cheng (2005) (descrito em breve), que tem o mesmo objetivo de comparar automaticamente dois produtos a partir de textos opinativos, mas apresenta uma solução com formato bem diferente da proposta por este estudo.

Tarefas similares podem ser adaptadas para se alcançar a solução do problema como definido neste trabalho. Por exemplo, Kim e Zhai (2009) (descrito em breve) estuda métodos de se identificar opiniões divergentes dentro de um conjunto de textos opinativos sobre uma entidade. Esse método é adaptado neste estudo para que se forme um resumo contrastivo entre duas entidades.

Este capítulo narra os estudos feitos em trabalhos anteriores, focando principalmente na parte prática da implementação dos métodos e nos resultados e conclusões obtidos nesses trabalhos. A parte teórica é detalhadamente apresentada no Capítulo 4, onde serão descritos os métodos investigados no presente estudo e apresentado um método novo.

A seguir, listam-se os trabalhos que são abordados neste capítulo.

Liu, Hu e Cheng (2005) fazem uma ferramenta de sumarização contrastiva indicativa. Ela não produz um resumo textual que compare duas entidades, mas sim contabiliza a frequência de opiniões positivas e negativas de cada entidade para cada tópico (assunto, aspecto, atributo) identificado. O formato de resumo usado é um gráfico de barras que indica esses números. Esse trabalho será resumido na Seção 3.1.

Lerman e McDonald (2009) descrevem um método de seleção de sentenças de textos opinativos sobre duas entidades a fim de gerar um resumo que as compare. O resumo é gerado de modo que ele mantenha (tanto quanto possível) a média e o desvio padrão das polaridades de cada aspecto de seu conjunto-fonte, porém priorizando as opiniões que divergem entre as entidades. Esse trabalho é brevemente descrito na Seção 3.2 e aprofundado na Seção 4.1, onde também se relatam sua implementação, adaptação e resultados obtidos com este projeto.

Kim e Zhai (2009) resolvem o problema de selecionar pares de opiniões divergentes dentro de um conjunto de textos opinativos sobre uma entidade. O objetivo é gerar um resumo representativo que contenha a maior quantidade possível de pares representativos. Dois algoritmos são propostos: um que seleciona sentenças representativas e depois tenta formar pares contrastivos com as sentenças selecionadas, e um que primeiro forma pares contrastivos de sentenças e depois os seleciona de maneira a obter um resumo com boa representatividade. Esse trabalho está resumido na Seção 3.3. Na Seção 4.2, o método será descrito com detalhes e será apresentada uma adaptação dele para resolver o problema da sumarização contrastiva como proposta neste

trabalho. Esse método já foi também estudado por Sousa (2018), que implementou o método e explorou variações dele.

Jin, Ji e Gu (2016) têm como objetivo obter um resumo que traga diferenças e similaridades encontradas em textos opinativos sobre um tópico específico de uma entidade. Eles fazem isso com base em medidas que estimam quão similares duas sentenças são; a partir daí, encontram pares de sentenças com alta similaridade (que indicam uma semelhança entre as duas entidades) e baixa similaridade (que indicam uma diferença). Esse trabalho é resumido na Seção 3.4 e seu método é detalhado no Seção 4.3.

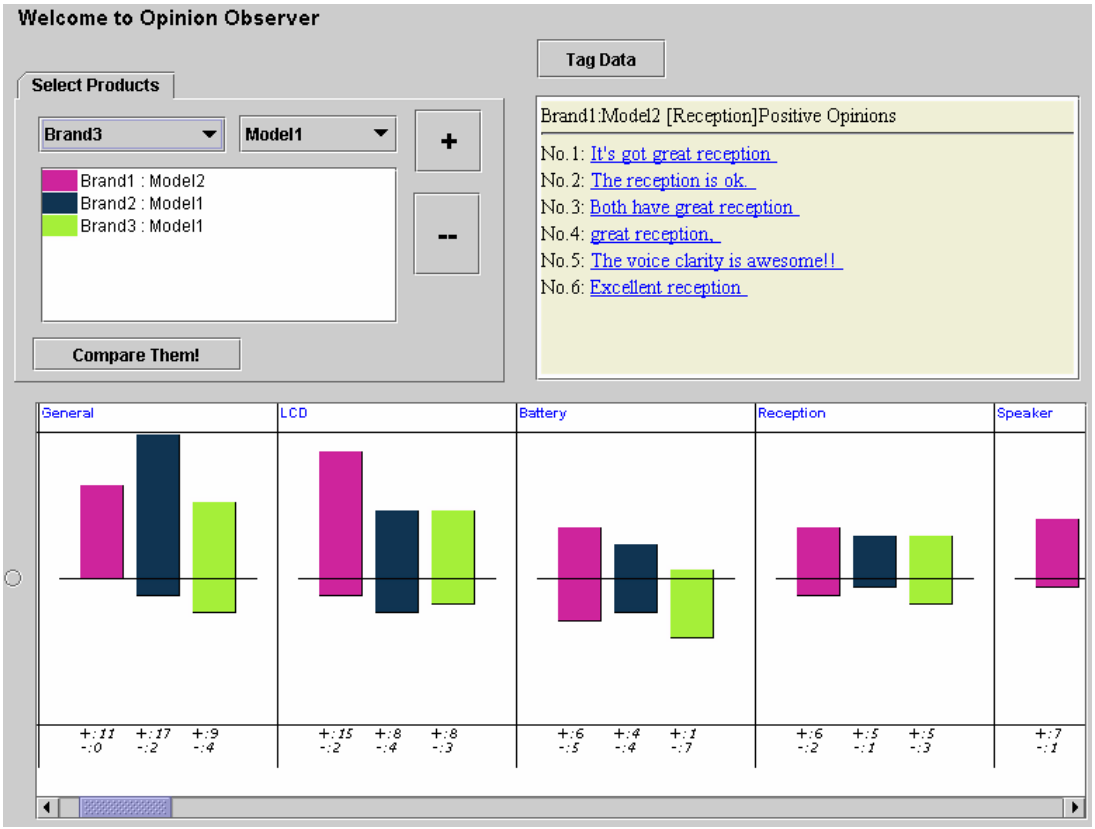
3.1 Sumarização contrastiva indicativa

Esta seção envolve o trabalho de Liu, Hu e Cheng (2005). Ele usa Hu e Liu (2004a) (que considera textos opinativos sobre produtos e descreve um método de identificar polaridades e aspectos a fim de separar as sentenças positivas das negativas para cada aspecto identificado) para gerar um sumário gráfico que contabiliza as sentenças positivas e negativas de dois ou mais produtos, separadas por aspecto; esse gráfico pode ser visto como um sumário contrastivo. O artigo é o mais antigo que se tem notícia a apresentar uma maneira automática de comparar dois produtos baseando-se em textos opinativos.

A comparação é feita identificando-se as opiniões de cada produto e contabilizando-as separadamente; não há, portanto, um confronto direto entre os produtos, ou seja, as opiniões sobre cada um não são individualmente comparadas umas com as outras, e somente a quantidade de opiniões positivas e negativas é levada em conta. Assim, o sumário gerado limita-se a trazer características do conjunto-fonte (quantidade de opiniões positivas e quantidade de opiniões negativas) e não propriamente conteúdo dele. Não se usam métodos de sumarização para escolher as informações mais relevantes, já que todas as informações são simplesmente contabilizadas.

Os pesquisadores implementam um sistema que auxilia o usuário a compreender as opiniões publicadas para comparar dois produtos. Esse sistema produz uma saída gráfica que permite ao usuário comparar diferentes aspectos de dois produtos de maneira rápida e intuitiva. A Figura 3.1 mostra a tela principal do sistema. Na figura, ocorre a comparação entre três produtos. Os produtos são selecionados na parte superior esquerda da janela. Na parte inferior, é mostrado o gráfico que indica a quantidade de opiniões positivas e negativas para cada aspecto de cada produto; a distância da linha horizontal preta até topo de cada barra é proporcional ao número de opiniões positivas, e até a parte de baixo de cada barra é proporcional ao número de opiniões negativas. O número de opiniões positivas e negativas é indicado embaixo de cada barra. Ao clicar na parte positiva ou negativa de uma barra, são mostradas as opiniões que a geraram na parte superior direita; todas as opiniões são mostradas, e não somente as mais relevantes.

Figura 3.1 – Tela principal do sistema implementado por Liu, Hu e Cheng (2005).



Fonte: Liu, Hu e Cheng (2005, p. 345)

3.1.1 Descrição teórica

O método usado para separar as opiniões positivas das negativas é descrito em Hu e Liu (2004a), onde a tarefa é executada em três passos principais:

1. Mineração de texto e identificação de aspectos de produtos sobre os quais os consumidores comentaram, com uso de técnicas de Mineração de Dados e Processamento de Linguagem Natural, como descrito em Hu e Liu (2004b).
2. Identificação de sentenças opinativas e suas polaridades. Sentenças que não contêm explicitamente um aspecto são descartadas. A identificação de polaridade é feita em três etapas:
   a) É listado um conjunto de adjetivos que são comumente usados para expressar opiniões. Isso é feito com técnicas de Processamento de Linguagem Natural.
   b) É identificada a polaridade de cada adjetivo do conjunto. Isso é feito com a WordNet (MILLER et al., 1990; FELLBAUM, 1998).
   c) É estimada a polaridade de cada sentença. Isso é feito com um algoritmo proposto no próprio artigo.

3. Sumarização do conjunto-fonte, com uma estrutura que contabiliza a quantidade de sentenças positivas e negativas de cada aspecto e lista todas as sentenças opinativas encontradas separadas por polaridade, como no Quadro 3.2.

Quadro 3.2 – Exemplo de estrutura de sumário de opinião (não contrastivo).

CÂMERA DIGITAL 1:
aspecto: Qualidade de foto
positivas: 253
<sentenças>
negativas: 6
<sentenças>
aspecto: Tamanho
positivas: 134
<sentenças>
negativas: 10
<sentenças>

Fonte: Hu e Liu (2004a, p. 168) (traduzido do inglês, formato adaptado)

3.1.2 Testes práticos

3.1.2.1 Conjunto de dados

Os conjuntos-fonte foram consistidos de resenhas de consumidores sobre cinco produtos eletrônicos: duas câmeras, um tocador de DVD, um tocador de MP3 e um telefone celular. Foram coletadas as 100 resenhas mais recentes de cada produto por meio das páginas virtuais Amazon e CNET.

3.1.2.2 Avaliação

Para a avaliação, as três tarefas (identificação de aspectos, de opiniões e de polaridade) foram executadas manualmente pelos autores a fim de comparar o resultado com aquele automaticamente obtido. Apesar de o sistema ignorar aspectos implícitos, eles foram considerados na execução manual: por exemplo, na sentença *'cabe bem no bolso'*, os autores identificaram o aspecto 'tamanho' como positivo.

A avaliação foi feita sob três perspectivas:

1. A efetividade de extração de aspectos: obteve-se cobertura de 80% e precisão de 72%.
2. A efetividade de extração de sentenças opinativas: obteve-se cobertura de 69% e precisão de 64%;
3. A efetividade de identificação de polaridade: obteve-se acurácia de 84%.

As medidas de avaliação informadas são as médias de cada uma delas calculada para todos os produtos.

A efetividade de extração de aspectos foi comparada contra o sistema de indexação automática FASTR[1]. Segundo os autores, ele foi escolhido para a comparação porque é muito conhecido e é publicamente disponível. Esse sistema forneceu cobertura de 17% e precisão de 3,1%.

3.1.3 Análise crítica

Quanto ao objetivo geral de comparar produtos por meio de textos opinativos, o método resumido nesta seção condiz com a proposta deste trabalho. No entanto, a forma como esse objetivo é atingido é dissonante dos propósitos aqui almejados. Em primeiro lugar, o método não apresenta uma forma de se escolher opiniões relevantes (que é o foco da sumarização contrastiva como definida neste projeto). Em segundo lugar, o maior esforço do método cai sobre a etapa de identificar opiniões, que neste trabalho é considerada tarefa pronta.

O método está reportado neste texto pela sua importância histórica (sendo o primeiro sistema conhecido a propor a comparação entre produtos por meio de textos opinativos) e também por permitir ilustrar um outro caminho para resolver o mesmo problema estudado nesta pesquisa. Por estar distante das rotas de investigação pretendidas aqui, esse trabalho não será subsequentemente explorado.

3.2 Sumarização contrastiva com distribuição de probabilidade

Lerman, Blair-Goldensohn e McDonald (2009) desenvolveram um sumarizador de opinião (não contrastivo) por meio de métodos estatísticos que consideram a probabilidade de um certo aspecto ter uma certa polaridade dentro do conjunto de entrada. Esse trabalho é usado por Lerman e McDonald (2009) para desenvolver um sumarizador contrastivo que gere sumários que destaquem diferenças entre duas entidades a partir de dois conjuntos de dados, cada um contendo textos opinativos sobre cada uma das entidades.

Esta seção fala brevemente sobre o trabalho de Lerman e McDonald. Mais adiante neste texto, o método proposto pelos autores será adaptado, estendido e avaliado.

[1] Disponível em perso.limsi.fr/jacquemi/FASTR/, acesso em 24/2/2018.

3.2.1 Descrição teórica

Para o trabalho desenvolvido por Lerman e McDonald, o tarefa da sumarização de opinião é vista como a resolução da equação

$$R = \arg\max_{|R| \leq k} L(R) \tag{3.1}$$

sendo L é uma função que atribui uma pontuação ao resumo R e k o limite de tamanho do resumo. A pontuação é maior quanto maior for a quantidade de informações desejáveis que o sumário contém. O problema da sumarização de opinião é então encontrar uma forma ideal para o cálculo de L, que atribua pontuações boas para resumos bons e pontuações ruins para resumos ruins. Uma vez definida tal função, o problema pode ser resolvido com técnicas de otimização.

O modelo proposto por Lerman, Blair-Goldensohn e McDonald usa uma abordagem probabilística para encontrar o resumo onde cada aspecto apareça com sentimentos que reflitam os sentimentos dele no conjunto-fonte: eles encontram a média e o desvio padrão da polaridade de cada aspecto e montam o resumo de forma que a distribuição de polaridades para cada aspecto no resumo seja o mais próximo possível da observada no conjunto-fonte, considerando que as polaridades de cada aspecto respeitam a distribuição normal[2].

O modelo de Lerman, Blair-Goldensohn e McDonald é adaptado por Lerman e McDonald para a resolução do problema da sumarização contrastiva.

Sejam:

e_1 : um produto qualquer;
e_2 : um produto concorrente de e_1;
E_1 : um conjunto de sentenças avaliativas sobre e_1;
E_2 : um conjunto de sentenças avaliativas sobre e_2;
R_1 : um subconjunto de E_1;
R_2 : um subconjunto de E_2;
k : a quantidade máxima de sentenças em R_1 e R_2.

Um **sumarizador contrastivo**, na visão desses autores, produz dois sumários R_1 e R_2 que incluem sentenças que enfatizam as diferenças nas opiniões sobre as entidades e_1 e e_2. Um resumo contrastivo R é visto então como um par de sumários R_1 e R_2, um sobre cada entidade. A ordem em que as sentenças aparecem no sumário não é considerada.

Partindo-se do sumarizador de opinião tradicional descrito em Lerman, Blair-Goldensohn e McDonald (2009), Lerman e McDonald (2009) produzem o método de geração de sumários contrastivos modificando-se a função objetivo L, que inicialmente só serve para a geração de

[2] Isso será detalhado na Seção 4.1.1

um sumário simples, de uma única entidade. São testadas três estratégias diferentes para a geração dos sumários R_1 e R_2:

1. Gerar os dois sumários de forma independente, sem modificar a função de pontuação;
2. Modificar a função L de forma que ela aumente a pontuação de pares de sumários se eles forem divergentes entre si;
3. Modificar a função L de forma que ela aumente a pontuação de cada sumário se ele for divergente do conjunto-fonte relativo à entidade oposta.

A terceira estratégia foi percebida como a mais vantajosa na fase de avaliação. Detalhes sobre a descrição teórica dessa técnica estão na Seção 4.1.

3.2.2 Testes práticos

3.2.2.1 Conjunto de dados

Para os experimentos, os autores usaram textos opinativos (extraídos de páginas como CNET, Epinions, PriceGrabber) sobre 56 produtos eletrônicos de 15 categorias (tocadores de MP3, câmeras, computadores, dispositivos de localização, etc). Cada produto tem pelo menos 4 resenhas; a média de textos por produto é 70.

3.2.2.2 Execução

O limite do sumário foi de 650 caracteres (aproximadamente quatro trechos de 160 caracteres). Foram feitos 89 sumários com cada estratégia.

O texto publicado não fornece exemplos de sumários gerados, o que permite entender que os sumários são formados simplesmente concatenando-se as sentenças selecionadas pelo método.

3.2.2.3 Avaliação

Cada sumário foi avaliado por três colaboradores. Os colaboradores recebiam um sumário contrastivo e eram solicitados a listar entre 1 e 3 diferenças entre os produtos. Depois, eles liam o conjunto inteiro de resenhas para decidir se as diferenças observadas de fato refletiam as opiniões coletadas. Ao final, eles davam pontuações indicando quão úteis os sumários foram para ajudá-los a encontrar essas diferenças.

A avaliação feita pelos usuários permitiu concluir que a terceira estratégia para o cálculo de L (descrita na Seção 3.2.1) se sobressaiu, especialmente quando os usuários foram solicitados a listar as diferenças entre os produtos.

3.2.3 Análise crítica

Lerman, Blair-Goldensohn e McDonald (2009) trazem um modelo interessante de sumarização de opinião, que considera as ocorrências de opiniões no conjunto-fonte como distribuições normais e tenta montar um resumo que seja fiel às distribuições de seu conjunto-fonte. Esse modelo é adaptado por Lerman, Blair-Goldensohn e McDonald (2009) para fazer sumarização contrastiva.

O fato de o modelo de sumarização contrastiva ser diretamente derivado de um modelo de sumarização comum é muito proveitoso, pois:

- Permite, com o exemplo, entender exatamente o que muda entre um sistema de sumarização comum e um contrastivo, inclusive servindo de inspiração para a eventual adaptação de outros modelos de sumarização comum que se deseja transformar em contrastivos;
- Permite comparar o modelo desenvolvido para sumarização contrastiva com o desenvolvido para sumarização comum, especialmente para averiguar se o primeiro de fato consegue comparar entidades melhor do que o segundo.

Por trazer um modelo interessante, por apresentar versões contrastiva e não contrastiva e por ser o único trabalho encontrado na literatura a definir a sumarização contrastiva como a proposta deste projeto, esses métodos serão investigados aprofundadamente e replicados.

Os dois trabalhos (especialmente Lerman, Blair-Goldensohn e McDonald (2009)) omitem algumas informações que seriam imprescindíveis para a replicação do método (por exemplo, alguns valores de ajuste de parâmetros e algumas decisões de modelagem). Portanto, as replicações demandarão que essas informações sejam deduzidas ou por meio de testes que indicarão a decisão que fornece melhores resultados, ou por meio de inferências sobre qual decisão é mais coerente dentro do modelo teórico. Essas deduções serão apontadas nos momentos oportunos.

3.3 Sumarização contrastiva com agrupamento de similaridade

Esta seção resume o trabalho de Kim e Zhai (2009). No trabalho, o objetivo principal é receber textos opinativos sobre uma entidade e identificar opiniões divergentes sobre ela. As opiniões identificadas são dispostas em um resumo aos pares, onde cada par é formado por duas opiniões que discordam sobre um certo assunto a respeito da entidade.

3.3.1 Descrição teórica

Kim a Zhai tratam a tarefa como um problema de otimização com base em funções heurísticas que estimam a representatividade e contrastividade de um resumo. Para resolver o problema, são propostos dois algoritmos gulosos.

Em um dos algoritmos propostos, faz-se um agrupamento para que sejam identificadas as sentenças similares, que supostamente tratam sobre um mesmo assunto. Então, escolhe-se uma sentença de cada grupo para entrar para o sumário. Assim, garante-se que o sumário será representativo (pois cada grupo será representado por uma sentença) e diversificado (porque as sentenças de um grupo supostamente são bem diferentes das de outro). São testadas duas estratégias de escolha da sentença que representará cada grupo no sumário. Ao final, formam-se pares contrastivos de sentenças, identificando, dentre as sentenças escolhidas, quais delas discordam entre si. Esse algoritmo é chamado de **R-First** (*representativity first*) pelo fato de priorizar a representatividade.

Em um outro algoritmo proposto por Kim e Zhai, o primeiro passo é a formação de pares contrastivos a partir das sentenças do conjunto-fonte. Esses pares são ranqueados do mais contrastivo para o menos contrastivo, segundo alguma função heurística que estime a contrastividade de pares de sentenças. Então, esses pares são selecionados um a um para preencher o sumário, começando do topo do ranque, até atingir o limite de tamanho do sumário. Porém, em vez de simplesmente selecionar os elementos do topo do ranque, a seleção também considera os elementos previamente selecionados para evitar redundância e maximizar a representatividade do sumário; algumas sentenças do topo podem ser descartadas por serem desnecessárias no sumário. Esse algoritmo é chamado de **C-First** (*contrastivity first*) porque prioriza a contrastividade.

Os algoritmos propostos por Kim e Zhai são detalhados na Seção 4.2.

Para ambos os algoritmos, o problema é visto como uma otimização cujo objetivo é encontrar o resumo R que maximize a função objetivo

$$f(R) = \lambda \mathrm{REP}(R) + (1 - \lambda)\mathrm{CONT}(R),$$

sendo $\mathrm{REP}(R)$ uma função que estima a representatividade de R, $\mathrm{CONT}(R)$ uma função que estima a contrastividade de R e λ um parâmetro que balanceia a importância entre as duas medidas.

No trabalho, a representatividade mede quão bem um resumo reflete as opiniões do conjunto-fonte, e a contrastividade mede quão bem cada par de sentenças do resumo representa um par contrastivo: se as sentenças do resumo forem muito similares a muitas sentenças do conjunto-fonte, tem-se alta representatividade; se, para todos os pares do resumo, um elemento do par contiver tópicos similares ao outro elemento do par, tem-se alta contrastividade.

A similaridade entre sentenças é medida com uma função de similaridade de termos: quanto mais termos similares uma sentença tiver a termos de outra sentença, mais similares essas duas sentenças são.

3.3.2 Testes práticos

3.3.2.1 Conjunto de dados

Para testar os algoritmos propostos, Kim e Zhai (2009) usaram um conjunto de opiniões sobre produtos obtido de trabalhos anteriores (HU; LIU, 2004a; HU; LIU, 2004b). Esses dados foram coletados na página Amazon e etiquetados quanto ao aspecto e polaridade de cada opinião. Cada caso de teste é composto por opiniões sobre um aspecto específico sobre um dos produtos, já previamente separados em positivos e negativos. Três exemplos de conjuntos de teste usados são: 14 opiniões sobre o design de um aparelho de MP3, 15 opiniões sobre a bateria de um telefone celular e 100 opiniões sobre um tocador de DVD. Para testar a generalidade dos métodos, os autores usam um outro conjunto de dados que não é sobre resenhas de produtos: trata-se de sentenças opinativas sobre o aspartame coletadas de diversas fontes da Web.

Para avaliar a performance dos algoritmos, foram feitos agrupamentos e pareamentos manuais no conjunto de dados usado nos testes. Dois auxiliares humanos trabalharam organizando as sentenças em grupos. Para cada caso de teste, eles deveriam identificar assuntos sobre os quais as sentenças falavam e etiquetá-las com esses assuntos. Por exemplo, em um caso de teste contendo opiniões sobre a bateria de um celular, possíveis assuntos seriam 'duração da carga', 'tempo de carregamento' e 'vida útil'. Essa anotação equivale ao agrupamento e ao alinhamento de grupos que são feitos nos algoritmos descritos na seção anterior: sentenças etiquetadas com um mesmo assunto pertencem (segundo os auxiliares) a um mesmo grupo; grupos de polaridades opostas etiquetados com um mesmo assunto formam um par. Grupos que (segundo os auxiliares) não poderiam ser alinhados foram descartados (como quando um assunto aparece apenas na parte positiva do conjunto, por exemplo).

Os autores argumentam que a sumarização contrastiva de opinião não faz sentido para produtos que tenham opiniões predominantemente positivas ou predominantemente negativas, em cujo caso um sumário simples seria o bastante. Então, quando esse era o caso, o produto era eliminado dos testes.

3.3.2.2 Execução

O limite de tamanho do sumário foi escolhido heuristicamente como $k = 1 + \log_2(|E_1| + |E_2|)$ (onde E_1 e E_2 são o número de sentenças no conjunto de dados de cada produto) a fim de

que ele seja maior para quanto mais sentenças houver na fonte, mas sature quando o conjunto de entrada for muito grande.

O Quadro 3.3 mostra um sumário obtido para um certo produto. Kim e Zhai (2009) enfatizam que o sumário ajuda o leitor a identificar diferentes perspectivas sobre um alvo; por exemplo, nos dois primeiros pares da tabela, fica claro que as opiniões são opostas por causa de diferentes experiências que as pessoas têm ao tentar fazer uma mesma coisa, também sugerindo que a maneira com que o produto foi usado em cada um dos casos é diferente, porque pessoas diferentes têm hábitos diferentes, capacidades diferentes, etc.

Quadro 3.3 – Exemplo de sumário contrastivo.

Sentenças positivas	**Sentenças negativas**
Ah, e a transferência de arquivos é bem rápida e fácil.	Tem que instalar um programa no computador para poder transferir arquivos.
A alavanca de ajuste é bem-feita e funciona muito bem.	A alavanca de ajuste parecia boa, mas é necessário puxar o equipamento com força na hora de ajustar.
Eu acho que se ficar sem mexer no aparelho (visor desligado), ele pode tocar por mais de 12 horas em volume baixo.	Tem 2 coisas que precisam melhorar. Primeiro, a duração da bateria. Ele funciona umas 6 horas sem problemas com uso moderado dos botões.

Fonte: Kim e Zhai (2009, p. 391) (tradução do original em inglês)

3.3.2.3 Avaliação

Os sumários automáticos são avaliados por Kim e Zhai (2009) de acordo com as anotações humanas quanto à precisão do alinhamento e à cobertura de aspectos. As anotações de cada auxiliar foram usadas em testes separados, e foi feita a média das duas para obter a avaliação geral.

A **precisão do alinhamento** é o percentual dos pares contidos no sumário que estão de acordo com o trabalho humano: para cada par contido no sumário, se ele existe em algum grupo pareado feito pelo anotador, então ela é contada como correta. Mede a contrastividade do sumário.

A **cobertura de aspecto** é o percentual de grupos alinhados pelo anotador que estão inclusos no sumário[3]. Mede a representatividade do sumário.

Ambos os métodos (R-First e C-First) obtiveram uma cobertura de aspectos bem maior do que a precisão de alinhamento, o que indica que é mais fácil atingir a representatividade do que a contrastividade. Os resultados do C-First ficaram melhores do que o R-First em quase todos os casos, confirmando que é mais importante priorizar a contrastividade.

[3] Grupos que não podem ser alinhados são desconsiderados (por exemplo, quando um grupo aparece somente nas opiniões positivas). Então, essa medida informa a quantidade de aspectos contidos no sumário em relação ao total que ele poderia conter.

3.3.3 Análise crítica

Existe uma grande diferença entre o problema resolvido por Kim e Zhai (2009) e o estudado neste trabalho: o método de Kim e Zhai é usado para encontrar opiniões contrastivas sobre uma mesma entidade, enquanto neste trabalho se deseja encontrar opiniões contrastivas de entidades diferentes. Todavia, é cogitada uma forma simples de resolver essa diferença: em vez de alimentar o sistema de Kim e Zhai com sentenças positivas e negativas de uma mesma entidade, pode-se colocar na entrada sentenças positivas de uma entidade e sentenças negativas de outra; assim, os pares contrastivos formados trarão diferenças entre as entidades, como pretendido.

Um ponto curioso do método é que ele não considera aspectos, mas sim todos os itens léxicos, para definir o assunto principal de uma sentença. Com isso, seria totalmente dispensável uma etapa de identificação de aspectos.

Quanto ao formato de saída, esse método se diferencia do de Lerman e McDonald (2009) (apresentado na seção anterior) pelo fato de parear as sentenças no resumo, ou seja, associar cada sentença a uma outra que (idealmente) contém uma opinião contrária à dela.

Por permitir testar uma forma de resolver o problema sem necessidade de identificação de aspectos e por proporcionar a possibilidade de uma adaptação inédita para resolver o problema de comparar entidades diferentes, esse método será investigado neste trabalho.

Embora cada caso de teste de Kim e Zhai contenha apenas sentenças sobre um aspecto específico de uma entidade, ele também pode ser executado sem problemas com todas as sentenças de uma entidade, como define a proposta deste projeto.

3.4 Sumarização contrastiva com ranqueamento de similaridade

Esta seção resumirá um trabalho que define medidas de similaridade para calcular características desejáveis em sumários comparativos. Esse trabalho foi publicado em Jin, Ji e Gu (2016).

Jin, Ji e Gu (2016) fazem um trabalho para identificar informações comparáveis entre produtos para textos opinativos em inglês. O processo se resume em: dados textos avaliativos sobre dois produtos, selecionar pares de sentenças (uma sobre cada produto) de tal forma que as duas sentenças do par sejam representativas e comparáveis. Com isso, espera-se que as sentenças extraídas explorem pontos similares (quer com opiniões opostas ou confluentes) a respeito de características similares dos dois produtos. Por exemplo, para uma câmera, uma característica pode ser a bateria, e um ponto de argumentação relacionado à bateria pode ser o tempo de duração; assim, ao comparar duas câmeras, é de alto valor encontrar duas sentenças (uma

sobre cada uma) que falem ambas sobre a duração da bateria, independentemente de elas concordarem ou discordarem.

Os autores definem maneiras de calcular a representatividade, a comparabilidade e a diversidade de resumos e descrevem a seleção de sentenças como um problema de otimização baseado em uma função de similaridade.

3.4.1 Descrição teórica

Para definir o problema da geração de sumário, sejam:

e_1 : um produto qualquer;
e_2 : um produto concorrente de e_1;
a : um aspecto comum a e_1 e e_2;
E_1 : um conjunto de sentenças avaliativas de e_1 que falam sobre a;
E_2 : um conjunto de sentenças avaliativas de e_2 que falam sobre a;gt.

Segundo esse método, o resumo ideal deve ser representativo, comparativo e diversificado. Essas medidas são definidas no nível da sentença: sentenças parecidas sobre uma mesma entidade são consideradas representativas umas das outras; sentenças parecidas sobre entidades diferentes são consideradas comparativas entre si; sentenças não parecidas sobre uma mesma entidade são consideradas de alta diversidade.

Para definir quão similares duas sentenças são, Jin, Ji e Gu usam o conceito de tópicos: quanto mais tópicos duas sentenças tiverem em comum, mais similares elas são. A função de similaridade é chamada neste texto de SIM.

Para resolver o problema, seria preciso encontrar um resumo com maiores valores possíveis de representatividade, comparatividade e diversidade. Porém, os pesquisadores informam que isso não é factível pois deveria usar uma solução por força bruta que pode requerer um tempo absurdamente grande de processamento. Para resolver isso, eles propõem três algoritmos onde cada um considera uma só medida:

- **R-First**: considera apenas a representatividade.
- **C-First**: considera apenas a comparabilidade.
- **D-First**: considera apenas a diversidade.

Em cada um dos três algoritmos, é calculado, para cada sentença disponível, o quanto ela favorece a medida considerada pelo algoritmo. Ao fim, são selecionadas para o resumo as melhores sentenças de acordo com o limite de tamanho.

3.4.2 Testes práticos

3.4.2.1 Conjunto de dados

Nos testes, foi usado um conjunto de dados contendo textos avaliativos sobre dois telefones celulares. Os textos foram extraídos da página Amazon. Foram usadas apenas sentenças que falam sobre o aspecto 'bateria' do celular. Totalizaram 154 sentenças sobre um celular e 105 de outro.

3.4.2.2 Execução

Uma vez identificadas as polaridades das sentenças, o método proposto foi aplicado para as quatro partes de sumário que são geradas.

O Quadro 3.4 mostra como exemplo o resultado do método executado com o algoritmo guloso que considera a comparabilidade. No artigo original, o quadro mostra as opiniões em uma lista; aqui, elas foram divididas em duas colunas (uma para cada produto) para facilitar a visualização dos pares.

Quadro 3.4 – Pares de sentenças extraídos pelo algoritmo de comparabilidade para os produtos P_1 e P_2.

Grupo	P_1	P_2
positivo vs. positivo	A bateria de fato dura bastante.	Funciona bem, a bateria dura bastante.
	A duração da bateria é muito boa.	A duração da bateria é bem boa.
negativo vs. negativo	A duração da bateria foi piorando com o tempo.	A duração da bateria não é melhor do que eu esperava.
	A duração da bateria é regular.	Eu sem querer drenei minha bateria.
negativo vs. positivo	A duração da bateria foi piorando com o tempo.	A duração da bateria também é excelente.
	A duração da bateria é regular.	A duração da bateria é boa.
positivo vs. negativo	A duração da bateria também é excelente.	A duração da bateria não é melhor do que eu esperava.
	A bateria dura o suficiente pela quantidade de processamento.	Eu sem querer drenei minha bateria.

Fonte: Jin, Ji e Gu (2016, p. 68) (traduzido do inglês, formato modificado)

3.4.2.3 Avaliação

Seja $S = (P \subset E_1, Q \subset E_2)$ o sumário obtido por algum dos métodos descritos, com $P = \{p_i\}_{i=1}^{K}$ e $Q = \{q_i\}_{i=1}^{K}$. Foram definidas algumas medidas para avaliar a performance dos experimentos sobre o resultado S:

1. **Comparabilidade da informação**: mede o quanto cada par de sentença selecionado trata dos mesmos assuntos.

2. **Representatividade da informação**: mede o quanto as sentenças selecionadas cobrem a informação mencionada na fonte.
3. **Diversidade da informação**: mede o quanto as sentenças selecionadas cobrem diferentes tópicos. Para estimá-la, define-se uma função de similaridade sobre um conjunto de sentenças como sendo a média das similaridades de todos os pares de sentenças contidos no conjunto.

Os autores descobriram que o algoritmo que usa a representatividade acaba selecionado sumários que possuem uma alta comparabilidade de informação, provavelmente porque essa otimização favorece a seleção de sentenças similares (é normal que as sentenças representativas tenham alta similaridade entre si, e sentenças similares são mais facilmente comparáveis).

Foi notado que o algoritmo de diversidade traz um ganho na representatividade, porque quando há sentenças diversificadas sendo selecionadas, é mais provável que haja muitos tópicos abordados nelas, o que favorece o aparecimento de vários tópicos mencionados na fonte.

Uma baixa representatividade de informação foi obtida para todos os três algoritmos, provavelmente porque a quantidade de pares selecionadas (três) impede que muitos tópicos sejam abordados.

Foram feitos experimentos comparando o efeito da quantidade de sentenças selecionadas sobre as três métricas de avaliação. A representatividade e a diversidade foram aumentando conforme o número de sentenças selecionadas aumentava. Já a comparabilidade diminui quando o número de sentenças aumenta muito; de fato, quanto mais sentenças há, mais difícil se torna a tarefa de se formar pares adequados de sentenças comparativas sem deixar nenhuma sentença com um par não tão adequado.

3.4.3 Análise crítica

O trabalho de Jin, Ji e Gu (2016) oferece métodos muito interessantes porque são bem distintos dos outros encontrados na literatura. Diferentemente de Lerman e McDonald (2009) e Kim e Zhai (2009), eles não usam agrupamentos ou distribuição de probabilidade para gerar o conteúdo do resumo. A sistema de ranqueamento que o método usa é bem pouco custoso, o que leva a crer que esse método tenha performance mais rápida do que os demais.

Embora a sumarização contrastiva de opinião tenha sido definida por Jin, Ji e Gu (2016) de maneira diferente da definição adotada neste trabalho, as diferenças são mínimas. Uma diferença é que os autores dizem que os sumários têm que conter opiniões concordantes sobre as entidades. Outra diferença é que os sumários de Jin, Ji e Gu (2016) são divididos em quatro partes, uma para cada combinação de polaridades das duas entidades. Essas diferenças podem ser suprimidas gerando-se, a partir do método de Jin, Ji e Gu, um sumário que contém apenas as duas partes que trazem as diferenças entre as entidades.

O trabalho original mostra resultados que não são muito conclusivos. As avaliações foram feitas em poucos casos de teste e as medidas usadas para avaliar são muito parecidas com as próprias medidas usadas para fazer o ranque, o que pode indicar que elas na verdade avaliam a capacidade do sistema de fazer o ranque, e não a qualidade do sumário. Além disso, o trabalho não compara o método com nenhum outro método anterior.

Por trazer uma estratégia inovadora e por ser o método mais recente encontrado na literatura, esse método será investigado neste projeto.

3.5 Considerações

Os trabalhos mostrados neste capítulo permitem perceber dois tipos diferentes de tarefas que fazem sumarização com foco em divergências de opiniões. Um deles é a sumarização gráfica (Seção 3.1), que se resume a contabilizar opiniões usando diretamente ferramentas de identificação de polaridade e exibir a contagem em gráficos. Outro é a sumarização textual (Seções 3.2, 3.3 e 3.4), que se preocupa em selecionar informações adequadas para fazerem parte do sumário.

A sumarização textual é mais desafiadora, porque além de ser necessário identificar as opiniões, deve-se selecionar as mais importantes para compor o sumário. Essa seleção é feita com várias técnicas diferentes, mas os trabalhos mostrados neste capítulo evidenciam uma mesma estrutura de fluxo: primeiro, define-se uma função heurística que atribua pontuações para sumários candidatos indicando quão bom um certo sumário é; uma vez definida a função, formula-se o problema de encontrar o melhor sumário como um problema de otimização cujo objetivo é encontrar o sumário que tem a melhor pontuação segundo a função; por fim, resolve-se o problema usando-se algoritmos gulosos que entregam uma solução subótima. A parte do problema que mais requer atenção é a formulação de funções para que o problema possa ser escrito como uma otimização. Encontrar uma função que permita distinguir sumários bons de sumários ruins é o núcleo da resolução do problema.

Observa-se que, enquanto a sumarização gráfica provoca uma perda de detalhes, a textual causa perda de informações. A perda de detalhes ocorre porque a sumarização gráfica, embora não desperdice informação (todas as opiniões encontradas são contabilizadas), tem um formato de saída que não permite compreender o que causa as opiniões a serem como são. Talvez por esse motivo os autores do trabalho relatado na Seção 3.1 incluíram no sumário uma listagem de todas as opiniões encontradas. A perda de informação ocorre na sumarização textual, porque, em geral, não é possível que todas as opiniões contidas no conjunto-fonte façam parte do sumário. No momento em que uma opinião não é inclusa no sumário, perde-se informação. Por outro lado, todos os detalhes das opiniões selecionadas são preservados (já que todos os

trabalhos retratados aqui fazem sumarização extrativa, onde as opiniões são copiadas como são do conjunto-fonte para o sumário).

Não se deve negligenciar que a fidelidade de detalhes vista na saída não pode ser um indicador da qualidade do sistema: embora as frases sejam inteiramente copiadas para o resumo, elas não são processadas assim, porque ocorrem várias simplificações na modelagem do conjunto-fonte antes de ele ser processado. Todos os trabalhos deste capítulo usam o nível 2 de detalhe para processar o texto, com exceção do mostrado na Seção 3.1, que usa o nível 1. Isso significa que, durante o processamento, eles não consideram o significado completo da opinião, mas somente quão positiva ou negativa a opinião é e a qual aspecto ela se refere.

A resolução do problema da sumarização depende fortemente de ferramentas auxiliares, pois o preprocessamento da informação é uma etapa inevitável: precisa-se, de alguma forma, converter o texto em uma representação que possa ser processada pelas funções heurísticas definidas. Então, além das simplificações de modelagem, outros fatores podem atrapalhar a tarefa, como ruídos no texto-fonte e falhas das ferramentas de preprocessamento. Muitas vezes esses erros não podem ser detectados apenas olhando-se para a saída; é preciso olhar também para o conjunto-fonte para saber se a saída o reflete de maneira satisfatória. No caso da sumarização contrastiva, olhar somente o sumário é o bastante para saber se ele permite comparar bem duas entidades e se ele cobre assuntos diversos. No entanto, o conjunto-fonte deve ser considerado para saber se o sumário é representativo. Lerman e McDonald (2009) seguem esses passos em seus experimentos, descritos na Seção 3.2.

Pelos trabalhos mostrados neste capítulo, pode-se perceber que a avaliação dos métodos ainda é feita de maneira muito diferenciada por cada autor: alguns avaliam com medidas automáticas (JIN; JI; GU, 2016), outros por meio de pesquisa de opinião (LERMAN; MCDONALD, 2009), e outros comparam o resultado obtido com aquele produzido por humanos (LIU; HU; CHENG, 2005). Não só a forma de avaliação é distinta como também os conjuntos usados para teste são completamente diferentes. Por esse motivo, não é possível saber como esses trabalhos se comparam uns com os outros quanto aos resultados que produzem.

Neste capítulo, foram elencados os três métodos que serão estudados com profundidade neste texto. Além desses, foi mostrado um outro para fins ilustrativos. Os três trabalhos escolhidos têm em comum o fato de permitirem selecionar, dentro de um conjunto de textos opinativos, as sentenças mais relevantes que permitem contrastar pontos de vista. Apesar disso, as definições de problema dos três trabalhos são diferentes entre si: apenas Lerman e McDonald (2009) definem o problema da maneira como ele é estudado neste projeto. Por esse motivo, os outros dois trabalhos serão adaptados para que todos tenham o mesmo formato. Os métodos usados nos três trabalhos são bem diferentes: um usa aproximação estatística, um usa agrupamento e outro usa ranqueamento com medidas de similaridade; um considera os aspectos das opiniões, outro considera todo item léxico da sentença e outro separa as opiniões por aspecto antes da

sumarização. Essas diferenças são interessantes pois indicam que os métodos escolhidos têm estratégias diversificadas.

Este trabalho permitirá comparar os métodos e entender os principais pontos fortes e fracos de cada um. O próximo capítulo descreve detalhadamente os métodos investigados. Além dos três métodos colhidos da literatura, é proposto um método novo, apresentado também no próximo capítulo.

CAPÍTULO

4

MÉTODOS

Este capítulo descreve detalhadamente os quatro métodos estudados neste trabalho. A Seção 4.1 descreve o método de Lerman e McDonald (2009), a Seção 4.2 descreve o método de Kim e Zhai (2009), a Seção 4.3 descreve o método de Jin, Ji e Gu (2016) e a Seção 4.4 descreve o método original. Os testes e resultados são apresentados no Capítulo 5.

4.1 Método 1: probabilidade

Esta seção descreve o método de sumarização contrastiva de opinião apresentado por Lerman e McDonald (2009), que é uma adaptação do algoritmo de sumarização de opinião publicado por Lerman, Blair-Goldensohn e McDonald (2009). A abordagem baseia-se em comparações estatísticas entre os resumos e os conjuntos-fonte, que ocorre após uma etapa de identificação automática de opiniões nos textos. A Seção 4.1.1 descreve o método usado por Lerman, Blair-Goldensohn e McDonald (2009) para desenvolver um sumarizador não contrastivo. A Seção 4.1.2 descreve o método de Lerman, Blair-Goldensohn e McDonald (2009) que adapata o método anterior para que ele gere sumários contrastivos.

4.1.1 Ranqueamento de sumários opinativos por distribuição de polaridade

A sumarização de opinião é vista por Lerman, Blair-Goldensohn e McDonald (2009) como a escolha de um sumário R a partir de várias opções possíveis para R. Isso pode ser feito com a resolução da equação

$$R = \arg\max_{|R| \leq K} L(R) \tag{4.1}$$

sendo L uma função que atribui uma pontuação ao resumo R e K é o limite de tamanho do resumo. A pontuação é maior quanto maior for a quantidade de informações desejáveis que o sumário contém baseado em alguma heurística que estime isso. Então, o problema pode ser resolvido com técnicas de otimização. Deve-se, portanto, definir uma função L que estime quão bom um sumário candidato é. A definição dessa função decorre a seguir, onde se fazem algumas definições preliminares.

É definida uma função de **polaridade** que mapeia cada item léxico w para um valor real:

$$\text{POLARIDADE}(w) \in [-1,1].$$

Essa função leva itens com polaridade positiva a valores altos e itens com polaridade negativa a valores baixos.

A **intensidade** de uma sentença s mede quão subjetiva essa sentença é, independentemente de polaridade. Ela é definida como a soma das polaridades (em valor absoluto) de todos os termos da sentença:

$$\text{INTENSIDADE}(s) = \sum_{w \in s} |\text{POLARIDADE}(w)|. \tag{4.2}$$

Experimentos iniciais mostraram que os sumários gerados frequentemente continham sentenças com baixa intensidade (o que é indesejável, pois sentenças com alta intensidade são mais fortemente comparáveis). Então Lerman, Blair-Goldensohn e McDonald (2009) decidiram remover as sentenças com baixa intensidade já no preprocessamento.

Com base nas duas funções já definidas, é definida uma função de **sentimento normalizado** sobre a sentença s, que mede a proporção entre sentimento léxico e intensidade contidos na sentença:

$$\text{SENT}(s) = \frac{\sum_{w \in s} \text{POLARIDADE}(w)}{\alpha + \text{INTENSIDADE}(s)}. \tag{4.3}$$

A constante α na equação serve para que se obtenha pontuações maiores (em valor absoluto) para sentenças que contêm sentimentos fortes do que para sentenças que têm vários sentimentos com a mesma polaridade.

Exemplo: Se $\text{POLARIDADE}(\text{bom}) = 0{,}5$ e $\text{POLARIDADE}(\text{aparelho}) = -0{,}06$, tem-se para a sentença 'bom aparelho':

$$\text{SENT}(\text{bom aparelho}) = \frac{0{,}5 - 0{,}06}{0{,}2 + (0{,}5 + 0{,}06)} = 0{,}6 \tag{4.4}$$

tendo sido usado $\alpha = 0{,}2$.

Em Lerman, Blair-Goldensohn e McDonald (2009), são testados três modelos diferentes, mas somente um deles é usado em Lerman e McDonald (2009) para fazer o sumarizador contrastivo, então apenas esse será descrito a seguir. O modelo proposto usa uma abordagem probabilística para encontrar o resumo onde cada aspecto aparece com sentimentos que reflitam os sentimentos dele no conjunto-fonte.

Sejam $\boldsymbol{A}$ o conjunto de todos os termos que representam aspectos encontrados no conjunto-fonte e $\boldsymbol{E}$ o multiconjunto[1] de todas as sentenças do conjunto fonte. O algoritmo para encontrar o melhor resumo por meio de modelos de probabilidade tem como primeiro passo a geração de uma lista que armazena os sentimentos relacionados a cada aspecto do conjunto-fonte, como descrito a seguir:

- Para cada aspecto $a \in A$, inicializa-se um multiconjunto $S_E^a = \{\}$.
- Para cada sentença $s \in E$:
  - Para cada termo $w \in s$, se $w \in A$, adiciona-se $\mathrm{SENT}(s)$ a S_E^w.

Isso gera multiconjuntos S_E^a, $\forall a \in A$, onde cada multiconjunto S_E^a contém os valores de $\mathrm{SENT}(s)$ para todas as sentenças de E onde o aspecto a ocorre.

Exemplo 2: Considere um conjunto E formado pelas seguintes sentenças:

- A tela é boa.
- A tela é ruim.
- O foco é bom.
- O foco é bom e a tela é excelente.

O conjunto A guarda a lista de aspectos contidos em E:

$$A = \{\text{tela}, \text{foco}\}$$

Suponha que a pontuação SENT das sentenças positivas seja sempre $+1$ e das negativas seja -1. Os multiconjuntos formados para cada item de A obtidos do conjunto E são:

- $S_E^{\text{tela}} = \{+1, -1, +1\}$
- $S_E^{\text{foco}} = \{+1, +1\}$

Para saber a pontuação de um resumo candidato R, criam-se, de forma análoga aos conjuntos S_E^a, multiconjuntos para os aspectos que ocorrem no resumo:

- Para cada aspecto $a \in A$, inicializa-se um multiconjunto $S_R^a = \{\}$.
- Para cada sentença $s \in R$:
  - Para cada termo $w \in s$, se $w \in A$, adiciona-se $\mathrm{SENT}(s)$ a S_R^w.

[1] Usam-se multiconjuntos (em vez de conjuntos simples) porque podem haver elementos repetidos.

Tem-se então multiconjuntos S_E^a que armazenam os valores SENT das ocorrências de cada aspecto $a \in A$ no conjunto-fonte E e multiconjuntos S_R^a que armazenam os valores SENT das ocorrências de cada aspecto $a \in A$ no resumo R.

A média e o desvio padrão de cada multiconjunto S_E^a são usadas para definir a distribuição normal da medida SENT para cada aspecto a do conjunto-fonte E. Da mesma forma, calcula-se a distribuição normal de cada conjunto S_R^a para o resumo R.

Para cada aspecto a contido no resumo R, compara-se a distribuição do valor SENT calculado com S_E^a (informações do conjunto-fonte) e aquela calculada com S_R^a (informações do resumo). Quanto mais próximas as distribuições forem, melhor é considerado o resumo, porque isso indica que o resumo recuperou as informações que refletem as ocorrências do conjunto-fonte segundo as distribuições de probabilidade.

Para que aspectos mais frequentes sejam favorecidos, cada distribuição normal é ainda multiplicada pela frequência relativa da ocorrência do respectivo aspecto no conjunto fonte.

A **distribuição de sentimento** de um aspecto a dentro de um conjunto C será denotada como

$$D_C^a = f_C^a \cdot \mathcal{N}(S_C^a)$$

onde $\mathcal{N}(X)$ é a distribuição normal obtida pelos valores de um multiconjunto X e f_C^x é a frequência relativa de x em C.

⌜Exemplo: Considere o caso descrito no Exemplo 2. Como há 3 ocorrências do aspecto 'tela' de um total de 5 opiniões no conjunto, tem-se $f_C^{\text{tela}} = 3/5 = 0{,}6$.

Com o multiconjunto $S_C^{\text{tela}} = \{+1, -1\}$, obtém-se média igual a 0 e desvio padrão aproximadamente igual a 1,4 para o aspecto 'tela'.

A distribuição de sentimento para o aspecto 'tela' pode ser obtida com

$$D_C^{\text{tela}} = 0{,}6 \cdot \mathcal{N}(0; 1{,}4^2), \tag{4.5}$$

onde $\mathcal{N}(\mu; \sigma^2)$ é a distribuição normal com média μ e desvio padrão σ. ⌟

A pontuação de um resumo candidato é calculada com o uso da entropia relativa. A distribuição de cada aspecto do resumo R é comparada com a distribuição de cada aspecto do conjunto-fonte E; quanto maior a divergência, menor fica a pontuação daquele resumo. Isso é representado como

$$L(R) = -\sum_{a \in A} \text{KL}(D_R^a, D_E^a), \tag{4.6}$$

onde $L(R)$ é a função que pontua o resumo R, $\text{KL}(d_1, d_2)$ é a entropia relativa calculada sobre as distribuições d_1 e d_2.

A entropia relativa é definida (KULLBACK; LEIBLER, 1951) como

$$\mathrm{KL}(d_1, d_2) = \sum \log \left(\frac{d_1(x)}{d_2(x)} \right) \cdot d_1(x).$$

O método de Lerman, Blair-Goldensohn e McDonald (2009) consiste então em resolver a Equação 4.1 usando a função L como na Equação 4.6 para encontrar um sumário R.

Para abreviar a notação, a similaridade relativa entre dois conjuntos C_1 e C_2 de opiniões será denotada como $\mathscr{S}(C_1, C_2)$ e definida como o oposto da somatória da entropia relativa dos aspectos dos conjuntos, como na Equação 4.6:

$$\mathscr{S}(C_1, C_2) = -\sum_{a \in A} \mathrm{KL}(D^a_{C_1}, D^a_{C_2}).$$

Agora, a função $L(R)$ pode passar a ser

$$L(R) = \mathscr{S}(E, R)$$

onde E é o conjunto-fonte e $R \subset E$ é um resumo candidato de E.

4.1.2 Adaptação para sumarização contrastiva

O modelo descrito na seção anterior construiu uma função L que atribui pontuações a resumos com base em sua comparação com seu respectivo conjunto-fonte. Com essa função, pode-se agora partir para a resolução do problema da sumarização contrastiva. Essa adaptação foi feita por Lerman e McDonald (2009).

Sejam:

e_1 : uma entidade;
e_2 : uma outra entidade (a ser comparada com e_1);
E_1 : um conjunto de sentenças avaliativas sobre e_1;
E_2 : um conjunto de sentenças avaliativas sobre e_2;
R_1 : um subconjunto de E_1;
R_2 : um subconjunto de E_2;

Um **sumarizador contrastivo** (como definido em Lerman e McDonald (2009)), deve selecionar dois subconjuntos: R_1 de E_1 e R_2 de E_2. Esses subconjuntos devem incluir sentenças que enfatizam as diferenças nas opiniões sobre as entidades. Um resumo contrastivo R das entidades e_1 e e_2 é visto então como o par

$$R = (R_1, R_2).$$

Com o sumarizador de opinião descrito na Seção 4.1.2, pode-se produzir um sumário contrastivo modificando-se a função de pontuação da Equação 4.6 de algumas maneiras diferentes a fim de se maximizar ou minimizar a similaridade entre os conjuntos-fonte e os resumos. Algumas estratégias possíveis (esquematizadas na Figura 4.1) são:

1. Gerar os dois sumários de forma independente, sem modificar a função de pontuação (Figura 4.1a):

$$L(R_1) = \mathscr{S}(E_1, R_1)$$
$$L(R_2) = \mathscr{S}(E_2, R_2)$$

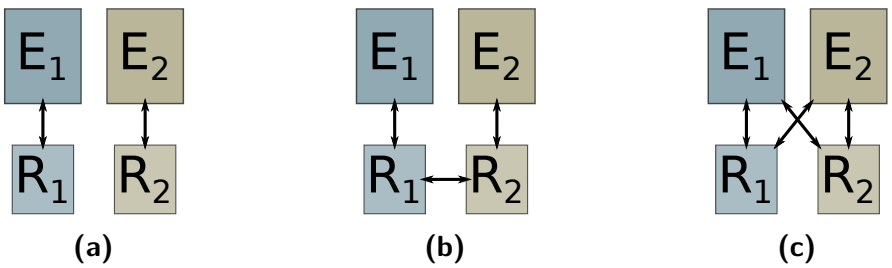
   Esse procedimento não garante que os dois sumários terão boa comparabilidade entre si, pois cada sumário não tem conhecimento do conteúdo do outro.

2. Modificar a pontuação L de forma que ela aumente a pontuação de pares de sumários se eles forem divergentes (Figura 4.1b). Por exemplo, ela poderia passar a ser

$$L(R) = \mathscr{S}(E_1, R_1) + \mathscr{S}(E_2, R_2) - \mathscr{S}(R_1, R_2)$$

   Assim, pares candidatos que forem muito diferentes serão favorecidos, pois a terceira parcela é o oposto da similaridade entre os dois lados do sumário. Esta estratégia tem um problema, exemplificado na Figura 4.2: suponha que dois aspectos a_1 e a_2 apareçam tanto em E_1 quanto em E_2 com a mesma frequência e polaridades diversificadas; suponha ainda que exista outro aspecto, a_3, que apareça com polaridade positiva em E_1 e negativa em E_2. Se o limite de tamanho do sumário for igual a uma sentença por entidade, o sumário com maior pontuação pode consistir em uma sentença de E_1 falando sobre a_1 e uma de E_2 falando sobre a_2, talvez ambas com a mesma polaridade. De fato, os dois sumários são altamente distintos (como a estratégia exige), porque os aspectos (no caso, somente um) mencionados em um não aparecem no outro, o que faz a entropia relativa desses sumários ficar muito grande. Entretanto, justamente por serem muito divergentes, essas sentenças não são comparáveis, pois falam sobre

Figura 4.1 – Três estratégias sugeridas por Lerman e McDonald (2009). Uma seta entre dois conjuntos indica que há comparação direta (isto é, cálculo da entropia relativa) entre eles.



Fonte: Original, baseada em Lerman e McDonald (2009, p. 115)

aspectos diferentes, e o pior é que havia uma excelente opção de sumário que foi desperdiçada: duas sentenças sobre a_3, uma de E_1 com alta polaridade e uma de E_2 com baixa polaridade. Isso ocorre porque o fato de um resumo ser divergente do outro resumo não garante que ele será comparável com a outra entidade. Se fosse exigido que o resumo R_1 divergisse de E_2 (em vez de R_2), a entropia relativa de a_3 teria sido maior do que a de a_1, porque a única ocorrência de a_3 em E_2 tem polaridade oposta da ocorrência de a_3 em E_1. A próxima estratégia faz isso para tentar favorecer soluções que escolham aspectos análogos ao a_3 em situações análogas a esse exemplo.

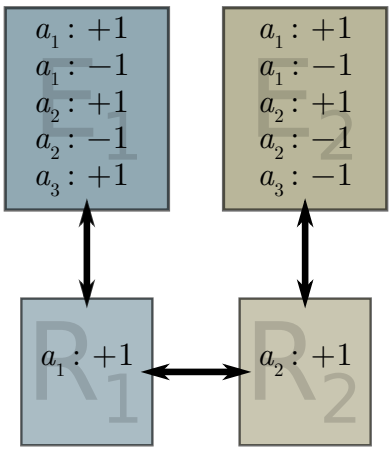
3. Modificar a pontuação L de forma que ela aumente a pontuação de cada sumário se ele for divergente do conjunto-fonte relativo à entidade oposta (Figura 4.1c). Ela passa a ser calculada com

$$\mathscr{L}(R) = \mathscr{S}(E_1, R_1) + \mathscr{S}(E_2, R_2) - \mathscr{S}(E_1, R_2) - \mathscr{S}(E_2, R_1)$$

Esta versão corrige o sumário indesejado exemplificado no item anterior; agora, aquele sumário tem uma pontuação muito baixa porque ambos os lados do sumário têm uma divergência da sua própria entidade igual à divergência do conjunto fonte da entidade oposta; como se deseja minimizar a divergência entre o sumário e a entidade correspondente e maximizar a divergência entre o sumário e a entidade oposta, a pontuação dele ficaria próxima a zero. Note que nesta configuração não se mede diretamente a divergência entre os dois lados do sumário.

Esses três itens serão referidos como Estratégias 1, 2 e 3.

Figura 4.2 – Solução indesejada causada pela Estratégia 2. Os números indicam as polaridades dos aspectos a_1, a_2 e a_3 em uma escala de -1 a $+1$; o par $a : b$ é uma ocorrência do aspecto a em uma sentença com polaridade b.



Fonte: Original

4.1.3 Decisões de projeto

Esta seção descreve como os métodos descritos na Seção 4.1.1 e na Seção 4.1.2 foram implementados quanto a especificações que não constam no trabalho que os originou.

4.1.3.1 Granularidade

No trabalho de Lerman, Blair-Goldensohn e McDonald (2009), o valor SENT é calculado considerando-se a polaridade de todas as palavras da sentença. No presente trabalho, como o conjunto de dados usado identifica as opiniões das sentenças por suas polaridades e aspectos, será usado o próprio valor de polaridade de cada opinião da sentença para calcular o valor SENT.

Exemplo: Considere a sentença $s =$ *'processador rápido mas esquenta demais e tem pouca memória'*. Se as opiniões identificadas nela são $(\text{processador}, +1)$, $(\text{processador}, -1)$ e $(\text{memória}, -1)$ (onde (a, p) é uma opinião que tem aspecto a e polaridade p), então tem-se

$$\text{SENT}(s) = \frac{+1-1-1}{0{,}2+(1+1+1)} = -0{,}3 \tag{4.7}$$

onde foi usado $\alpha = 0{,}2$.

4.1.3.2 Escolha de parâmetros

Constante α Foram feitos testes com vários valores entre 0 e 2. Observou-se que o valor $\alpha = 0{,}2$ fornecia os melhores resultados (embora sem diferença significativa), então esse será o valor usado.

Intensidade baixa Lerman, Blair-Goldensohn e McDonald (2009) afirmam que tiveram que remover sentenças de baixa intensidade, mas não relatam o limiar usado. Nesta implementação, foram descartadas apenas as sentenças com intensidade igual a 0 (isto é, sentenças que não contêm opinião).

4.1.3.3 Funções estatísticas

Distribuições de probabilidade As distribuições de probabilidade $\varphi : X \to Y$ (que matematicamente têm domínio infinito $X = [-\infty, +\infty]$) foram representadas com domínio X no intervalo $[-1{,}5, +1{,}5]$ e discretizadas com passos de $0{,}1$ ($X = \{x \in [-1{,}5, +1{,}5], x = 0{,}1k, k \in \mathbb{Z}\}$). Foi observado que usar intervalos maiores geraria ruído nos cálculos, pois as informações relevantes das distribuições de probabilidade tendem a ficar próximas do intervalo $[-1, +1]$ que contém os possíveis valores da função SENT. Também foram testados passos mais finos de discretização,

mas não se observou vantagem. Ao final, para cada valor da distribuição de probabilidade foi somado um valor infinitesimal para evitar divisões por zero no cálculo da entropia.

Desvio padrão Estabeleceu-se que o desvio padrão mínimo de uma distribuição deveria ser igual a $0{,}2$; caso o desvio fosse menor que esse valor, ele era forçado a valer $0{,}2$. Isso foi feito porque em casos de desvios padrões muito baixos, a entropia relativa tendia a ficar sempre grande demais[2], e o algoritmo acabava rejeitando soluções parciais que teriam potencial para se tornar uma boa solução final. De fato, fazer essa modificação melhorou consideravelmente os resultados.

Essas duas concepções da parte estatística (limitar o intervalo das distribuições e estipular um desvio padrão mínimo) foram uma das partes mais complicadas da implementação. Elas não são óbvias e não estão descritas no trabalho original, e sem elas os restultados ficavam insatisfatórios.

Divergência entre distribuições de probabilidade Além do uso da entropia relativa ($\mathrm{KL}(d_1, d_2) = \sum \log\left(\frac{d_1(x)}{d_2(x)}\right) \cdot d_1(x)$, como usada pelos autores originais) para calcular a divergência entre duas distribuições de probabilidade, foram testadas outras medidas:

- Distância de Hellinger: $\mathrm{H}(d_1, d_2) = \frac{1}{\sqrt{2}} \left(\sum \left(\sqrt{d_1(x)} - \sqrt{d_2(x)} \right)^2 \right)^{\frac{1}{2}}$.
- Soma das diferenças entre as funções: $\mathrm{D}(d_1, d_2) = \sum |d_1(x) - d_2(x)|$.
- Diferença entre as médias das funções: $\mathrm{D}(d_1, d_2) = |\mu(d_1) - \mu(d_2)|$, onde $\mu(d)$ é a média da distribuição normal d.

Algumas medidas se saíram bem melhor do que a entropia relativa em alguns casos de teste, porém a entropia relativa se mostrou mais estável no sentido de não fornecer piores casos muito ruins. A entropia relativa não é comutativa, então foi testada com as duas diferentes combinações de parâmetros: $\mathrm{KL}(D_R^a, D_E^a)$ e $\mathrm{KL}(D_E^a, D_R^a)$ (D_R^a é a distribuição do aspecto a no resumo e D_E^a é sua distribuição no conjunto-fonte correspondente). Alguns casos de teste se saíram melhor com uma das formas, e outros com outra, mas a diferença só foi significativa em poucos casos. Optou-se por usar a primeira forma em todos os testes reportados neste texto.

Falta de dados No caso em que algum aspecto não ocorria em um sumário candidato, a distribuição de probabilidade dele era considerada como a função constante igual a 0.

[2] Um desvio padrão próximo a zero fornece uma distribuição normal com um pico muito acentuado em um ponto do intervalo e valores próximos a zero em outras partes do domínio, o que a faz ser mais próxima da distribuição constante igual a zero (obtida quando nenhuma opinião daquele aspecto é selecionada) do que de outras distribuições que tenham desvio padrão pequeno e média quase igual à dela.

4.1.3.4 Otimização

A otimização foi feita usando-se uma estratégia gulosa: a primeira sentença do resumo (ou o primeiro par, no caso do método contrastivo) é escolhida testando-se todos os possíveis resumos de tamanho 1 e escolhendo-se o melhor (de acordo com a função objetivo); uma vez encontrado o melhor resumo de tamanho n, procura-se o melhor resumo de tamanho $n+1$ testando-se todas as possibilidades de inserção de uma sentença (ou par de sentenças) ao melhor resumo de tamanho n.

Para a otimização do sumário contrastivo, foi testada uma outra estratégia que permite à otimização inserir uma sentença de apenas um dos conjuntos (em vez de ser obrigada a inserir sempre um par com uma sentença de cada conjunto) caso isso gere uma pontuação melhor (podendo até gerar sumários onde os lados têm quantidades diferentes de sentenças). Porém, não se observou diferença significativa, e a estratégia foi preterida.

Cada elemento candidato só é inserido durante a otimização se houver espaço para ele no sumário de acordo com um limite de palavras predefinido. O algoritmo termina quando não puderem ser inseridos mais itens sem desrespeitar o limite.

4.2 Método 2: agrupamento

Esta seção trata de um estudo sobre sumarização contrastiva de opinião publicado em Kim e Zhai (2009). O estudo apresenta métodos que permitem a geração de um resumo de textos opinativos que realça opiniões divergentes sobre uma certa entidade. A Seção 4.2.1 resume os métodos desenvolvidos, trazendo todas as informações pertinentes contidas no trabalho original. A Seção 4.2.2 apresenta uma adaptação inédita do trabalho que permite executar a sumarização contrastiva de duas entidades diferentes. A Seção 4.2.3 descreve decisões de projeto (não especificadas no método) que foram tomadas para a implementação dos algoritmos.

4.2.1 Agrupamento de sentenças para formação de sumário

O principal objetivo da tarefa proposta em Kim e Zhai (2009) é gerar um resumo a partir de textos opinativos (escritos por várias pessoas) sobre alguma entidade de maneira que o resumo permita identificar opiniões divergentes sobre ela. As opiniões aparecem no resumo em pares: cada par é formado por uma opinião positiva e uma negativa; idealmente, ambas as opiniões em um par falam sobre o mesmo assunto de maneira discordante.

A aplicação que Kim e Zhai (2009) usam para ilustrar o problema e testar os métodos é a possibilidade de encontrar opiniões divergentes de compradores a respeito de um certo produto. Como as pessoas têm hábitos de uso diferentes, é normal que existam opiniões contrastantes em

um conjunto de várias avaliações sobre um produto. Encontrar essas opiniões automaticamente e exibi-las de maneira amigável pode ajudar um potencial comprador a entender como aquele produto se sai perante diferentes experiências de uso.

Em Kim e Zhai (2009), a tarefa da sumarização contrastiva de opinião foi formulada como um problema de otimização considerando a representatividade e a contrastividade dos sumários produzidos. O conjunto de entrada é formado por sentenças opinativas previamente identificadas como positivas ou negativas. São propostos dois algoritmos gulosos para resolver de maneira subótima o problema.

Em um dos algoritmos, a ideia básica é fazer um agrupamento de sentenças de modo a identificar opiniões que tratam sobre um mesmo assunto. Se se deseja fazer um sumário de tamanho k (ou seja, com k pares contrastivos), então realiza-se um agrupamento de modo a dividir o conjunto de opiniões positivas em k grupos de forma que sentenças parecidas (ou seja, que falem sobre um mesmo assunto) fiquem dentro de um mesmo grupo; idem para as negativas. Deve-se então escolher uma sentença de cada grupo para formar o sumário; assim, cada grupo estará bem representado nele, de onde se supõe que os principais assuntos terão sido cobertos. Além disso, para formar os pares contrastivos, deve-se descobrir um alinhamento adequado entre os grupos do conjunto positivo e os do conjunto negativo; isto é, para cada grupo do conjunto positivo, deve-se encontrar um grupo do conjunto negativo que mais se assemelhe a ele, de forma que os pares formados com uma opinião de cada um dos grupos falem sobre um mesmo assunto.

No outro algoritmo, o alinhamento é feito entre sentenças (e não entre grupos de sentenças). Os pares de sentenças são formados e são selecionados um a um para obter um sumário que cubra os principais assuntos do conjunto-fonte.

Kim e Zhai (2009) chamam de **par contrastivo** de sentenças a um conjunto de duas sentenças opinativas sobre um mesmo assunto[3] tais que uma tem a polaridade oposta da outra. Eles definem a tarefa como segue.

Sejam $\boldsymbol{X} = \{x_1, ..., x_n\}$ um conjunto de opiniões positivas e $\boldsymbol{Y} = \{y_1, ..., y_m\}$ um conjunto de opiniões negativas sobre um mesmo tópico. A tarefa da **sumarização contrastiva de opinião** é gerar $\boldsymbol{k}$ pares contrastivos para compor um resumo $\boldsymbol{R} = \{(u_i, v_i), u_i \in X, v_i \in Y\}_{i=1}^{k}$, de modo que $U = \{u_i\}_{i=1}^{k} \subset X$ represente bem as opiniões de X e $V = \{v_i\}_{i=1}^{k} \subset Y$ represente bem as opiniões de Y (KIM; ZHAI, 2009).

Para a resolução do problema, são definidas duas funções de similaridade: a similaridade de conteúdo e a similaridade contrastiva.

A função de **similaridade de conteúdo** entre duas sentenças (definida apenas para sentenças de mesma polaridade) mede o quanto essas sentenças tratam sobre tópicos similares. A

[3] No nível de granularidade usado pelos autores, as duas sentenças devem ser sobre um mesmo aspecto, e o assunto de ambas deve ser uma mesma característica desse aspecto.

similaridade de conteúdo entre s_1 e s_2 é denotada[4] por $\mathrm{SIM}(s_1, s_2)$ e seu valor está no intervalo $[0,1]$. Essa função é calculada sobre cada termo das sentenças, e é definida como

$$\mathrm{SIM}(s_1, s_2) = \frac{\sum_{u \in s_1} \max_{v' \in s_2} \omega(u, v') + \sum_{v \in s_2} \max_{u' \in s_1} \omega(u', v)}{|s_1| + |s_2|}. \tag{4.8}$$

A função $\omega(u,v)$ é uma função de similaridade de termos. Os autores experimentam duas variações diferentes para essa função:

1. **Sobreposição de palavras**: $\omega_1(u,v) = 1 \Leftrightarrow u = v$ e $\omega_1(u,v) = 0 \Leftrightarrow u \neq v$.
2. **Similaridade semântica**: $\omega_2(u,v) = 1 \Leftarrow u = v$ e $\omega_2(u,v) = \gamma \mathrm{sim}(u,v) \Leftarrow u \neq v$, com γ sendo um parâmetro à escolha e $\mathrm{sim}(u,v)$ sendo alguma medida de similaridade semântica. Os autores usaram a similaridade semântica da WordNet (PEDERSEN; PATWARDHAN; MICHELIZZI, 2004).

A ideia da Equação 4.8 é: para cada palavra de uma sentença, encontrar a palavra da outra sentença com a qual ela melhor se relaciona e encontrar a pontuação da similaridade entre elas (repetir isso nas duas direções, ou seja, primeiro de s_1 para s_2 e depois de s_2 para s_1); então, somar essa pontuação para todas as palavras de cada sentença; ao fim, normalizar a soma dividindo-a pelo total de palavras nas duas sentenças.

A **similaridade contrastiva**[5] $\mathrm{SIMC}(s_1, s_2)$ mede o quanto duas sentenças s_1 e s_2 de polaridades opostas tratam sobre o mesmo tópico. A ideia é usá-la para identificar sentenças que trazem opiniões divergentes sobre um mesmo assunto. Para isso, removem-se os adjetivos e as palavras de negação das sentenças e aplica-se sobre as sentenças modificadas a função de similaridade exatamente como na Equação 4.8. Removendo-se esses elementos, as sentenças perdem seus elementos opinativos, e isso permite que se compare apenas o assunto discutido em cada uma delas. Como já se sabe previamente que uma das sentenças é negativa e a outra é positiva, se elas falarem sobre um mesmo assunto, pode-se concluir que elas formam um par contrastivo[6].

No nível do sumário, definem-se as funções de representatividade e contrastividade sobre o resumo $R = \{(u_i, v_i)\}_{i=1}^{k}, u_i \in X, v_i \in Y$.

A **representatividade**[7] $\mathrm{REP}(R)$ do resumo R feito a partir dos conjuntos de sentenças positivas X e negativas Y mede quão bem esse resumo reflete as opiniões contidas em X e Y. Seu valor é a média da similaridade de cada sentença dos conjuntos-fonte com a sentença do resumo

[4] A similaridade de conteúdo é denonada por ϕ no texto de Kim e Zhai (2009).
[5] A similaridade contrastiva é denonada por ψ no texto de Kim e Zhai (2009)
[6] Esse trabalho preocupa-se apenas em separar as opiniões positivas e negativas; assim, as sentenças *'a interface é bonita'* e *'a interface é lenta'* podem formar um par contrastivo, mesmo não sendo 'bonita' um antônimo de 'lenta'.
[7] A representatividade é denonada por r no texto de Kim e Zhai (2009).

mais similar a ela:

$$\mathrm{REP}(R) = \frac{1}{|X|}\sum_{x\in X}\max_{i\in[1,k]}\mathrm{SIM}(x,u_i) + \frac{1}{|Y|}\sum_{y\in Y}\max_{i\in[1,k]}\mathrm{SIM}(y,v_i).$$

Assim, se para toda sentença de X houver pelo menos uma sentença em R muito similar a ela, pode-se dizer que o resumo representa bem X, e sua função de representatividade terá um valor alto; o mesmo ocorre para Y.

A **contrastividade**[8] $\mathrm{CONT}(R)$ mede quão bem cada par (u_i, v_i) de R representa um par contrastivo, ou seja, o quanto cada u_i se relaciona com seu v_i correspondente. É definida como a média da similaridade contrastiva dos pares contidos em R:

$$\mathrm{CONT}(R) = \frac{1}{k}\sum_{i=1}^{k}\mathrm{SIMC}(u_i, v_i).$$

Um bom resumo contrastivo de opinião deve ter tanto uma alta representatividade quanto uma alta contrastividade. O problema de encontrar o resumo ótimo R^* pode então ser formulado com a seguinte **função objetivo**:

$$f(R) = \lambda\mathrm{REP}(R) + (1-\lambda)\mathrm{CONT}(R)$$

e resolvido com

$$R^* = \arg\max_R f(R),$$

onde $\boldsymbol{\lambda} \in [0,1]$ é um parâmetro para controlar a importância relativa das duas medidas[9].

São propostos dois algoritmos gulosos para resolver a otimização. Um deles maximiza primeiro a representatividade, definindo quais sentenças entrarão para o sumário, e depois forma pares com essas sentenças de forma que sua contrastividade seja maximizada; isto é, primeiro se descobre como selecionar um conjunto representativo de sentenças, e só então essas sentenças são associadas em pares. O outro maximiza primeiro a contrastividade: ele calcula a similaridade contrastiva entre todos os possíveis pares e seleciona os pares que têm maior pontuação de maneira que a seleção maximize a representatividade.

O algoritmo que prioriza a representatividade é chamado de **R-First**. Para executá-lo, usa-se um algoritmo de agrupamento qualquer[10] para separar as sentenças de cada um dos conjuntos X e Y em k grupos (k é a quantidade de pares que se deseja ter no sumário) de acordo com a similaridade entre as sentenças (conforme a Equação 4.8). Os grupos de X serão identificados

[8] A contrastividade é denonada por c no texto de Kim e Zhai (2009).

[9] Os autores colocam como intervalo aberto, $\lambda \in (0,1)$, mas pode-se usar $\lambda = 1$ ou $\lambda = 0$ para considerar apenas uma das duas métricas.

[10] Kim e Zhai (2009) usam agrupamento hierárquico aglomerativo.

como $\boldsymbol{U}_{i=1}^{k} = U_1, U_2, ..., U_k$, e os de Y como $\boldsymbol{V}_{i=1}^{k} = V_1, V_2, ..., V_k$, com $\bigcup_{i=1}^{k} U_i = X$ e $\bigcup_{i=1}^{k} V_i = Y$ e $\bigcap_{i=1}^{k} U_i \cup V_i = \emptyset$.

Admite-se que a similaridade entre uma sentença qualquer e uma sentença de um outro grupo é sempre menor do que a similaridade dessa sentença com outra do mesmo grupo. Para obter uma boa representatividade, selecionam-se as sentenças que são centroides de cada grupo; de fato, essa seleção maximiza $\mathrm{REP}(R)$. Neste ponto, já foram escolhidas as k sentenças positivas e as k sentenças negativas que vão formar o sumário; deve-se agora encontrar a melhor forma de alinhá-las em pares para que cada par tenha alta contrastividade, a fim de maximizar $\mathrm{CONT}(R)$. Como k é geralmente pequeno, isso pode ser feito por força bruta. Essa estratégia pode ser melhorada, como descrito a seguir.

Ao se selecionar os centroides de cada grupo para maximizar $\mathrm{REP}(R)$, ignorou-se que podem haver outras sentenças (que não as centroides) que beneficiem a função $\mathrm{CONT}(R)$; de fato, embora os centroides representem bem o grupo a que pertencem, nada garante que eles são o melhor par contrastivo possível dos seus respectivos grupos. Assim, é possível que haja uma melhor escolha de sentenças para maximizar a função objetivo. Descreve-se a seguir como identificá-la.

Aproveita-se a escolha de pares contrastivos (feita anteriormente com os centroides) para alinhar os grupos: se um par de centroides $(u_c, v_c) \in U_c \times V_c$ foi formado, então o grupo U_c pode ser pareado com V_c. Afinal, as sentenças dentro de um grupo são todas similares entre si; se uma sentença de um grupo U_c forma um bom par com outra sentença de outro grupo V_c, pode-se admitir que qualquer dupla de $U_c \times V_c$ é um bom par contrastivo.

Mantendo-se o alinhamento de grupos feito a partir dos centroides, sejam $(U_1, ..., U_k)$ e $(V_1, ..., V_k)$ os conjuntos desses grupos de modo que U_i é pareado com V_i. A partir dos k pares $(u_i, v_i) \in U_i \times V_i$ contidos em um sumário candidato R e com as funções de similaridade de conteúdo (SIM) e de contrastividade (SIMC) definidas anteriormente, pode-se fazer uma nova **função objetivo**

$$g(R) = \sum_{i=1}^{k} g_i(u_i, v_i),$$

onde

$$g_i(u_i, v_i) = \lambda \left(\frac{1}{|X|} \sum_{x \in U_i} \mathrm{SIM}(x, u_i) + \frac{1}{|Y|} \sum_{y \in V_i} \mathrm{SIM}(y, v_i) \right) + \frac{(1-\lambda)}{k} \mathrm{SIMC}(u_i, v_i).$$

As duas primeiras parcelas são divididas por $|X|$ e $|Y|$ para se obter a média das pontuações em relação aos conjuntos-fonte (ao se calcular $g(R)$, cada elemento da fonte é computado uma vez na somatória). A última parcela é dividida por k pois esta computa apenas os pares que estão no sumário candidato.

A função $g_i(u_i, v_i)$, se maximizada, escolhe o melhor u_i de U_i e o melhor v_i de V_i que favoreçam a representatividade e contrastividade. A solução ótima $R^* = \{(u_i^*, v_i^*)\}_{i=1}^k$ para se encontrar o melhor par $(u_i, v_i) \in U_i \times V_i$ para cada par de grupos (U_i, V_i) é então dada pela otimização independente de cada par:

$$(u_i^*, v_i^*) = \arg\max_{u_i, v_i} g_i(u_i, v_i).$$

Os pares de grupos $\{(U_i, V_i)\}_{i=1}^k$ já estão fixos, portanto deve-se aplicar a equação acima k vezes, uma para cada par. Por força bruta, deve-se testar cada $(u_i, v_i) \in U_i \times V_i$ de cada par (U_i, V_i). Mas nem todas as combinações precisam ser testadas: os pares que têm contrastividade menor do que o centroide do grupo jamais vão fornecer uma solução melhor, então podem ser ignorados[11]. Pode-se então ordenar os pares (u_i, v_i) de cada par de grupos (U_i, V_i) em ordem decrescente de contrastividade e selecionar apenas os pares que estiverem no topo da ordenação com contrastividade maior do que a do centroide. Ao final, obtém-se o sumário ótimo R^*.

Para o algoritmo que prioriza a contrastividade, chamado de **C-First**, primeiro se computa a similaridade contrastiva $\mathrm{SIMC}(u,v)$ para toda combinação de $u \in X$ e $v \in Y$. Então, ordenam-se esses pares em ordem decrescente de similaridade contrastiva. Para montar o resumo, poder-se-ia simplesmente selecionar as k sentenças mais contrastivas[12]. Porém, isso ao se fazer isso, ignora-se completamente a representatividade. Para melhorar a aproximação, deve-se sacrificar um pouco da contrastividade em prol da representatividade.

O algoritmo guloso proposto pelos autores começa escolhendo aquele par que tem a maior contrastividade e inserindo-o no sumário (que inicialmente é um conjunto vazio). Suponha que em um certo ponto já tenham sido escolhidos $n-1$ pares, formando o resumo R_{n-1}, e deseja-se selecionar o próximo par para formar o resumo R_n de tamanho n. Esse par deve ser selecionado de modo a aumentar a função objetivo do resumo o máximo possível. Seja $(u,v) \in X \times Y$ um par qualquer candidato a entrar para o sumário. O n-ésimo par ótimo (u_n^*, v_n^*) é escolhido pela seguinte equação:

$$(u_n^*, v_n^*) = \arg\max_{u,v} \left(\lambda \mathrm{REP_a}(u,v) + (1-\lambda)\mathrm{CONT_a}(u,v)\right),$$

com a função de **contrastividade agregada**[13] indicando quanto o par candidato contribui para aumentar a contrastividade do sumário:

$$\mathrm{CONT_a}(u,v) = \frac{1}{k}\mathrm{SIMC}(u,v)$$

[11] De fato, os centroides são os elementos que formam o par com melhor representatividade (isso decorre diretamente das definições de centroide e de representatividade). Qualquer outro par formado tem representatividade inferior (ou igual) a ele. Então, se esse outro par, além de ter representatividade menor, ainda tiver a contrastividade menor, é impossível que sua função g_i tenha valor mais alto do que o valor dela para o par de centroides.

[12] Eventualmente esquivando-se de repetições (cada sentença aparece em vários pares, já que se consideram todas as combinações).

[13] Essa função é denotada por c em Kim e Zhai (2009) e não recebe um nome.

e a função de **representatividade agregada**[14] indicando quanto o par candidato contribui para aumentar a representatividade do sumário:

$$\mathrm{REP_a}(u,v) = \frac{1}{|X|}\sum_{x\in X_u}\mathrm{SIM}(x,u) + \frac{1}{|Y|}\sum_{y\in Y_u}\mathrm{SIM}(y,v)$$

onde X_u é o conjunto das sentenças em X que são mais similares a u do que quaisquer outras sentenças que já foram escolhidas para o sumário $R_{n-1} = \{(u_j, v_j)\}_{j=1}^{n-1}$. O conjunto Y_u é análogo a X_u:

$$X_u = \{x \in X \mid \mathrm{SIM}(x,u) > \mathrm{SIM}(x,u_j)\ \forall j \in [1, n-1]\}$$

$$Y_u = \{y \in Y \mid \mathrm{SIM}(y,v) > \mathrm{SIM}(y,v_j)\ \forall j \in [1, n-1]\}.$$

Como informa a descrição acima, ao selecionar um novo par, o algoritmo considera não só a contrastividade dele mas também a representatividade que esse par agrega ao sumário. As sentenças dos conjuntos X e Y que são similares a sentenças já escolhidas são desconsideradas porque elas já estão bem representadas no sumário.

Para evitar testar todas as combinações de pares $(u,v) \in X \times Y$, pode-se definir um limiar de contrastividade a partir do qual o par nem chega a ser candidato.

O algoritmo termina após ter escolhido k pares.

4.2.2 Adaptação para sumarização de duas entidades

As estratégias da seção anterior foram adaptadas no presente estudo para a resolução de um problema ligeiramente diferente: fazer um sumário que permita comparar dois produtos diferentes a partir de textos opinativos. Um sumário assim é útil a um potencial comprador que quer descobrir qual de dois produtos é a melhor opção para ele baseado na experiência de uso de outras pessoas. Também pode ser útil para fabricantes entenderem como seus diferentes produtos se comparam entre si e contra produtos da concorrência de acordo com a opinião popular. Esta seção descreve o método elaborado e a sua implementação.

O método original consiste em gerar um sumário a partir de dois conjuntos X e Y de opiniões sobre uma mesma entidade, cada um contendo opiniões de uma certa polaridade. São descritas duas estratégias, R-First e C-First. Para gerar um sumário que exiba pares contrastivos de opiniões sobre um produto, cada uma dessas estratégias é implementada como um algoritmo que recebe como entrada os conjuntos X e Y de opiniões positivas e negativas sobre esse produto.

[14] Essa função é denotada por r em Kim e Zhai (2009) e não recebe um nome.

Na adaptação, a entrada é formada por quatro conjuntos: X_A: opiniões positivas sobre uma entidade A; Y_A: opiniões negativas sobre uma entidade A; X_B: opiniões positivas sobre uma entidade B; Y_B: opiniões negativas sobre uma entidade B.

Partindo das implementações dos algoritmos R-First e C-First como descritas no trabalho original, a adaptação para a comparação de duas entidades A e B decorre como segue:

1. Escolha um dos algoritmos (R-First ou C-First).
2. Execute o algoritmo usando como entrada os conjuntos X_A e Y_B; isso gerará um sumário contrastivo das opiniões positivas de A com as opiniões negativas de B.
3. Execute o algoritmo usando como entrada os conjuntos X_B e Y_A; isso gerará um sumário contrastivo das opiniões positivas de B com as opiniões negativas de A.

A saída (isto é, o resumo) é formada pela concatenação das saídas obtidas após as duas execuções do algoritmo, uma com cada possível combinação de conjuntos contrastivos das duas entidades (passos 2 e 3). A saída de cada uma dessas duas execuções será chamada de uma **parte** do sumário.

4.2.3 Decisões de projeto

Neste trabalho, partiu-se da implementação feita por Sousa (2018), que replicou o método descrito em Kim e Zhai (2009). Esta seção descreve alguns pontos relacionados à adaptação.

Agrupamento Para fazer o agrupamento de sentenças, foi usado o algoritmo de agrupamento aglomerativo[15]. Foram testados dois critérios de ligação: completo (usa a máxima distância entre todos os elementos de dois conjuntos) e média (usa a distância média dos elementos dos conjuntos). Testes iniciais baseados em avaliação empírica e também nas métricas de avaliação definidas em Kim e Zhai (2009) apontaram que usar a distância máxima é a melhor escolha, e portanto ela será usada em todos os testes.

Escolha do representante do grupo Kim e Zhai (2009) relatam duas estratégias de escolha para o par que melhor representa um grupo (no método R-First): a primeira escolhe sempre o centroide do grupo; a segunda usa força bruta para encontrar um representante melhor. Nesta implementação, foi usada a estratégia da força bruta porque é a que tem capacidade de fornecer melhores resultados.

Limiar de contrastividade No método C-First, pares que tivessem contrastividade igual a zero foram descartados da seleção. Isso reduziu significativamente a quantidade de elementos a serem testados, e, consequentemente, o tempo de execução do algoritmo.

[15] Foi usado o método AgglomerativeClustering da biblioteca sklearn.cluster (scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html)

Similaridade de termos À luz do que foi relatado em Kim e Zhai (2009), foi usada a medida de similaridade de termos mais simples (sobreposição de palavras), já que usar a similaridade semântica não parece trazer benefício algum.

Escolha de parâmetros Foi usado $\lambda = 0{,}5$ em todos os testes, pois experimentos iniciais indicaram que não há motivo para usar valor diferente.

Preprocessamento de sentenças Os algoritmos de sumarização (em todas as suas etapas, como agrupamento e alinhamento) são executados nas sentenças após elas passarem por um preprocessamento que visa eliminar ruídos e características indesejáveis e modelar a sentença de acordo com as necessidades do problema. Os procedimentos feitos foram:

- Normalização: as letras das sentenças foram todas convertidas para caracteres minúsculos e suas pontuações foram removidas. A acentuação foi mantida.
- Palavras-vazias: foram removidas palavras que não contribuem para a interpretação automática da sentença. Foram considerados úteis somente substantivos, advetivos e verbos[16]. Foram testadas outras duas estratégias, que se saíram inferiores: (1) manter todas as palavras e (2) remover apenas as palavras contidas no conjunto 'stopwords' para o idioma português fornecido pelo NLTK[17].
- Lematização: cada palavra foi substituída pelo seu lema[18].
- Palavras de negação: para calcular a similaridade contrastiva, foram consideradas (de maneira insensível a acentuação) as seguintes palavras de negação: 'jamais', 'nada', 'nem', 'nenhum', 'ninguém', 'nunca', 'não', 'ñ', 'tampouco', 'longe'.

Compressão Para escolher o tamanho do sumário, foi usada a sugestão de Kim e Zhai (2009): para fazer a primeira parte do sumário, considerando os conjuntos X_A e Y_B, o número de pares escolhidos é igual a $1 + \lfloor \log_2(|X_A| + |Y_B|) \rfloor$; análogo para a segunda parte. Note que a execução do R-First depende de antemão da escolha do tamanho do resumo, pois é esse tamanho que determina como ocorre o agrupamento de sentenças similares, indicando a quantidade de grupos. Por isso, não é possível, nesse método, limitar o sumário pelo número de palavras ou de caracteres (o algoritmo não pode simplesmente parar quando atingir um limite de palavras porque o número de sentenças a entrarem para o sumário já é predefinido e fixo).

[16] As classes gramaticais foram identificadas com o método nlpnet.POSTagger nilc.icmc.usp.br/nlpnet/intro.html

[17] pythonspot.com/nltk-stop-words

[18] Foi usado o lematizador do NLTK nltk.org/howto/stem.html.

4.3 Método 3: similaridade

Esta seção resume o trabalho de Jin, Ji e Gu (2016), que define medidas de similaridade entre sentenças e a partir delas estimam quanto valor cada sentença pode agregar a um sumário contrastivo.

4.3.1 Ranqueamento de sentenças por similaridade de tópicos

O trabalho de Jin, Ji e Gu (2016) identifica itens comparáveis em textos opinativos sobre produtos. Para isso, são definidas medidas de similaridade que relacionam cada sentença ao conjunto-fonte quanto à representatividade, contrastividade e diversidade.

Para definir o problema da geração de sumário, sejam:

e_1 : um produto qualquer;
e_2 : um produto concorrente de e_1;
a : um aspecto comum a e_1 e e_2;
E_1 : um conjunto de sentenças avaliativas de e_1 que falam sobre a;
E_2 : um conjunto de sentenças avaliativas de e_2 que falam sobre a;
L_1 : a quantidade de sentenças em E_1;
L_2 : a quantidade de sentenças em E_2;
R_1 : um subconjunto ordenado de E_1;
R_2 : um subconjunto ordenado de E_2;
K : a quantidade de sentenças em R_1, igual à quantidade de sentenças em R_2.

Para que R_1 e R_2 formem um sumário contrastivo de e_1 e e_2 sobre o aspecto a, segundo o trabalho, é necessário que as sentenças nesses dois conjuntos sejam:

1. **Representativas**: elas devem indicar as principais informações contidas no texto-fonte. Para isso, a similaridade entre R_1 e E_1 deve ser a maior possível. Como $R_1 \subset E_1$, pode ser mais adequado considerar a similaridade entre R_1 e $E_1 - R_1$. O mesmo ocorre com R_2 e E_2.
2. **Comparativas**: as sentenças nos dois conjuntos devem falar sobre tópicos similares. Para isso, a similaridade entre R_1 e R_2 deve ser a maior possível.
3. **Diversificadas**: os dois subconjuntos devem incluir tópicos variados. Isso pode ser obtido minimizando-se a similaridade entre as sentenças de R_1; neste caso, o valor a ser minimizado é a soma das similaridades entre todos os possíveis pares de sentenças em R_1. O mesmo ocorre com R_2.

O sumário é visto como um conjunto de pares de sentenças $\{(R_1(i), R_2(i))\}_{i=1}^{K}$, onde $R(i)$ é a i-ésima sentença do conjunto R; assim, a i-ésima sentença de R_1 é relacionada à i-ésima sentença de R_2. É desejável que essas duas sentenças sejam altamente contrastivas.

A pergunta que o trabalho quer responder é: dados os conjuntos de sentenças E_1 e E_2 e escolhido um inteiro K, como selecionar os dois subconjuntos $R_1 \subset E_1$ e $R_2 \subset E_2$ de tamanho K de modo que eles satisfaçam aos três itens acima?

Para elaborar essa pergunta como um problema de otimização (e para todas as equações de otimização desta seção), considere dois conjuntos ordenados $X = (x_i \in E_1)_{i=1}^{K}$ e $Y = (y_i \in E_2)_{i=1}^{K}$, assumindo $x_i \neq x_j \forall i \neq j$ e $y_i \neq y_j \forall i \neq j$. Os conjuntos X e Y são candidatos a formar um sumário contrastivo $\{(x_i, y_i)\}_{i=1}^{K}$; o conjunto X é uma possível escolha para R_1 e Y é uma possível escolha para R_2.

Os três requisitos acima podem ser reescritas simplificadamente[19] como:

$$\begin{cases} \begin{cases} R_1 = \arg\max\limits_{X} \text{similaridade}(X, E_1 - X) \\ R_2 = \arg\max\limits_{Y} \text{similaridade}(Y, E_2 - Y) \end{cases} \\ R_1, R_2 = \arg\max\limits_{(X,Y)} \text{similaridade}(X, Y) \\ \begin{cases} R_1 = \arg\min\limits_{X} \text{similaridade}(X) \\ R_2 = \arg\min\limits_{Y} \text{similaridade}(Y) \end{cases} \end{cases}$$

Uma função de similaridade SIM é definida no nível da sentença de tal modo que seu valor seja maior para quanto mais tópicos em comum as sentenças tiverem. Sejam s_1 e s_2 duas sentenças e T_1 e T_2 os tópicos abordados em s_1 e s_2, respectivamente. A similaridade é definida como a quantidade de tópicos que as duas sentenças têm em comum dividida pela quantidade total de tópicos em ambas:

$$\text{SIM}(s_1, s_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \tag{4.9}$$

Os autores também testam uma outra forma de calcular a similaridade, que considera repetições ao contar o total de tópicos nas duas sentenças caso um mesmo tópico esteja em ambas:

$$\text{SIM}(s_1, s_2) = \frac{|T_1 \cap T_2|}{|T_1| + |T_2|} \tag{4.10}$$

Uma solução por força bruta levaria $\binom{L_1}{K}\binom{L_2}{K}K!$ comparações, porque existem $\binom{L_1}{K}K!$ possíveis seleções[20] para X, e para cada uma destas existem $\binom{L_2}{K}K!$ seleções para Y; mas ainda deve-se descontar a permutação dos pares, porque a ordem em que os pares aparecem no sumário não

[19] O problema completo deve ser escrito no nível da sentença, o que é feito nas etapas posteriores.

[20] Lembrando que é um conjunto ordenado, por isso o $K!$ conta as permutações dos elementos selecionados.

é importante. Existem $K!$ permutações de pares em um sumário de tamanho K. Por isso, o número total de sumários candidatos[21] é $\frac{1}{K!} \times \binom{L_1}{K} K! \times \binom{L_2}{K} K! = \binom{L_1}{K}\binom{L_2}{K} K!$. Essa quantidade de comparações não seria computacionalmente viável: para os valores razoáveis $L_1 = L_2 = 100$ e $K = 5$, esse número ultrapassaria 10^{17}.

Para encontrar soluções subótimas, as restrições são relaxadas. São propostos três algoritmos gulosos, onde cada um deles considera apenas uma das restrições.

O primeiro algoritmo, **R-First**, considera apenas a representatividade. Reescrevendo o primeiro requisito no nível da sentença, pode-se encontrar soluções $\tilde{R_1}$ e $\tilde{R_2}$ resolvendo-se

$$\tilde{R_1} = \arg\max_X \left(\sum_{t=1}^{L_1-K} \sum_{k=1}^{K} \mathrm{SIM}(x_k, u_t) \right)$$

$$\tilde{R_2} = \arg\max_Y \left(\sum_{t=1}^{L_2-K} \sum_{k=1}^{K} \mathrm{SIM}(y_k, v_t) \right)$$

onde x_i é a i-ésima sentença de X, y_i é a i-ésima sentença de Y, u_i é a i-ésima sentença de $E_1 - X$ e v_i é a i-ésima sentença de $E_2 - Y$.

Para encontrar $\tilde{R_1}$ com essa equação por força bruta, deve-se analisar todas as $\binom{L_1}{K}$ possibilidades de escolha para X. Para cada uma dessas possibilidades, deve-se ainda calcular a similaridade entre cada uma de suas K sentenças e cada uma das $L_1 - K$ sentenças do conjunto $E_1 - R_1$. Por isso, são feitas $\binom{L_1}{K} \times K \times (L_1 - K)$ comparações. A contagem é análoga para $\tilde{R_2}$. Como a escolha de R_1 e R_2 são independentes, há $\binom{L_1}{K} K(L_1 - K) + \binom{L_2}{K} K(L_2 - K)$ comparações para se fazer[22]; para $L_1 = L_2 = 100$ e $K = 5$, esse número é próximo a 10^{11}.

O segundo algoritmo, **C-First**, leva em conta somente a comparabilidade. Ele consiste em resolver

$$\tilde{R_1}, \tilde{R_2} = \arg\max_{(X,Y)} \sum_{k=1}^{K} \mathrm{SIM}(x_k, y_k).$$

Por força bruta, seria preciso formar todos os pares possíveis (X,Y) e calcular a similaridade entre suas sentenças correspondentes, o que usaria um total de $\binom{L_1}{K}\binom{L_2}{K} K$ comparações. Mas usando-se um algoritmo guloso, pode-se calcular a similaridade de cada uma das K sentenças de E_1 contra cada uma das sentenças em E_2 e simplesmente selecionar as K sentenças com maior similaridade. Isso gasta $L_1 \times L_2$ comparações para computar (excluindo-se o passo final que seleciona as sentenças, que pode ser feito com algum algoritmo de ordenação).

O terceiro algoritmo, **D-First**, considera a diversidade. Para isso, deve-se resolver

$$\tilde{R_1} = \arg\min_X \sum_{i=1}^{K} \sum_{j=1 j \neq i}^{K} \mathrm{SIM}(x_i, x_j)$$

[21] O artigo contabilizou $\binom{L_1}{K}\binom{L_2}{K} K! K!$ combinações, mas não diz como fez a contagem.

[22] O artigo usou um algoritmo guloso (não descrito) que gastou $\binom{L_1}{K} K + \binom{L_2}{K} K$ comparações.

$$\tilde{R_2} = \arg\min_{Y} \sum_{i=1}^{K} \sum_{j=1 j\neq i}^{K} \mathrm{SIM}(y_i, y_j).$$

Pode-se calcular a similaridade de cada uma das sentenças de E_1 contra cada uma das outras sentenças em E_1, o que leva menos de $L_1 \times L_1$ comparações (menos porque cada sentença não precisa ser comparada consigo mesma e nem com as sentenças que já foram previamente comparadas com ela, pois a similaridade é uma operação comutativa). Depois, basta selecionar os K pares de sentenças que têm menor similaridade. Para resolver para E_1 e E_2, o total de comparações não ultrapassa $L_1^2 + L_2^2$.

4.3.2 Decisões de projeto

Alguns detalhes da implementação feita neste trabalho são diferentes da implementação de Jin, Ji e Gu (2016) para atender melhor aos objetivos do presente estudo. Esta seção os descreve.

Tópicos Os tópicos de uma sentença são usados por Jin, Ji e Gu (2016) para definir as opiniões selecionadas para o sumário, mas o trabalho não define formalmente o que é um tópico. Na implementação deste trabalho de mestrado, são considerados como tópicos as opiniões contidas na sentença, rotuladas com aspecto e polaridade. Esses aspectos são identificados diretamente no conjunto de dados; especificações sobre eles estão no Seção 5.1.

Granularidade Em Jin, Ji e Gu (2016), os exemplos mostrados fazem um sumário de apenas um único aspecto da entidade: da entidade 'celular', são selecionadas todas as opiniões sobre bateria e então é feito um sumário sobre a bateria do celular considerando tópicos como 'duração da bateria', 'tempo de carregamento', etc. Neste trabalho, os sumários serão feitos para a entidade toda, considerando todas as opiniões sobre todos os aspectos, pois este é o objetivo do trabalho. Como explicado no parágrafo anterior, os aspectos da entidade serão usados para selecionar as opiniões do sumário, e não os tópicos sobre os aspectos, como no original.

4.4 Método 4: ranqueamento

Esta seção apresenta um método desenvolvido neste trabalho de mestrado para a sumarização contrastiva de opinião. O método é baseado em ranqueamento de opiniões: dados dois conjuntos de opiniões, ele calcula quais opiniões têm maior valor para serem adicionadas ao sumário; depois, ele seleciona sentenças que contêm as opiniões mais valiosas e as insere no sumário.

4.4.1 Definições básicas

Uma **opinião** é definida como uma dupla (a, p) onde a é um **aspecto** (assunto principal da opinião) e p é a polaridade correspondente ao aspecto. A **polaridade** será representada por um número no intervalo $[-1, +1]$ com $p < 0$ se a opinião sobre o aspecto for negativa, $p > 0$ se positiva, e $p = 0$ se neutra. Quando suficiente, a polaridade será indicada apenas pelo sinal: $(a, +)$ em vez de $(a, +1)$.

Define-se a função $\mathrm{OP}(s)$ que extrai as opiniões contidas em uma sentença s. Essa função tem como resultado um conjunto contendo todas as opiniões identificadas em s.

⌜Exemplo 5: Considere as sentenças $s_1 =$ *'a tela é nítida e tem boa resolução mas a bateria não dura'* e $s_2 =$ *'a bateria dura bastante mas demora para carregar, alto-falante ruim'*. Uma possível[23] extração de opinião dessas sentenças é:

$$\mathrm{OP}(s_1) = \{(\text{tela}, +), (\text{bateria}, -)\}$$
$$\mathrm{OP}(s_2) = \{(\text{bateria}, +), (\text{bateria}, -), (\text{áudio}, -)\}$$

⌟

A função OP também é definida sobre um conjunto de sentenças. Seja $S = \{s_i\}_{i=1}^{n}$ um conjunto de sentenças. Define-se $\mathrm{OP}(S)$ como o multiconjunto formado pelas opiniões contidas em cada sentença de s, incluindo opiniões repetidas.

$$\mathrm{OP}(S) = \bigcup_{i=1}^{n} \mathrm{OP}(s_i)$$

⌜Exemplo: Considere as sentenças s_1 e s_2 do Exemplo 5. Se $S = \{s_1, s_2\}$, tem-se:

$$\begin{aligned}\mathrm{OP}(S) &= \mathrm{OP}(s_1) \cup \mathrm{OP}(s_2) \\ &= \{(\text{tela}, +), (\text{bateria}, -), (\text{bateria}, +), (\text{bateria}, -), (\text{áudio}, -)\}\end{aligned}$$

⌟

Sejam:

e_1 : uma entidade;
e_2 : outra entidade, a ser comparada com e_1;
E_1 : um conjunto de sentenças opinativas sobre e_1;
E_2 : um conjunto de sentenças opinativas sobre e_2;
O_1 : o multiconjunto das opiniões contidas em E_1 (definido por $\mathrm{OP}(E_1)$);
O_2 : o multiconjunto das opiniões contidas em E_2 (definido por $\mathrm{OP}(E_2)$);

[23] Podem haver outras soluções porque a identificação de opiniões é por si uma tarefa subjetiva. Além do mais, a definição de quais elementos são aspectos não é única e pode variar em cada situação.

R_1 : um subconjunto de E_1;
R_2 : um subconjunto de E_2.

A tarefa da **sumarização contrastiva de opinião** é encontrar os conjuntos R_1 e R_2 de forma que $R = (R_1, R_2)$ seja um resumo que permita comparar as duas entidades considerando suas principais diferenças. O conjunto R_1 deve conter opiniões relevantes de O_1, e R_2 de O_2. Os resumos R_1 e R_2 serão chamados de **lados** do resumo R.

Um **par contrastivo** é um conjunto de opiniões (o_1, o_2) tal que o_1 e o_2 têm aspectos iguais e polaridades opostas. Isto é, se $o_1 = (a_1, p_1)$ e $o_2 = (a_2, p_2)$, então o_1 e o_2 podem formar um par contrastivo se $a_1 = a_2$ e $p_1 \times p_2 < 0$.

Uma **opinião oposta** ou **divergente** a uma opinião o é uma opinião que tem o mesmo aspecto de o e polaridade oposta à de o; é, portanto, uma opinião que poderia formar um par contrastivo com o.

4.4.2 Seleção de opiniões

Esta seção expõe as ideias usadas para se selecionar as opiniões que se julgam convenientes para constar no sumário.

O algoritmo começa com a identificação de pares contrastivos que podem ser formados a partir de O_1 e O_2, que são os pares contrastivos (o_1, o_2) em que $o_1 \in O_1$ e $o_2 \in O_2$. Será rotulado $\boldsymbol{C}(O_1, O_2)$ o conjunto de todos esses pares.

⌜Exemplo: Sejam

$$O_1 = \{(\mathrm{A}, -), (\mathrm{A}, +), (\mathrm{B}, -), (\mathrm{B}, -), (\mathrm{B}, +), (\mathrm{C}, +), (\mathrm{D}, +)\},$$

$$O_2 = \{(\mathrm{A}, +), (\mathrm{A}, +), (\mathrm{B}, +), (\mathrm{B}, +), (\mathrm{B}, +), (\mathrm{C}, +), (\mathrm{D}, -)\}.$$

Tem-se

$$C(O_1, O_2) = \{((\mathrm{A}, -), (\mathrm{A}, +)), ((\mathrm{B}, -), (\mathrm{B}, +)), ((\mathrm{D}, +), (\mathrm{D}, -))\}.$$

⌟

A ideia é usar o conjunto $C(O_1, O_2)$ para escolher as sentenças que entrarão para o resumo: para cada par contrastivo (o_1, o_2) de C, escolhe-se uma sentença de E_1 que contenha a opinião o_1 e uma de E_2 que contenha a opinião o_2 e inserem-se essas sentenças no resumo. Nota-se que são indicadas apenas quais opiniões devem estar no sumário, porém o sumário não é formado por opiniões, mas sim por sentenças que contêm opiniões; para a mesma opinião, é possível que haja mais de uma escolha de sentença que a contenha.

Um resumo contrastivo ideal conteria todos os pares contrastivos de $C(O_1, O_2)$ (ou seja, $C(\mathrm{OP}(R_1), \mathrm{OP}(R_2) = C(O_1, O_2))$. Mas o limite de tamanho do sumário pode impedir que isso

aconteça. Deve-se então encontrar uma forma para ranquear os itens de C para que as opiniões mais relevantes sejam priorizadas, e as menos relevantes sejam inseridas no sumário apenas se houver espaço sobrando.

4.4.3 Ranqueamento de opiniões

Esta seção discute como decidir quais opiniões são mais importantes para serem inseridas no resumo. Para isso, os pares contrastivos do conjunto C definido na seção anterior devem ser ranqueados, isto é, colocados em ordem de relevância. São consideradas relevantes as opiniões que favorecem a representatividade do resumo. Isso será definido por uma pontuação atribuída a cada opinião. Duas estratégias são propostas: uma que pontua cada par contrastivo de forma una, e uma que separa cada opinião do par e a pontua de maneira independente da outra. Elas serão descritas nas próximas subseções; por enquanto, o método de ranqueamento será definido destituidamente dessa função.

Pontuar uma opinião significa atribuir a ela um valor que expresse o quanto ela faz jus a ser escolhida para um sumário. Será chamada de $\boldsymbol{L}(o)$ a função que atribui pontuação a uma opinião o. Após pontuar todas as opiniões de C, é montada uma fila de prioridade para guiar a construção do sumário, onde as opiniões com maior pontuação aparecem primeiro.

Escrevendo o conjunto de pares contrastivos de maneira genérica como $C = \{(o_1^i, o_2^i)\}_{i=1}^n$, definem-se os conjuntos $\boldsymbol{C_1}$, que contém as opiniões de C que pertencem a O_1, e $\boldsymbol{C_2}$, que contém as opiniões de C que pertencem a O_2:

$$C_1 = \{o_1 \mid \exists (o_1, o_2) \in C\}$$

$$C_2 = \{o_2 \mid \exists (o_1, o_2) \in C\}$$

⌜Exemplo: Seja

$$C(O_1, O_2) = \{((\mathrm{A}, -), (\mathrm{A}, +)), ((\mathrm{B}, -), (\mathrm{B}, +)), ((\mathrm{D}, +), (\mathrm{D}, -))\}.$$

Tem-se:

$$C_1 = \{(\mathrm{A}, -), (\mathrm{B}, -), (\mathrm{D}, +)\}$$

$$C_2 = \{(\mathrm{A}, +), (\mathrm{B}, +), (\mathrm{D}, -)\}$$

⌟

Com a função de pontuação $L(o)$ (a ser definida adiante), obtêm-se as filas de prioridade $\boldsymbol{Q_1^c}$ para a entidade e_1 e $\boldsymbol{Q_2^c}$ para a entidade e_2:

$$Q_1^c = (o_1, o_2, ..., o_n), o_i \in C_1, L(o_i) \leq L(o_{i-1}) \forall i > 1$$

$$Q_2^c = (o_1, o_2, ..., o_n), o_i \in C_2, L(o_i) \leq L(o_{i-1}) \forall i > 1$$

Essa definição simplesmente forma as filas de prioridade como tuplas (uma para cada entidade) onde uma opinião sempre está posicionada à esquerda das opiniões que têm pontuação inferior à dela.

Uma vez formadas as filas de prioridade, o sumário contrastivo é gerado de maneira independente para cada entidade. Primeiro, consulta-se a fila para saber qual opinião tem maior relevância; busca-se no conjunto de dados uma sentença onde essa opinião ocorre e insere-se esse sentença no sumário; coloca-se no final da fila todas as opiniões que estiverem na sentença escolhida (para evitar redundância no sumário, que deve ocorrer somente se sobrar espaço no sumário); repete-se o procedimento, agora com a fila modificada.

Para formar o lado R_1 da entidade e_1 a partir do conjunto de sentenças E_1 e da fila de prioridade Q_1^c, faz-se:

1. Inicializa-se o sumário R_1 como um conjunto vazio.
2. Rotula-se o_1^c o primeiro elemento de Q_1^c.
3. Insere-se no sumário uma sentença s de E_1 que contenha a opinião o_1^c e que caiba no sumário e que ainda não esteja no sumário (se essa sentença existir).
4. Para cada opinião o_i^s contida em $\mathrm{OP}(s)$, se $o_i^s \in Q_1^c$, remove-se o_i^s de Q_1^c e adiciona-se o_i^s como último elemento de Q_1^c.
5. Volta-se ao passo 2 e repete-se o procedimento enquanto uma nova sentença puder ser adicionada.

Para formar o lado da entidade e_2, usa-se algoritmo análogo.

As próximas seções descrevem o cálculo da pontuação das opiniões com duas estratégias: a pontuação conjugada (que pontua opiniões de um par contrastivo de maneira conjunta) e a pontuação independente (que separa as opiniões de um par contrastivo para pontuá-los).

4.4.3.1 Pontuação conjugada

Na pontuação conjugada, opiniões que pertencem a um mesmo par do conjunto C recebem uma mesma pontuação. Isto é, considerando o par $(o_1, o_2) \in C$, o cálculo de $L(o_1)$ e o de $L(o_2)$ serão feitos a partir de uma função $L(o_1, o_2)$ que considera simultaneamente ambos os elementos do par: $L(o_1) = L(o_2) = L(o_1, o_2)$. Esta seção mostrará como calcular $L(o_1, o_2)$.

Sejam c_1 a quantidade de elementos de O_1 iguais a o_1 e c_2 a quantidade de elementos de O_2 iguais a o_2. Sejam f_1 a frequência relativa de elementos iguais a o_1 em O_1 ($f_1 = \frac{c_1}{|O_1|}$) e f_2 a frequência relativa de elementos iguais a o_2 em O_2. Foram testadas as seguintes heurísticas para o cálculo de L:

1. $L(o_1, o_2) = \min(c_1, c_2)$: a quantidade máxima de pares contrastivos iguais a (o_1, o_2) que podem ser formados sem repetir sentenças de E_1 e E_2;
2. $L(o_1, o_2) = c_1 \times c_2$: a quantidade de combinações de sentenças de E_1 e E_2 que podem formar um par contrastivo igual a (o_1, o_2);
3. $L(o_1, o_2) = \frac{1}{2}(f_1 + f_2)$: a média da frequência das opiniões do par em seu respectivo conjunto-fonte;
4. $L(o_1, o_2) = \max(f_1, f_2)$: a frequência da opinião mais frequente do par em seu respectivo conjunto-conte;
5. $L(o_1, o_2) = r$, onde r é um número aleatório: nenhum critério, para que a fila de prioridade seja aleatória.

Testes iniciais mostraram que usar $L(o_1, o_2) = c_1 \times c_2$ ou $L(o_1, o_2) = \min(c_1, c_2)$ fornece melhores resultados, especialmente quando o limite de tamanho do sumário é pequeno (e apenas algumas poucas opiniões do topo da fila de prioridade conseguem caber nele). De maneira global, não se observou diferença entre as duas escolhas (exceto em alguns casos de teste); optou-se então por usar sempre $L(o_1, o_2) = c_1 \times c_2$.

Exemplo: Considere os conjuntos de opiniões

$$O_1 = \{(\mathsf{A}, -), (\mathsf{A}, +), (\mathsf{B}, -), (\mathsf{B}, -), (\mathsf{B}, +), (\mathsf{C}, +), (\mathsf{D}, +)\}$$
$$O_2 = \{(\mathsf{A}, +), (\mathsf{A}, +), (\mathsf{B}, +), (\mathsf{B}, +), (\mathsf{B}, +), (\mathsf{C}, +), (\mathsf{D}, -)\}$$

que levam à formação dos pares contrastivos

$$p_A = ((\mathsf{A}, -), (\mathsf{A}, +))$$
$$p_B = ((\mathsf{B}, -), (\mathsf{B}, +))$$
$$p_D = ((\mathsf{D}, +), (\mathsf{D}, -))$$

de forma que

$$C(O_1, O_2) = \{p_A, p_B, p_D\}.$$

Denotando $c_1(i)$ a quantidade de vezes que o item i ocorre em O_1, e $c_2(i)$ a quantidade de vezes que o item i ocorre em O_2, tem-se, para os itens de C:

$$L(p_A) = L((\mathsf{A}, -), (\mathsf{A}, +)) = c_1((\mathsf{A}, -)) \times c_2((\mathsf{A}, +)) = 1 \times 2 = 2$$
$$L(p_B) = L((\mathsf{B}, -), (\mathsf{B}, +)) = c_1((\mathsf{B}, -)) \times c_2((\mathsf{B}, +)) = 2 \times 3 = 6$$
$$L(p_D) = L((\mathsf{D}, +), (\mathsf{D}, -)) = c_1((\mathsf{D}, +)) \times c_2((\mathsf{D}, -)) = 1 \times 1 = 1$$

e as filas de prioridade ficam

$$Q_1^c = ((\mathsf{B},-),(\mathsf{A},-),(\mathsf{D},+))$$
$$Q_2^c = ((\mathsf{B},+),(\mathsf{A},+),(\mathsf{D},-)).$$

Isso indica que B é o aspecto mais relevante da fonte, e opiniões sobre ele serão as primeiras a serem escolhidas para o resumo. ⌟

4.4.3.2 Pontuação independente

Agora será proposta outra estratégia para pontuar opiniões. Nesta estratégia, ao contrário da proposta na seção anterior, cada opinião o_1 de O_1 recebe uma pontuação independente de seu par contrastivo (o_1, o_2) de C; idem para as opiniões o_2 de O_2.

Considere $o \in O$, onde $O \in \{O_1, O_2\}$. Seja c a quantidade de elementos de O iguais a o. Para o cálculo de L, será usado $L(o) = c$. Em outras palavras, a pontuação de cada opinião será igual à frequência absoluta da ocorrência dessa opinião em seu conjunto-fonte.

⌜Exemplo: Sejam

$$O_1 = \{(\mathsf{A},-),(\mathsf{A},+),(\mathsf{B},-),(\mathsf{B},-),(\mathsf{B},+),(\mathsf{C},+),(\mathsf{D},+)\}$$
$$O_2 = \{(\mathsf{A},+),(\mathsf{A},+),(\mathsf{B},+),(\mathsf{B},+),(\mathsf{B},+),(\mathsf{C},+),(\mathsf{D},-)\}$$

de forma que

$$C(O_1, O_2) = \{((\mathsf{A},-),(\mathsf{A},+)),((\mathsf{B},-),(\mathsf{B},+)),((\mathsf{D},+),(\mathsf{D},-))\}.$$

Denotando $c_1(i)$ a quantidade de vezes que o item i ocorre em O_1, e $c_2(i)$ a quantidade de vezes que o item i ocorre em O_2, tem-se, para os itens de C_1:

$$L((\mathsf{A},-)) = c_1((\mathsf{A},-)) = 1$$
$$L((\mathsf{B},-)) = c_1((\mathsf{B},-)) = 2$$
$$L((\mathsf{D},+)) = c_1((\mathsf{D},+)) = 1$$

e para os itens de C_2:

$$L((\mathsf{A},+)) = c_2((\mathsf{A},+)) = 2$$
$$L((\mathsf{B},+)) = c_2((\mathsf{B},+)) = 3$$
$$L((\mathsf{D},-)) = c_2((\mathsf{D},-)) = 1$$

e as filas de prioridade ficam

$$Q_1^c = ((\mathrm{B},-),(\mathrm{D},+),(\mathrm{A},-))$$
$$Q_2^c = ((\mathrm{B},+),(\mathrm{A},+),(\mathrm{D},-))$$

O empate entre $(\mathrm{D},+)$ e $(\mathrm{A},-)$ foi resolvido de maneira arbitrária. ⌟

4.4.4 Maximização da representatividade

O ranqueamento de pares contrastivos (como estabelecido anteriormente) visa maximizar a representatividade do sumário no momento em que favorece as opiniões mais frequentes. Todavia, essa estratégia parece ainda ser muito voltada à contrastividade. De fato, só são selecionadas para o sumário opiniões que tenham a possibilidade de formar pares contrastivos. Se uma opinião for muito frequente em um dos conjuntos-fonte mas não houver uma opinião oposta a ela no conjunto-fonte da entidade concorrente, ela não tem chance alguma de entrar para o sumário.

Será proposta nesta seção uma estratégia para valorizar mais as opiniões frequentes de cada conjunto-fonte independentemente dos pares contrastivos. Para isso, serão feitas outras filas de prioridade (a serem usadas em conjunto com as filas Q_1^c e Q_2^c já definidas) que consideram somente as ocorrências de cada opinião em seu próprio conjunto-fonte. Essas novas filas serão denotadas por $\boldsymbol{Q_1^r}$ e $\boldsymbol{Q_2^r}$:

$$Q_1^r = (o_1, o_2, ..., o_{n_1}), o_i \in O_1, L(o_i) \leq L(o_{i-1}) \forall i \neq 1$$

$$Q_2^r = (o_1, o_2, ..., o_{n_2}), o_i \in O_2, L(o_i) \leq L(o_{i-1}) \forall i \neq 1$$

Essa definição se assemelha à das filas Q_1^c e Q_2^c, com a distinção que são agora consideradas todas as opiniões de O_1 e O_2, e não somente as que estão em C. A função $L(o)$ de pontuação é definida simplesmente como a frequência de o em seu conjunto-fonte.

Então, para gerar cada lado do sumário (por exemplo, R_1), usam-se as duas filas, Q_1^c e Q_1^r, de maneira alternada: primeiro consulta-se Q_1^c para saber qual elemento essa fila prioriza; adiciona-se ao sumário uma opinião igual a esse elemento; remove-se esse elemento da fila e coloca-se esse elemento no final da fila (para o eventual caso de todos os elementos da fila já estiverem representados no sumário e sobre espaço para repetir elementos); move-se esse elemento para o final também na outra fila, Q_1^r (pois ele já está representado no sumário e não há necessidade de repeti-lo, a menos que sobre espaço); alterna-se a consulta para Q_1^r, escolhendo-se o elemento que essa fila prioriza e repetindo-se o procedimento.

O algoritmo para a geração do sumário R_1 para a entidade e_1 (e para a entidade e_2 é análogo) a partir do conjunto de sentenças E_1 e das filas de prioridade Q_1^c e Q_1^r é:

1. Inicializa-se o sumário R_1 como um conjunto vazio.
2. Rotula-se o_1^c o primeiro elemento de Q_1^c.
3. Insere-se no sumário uma sentença s_c de E_1 que contenha a opinião o_1^c e que caiba no sumário e que ainda não esteja no sumário (se essa sentença existir).
4. Para cada opinião o_i^s de $\mathrm{OP}(s_c)$, se $o_i^s \in Q_1^c$, remove-se o_i^s de Q_1^c e adiciona-se o_i^s ao final de Q_1^c.
5. Rotula-se o_1^r o primeiro elemento de Q_1^r.
6. Insere-se no sumário uma sentença s_r de E_1 que contenha a opinião o_1^r e que caiba no sumário e que ainda não esteja no sumário (se essa sentença existir).
7. Para cada opinião o_i^r de $\mathrm{OP}(s_r)$, se $o_i^r \in Q_1^r$, remove-se o_i^r de Q_1^r e adiciona-se o_i^r ao final de Q_1^r.
8. Volta-se ao passo 2 e repete-se o procedimento enquanto uma nova sentença puder ser adicionada.

4.4.5 Estratégias

Em seções anteriores (4.4.3.1 e 4.4.3.2), foram apresentadas duas formas para se calcular a pontuação usada para ranquear a lista de pares contrastivos possíveis: a conjugada e a independente. Cada uma dessas escolhas concebe uma estratégia de ranqueamento diferente. Para cada uma dessas estratégias, pode-se usar ou não a fila de prioridade adicional para a maximização da representatividade (como enunciado na Seção 4.4.4). As quatro combinações de estratégia serão testadas. As estratégias serão referidas por números como listado na Tabela 4.1.

Tabela 4.1 – Estratégias usadas para o ranqueamento de opiniões.

estratégia	uso de Q_r	cálculo de L
1	não	conjugada
2	não	independente
3	sim	conjugada
4	sim	independente

4.4.6 Aprimoramento

Na tentativa de obter um sumário mais informativo e com menos texto irrelevante, foi testada uma forma de priorizar sentenças de acordo com seu número de palavras. Com a hipótese que sentenças curtas demais são pouco úteis por serem muito genéricas e sentenças longas demais tendem a divagar e conter outros assuntos que não a opinião de interesse, estabeleceu-se que seriam priorizadas as sentenças cujo número de palavras descontando-se palavras vazias fosse o

mais próximo possível de 5. Esse critério foi usado como desempate nos casos em que há mais de uma sentença no conjunto-fonte contendo a opinião de interesse que se deseja adicionar ao sumário. O tamanho ideal de sentenças foi escolhido heuristicamente a fim de se encontrar sentenças curtas o suficiente para aproveitar melhor o espaço limitado do sumário, mas longas o suficiente para serem informativas. O aprimoramento será chamado de **Ranqueamento+**.

CAPÍTULO

5

RESULTADOS

Este capítulo relata as atividades práticas executadas neste trabalho. O objetivo geral dessas atividades é testar e avaliar os métodos estudados. Para isso, foi elaborado um conjunto de dados contendo comentários opinativos sobre produtos, que está descrito na Seção 5.1. Foram definidas métricas para aferir a qualidade dos sumários e cada método foi avaliado de acordo com elas, como reportado na Seção 5.2. A Seção 5.3 exibe alguns dos sumários obtidos para ilustrar o funcionamento de cada método. A Seção 5.4 faz uma análise das atividades desenvolvidas com base nos resultados obtidos.

5.1 Conjunto de dados

Esta seção descreve o conjunto de dados usado neste projeto. Para testar e avaliar os métodos de sumarização contrastiva, são usados conjuntos contendo avaliações de compradores sobre celulares e câmeras. Foram coletadas opiniões sobre quatro produtos diferentes, totalizando 542 avaliações (cerca de 13 mil palavras). Dos conjuntos de dados desses quatro produtos, delimitaram-se subconjuntos para totalizar oito pares de conjuntos de dados, onde um par de conjunto de dados contém textos sobre duas entidades comparáveis (dois produtos do mesmo tipo). Os conjuntos foram construídos visando obter uma diversidade de testes, porquanto eles têm características diferentes uns dos outros. Os conjuntos de dados tiveram suas opiniões manualmente anotadas por meio da identificação de polaridades e aspectos.

5.1.1 Conjuntos de dados usados na literatura

A tarefa da sumarização contrastiva de opinião aparece na literatura geralmente voltada para a comparação entre opiniões de compradores sobre produtos (JIN; JI; GU, 2016; LERMAN; MCDONALD, 2009; KIM; ZHAI, 2009; LIU; HU; CHENG, 2005), embora tambem se encontrem

trabalhos que a executam para outros tipos de texto, como assuntos controversos (GUO et al., 2015). Alguns trabalhos estudam a tarefa de sumarização contrastiva para textos não opinativos e usam outros tipos de texto em seus conjuntos de dados, como artigos jornalísticos sobre assuntos polêmicos (PARK; LEE; SONG, 2011; WANG et al., 2012).

Liu, Hu e Cheng (2005) fizeram um conjunto de dados manualmente etiquetado com comentários em inglês sobre 15 produtos eletrônicos. As opiniões foram coletadas do site Epinions. Os autores não dão outras informações sobre o conjunto de dados.

Lerman e McDonald (2009) coletaram dados de comentários em inglês sobre 56 produtos eletrônicos de 15 tipos (câmeras, computadores, sistemas de GPS, tocadores de MP3, etc) de várias fontes (CNet, Epinions, PriceGrabber, etc). Cada produto tem um mínimo de 4 comentários e a média de comentários por produto é 70.

Kim e Zhai (2009) usaram um conjunto de dados com comentários etiquetados sobre 12 produtos coletados do Amazon. Para testar a generalidade do método, também foi usado um conjunto extra formado por 50 comentários favoráveis e 50 contrários ao uso de aspartame.

A proposta deste estudo inclui a decisão de fazer processamento de textos escritos em língua portuguesa. Por esse motivo, não serão usados conjuntos de dados publicados na literatura que contenham textos em outras línguas.

O conjunto criado por Condori e Pardo (2017) com textos em português mostrou-se adequado para a tarefa de sumarização de opinião. Entretanto, para a sumarização contrastiva de opinião, buscam-se conjuntos maiores (isto é, com mais sentenças opinativas) onde haja maior possibilidade de confrontar conjuntos concorrentes. O conjunto formado por Condori e Pardo traz muitas possibilidades de seleção de sentenças representativas, o que atende à tarefa estudada por eles, porém oferece poucas possibilidades de sentenças contrastivas, de que carece o presente trabalho. Por esse motivo, não será usado o conjunto de dados feito por Condori e Pardo.

5.1.2 Construção do conjunto de dados

Por não ter encontrado trabalhos publicados que apresentam conjuntos de dados em português que atendessem a esta tarefa, optou-se por criar um. Esta seção descreve o conjunto de dados usado neste projeto.

5.1.2.1 Planejamento

Como fizeram outros trabalhos (LIU; HU; CHENG, 2005; LERMAN; MCDONALD, 2009; KIM; ZHAI, 2009; JIN; JI; GU, 2016), decidiu-se avaliar os métodos com textos opinativos sobre produtos eletrônicos. Essa escolha deve-se principalmente à facilidade de coletar esse tipo de

texto (dada sua abundância em algumas páginas da Internet) e também por se acreditar que a identificação de aspectos é uma tarefa mais bem definida do que em outros casos (por exemplo, avaliações de serviços, resenhas de livros e discussões políticas) e já estudada com profundidade (VARGAS; PARDO, 2018).

Para avaliar os métodos de sumarização, seria pensável fazer um conjunto com sumários ideais de cada conjunto-fonte (como fizeram Condori e Pardo (2017) e outros trabalhos que usam a medida ROUGE[1]). Decidiu-se não seguir esta ideia por causa da dificuldade em definir o que é um sumário ideal, como ocorre com toda tarefa não exata. Especialmente para a sumarização contrastiva de opinião, seria difícil fazer manualmente resumos dos conjuntos-fonte e defini-los como padrões a serem seguidos, pois:

- Existem muitas possibilidades a serem analisadas: para conjuntos-fonte E_1 e E_2 de tamanhos n_1 e n_2, se o resumo R tiver tamanho igual a k, existem $\binom{n_1}{k}\binom{n_2}{k}$ possibilidades de escolha de itens de E_1 e E_2 para preencher R; se $n_1 = 50$, $n_2 = 50$ e $k = 5$, o número de possíveis sumários ultrapassa 10^{12}. Isso é um problema pois é difícil para um humano analisar essa grande quantidade de possibilidades, mesmo usando algoritmos gulosos ou a própria intuição para escolher as sentenças que devem formar o sumário.
- Podem haver vários sumários igualmente bons porém muito diferentes entre si, onde seria complicado escolher apenas um como o ideal;
- A escolha de um sumário ideal seria bastante subjetiva: pessoas diferentes teriam opiniões diferentes sobre o que é um sumário ideal. Por exemplo, em Tadano, Shimada e Endo (2010), três participantes receberam 450 sentenças com opiniões sobre um videogame para selecionar 50 para formar um resumo simples (não contrastivo), e a concordância calculada entre os anotadoes (ROUGE-1) foi de 0,48, número baixo o suficiente para provar que a tarefa é difícil.

O conjunto de dados usado é então formado por sentenças opinativas que foram manualmente anotadas quanto às opiniões que expressam (sendo cada opinião identificada por um aspecto e uma polaridade). A avaliação é feita não comparando os resumos obtidos com resumos ideais, mas sim estimando a qualidade dos sumários obtidos de acordo com métricas que foram definidas e são descritas adiante neste texto.

[1] A ROUGE é uma métrica de avaliação que compara sumários produzidos automaticamente com sumários de referência, fazendo a pontuação com base em quantidade de n-gramas que os sumários têm em comum (LIN, 2004).

5.1.2.2 Descrição

Esta seção descreve as principais características do conjunto de dados usado. Detalhes sobre sua construção e características adicionais estão descritos no Apêndice A e em Silva e Pardo (2019).

Os dados coletados são opiniões sobre quatro produtos extraídas do site Buscapé[2]. São dois tipos de produtos: celulares e câmeras. Após a extração e anotação, o conjunto de dados foi estendido por meio da criação de outros subconjuntos de teste que contêm sentenças dos subconjuntos originais.

Os comentários extraídos foram separados em sentenças. A Tabela 5.1 mostra a contagem de sentenças e palavras de cada subconjunto e a quatidade de aspectos diferentes encontrados em cada um. A tabela também mostra os nomes dos subconjuntos referentes a cada entidade.

Tabela 5.1 – Estatísticas do conjunto de dados.

tipo	nome		entidade	aspectos	sentenças	palavras
celular	**D1**	D1a	Motorola Moto G5 Plus	15	269	3767
		D1b	Galaxy S7	14	253	3462
câmera	**D2**	D2a	Canon EOS Rebel T5	13	68	1108
		D2b	Canon PowerShot SX520 HS	15	52	594
celular	**D3**	D3a	*(subconjunto de D1a)*	11	150	1948
		D3b	*(subconjunto de D1b)*	10	109	1508
celular	**D4**	D4a	*(subconjunto de D1a)*	13	43	518
		D4b	*(cópia de D1b)*	14	253	3462
câmera	**D5**	D5a	*(subconjunto de D2a)*	12	39	686
		D5b	*(subconjunto de D2b)*	10	30	277
câmera	**D6**	D6a	*(subconjunto de D2a)*	8	29	422
		D6b	*(subconjunto de D2b)*	11	22	317
câmera	**D7**	D7a	*(subconjunto de D2a)*	4	31	636
		D7b	*(subconjunto de D2b)*	4	25	261
celular	**D8**	D8a	*(subconjunto de D1a)*	12	39	572
		D8b	*(subconjunto de D1b)*	12	32	284

5.1.2.3 Identificação de opiniões

As polaridades são indicadas como:

- **positiva**: se a opinião reflete algo bom, desejável;
- **negativa**: se a opinião reflete algo ruim, indesejável.

Trechos que não contêm opiniões positivas ou negativas não foram usados.

[2] www.buscape.com.br

Os aspectos foram identificados pelos anotadores seguindo sua listagem e definição mostradas no Apêndice A. Opiniões que não se referem a um aspecto específico (por exemplo, *'o produto é bom'*) são ditas **genéricas** e receberam um rótulo especial.

Quadro 5.1 – Exemplos da identificação manual de opiniões.

sentença	opiniões
Muito rápido!	desempenho +
Lindo.	design +
Ótimo celular.	produto +
Não compre.	*genérico* −
Bom mas pode melhorar.	*genérico* +
Boa velocidade na reprodução de vídeos e bom de espaço de memória.	desempenho + armazenamento +
A tela não tem o melhor contraste de cores (eu descobri que prefiro tela super AMOLED), mas a nitidez é imbatível (para ler textos por exemplo)	tela − tela +
Bateria muito boa que aguenta até mais de um dia sem recarregar com minha experiência de uso, display excelente e com cores vívidas, além de um processador muito bom que otimiza demais o uso no dia-a-dia.	bateria + tela + desempenho +

Tabela 5.2 – Estatísticas das opiniões positivas e negativas manualmente identificadas no conjunto de dados.

conjunto	aspectos	positivas	negativas	opiniões
D1a	15	247	90	337
D1b	14	242	89	331
D2a	13	44	11	55
D2b	15	43	8	51
D3a	11	103	55	158
D3b	10	52	61	113
D4a	13	41	11	52
D4b	14	242	89	331
D5a	12	27	10	37
D5b	10	26	3	29
D6a	8	17	1	18
D6b	11	17	5	22
D7a	4	33	6	39
D7b	4	22	4	26
D8a	12	47	10	57
D8b	12	24	14	38

5.2 Avaliação

Esta seção relata os resultados obtidos nos experimentos. Os métodos descritos no Capítulo 4 foram executados sobre o conjunto de dados descrito na Seção 5.1 e avaliados sob as medidas definidas na Seção 5.2.1. Primeiro, foram feitos testes com cada variação de cada método; esses resultados estão reportados na Seção 5.2.2. Depois, os métodos foram comparados entre si a partir da melhor variação de cada um; a Seção 5.2.3 reporta os resultados. Foi feito um experimento para investigar a percepção dos humanos sobre os sumários gerados; ele está

reportado na Seção 5.2.4. A subseção inicial desta seção apresenta as métricas usadas para avaliar.

5.2.1 Métricas

Para avaliar os sumários obtidos, usou-se a etiquetação manual do conjunto de dados. Com essa etiquetação, pode-se comparar as opiniões contidas em um sumário gerado com as contidas no conjunto-fonte de acordo com uma classificação de opiniões tida como ideal pelos colaboradores que a fizeram.

Para aferir melhor a identificação de aspectos específicos, não foram consideradas na avaliação as opiniões marcadas como genéricas pelos colaboradores. As opiniões marcadas pelos anotadores como alheias foram desconsideradas pelos próprios métodos de sumarização, pois elas são indesejáveis no sumário por não contribuírem com informações sobre o produto.

Para a avaliação, foram definidas três medidas: uma que avalia a representatividade, outra que avalia a contrastividade e outra que avalia a diversidade.

O **percentual de representatividade** (indicado nas tabelas pela letra **R**) é a porcentagem de opiniões do conjunto-fonte que estão representadas em seu sumário. Para cada opinião do conjunto fonte, se houver alguma opinião do sumário igual a ela (mesmo aspecto e mesma polaridade), então essa opinião está representada no sumário. Seja R o resumo gerado a partir do conjunto-fonte E. Seja c a quantidade de opiniões de E que estão representadas em R: para cada opinião de E, se houver alguma opinião em R com o mesmo aspecto e mesma polaridade dela, então ela está representada em R. $|E|$ denota a quantidade de opiniões em E. Define-se o percentual de representatividade de R como

$$\Pr(R) = \frac{c}{|E|}$$

O **percentual de contrastividade** (indicado nas tabelas pela letra **C**) considera os possíveis pares contrastivos que podem ser formados a partir dos dois conjuntos-fonte alvos da sumarização. Um par contrastivo (o_1, o_2) contendo as opiniões $o_1 \in E_1$ e $o_2 \in E_2$ pode ser formado a partir dos dois conjuntos-fonte E_1 e E_2 se o aspecto de o_1 for igual ao aspecto de o_2 e a polaridade de o_1 for oposta à de o_2. Seja C o conjunto de todos os possíveis pares contrastivos que podem ser formados a partir de E_1 e E_2 (sem repetições[3]), como o definido na Seção 4.4.3. Para avaliar os resumos R_1 e R_2 (gerados respectivamente a partir de E_1 e E_2), seja c_1 a quantidade de pares $(o_1, o_2) \in C$ tais que $o_1 \in R_1$ e c_2 a quantidade de pares $(o_1, o_2) \in C$ tais que

[3] Seria possível considerar as repetições para valorizar os pares contrastivos mais frequentes, contudo já é papel do percentual de representatividade verificar se as opiniões mais frequentes estão no sumário.

$o_2 \in R_2$. O percentual de contrastividade do sumário contrastivo $R = (R_1, R_2)$ é definido como

$$\mathrm{Pc}(R_1, R_2) = \frac{\frac{1}{2}(c_1 + c_2)}{|C|}$$

O **percentual de diversidade** (indicado nas tabelas pela letra **D**) é a quantidade de opiniões diferentes contidas no sumário em relação à quantidade de opiniões diferentes contidas no conjunto-fonte correspondente. Considere o resumo R formado a partir do conjunto-fonte E. Seja C_R o conjunto de todas as opiniões contidas em R (sem repetições) e C_E o conjunto de todas as opiniões contidas em E (sem repetições). O percentual de diversidade de R é

$$\mathrm{Pd}(R) = \frac{|C_R|}{|C_E|}$$

Toda a avaliação é feita considerando-se apenas o aspecto e a polaridade de cada opinião (de acordo com a etiquetação manual); se duas opiniões têm mesmo aspecto e mesma polaridade, elas são consideradas a mesma opinião independentemente de as sentenças serem diferentes.

Para todos os cálculos das medidas de avaliação, foram ignoradas as opiniões neutras, pois de todo modo elas não são desejáveis no sumário.

Para avaliar a representatividade do sumário contrastivo inteiro, considerando as duas entidades, usa-se a média simples do percentual de representatividade calculado para cada lado do resumo. O mesmo ocorre com o percentual de diversidade.

Exemplo: Suponha que as opiniões contidas nos conjuntos-fonte sejam

$$E_1 = \{(\text{tela}, +), (\text{tela}, +), (\text{tela}, -), (\text{bateria}, +), (\text{design}, +)\},$$
$$E_2 = \{(\text{tela}, +), (\text{tela}, -), (\text{tela}, -), (\text{bateria}, -), (\text{design}, +)\}.$$

O conjunto de todos os possíveis pares contrastivos que podem ser formados a partir de E_1 e E_2 é

$$C = \{((\text{tela}, +), (\text{tela}, -)), ((\text{tela}, -), (\text{tela}, +)), ((\text{bateria}, +), (\text{bateria}, -))\}$$

Se os resumos forem

$$R_1 = \{(\text{tela}, +), (\text{bateria}, +)\},$$
$$R_2 = \{(\text{tela}, -), (\text{design}, +)\},$$

tem-se o seguinte percentual de contrastividade:

$$\mathrm{Pc}(R_1, R_2) = \frac{\frac{1}{2}(c_1 + c_2)}{|C|} = \frac{\frac{1}{2}(2+1)}{3} = 50\%.$$

Isso significa que metade dos pares de C estão representados em R_1 e R_2; de fato, dos três pares de C, existe um par totalmente representado $(((\text{tela},+),(\text{tela},-))$, pois $(\text{tela},+) \in R_1$ e $(\text{tela},-) \in R_2)$ e um par representado parcialmente $(((\text{bateria},+),(\text{bateria},-))$, pois $(\text{bateria},+) \in R_1$ mas $(\text{bateria},-) \notin R_2)$.

A média do percentual de representatividade dos resumos R_1 e R_2 fica

$$\Pr(R_1,R_2) = \frac{1}{2}\left(\frac{q_1}{|E_1|} + \frac{q_2}{|E_2|}\right) = \frac{1}{2}\left(\frac{3}{5} + \frac{3}{5}\right) = 60\%.$$

O percentual de diversidade de R_1 fica

$$\text{Pd}(R_1) = \frac{|\{(\text{tela},+),(\text{bateria},+)\}|}{|\{(\text{tela},+),(\text{tela},-),(\text{bateria},+),(\text{design},+)\}|} = \frac{2}{4} = 50\%.$$

⌟

A **pontuação de um sumário** será definida pela média harmônica das medidas Pr, Pc e Pd. Ela será representada pela letra H. A média harmônica foi escolhida porque, das três médias pitagóricas, é a que mais enfatiza valores baixos dentro de um conjunto, e considera-se que sumários que tenham qualquer uma das três medidas baixa demais são ruins, mesmo que as outras medidas sejam altas.

$$\text{H}(R) = 3 \times \left((\text{Pr}(R))^{-1} + (\text{Pc}(R_1,R_2))^{-1} + (\text{Pd}(R))^{-1}\right)^{-1}$$

Para eliminar o efeito da ordem das sentenças no conjunto de dados, essa ordem foi embaralhada antes de cada execução. Como isso faz os resultados serem não determinísticos, cada caso de teste foi executado 100 vezes, e os resultados reportados são a média dos valores encontrados para essas execuções descartando-se os 10 piores e os 10 melhores resultados.

5.2.2 Testes iniciais

Esta seção mostra os experimentos preliminares feitos em cada variação dos métodos descritos no Capítulo 4. Os métodos descritos no Capítulo 4 foram inicialmente executados sobre os quatro maiores conjuntos de dados descritos na Seção 5.1 (D1, D2, D3 e D4) e avaliados sob as medidas definidas na Seção 5.2.1. Os testes foram feitos para sumários de tamanho 100, sendo o tamanho do sumário medido pela quantidade máxima de palavras que cada um de seus lados pode conter. Cada caso de teste foi repetido 100 vezes e as pontuações mostradas são a média das obtidas nas 100 execuções. As próximas seções relatam os resultados de cada método.

5.2.2.1 Método 1: estatístico

O método de Lerman e McDonald (2009) (descrito na Seção 4.1.2) é uma variação do método de Lerman, Blair-Goldensohn e McDonald (2009), feito com o intuito de se obter sumários que realcem a contrastividade. Foram testadas deste método a Estratégia 1 (que é uma adaptação direta do método publicado em Lerman, Blair-Goldensohn e McDonald (2009) e não performa uma otimização no contraste do sumário) e a Estratégia 3 (que é, segundo Lerman e McDonald (2009), a que fornece melhores resultados). A Tabela 5.3 mostra a comparação das duas estratégias. Vê-se que, como esperado, a Estratégia 3 obteve contrastividade significativamente melhor para todos os casos de teste. Além disso, todos encontraram uma média harmônica melhor com a Estratégia 3 (o que se deve principalmente à melhora do percentual de contrastividade); isso confirma a hipótese de que que os métodos de sumarização contrastiva são mais adequados para comparar entidades do que os métodos de sumarização simples.

Tabela 5.3 – Avaliação dos métodos de Lerman e McDonald (2009) para sumários de tamanho 100.

	D1				D2				D3				D4			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
Estratégia 1	61	30	40	**40**	80	36	51	**50**	54	43	56	**50**	74	40	49	**51**
Estratégia 3	60	48	47	**51**	84	70	56	**68**	52	70	55	**58**	77	50	51	**57**

5.2.2.2 Método 2: agrupamento

O método publicado por Kim e Zhai (2009) tem duas variações: R-First e C-First. Ambas foram adaptadas no presente trabalho para gerar sumários de duas entidades (originalmente, o método faz sumários contrastivos de uma única entidade). Elas foram executadas e estão comparadas na Tabela 5.4. Concluiu-se que o método que desempenha melhor nos conjuntos de dados usados é o C-First. Nos conjuntos de dados usados pelos autores originais, a conclusão foi a mesma.

Tabela 5.4 – Avaliação dos métodos de Kim e Zhai (2009) para sumários de tamanho 100.

	D1				D2				D3				D4			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
R-First	60	34	31	**38**	61	48	36	**45**	52	49	40	**46**	60	33	32	**39**
C-First	71	42	36	**45**	37	44	30	**36**	64	64	48	**58**	65	39	36	**44**

Nota-se que o caso de teste mais bem sucedido foi com o conjunto D3, que é o conjunto mais balanceado quando às polaridades. Isso é condizente com a observação feita em Kim e Zhai (2009) de que não havia pretensão de usar o método em conjuntos com opiniões predominantemente positivas ou negativas.

5.2.2.3 Método 3: similaridade

O método publicado por Jin, Ji e Gu (2016) tem três variações: R-First, C-First e D-First. As três foram executadas e os resultados são mostrados na Tabela 5.5. O método D-First se mostrou superior aos outros dois.

Tabela 5.5 – Avaliação dos métodos de Jin, Ji e Gu (2016) para sumários de tamanho 100.

	D1				D2				D3				D4			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
R-First	64	37	34	**41**	69	45	39	**48**	64	55	52	**57**	66	36	37	**43**
C-First	62	37	35	**42**	72	55	45	**55**	65	53	53	**56**	69	45	44	**50**
D-First	67	41	38	**45**	73	51	44	**54**	72	63	57	**64**	72	47	47	**53**

5.2.2.4 Método 4: ranqueamento

O método do ranqueamento foi testado em suas quatro estratégias. Nota-se da Tabela 5.6 que não existe uma diferença significativa entre as performances dos quatro. Foram feitos testes com sumários de tamanhos menores (50 e 75) e maiores (125 e 150) e também não foi observada diferença. Assim, optou-se por usar a Estratégia 3 para os próximos testes porque ela obteve uma leve vantagem com sumários grandes (150 palavras).

Tabela 5.6 – Avaliação dos métodos originais para sumários de tamanho 100.

	D1				D2				D3				D4			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
Estratégia 1	73	49	43	**52**	65	79	51	**62**	71	89	61	**72**	79	61	53	**62**
Estratégia 2	79	50	44	**54**	76	81	55	**69**	68	87	59	**70**	80	58	51	**61**
Estratégia 3	80	50	46	**55**	81	69	58	**67**	76	71	62	**69**	82	55	52	**60**
Estratégia 4	81	48	44	**54**	84	66	59	**67**	74	65	59	**65**	83	54	53	**61**

5.2.3 Comparação entre os métodos

Para comparar os métodos descritos no Capítulo 4, eles foram executados sobre o conjunto de dados descrito na Seção 5.1 e avaliados sob as medidas definidas na Seção 5.2. Os testes foram feitos para sumários de tamanho 100, sendo o tamanho do sumário medido pela quantidade máxima de palavras que cada um de seus lados pode conter.

Foi testada a melhor versão de cada método, de acordo com experimentos relatados na seção anterior:

- **Estatístico**: método de Lerman e McDonald (2009) na variação chamada de Estratégia 3, que considera a contrastividade entre cada lado do sumário e o conjunto-fonte da entidade oposta;

- **Agrumapento**: método de Kim e Zhai (2009) na versão C-First, que prioriza a contrastividade do sumário, adaptado para sumarizar duas entidades;
- **Similaridade**: método de Jin, Ji e Gu (2016) na estratégia D-First, que prioriza a diversidade do sumário;
- **Ranquemaneto**: método inédito desenvolvido neste trabalho, na Estratégia 3, que pontua sentenças condiderando pares contrastivos e considera a representatividade além da contrastividade;
- **Ranqueamento+**: método inédito em uma variação da Estratégia 3 que prioriza sentenças de acordo com seu número de palavras.

É válido questionar se uma escolha qualquer de sentenças do conjunto-fonte não pode ser suficiente para se obter um sumário contrastivo, dispensando algoritmos específicos para a sumarização. Portanto, será avaliado também um método ingênuo que forma o resumo através da seleção de sentenças por sorteio. Esse método será chamado de 'método **aleatório**'.

A seguir serão mostradas duas tabelas com os resultados da avaliação de conteúdo para sumários gerados com restrição de tamanho de 200 palavras (100 para cada entidade). Nas tabelas, o melhor resultado de cada coluna está sublinhado, o segundo melhor está sublinhado tracejado, o terceiro melhor está sublinhado pontilhado e o pior está sublinhado ondulado; são considerados empatados valores com diferença de até dois pontos.

A Tabela 5.7 mostra os resultados para os quatro maiores conjuntos de dados e a Tabela 5.8 mostra a pontuação geral de cada caso de teste e a média da pontuação de cada teste: a penúltima coluna mostra a média aritmética das pontuações de cada método e a última mostra o teste t de Student pareado unicaudal considerando o método em questão e o método da linha anterior; os métodos estão ordenados pela pontuação, do pior para o melhor avaliado.

Tabela 5.7 – Avaliação dos métodos para sumários de tamanho 100.

	D1				D2				D3				D4			
método	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
aleatório	56	29	26	**33**	70	25	30	**34**	52	42	41	**45**	64	34	33	**40**
agrupamento	71	42	36	**45**	37	44	30	**36**	64	64	48	**58**	65	39	36	**44**
similaridade	67	41	38	**45**	73	51	44	**54**	72	63	57	**64**	72	47	47	**53**
estatístico	60	48	47	**51**	84	70	56	**68**	52	70	55	**58**	77	50	51	**57**
ranqueamento	81	50	46	**55**	81	69	58	**67**	76	71	62	**69**	82	55	52	**60**
ranqueamento+	86	63	58	**67**	91	78	67	**77**	95	92	89	**92**	91	67	66	**73**

O método aleatório claramente é um fracasso, pois forneceu resultados muito aquém dos demais. A adaptação do método do agrupamento parece infrutífera, tendo se sobressaído pouco em relação ao método aleatório. O método da similaridade teve um desempenho melhor do que o agrupamento, se destacando principalmente nos conjuntos D3, D7 e D8, onde ficou entre os três melhores métodos. O método estatístico mostrou-se superior ao do agrupamento, com

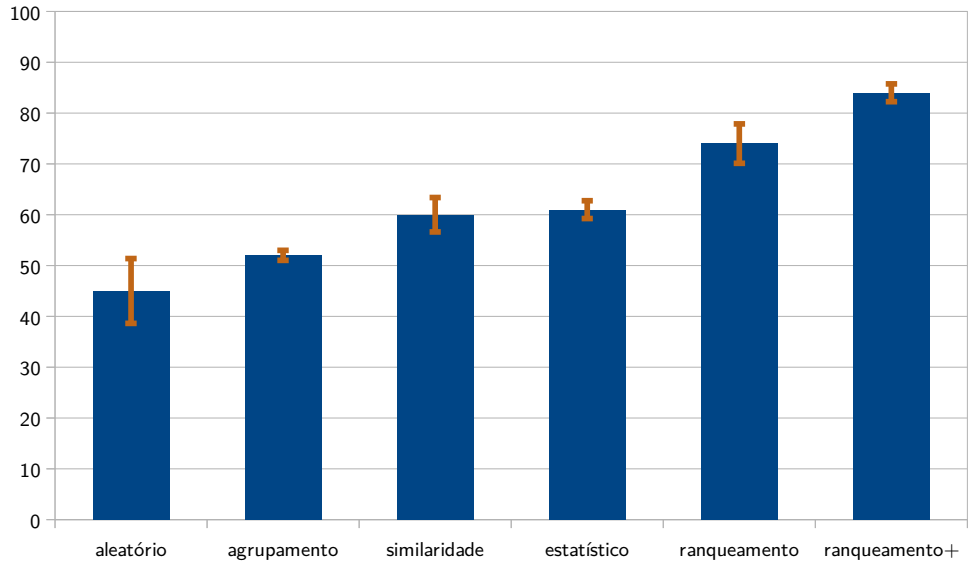
Tabela 5.8 – Avaliação geral dos métodos para sumários de tamanho 100.

método	D1	D2	D3	D4	D5	D6	D7	D8	média	teste t
aleatório	33	34	45	40	48	49	62	51	**45**	
agrupamento	45	36	58	44	59	59	63	53	**52**	0,003
similaridade	45	54	64	53	59	67	79	60	**60**	0,005
estatístico	51	68	58	57	70	88	57	42	**61**	0,412
ranqueamento	55	67	69	60	76	93	94	80	**74**	0,026
ranqueamento+	67	77	92	73	79	95	100	85	**84**	0,003

diferença enorme em três dos casos (D2, D5 e D6), mas encontrou um estorvo nos conjuntos D7 e D8, o que o deixou com média quase igual ao seu anterior no ranking. O método inédito saiu-se melhor ou igual aos anteriores em todos os casos, sendo a diferença (em relação ao método anterior no ranking) mais forte nos conjuntos D3, D7 e D8. O método obteve média de 13 pontos percentuais a mais do que o anterior de acordo com esta avaliação. Essa diferença se deve principalmente à melhora da representatividade, distintamente notável nos três primeiros conjuntos. O aprimoramento do método inédito conseguiu subir mais 10 pontos e foi o melhor colocado em todos os conjuntos de dados.

A Tabela 5.9 mostra o detalhamento da avaliação dos métodos, incluindo as pontuações obtidas para cada um os três critérios. A tabela também mostra os desvios padrões dentro de cada conjunto de teste (que é formado por 100 execuções, excluindo-se as 10 melhores e as 10 piores). O Gráfico 5.1 mostra a média das pontuações e a média dos desvios padrões de cada método. Desvios padrões pequenos indicam que o método gera sumários muito similares independentemente da ordem das sentenças no conjunto-fonte.

Gráfico 5.1 – Pontuações dos métodos, de acordo com a coluna em negrito da Tabela 5.8.



O gráfico Gráfico 5.2 mostra uma visualização de todos os testes executados (100 execuções para cada um dos 8 conjuntos de dados para cada um dos 6 métodos). Cada quadradinho no gráfico representa um teste, associando os conjuntos de dados (cada um em uma fileira vertical,

Tabela 5.9 – Avaliação detalhada dos métodos para sumários de tamanho 100.

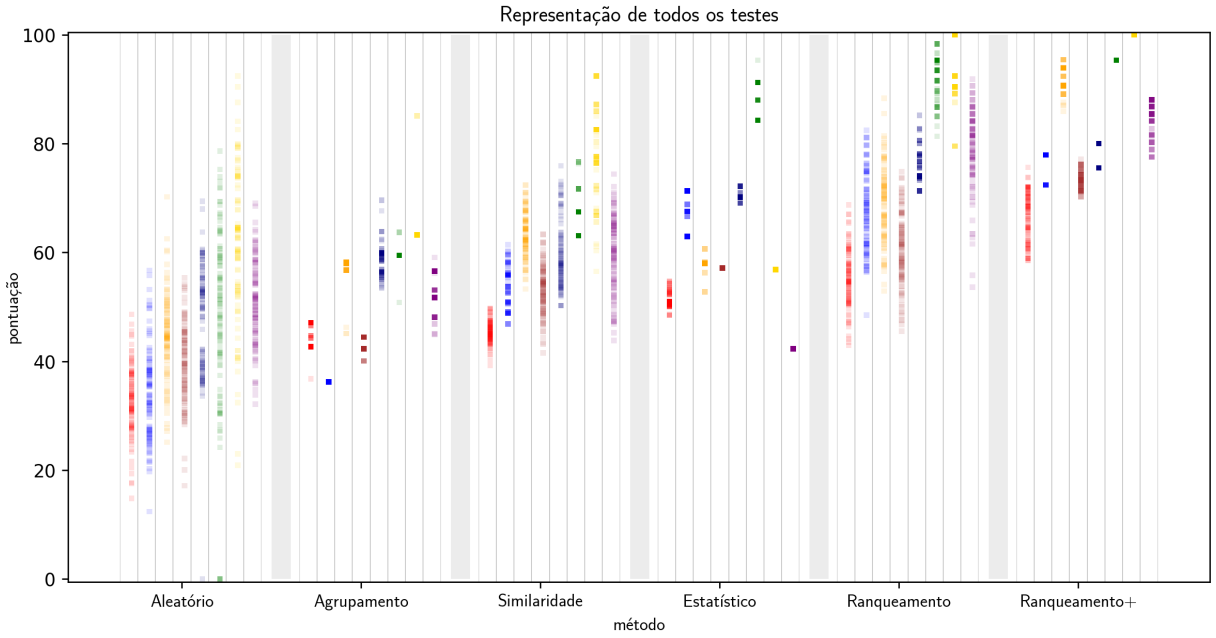
	D1				D2				D3				D4			
método	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
aleatório	56 ±6	29 ±4	26 ±3	**33 ±4**	70 ±4	25 ±6	30 ±4	**34 ±6**	52 ±6	42 ±6	41 ±5	**45 ±5**	64 ±5	34 ±5	33 ±4	**40 ±5**
agrupamento	71 ±2	42 ±2	36 ±2	**45 ±2**	37 ±0	44 ±0	30 ±0	**36 ±0**	64 ±0	64 ±2	48 ±0	**58 ±0**	65 ±1	39 ±1	36 ±1	**44 ±1**
similaridade	67 ±1	41 ±1	38 ±1	**45 ±1**	73 ±2	51 ±5	44 ±3	**54 ±3**	72 ±2	63 ±4	57 ±3	**64 ±3**	72 ±3	47 ±3	47 ±3	**53 ±3**
estatístico	60 ±1	48 ±1	47 ±1	**51 ±1**	84 ±5	70 ±3	56 ±3	**68 ±3**	52 ±0	70 ±0	55 ±0	**58 ±0**	77 ±0	50 ±0	51 ±0	**57 ±0**
ranqueamento	81 ±3	50 ±3	46 ±3	**55 ±3**	81 ±10	69 ±6	58 ±4	**67 ±5**	76 ±5	71 ±6	62 ±5	**69 ±5**	82 ±3	55 ±5	52 ±5	**60 ±5**
ranqueamento+	86 ±6	63 ±4	58 ±3	**67 ±4**	91 ±5	78 ±0	67 ±1	**77 ±2**	95 ±1	92 ±3	89 ±2	**92 ±2**	91 ±1	67 ±1	66 ±2	**73 ±1**

	D5				D6				D7				D8			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
aleatório	69 ±5	42 ±13	44 ±4	**48 ±7**	76 ±5	42 ±15	45 ±7	**49 ±10**	77 ±8	53 ±11	62 ±8	**62 ±9**	70 ±5	45 ±6	46 ±5	**51 ±5**
agrupamento	62 ±5	77 ±6	46 ±1	**59 ±2**	53 ±0	100 ±0	46 ±0	**59 ±0**	70 ±0	62 ±0	58 ±0	**63 ±0**	71 ±2	52 ±2	43 ±3	**53 ±3**
similaridade	73 ±4	55 ±10	53 ±4	**59 ±4**	82 ±2	68 ±4	56 ±3	**67 ±2**	84 ±6	76 ±8	78 ±6	**79 ±6**	74 ±5	54 ±6	56 ±5	**60 ±5**
estatístico	76 ±1	75 ±0	62 ±1	**70 ±1**	93 ±1	89 ±8	81 ±2	**88 ±3**	60 ±0	62 ±0	50 ±0	**57 ±0**	45 ±0	46 ±0	37 ±0	**42 ±0**
ranqueamento	74 ±5	100 ±0	62 ±3	**76 ±2**	96 ±1	99 ±4	86 ±4	**93 ±3**	96 ±4	92 ±6	95 ±4	**94 ±4**	89 ±2	78 ±6	74 ±5	**80 ±4**
ranqueamento+	82 ±7	100 ±0	63 ±0	**79 ±2**	97 ±0	100 ±0	89 ±0	**95 ±0**	100 ±0	100 ±0	100 ±0	**100 ±0**	91 ±1	85 ±2	80 ±4	**85 ±3**

representados por cores diferentes) executados pelos métodos (agrupados no eixo horizontal) à pontuação geral do sumário obtido (eixo horizontal). Os quadradinhos são translúcidos para permitir visualizar a densidade quando muitos testes obtêm pontuação parecida. O gráfico permite analisar alguns pontos:

- Três métodos apresentaram poucas pontuações distintas dentro de cada conjunto de dados. Isso indica que eles fazem escolhas um pouco mais determinísticas do que os demais. De fato, os três métodos onde isso ocorreu são condicionados a critérios mais restritivos do que os outros: o método do agrupamento considera todas as palavras da sentença (e não os aspectos das opiniões), e como há poucas sentenças idênticas (em oposição a aspectos idênticos), as decisões do algoritmo variam menos; o método do ranqueamento aprimorado favorece sentenças de acordo com o número de palavras delas, o que torna o sorteio de sentenças menos caótico; o método estatístico favorece sentenças com muitos aspectos identificados (pois é um algoritmo guloso que tenta, a cada iteração, aproximar o sumário do conjunto-fonte), e sentenças com muitos aspectos são raras, o que faz ele escolher sempre as mesmas em alguns casos.
- Nota-se uma variabilidade grande entre as pontuações de cada conjunto de dados no método aleatório: alguns conjuntos tiveram pontuações bem melhores que outros mesmo sem critério de escolha. Isso indica que conjuntos diferentes têm níveis diferentes de dificuldade para se gerar um resumo contrastivo de tamanho fixo.

Gráfico 5.2 – Representação gráfica das pontuações de todos os testes executados.



O Gráfico 5.3 mostra os métodos ordenados do melhor para o pior em cada conjunto de teste, e suas respectivas pontuações. Uma linha tracejada entre duas células indica empate (diferença de até dois pontos). Esse gráfico fornece uma visualização que elimina as diferenças de pontuações causadas pelas dificuldade dos conjuntos de teste da análise (alguns conjuntos de teste são naturalmente mais fáceis de serem sumarizados devido às suas características, especialmente tamanho), pois representa apenas a posição de cada um dos métodos dentro dos conjuntos de teste.

Gráfico 5.3 – Métodos ordenados por pontuação em cada conjunto de teste.

	1°	2°	3°	4°	5°	6°
D1	ranqueamento+ 67	ranqueamento 55	estatístico 51	similaridade 45	agrupamento 45	aleatório 33
D2	ranqueamento+ 77	ranqueamento 67	estatístico 68	similaridade 54	agrupamento 36	aleatório 34
D3	ranqueamento+ 92	ranqueamento 69	similaridade 64	estatístico 58	agrupamento 58	aleatório 45
D4	ranqueamento+ 73	ranqueamento 60	estatístico 57	similaridade 53	agrupamento 44	aleatório 40
D5	ranqueamento+ 79	ranqueamento 76	estatístico 70	similaridade 59	agrupamento 59	aleatório 48
D6	ranqueamento+ 95	ranqueamento 93	estatístico 88	similaridade 67	agrupamento 59	aleatório 49
D7	ranqueamento+ 100	ranqueamento 94	similaridade 79	agrupamento 63	aleatório 62	estatístico 57
D8	ranqueamento+ 85	ranqueamento 80	similaridade 60	agrupamento 53	aleatório 51	estatístico 42

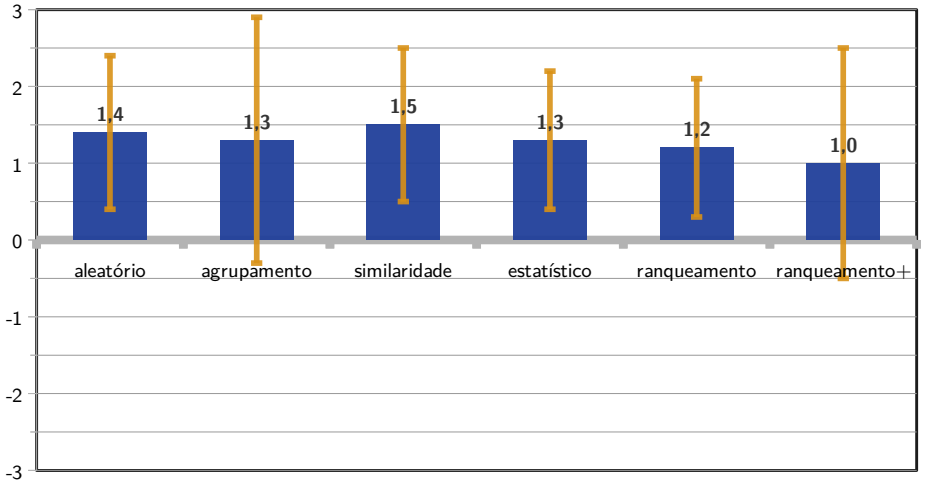
5.2.4 Percepção humana

Para entender a percepção humana sobre a utilidade dos sumários, foram convidadas 7 pessoas para avaliá-los. Cada método foi avaliado 13 vezes quanto ao resultado obtido para diferentes conjuntos de dados. Cada pessoa foi instruída a atribuir uma nota para o sumário entre -3 e 3 de acordo com quanto ela acha que o sumário ajuda a entender as diferenças entre os produtos. O Apêndice B mostra a folha de instruções que os colaboradores receberam. A Tabela 5.10 mostra as pontuações obtidas. O Gráfico 5.4 representa a média e desvio padrão das pontuações após convertê-los para uma escala de 0 a 100. Os resultados deste experimento sugerem que as pessoas não percebem diferença entre os métodos implementados.

Tabela 5.10 – Pontuações atribuídas pelos colaboradores aos sumários.

		método					
conjunto	colaborador	aleatório	agrupamento	similaridade	estatístico	ranqueamento	ranqueamento+
D1	a	3	3	2	3	1	3
	b	3	0	2	1	2	1
	f	1	2	0	0	1	−2
D2	e	1	2	2	1	2	2
D3	d	1	−2	1	2	0	−1
	e	1	0	2	2	1	1
D4	d	2	1	2	1	2	3
	g	0	2	0	2	0	−1
D6	c	1	3	2	0	1	2
	f	0	3	2	1	1	1
D7	a	3	−1	3	2	3	2
	b	1	2	0	0	0	1
	g	1	2	1	2	1	1
	média	**1,4**	**1,3**	**1,5**	**1,3**	**1,2**	**1,0**
	desvio padrão	1,0	1,6	1,0	0,9	0,9	1,5
	moda	1	2	2	2	1	1
	mediana	1	2	2	1	1	1

Gráfico 5.4 – Médias das pontuações atribuídas pelos colaboradores aos sumários.



5.3 Sumários obtidos

Esta seção mostra alguns resumos gerados pelos métodos para o conjunto de dados D1. Todos foram gerados com limite de 200 palavras (aproximado para o método do agrupamento, que não permite escolher um limite de palavras, mas sim de sentenças). Dos vários sumários gerados por cada método (que são não determinísticos), o sumário exibido aqui é aquele que mais se aproxima da pontuação média do respectivo método (reportada na Seção 5.2.3).

Quadro 5.2 – Sumário obtido com o método aleatório.

Celular A	Celular B
Aparelho muito bom androide e bom despacha bem rapido , só a câmera não é tudo aquilo que é anunciado me decepcionou.	O hardware é potente e passa a impressão de que durará vários anos sem se tornar obsoleto.
Na realiza eu comprei para o meu filho, segundo ele, o aparelho é ótimo, atende quase todas às necessidades, único detalhe contra é que não acita cartão de memória então é tudo de bom, cinco estrelas.	Estava muito satisfeito com o antigo, porém esse aparelho conseguiu me surpreender em todos os aspectos.
Comparado com outras marcas, atualmente achei que este celular tem o melhor custo-benefício.	Tem ótima resposta (tempo), capacidade de armazenamento muito bom...
Ótimo smartphone! Bateria de longa duração, boa tela, ótima câmera.	PRODUTO COM O PROCESSADOR MUITO RAPIDO.
O celular realmente é bom mas o fone de ouvido deixa a desejar, machuca o ouvido e incomoda bastante.	O S6 tinha Infravermelho e podia emular controles remotos, o S7 não tem mais.
	Smartphone bonito, elegante, com ótima performance e que tira fotos com qualidade.
	Troquei um iPhone 5S pelo Galaxy S7 e estou muito satisfeita.

Quadro 5.3 – Sumário obtido com o método do agrupamento.

Celular A	Celular B
Otimo custo-beneficio.	Podia ter melhor custo-benefício.
Muito bom o produto, ótimo custo beneficio.	Poderia custar mais barato.
Muito bom custo/benefício.	O problema pra mim é a bateria, custo e não ser dual chip.
Câmera muito boa.	Além disso, é caro demais.
Design muito bonito, rápido e com uma boa duração de bateria.	A bateria poderia durar mais e também ter um pouco mais de memória.
Bateria já estava ruim.	Excelente bateria.
Peca pela pouca duração da bateria.	Rápido e com boa duração de bateria.
Um bom aparelho porém baixo desempenho.	O aparelho é excelente, o desempenho me surpreendeu muito.
A câmera não é das melhores, nem o design.	É rápido, muito bonito, funcional, ótima câmera.
O celular é bom, mas bateria dura pouco e a camera deixa bastante a desejar.	Até agora não tive nenhum problema com o celular, apenas achei que a bateria fosse durar um pouco mais, mas fora isso o celular é excelente e o acabamento dele é ótimo.

Quadro 5.4 – Sumário obtido com o método estatístico.

Celular A	Celular B
Bonito e leve. Vale a pena principalmente pelo Android puro e pelas Moto Ações que facilitam a usabilidade.Ponto fraco por possuir apenas um alto-falante e para design, as partes em plástico o tornam frágil. Tendo em vista seu custo, vale muito a pena...apenas não gostei muito da câmera, mas quando ao restante estou satisfeita! Achei o Moto G5 um pouco mais lento e ocasionalmente trava e reinicia sozinho. Também percebi um bug que quando se carrega o celular, dependendo da capa de proteção não dá pra usar a tela touch (a voltagem interfere na detecção dos toques)	Por hora estou gostando muito, exceto a bateria pois não vi melhoria.O material parece de qualidade, apesar de ser um ímã para impressão digital e ser um pouco pesado.A câmera esta me surpreendendo a cada dia. Um ótimo Smartphone com ótimo desempenho, peca um pouco na duração de bateria e custo benefício. Alta capacidade de memória, tela amoled e quad hd dispensa comentários. Velocidade de processamento boa e design agradável.O único defeito é que a películas protetoras disponíveis no mercado não protegem toda a tela por conta do design ligeiramente curvo. Câmera, sensor de digital, desing, desempenho e bateria excelentes.

Quadro 5.5 – Sumário obtido com o método do ranqueamento aprimorado.

Celular A	Celular B
Melhor celular na sua faixa de preço. Ótima memória, muito bom desempenho. Um bom aparelho porém baixo desempenho. Boa durabilidade da bateria e câmeras perfeitas. A câmera não é das melhores, nem o design. Lindo e inovador. Peca pela pouca duração da bateria. A resolução da câmera é perfeita, o audio é muito bom. A tv não tem boa recepção. Muito bom o manuseio, configuração ótima. Também percebi um bug que quando se carrega o celular, dependendo da capa de proteção não dá pra usar a tela touch (a voltagem interfere na detecção dos toques) Aparelbo com problema. Otimo custo-beneficio.	Em seu preço original de $4200,00 não mesmo. Não trava nunca e isso é essencial. Aparelho premium, com câmera impressionante. É rápido, muito bonito, funcional, ótima câmera. Design lindo, mas bateria poderia ser melhor. Rápido e com boa duração de bateria. O leitor de digital é muito rápido e eficiente. A tela e a câmera têm muita qualidade. O S6 tinha Infravermelho e podia emular controles remotos, o S7 não tem mais. Só não deixe cair, quebra fácil. O prata suja bastante na parte de trás do aparelho. Top de linha, excelente custo beneficio. É à prova d'água mesmo.

5.4 Análise

Esta seção traz considerações sobre os elementos apresentados ao longo deste capítulo.

5.4.1 Dados usados

Buscou-se usar conjuntos de dados com características diversificadas para cada caso de teste. As principais características em que se buscou variabilidade são: tamanho absoluto do conjunto de dados, tamanho do conjunto de dados em relação ao tamanho do outro conjunto do mesmo par, frequência relativa de cada polaridade e quantidade de aspectos (sendo essas duas últimas obtidas por meio da identificação manual). Na busca por essa diversificação, fez-se uma manipulação dos conjuntos de dados orgânicos para se criar pares de conjuntos artificiais (que todavia simulam cenários que podem ser reais).

Foi de muito interesse fazer testes diversificados, em que cada par de entidade tem características diferentes dos outros quanto à quantidade de opiniões e a proporção entre opiniões positivas e negativas (como mostra a Tabela 5.2); além disso, a proporção entre o tamanho do conjunto de dados de uma entidade em relação à entidade oposta é bastante diferente entre os pares de entidades (ver Tabela 5.1). Buscar conjuntos de dados novos que fossem diferenciados e (principalmente) identificar manualmente suas opiniões seria muito laboroso, por isso foi conveniente fazer a manipulação para gerar conjuntos novos a partir dos já existentes.

5.4.2 Critério de avaliação

As métricas propostas para avaliação foram úteis para avaliar a composição do sumário: o percentual de representatividade indica se as opiniões importantes aparecem no sumário; o percentual de contrastividade indica se o sumário apresenta as diferenças entre as duas entidades; o percentual de diversidade indica se um sumário mostra opiniões diferentes entre si, evitando redundância.

A avaliação proposta é normalizada pelas expectativas instituídas pelos conjuntos-fonte. Por exemplo, vê-se que a avaliação dos testes em D4 foram bem mais satisfatórias do que as de D1. Isso ocorreu porque D4 é um subconjunto de D1. Quando se diminui um conjunto de dados, diminuem-se as expectativas sobre um resumo dele (por exemplo, existem menos pares contrastivos a serem formados). Além disso, como se mantém o limite de palavras do sumário, a compressão de D4 é menor que a de D1, o que também colabora para que a avaliação de D4 a seja melhor que a de D1.

A avaliação de utilidade foi feita pedindo-se para pessoas pontuarem os sumários de acordo com quanto elas acreditam que o sumário ajuda a comparar os produtos. Na avaliação, os

desvios padrões das pontuações atribuídas a cada método foram muito altos, o que levou a concluir que as pessoas não perceberam diferença significativa entre os métodos, e cada uma delas pode ter usado critérios diferentes para julgar os sumários.

Embora não tenha sido medida, a informatividade do sumário (quantidade de informação dos textos-fonte que é preservada no resumo (MANI, 2001a)) foi respeitada graças a algumas decisões feitas no projeto. Por exemplo, visando maximizar as três medidas de avaliação definidas, seria possível sempre colocar no sumário a menor sentença possível. De fato, isso economizaria muito espaço e aumentaria o número de opiniões no sumário, o que facilitaria às métricas atingir um valor alto. Porém, isso faria haver no resumo sentenças pouco explicativas do tipo *'câmera ruim'* e *'gostei da tela'*, e um sumário apenas com sentenças assim seria pouco informativo. Optou-se então por selecionar as sentenças de forma aleatória para pegar ora sentenças longas (que tendem a ser mais descritivas), ora sentenças curtas (que economizam espaço e facilitam a leitura). Também para melhorar a informatividade, foram eliminadas as sentenças curtas que não contêm um aspecto específico. Além dessa, outras justificativas mitigam a supressão do cálculo da informatividade:

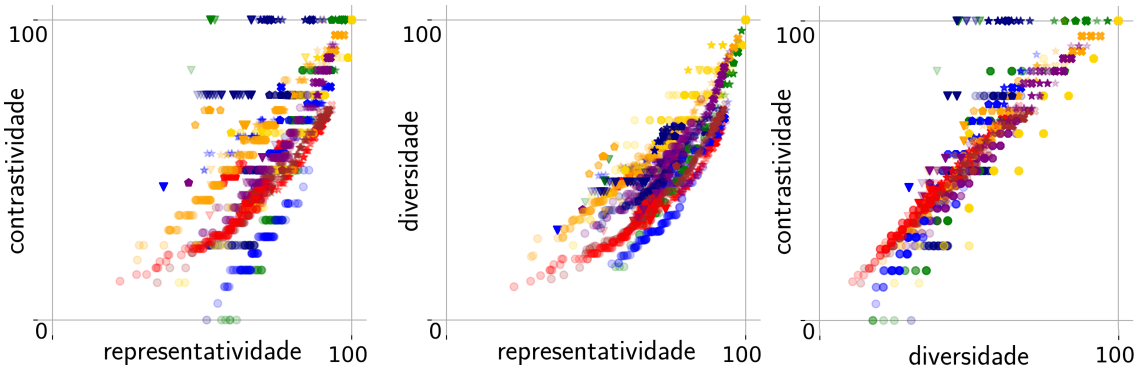
- A representatividade já mede, em certo nível, a informatividade do sumário (ainda que de forma simplificada, pois considera apenas aspectos e polaridades);
- Uma forma de medir a informatividade é por meio da geração manual de sumários: voluntários tomam o texto-fonte e fazem sumários que julgam ideal (CONDORI; PARDO, 2017). Essa tarefa seria muito complicada no caso dos sumários contrastivos (comparado a sumários simples), pois ia requerer uma análise de muitas possibilidades, o que humanos são ruins em fazer.

A avaliação proposta aqui não é de forma alguma a única forma de aferir a qualidade dos sumários. Para aceitar os resultados encontrados aqui, é preciso que se concorde com os critérios de avaliação adotados. É possível que alguém defina critérios de qualidade de sumários contrastivos diferentes dos apresentados aqui e, ao repetir os experimentos, encontre resultados diferentes. Isso depende do que se considera como um sumário contrastivo ideal.

5.4.2.1 Correlação entre as métricas

Foram usadas três métricas para avaliar o sumário, que supostamente consideram características distintas. Para saber se as métricas realmente são independentes e não existe uma correlação entre elas, foram feitas as visualizações mostradas no Gráfico 5.5. Nos gráficos, cada marcador se refere a um teste executado e mostra a relação entre duas métricas de avaliação obtidas no teste. Conjuntos de dados foram identificados com cores diferentes e métodos foram identificados por formatos diferentes de marcador. Como se vê, não há uma correlação clara entre nenhuma combinação de métrica (isso é, dada uma métrica, não é possível prever a outra), o que sugere que elas são de fato independentes e não há redundância ao usar as três

métricas para caracterizar a qualidade de um sumário. Apesar disso, como esperado, cada métrica sempre tende a melhorar quando qualquer outra métrica melhora.



(a) Relação entre representatividade e contrastividade.

(b) Relação entre representatividade e diversidade.

(c) Relação entre diversidade e contrastividade.

Gráfico 5.5 – Relações entre as métricas de avaliação.

5.4.3 Método original

De acordo com a definição do problema e a avaliação proposta, o método inédito apresentado se mostrou eficiente em todos os testes e superou outros métodos anteriormente publicados. Suas quatro variações tiveram performances similares. Mesmo com resultados similares, cada variação tem suas próprias características, que serão analisadas a seguir.

5.4.3.1 Estratégias

Nesta seção serão feitos comentários sobre as diferenças entre as quatro variações do método que foram observadas através dos experimentos.

Pontuação independente O uso da pontuação independente pontua cada opinião de acordo somente com sua frequência em seu conjunto-fonte de origem. Assim, as opiniões mais frequentes de cada grupo têm sempre maior prioridade (ao contrário da pontuação conjugada, que considera os dois elementos de cada par contrastivo para estimar a relevância de ambas as opiniões). Portanto, espera-se que essa abordagem maximize a representatividade do resumo. De fato, isso é observado (por exemplo) nas duas primeiras linhas da Tabela 5.6: a segunda estratégia, que usa pontuação conjugada, teve representatividade maior em quase todos os casos.

Pontuação conjugada Usando-se a pontuação conjugada, as filas de prioridade para cada entidade ficam ordenadas de forma que o n-ésimo elemento da fila de e_1 possa formar um par contrastivo com o n-ésimo elemento de e_2 (pois ambas as opiniões de cada par contrastivo recebem a mesma pontuação e consequentemente tomam o mesmo lugar no ranking). Assim

sendo, é esperado que a n-ésima sentença de um lado do sumário tenha uma opinião oposta à n-ésima sentença do lado oposto, o que pode favorecer a leitura do sumário para entender as diferenças entre as entidades. Isso pode ser visualizado no Quadro 5.6 e (mais notavelmente, por ser usada a Estratégia 1, que considera somente a contrastividade) no Quadro 5.7. O pareamento não é perfeito por causa das sentenças que contêm mais de uma opinião.

Quadro 5.6 – Resumo obtido com a segunda estratégia do método do ranqueamento para o conjunto D1.

Celular A	Celular B
Otimo custo beneficio.	Um ótimo smartphone, porém com baixo custo-benefício (como os demais high-ends)
Os gestos ajudam bastante no dia a dia, e o aparelho é bem rápido, e o leitor de digitais dá uma segurança a mais.	Aparelho muito rápido, bonito, leve e com uma câmara ótima.
Creio que a única coisa negativa são os “apenas” 2 GB de memória, se fossem 3 GB o aparelho seria imbativel, mas na sua faixa de preço ainda está valendo a pena.	Gostei muito do s7, desempenho excelente, apenas a bateria ruim.
Um celular excelente, rápido, câmera boa e funcionalidade boa.	Rápido, com fotos de alta resolução, ótima bateria, display de alta resolução.Até agora, não vi pontos negativos.
O celular é bom, mas bateria dura pouco e a camera deixa bastante a desejar.	Sistema muito limpo, fluido, camera muito boa e oferece atualizações frequentes.
Bonito e leve.	Produto maravilhoso, pra quem reclama da fragilidade recomendo que use uma pelicula de vidro nele e uma capa, pois todo aparelho é quebravel, o moto force se sobressai, mas é feio, então pelicula nele, vale a pena.
Celular de qualidade, lindo e com uma bateria super durável.	Excelente custo beneficio...
Pratico.	
Aparelbo com problema.	

Quadro 5.7 – Resumo obtido com a primeira estratégia do método do ranqueamento aprimorado para o conjunto D1, com limite de tamanho 75 e tamanho ideal de sentença igual a 5.

Celular A	Celular B
Muito bom o produto, ótimo custo beneficio.	O problema pra mim é a bateria, custo e não ser dual chip.
Um bom aparelho porém baixo desempenho.	Mostrou-se muito rápido, em meu uso nunca travou.
A câmera não é das melhores, nem o design.	A câmera é ótima para fotos em alta velocidade.
Boa durabilidade da bateria e câmeras perfeitas.	Muito rápido! Não trava! Bateria dura muito!
Muito bom deixa a desejar só a bateria e designer.	A rapidez do leitor de digital também é um ponto positivo.
Não tem a opção de excluir uma unica ligação feita ou recebida.	É rápido, muito bonito, funcional, ótima câmera.
Celular de qualidade, lindo e com uma bateria super durável.	O prata suja bastante na parte de trás do aparelho.
A resolução da câmera é perfeita, o audio é muito bom.	Produto de otima qualidade, produto feito para durar.
Aparelbo com problema.	Imagem top.

Ranking de representatividade O uso de uma fila de prioridade extra para maximizar a representatividade ajudou a gerar resumos que reflitam melhor as opiniões frequentes dos conjuntos-fonte, mas além disso resolveu um problema crucial: sem ela, os sumários gerados trariam apenas opiniões que podem formar pares contrastivos entre as entidades, porém é

possível que haja um número muito pequeno desses pares. No caso extremo onde nenhum par contrastivo pode ser formado, o resumo gerado seria vazio; se houver apenas um par contrastivo possível, o resumo ficaria repetindo opiniões até preencher o limite de tamanho ou esgotar as possibilidades de escolha, pois cada fila de prioridade conteria apenas uma opinião. Com a fila de representatividade, o resumo conteria as opiniões mais frequentes de cada conjunto-fonte independentemente da possibilidade de formar pares contrastivos. Não usar uma fila de representatividade seria viável apenas no caso em que se deseje obter um resumo totalmente comparativo, com interesse somente em opiniões que contrapõem as duas entidades. Nos testes feitos neste trabalho, esse problema não foi observado porque coincidentemente sempre houve um número suficiente de pares contrastivos entre as entidades.

5.4.3.2 Não determinismo

O algoritmo de sumarização é inteiramente baseado nas opiniões identificadas em cada sentença: ele indica, em cada iteração, uma opinião que é desejável ter no sumário para que se possa então escolher uma sentença que contenha essa opinião para ser inserida no sumário. Muitas vezes, existe mais de uma sentença no conjunto-fonte que contém a opinião indicada. Nesse caso, escolhe-se uma sentença arbitrariamente. O fato de a escolha ser aleatória, além de fazer a saída ser não determinística, afeta a performance do método (de acordo com a avaliação usada) principalmente por motivos relacionados a duas características da sentença selecionada:

- Tamanho da sentença: é possível que as sentenças empatadas tenham uma variedade grande de tamanhos. Escolher uma sentença muito grande ao invés de uma pequena significa deixar menos espaço no sumário para as iterações vindouras. A Seção 4.4.6 sugeriu um meio que pode amenizar o não determinismo causado por isso.
- Outras opiniões na sentença: é possível que a sentença escolhida contenha outras opiniões que não a indicada pelo algoritmo. Quando mais de uma sentença pode ser escolhida (por conter a opinião de interesse), sendo a escolha aleatória, não se pode determinar quais dessas opiniões intrusas entrarão no sumário, pois sentenças diferentes podem ter opiniões diferentes. Este caso será discutido na Seção 6.4.2.

CAPÍTULO

6

CONCLUSÃO

6.1 Considerações gerais

A ideia do método proposto neste trabalho e suas variações investigadas permitiram a obtenção de algoritmos inovadores para gerar sumários contrastivos sobre duas entidades. Eles foram testados com diferentes conjuntos de dados, avaliados quanto ao teor de características desejáveis contidas em seus resumos gerados e comparados com outros métodos publicados anteriormente.

Os conjuntos de dados usados pareceram suficientes para testar o método. Foram testados conjuntos que refletem situações comuns do uso real e conjuntos que simulam situações atípicas. Tendo características e tamanhos bem diversificados, os oito conjuntos puderam pôr em prova a performance dos métodos em diferentes cenários. Seria ideal haver mais conjuntos de dados de casos reais, porém a etapa da identificação manual de opinião é trabalhosa demais para valer o esforço.

A avaliação foi feita com heurísticas que estimam a representatividade, comparabilidade e diversidade do sumário baseadas na identificação manual de opiniões feita no conjunto de dados. Para concordar com os resultados obtidos neste trabalho, é preciso que se concorde com os critérios de avaliação adotados, que não são únicos: é possível que alguém considere outras características como importantes em sumários contrastivos e, ao medir a qualidade dos sumários de acordo com elas, encontre resultados que divergem dos apresentados aqui.

Além do método que decide quais opiniões devem entrar para o sumário, foram também testadas formas de se selecionar as sentenças mais relevantes. Como não era o foco deste trabalho, esse estudo se limitou a alguns testes empíricos. As observações feitas sobre essa etapa ao longo do texto são convincentes para argumentar que esse é um processo importante,

e uma implementação para uso prático deveria estudar a melhor maneira de executá-lo de acordo com as características desejáveis para o resumo.

Os testes feitos neste trabalho usaram a anotação manual de opinião, supondo essa etapa resolvida. Para um sistema de uso prático, a anotação deveria ser automática. Isso demandaria o uso de uma ferramenta competente para essa tarefa (como o estudo de Avanço, Brum e Nunes (2016), que encontrou métodos que obtiveram acurácia superior a 90% para identificação de polaridade em textos em português); usar uma anotação que não funcione bem pode comprometer os resultados.

6.2 Hipóteses investigadas

A hipótese que os métodos têm performances diferentes foi provada nas avaliações, onde se pode ver com clareza que existe uma enorme diferença de pontuação entre os métodos (Gráfico 5.1, página 110). Os testes também mostraram que os métodos que têm performance melhor são melhores em todos os casos de teste, provando errada a conjectura que métodos diferentes se sobressairiam em diferentes cenários.

A hipótese que métodos de sumarização contrastiva são mais adequados para comparar entidades do que métodos de sumarização simples foi verificada na Seção 5.2.2.1, onde o mesmo método foi aplicado para gerar sumários independentes (não contrastivos) e contrastivos e descobriu-se que considerar a contrastividade eleva significativamente as métricas que medem a qualidade de resumos contrastivos.

A avaliação humana permitiu investigar a hipótese de utilidade do sumário. Nessa avaliação, todos os métodos obtiveram pontuação próxima a 1,2 (em escala de -3 a 3). Seria o suficiente para concluir que os métodos são úteis, até porque as pontuações dadas pelas pessoas são majoritariamente positivas (Tabela 5.10, página 113). Todavia, os desvios padrões calculados são muito altos, o que faz com que os dados sejam inconclusivos. De fato, houve baixa concordância entre os anotadores mesmo para sumários idênticos (em um dos casos, o sumário do conjunto D1 recebeu notas 3, 1 e -2 das três pessoas que o avaliaram). Assim, a única conclusão palpável que esse experimento trouxe é que existe uma alta discordância entre pessoas na tarefa de avaliar a utilidade de um sumário, o que não era inesperado, por se tratar de uma tarefa muito subjetiva.

6.3 Contribuições

Como principais contribuições deste trabalho, consideram-se:

1. A avaliação de métodos de sumarização contrastiva da literatura, que até então não tinham sido comparados entre si (Capítulo 5);
2. A adaptação de um trabalho da literatura que trata uma tarefa similar para que ele execute a sumarização contrastiva como definida aqui (Seção 4.2);
3. A comparação entre dois métodos de sumarização muito similares, um que considera a contrastividade e outro que não considera, para entender a diferença entre eles para se comparar entidades (Seção 5.2.2.1);
4. O desenvolvimento de um método novo de sumarização contrastiva, que superou os métodos já existentes, considerando as métricas propostas neste trabalho (Seção 4.4);
5. A construção de um conjunto de dados para sumarização contrastiva com textos em português, anotado e estendido manualmente (Seção 5.1).

6.4 Trabalhos futuros

Sugerem-se nesta seção algumas modificações que poderiam ser feitas no método original para melhorar a sua qualidade em uma implementação para uso real.

6.4.1 Pré-processamento textual

O pré-processamento é uma parte crítica da execução, e, em uma implementação para uso prático, valeria a pena investir tempo para descobrir como essa etapa pode ser feita da melhor maneira possível. Algumas ações auxiliares (não usadas neste projeto) que podem ser feitas nessa etapa são a normalização textual, que permitiria corrigir erros ortográficos e outros ruídos presentes no texto (BERTAGLIA, 2017) e uma estimação da importância das opiniões, que permitiria manter apenas os comentários mais relevantes (SOUSA, 2015).

A segmentação do texto em sentenças não foi perfeita mas serviu o propósito de se obter pequenos fragmentos de texto. Ela poderia ter sido melhorada com um algoritmo que identifique o caso em que o autor do texto não coloca um espaço após a pontuação e identifique o caso em que pontos são usados em abreviaturas. Mas a segmentação foi satisfatória, e deixar algumas sentenças maiores do que deveriam (por conterem pontuação sem espaço) foi interessante para ver como o algoritmo lida com a presença de sentenças grandes: em muitos casos, ele as prefere por conterem várias opiniões, o que contribui muito com a representatividade (e às vezes contrastividade) do resumo.

6.4.2 Opiniões extras na sentença

Idealmente, seriam inseridas no sumário apenas as opiniões indicadas pelo algoritmo, mas algumas sentenças têm mais de uma opinião, o que pode favorecer ou prejudicar a performance se forem inseridas no resumo, como exemplificado pelos seguintes casos:

- Se todas as opiniões contidas em uma sentença escolhida seriam indicadas proximamente pelo algoritmo, ele atinge o seu objetivo para a iteração atual e (por sorte) para as próximas, eventualmente economizando espaço por ter encontrado uma única sentença que traz várias opiniões de interesse.
- Se alguma opinião contida na sentença já tinha entrado no sumário, ocorre redundância, e não é desejável incluir opiniões repetidas (a menos que sobre espaço no resumo após todas as opiniões de interesse terem sido inclusas).
- Se alguma opinião contida na sentença não seria nunca indicada pelo algoritmo, ocorre a presença de informação irrelevante no sumário.

A observação desses pontos permite formular aprimoramentos para o método, para que a seleção de sentença seja feita considerando todas as opiniões contidas em uma sentença candidata, atribuindo pontuação maior às sentenças que contêm opiniões que favorecem o método.

6.4.3 Resumo quantitativo

O formato de resumo gerado pelo método é informativo extrativo: ele seleciona os trechos de texto que considera mais relevantes e os concatena para formar o resumo. Do ponto de vista do usuário, fica obscuro saber quais opiniões apresentadas no resumo são as mais relevantes e mais frequentes do conjunto-fonte, pois:

- algumas sentenças do sumário podem conter também informação irrelevante (ou seja, informação que não se desejaria inserir no sumário, mas foi inserida pelo fato de a sentença ter alguma opinião de interesse);
- algumas sentenças podem ter sido selecionadas apenas pela possibilidade de formar pares contrastivos, sem que necessariamente contenham opiniões frequentes;
- o sumário pode conter informação redundante (quando sobra espaço no sumário e faltam opções de sentenças diversificadas para se inserir), mas o fato de uma opinião aparecer repetida no sumário não necessariamente reflete à frequência dessa opinião no conjunto-fonte;
- o sumário não indica quão frequentes as opiniões são.

Propõe-se aqui um formato de resumo contrastivo que exibe estatísticas sobre os pares contrastivos detectados no conjunto-fonte. O Quadro 6.1 mostra um exemplo real feito para o

conjunto D1. Ele mostra todos os pares contrastivos detectados, ordenados por importância[1], onde a importância é estimada pela função de pontuação definida na Seção 4.4.3.1. Para destacar as opiniões mais frequentes, usou-se em cada par contrastivo um tamanho de fonte grosseiramente proporcional à sua importância.

Quadro 6.1 – Sumário indicativo baseado nas estatísticas de pares contrastivos do conjunto D1.

aspecto	Celular A	Celular B	importância
preço	+	−	1170
produto	−	+	1056
desempenho	−	+	616
câmera	−	+	530
produto	+	−	368
bateria	+	−	360
bateria	−	+	280
outro	−	+	143
design	−	+	115
design	+	−	105
outro	+	−	98
resistência	−	+	70
tela	−	+	54
preço	−	+	39
desempenho	+	−	34

Exemplificando o uso, o formato de sumário proposto no Quadro 6.1 informa ao leitor que a principal diferença entre os produtos (segundo as opiniões coletadas) é que o preço do Celular A é melhor do que o do Celular B. Todavia, como informa a segunda linha, o Celular B é, de maneira geral, superior ao Celular A. As duas próximas linhas confirmam e detalham a informação contida na segunda: muitas pessoas acham que o desempenho e a câmera do Celular B são melhores do que os do Celular A. Apenas uma pequena parte das pessoas acha que o Celular A é melhor do que o Celular B, como informa a linha seguinte. As duas próximas linhas indicam que a avaliação da bateria é controversa, pois mostram duas opiniões divergentes sobre ela com quase o mesmo tamanho de fonte tipográfica. Com tudo isso, um comprador que não se importe com o preço pode usar essas informações e optar pelo Celular B.

Além de permitir identificar facilmente os pares contrastivos mais importantes, esse formato permite uma avaliação geral de cada entidade: pode-se olhar para as colunas que contêm as indicações de polaridades a fim de se descobrir qual das duas têm mais pontos positivos e quão importantes esses pontos positivos são.

Sugere-se usar um resumo indicativo como o definido aqui em conjunto com um resumo informativo para que o usuário tenha tanto uma visão geral das opiniões quanto uma amostra daquelas que se julga mais relevantes e informativas.

[1] Por ser uma informação que requer conhecimento teórico do método para ser interpretada, não se recomenda exibir o valor de importância para usuário (que foi incluído no Quadro 6.1), mas apenas usá-lo para ordernar as opiniões no quadro.

6.5 Uso prático

Esta seção destaca pontos que devem ser levados em conta em possíveis implementações para uso real dos sistemas estudados neste trabalho.

6.5.1 Eficiência

Uma implementação para uso prático traria uma preocupação quanto à eficiência da implementação, especialmente quanto ao tempo de execução. Nos experimentos feitos em um computador pessoal comum, não se observou nenhum caso que tenha levado mais de 0,1 segundo para o algoritmo concluir sua execução.

O algoritmo aqui proposto tem a vantagem de confrontar os dois conjuntos de entrada apenas uma vez (para formar os pares contrastivos), ao contrário de outros (como Lerman e McDonald (2009)) em que a otimização deve ser feita com um algoritmo que a cada nova inserção no sumário deve verificar todos os possíveis pares que podem ser formados pelos dois conjuntos. Além disso, o algorimo proposto aqui não usa recursos tão computacionalmente custosos quanto as distribuições de probabilidade usadas por Lerman e McDonald (2009) ou o agrupamento de similaridade usado por Kim e Zhai (2009). Assim como é o método de ranqueamento é o método de similaridade: ele também seleciona apenas as sentenças mais valiosas para o sumário, porém ele faz muitas simplificações (para se esquivar de um problema de otimização mais rigoroso), o que o deixa com resultados mais modestos.

A Tabela 6.1 mostra o tempo de execução para gerar um resumo para cada conjunto de dados usando cada um dos métodos. O tempo é indicado em segundos e foi obtido a partir da média de 100 execuções de cada caso de teste em um computador pessoal comum. A parte de leitura e pré-processamento de dados demorou 7 segundos para o conjunto de dados maior.

Tabela 6.1 – Tempo de execução dos algoritmos (em segundos) em cada conjunto de dados.

método	D1	D2	D3	D4	D5	D6	total
aleatório	0,04	0,00	0,01	0,02	0,00	0,00	**0,07**
agrupamento	13,35	0,02	1,72	0,50	0,01	0,00	**15,6**
similaridade	0,15	0,01	0,02	0,07	0,00	0,00	**0,25**
estatístico	40,46	1,78	9,10	4,54	0,47	0,21	**56,56**
ranqueamento	0,07	0,01	0,02	0,03	0,00	0,01	**0,14**
ranqueamento+	0,07	0,01	0,01	0,03	0,00	0,01	**0,13**

O método aleatório é naturalmente o mais rápido, pois ele não usa nenhum algoritmo para selecionar sentenças adequadas, e sim executa um sorteio para escolhê-las. Por ser um método aparentemente inútil, é irrelevante comparar o seu tempo de execução com os dos demais métodos.

6.5.2 Dificuldades da área

Algumas dificuldades são peculiares a tarefas de análise de opinião:

- Para identificar se uma opinião é positiva ou negativa, não é suficiente analisar palavras isoladas do texto. Por exemplo, embora o adjetivo 'bom' seja positivo, não são positivas as sentenças *'queria saber se esse produto é mesmo bom'*, *'não sei se esse produto é bom'*, *'esse produto não é muito bom'*, *'foi bom ter devolvido o produto'*.
- O próprio ato de identificar o que é uma opinião pode ser complicado em alguns casos. Muitas vezes essa identificação depende fortemente de um contexto, o que a torna um atributo difícil de ser modelado e automatizado. Por exemplo, a sentença *'a bateria desse celular não dura um dia'* pode ser somente um fato objetivo. No entanto, ao encontrar essa frase em uma resenha de um aparelho de telefone celular, o que ela significa é que, de acordo com a experiência do autor da frase, a bateria dura um tempo insuficiente, e isso é algo indesejado. Fazer essa interpretação é fácil para uma pessoa desde que ela saiba o que é um telefone celular, saiba a função da bateria em um telefone celular, saiba que um dia de duração para uma bateria de celular é um tempo abaixo da média para os padrões de uso atuais, e que isso é algo indesejado. Transferir esse tipo de conhecimento para uma máquina pode ser muito complicado.
- Por causa de seu objetivo fundamental (o de analisar informações publicadas por diversas pessoas na Web), os textos processados não passam por uma revisão profissional (como os textos publicados em veículos convencionais de informação). Por esse motivo, precisa-se considerar que os textos usados podem conter erros de gramática e ortografia, abreviaturas e abreviações não oficiais, gírias, e outras características que distanciam o texto daqueles escritos visando usar uma norma padrão.
- Muitos textos opinativos incluem um linguajar restrito a um grupo de pessoas. Por exemplo, na rede social Twitter, é comum referir-se ao Facebook como 'a outra rede social' ou 'a rede social ao lado', como na Figura 6.1. Um minerador de opinião automático (ou talvez até um ser humano) que esteja levantando informações sobre o Facebook ignoraria completamente esses termos se não fosse previamente informado para reconhecê-los.
- Outra característica comum em textos opinativos (em contraste aos objetivos) é a presença de ironia[2]. O uso dessa figura de linguagem pode confundir até atentos leitores humanos se não for empregada de maneira prudente pelo autor. Para uma máquina, essa confusão tende a ser ainda mais frequente. Por exemplo, na sentença

[2] Tem-se usado indistintamente os termos 'ironia' e 'sarcasmo' (ao menos na língua inglesa) para referir-se ao mesmo fenômeno: simplificadamente, um falante F profere algo a com significado s a um ouvinte O dentro de um contexto C com a intenção que O ao receber a compreenda s' com $s' \neq s$; para que ocorra com sucesso, falante e ouvinte devem compartilhar conhecimento sobre o contexto C para saber que s é absurdo ou inapropriado dentro de C, o que deve gerar em O um processo abdutivo onde ele julga que F na verdade quer dizer s' (EISTERHOLD; ATTARDO; BOXER, 2006).

Figura 6.1 – Usuário do Twitter chamando o Facebook pelo apelido popular 'a outra rede social'.

Fonte: Captura de tela de twitter.com/isTHEREaMIND/status/962038540568686592

> *'esse aparelho é tão bom que não durou nem um mês'*, para identificar a ironia, é preciso saber que não durar nem um mês é algo ruim, portanto oposto de bom. Uma máquina desavisada poderia interpretar e inclusive aprender, erradamente, que durar pouco é algo bom.

Essas dificuldades podem depender muito do domínio (tipo de texto) com o qual se lida. Elas foram supostas resolvidas neste trabalho para manter o foco de investigar a sumarização contrastiva de maneira independente de possíveis erros ocorridos na etapa inicial do processamento dos textos.

6.6 Disponibilização

Este trabalho de mestrado é parte do Projeto Opinando. Os códigos-fonte usados nos experimentos estão disponíveis na página do projeto, sites.google.com/icmc.usp.br/opinando, e no repositório github.com/raphsilva/contrastive-summarization. Também foram disponibilizados os resultados dos experimentos, incluindo sumários obtidos e valores das métricas de avaliação para cada uma das 100 execuções de cada método sobre cada conjunto de teste.

6.7 Considerações finais

Esta pesquisa foi relevante por se tratar de um trabalho inédito que reúne e compara métodos de sumarização contrastiva de opinião a partir de testes diversificados. Ela permite tanto o avanço do conhecimento na área quanto o desenvolvimento de ferramentas que pessoas e empresas demandam.

Do ponto de vista acadêmico, espera-se que este trabalho contribua com a pesquisa em Processamento de Linguagem Natural, somando a ela ideias novas que poderão ser futuramente utilizadas em trabalhos derivados, e dando continuidade aos esforços voltados a processamento de textos opinativos que têm se intensificado nos últimos anos.

Do ponto de vista prático, a pesquisa permitirá que uma análise feita sobre textos encontrados na Web possa resultar rapidamente em uma comparação entre dois produtos similares e dizer quais aspectos de um são melhores do que aspectos de outro. Tal análise tem importância para empresas, porque permitiria a elas saber em curto tempo quais são os pontos fortes e as fraquezas ou até defeitos de seus produtos, e ainda saber quais produtos concorrentes são piores ou melhores, o que permitiria esboçar um nítido mapa da colocação da empresa no mercado do ponto de vista de seus clientes. Para o consumidor, seria útil ter uma ferramenta que use dados atualizados frequentemente (a partir de blogs, fóruns, etc.) e as traduza em um conjunto sintético de informações para que ele possa decidir sobre quais produtos ou serviços são melhores para ele.

Apesar de essa proposta de pesquisa ser inédita para o português, ela dá continuidade às iniciativas mais recentes de sumarização de opinião e análise de sentimentos, trazendo conceitos inovadores e contribuindo para o avanço da fronteira de conhecimento.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. (Ed.). **Mining Text Data**. Boston, MA: Springer US, 2012. ISBN 978-1-4614-3223-4. Disponível em: link.springer.com/book/10.1007/978-1-4614-3223-4. Citado na página 24.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 14724:2011**: Informação e documentação — trabalhos acadêmicos — apresentação. Rio de Janeiro, 2011. 15 p. Citado na página 23.

AVANÇO, L. V.; BRUM, H. B.; NUNES, M. Improving opinion classifiers by combining different methods and resources. **XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)**, p. 25–36, 2016. Citado na página 122.

BEINEKE, P.; HASTIE, T.; MANNING, C.; VAITHYANATHAN, S. An exploration of sentiment summarization. In: SHANAHAN, J. G.; WIEBE, J.; QU, Y. (Ed.). **Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications**. Stanford, US: [s.n.], 2004. Disponível em: nlp.stanford.edu/~manning/papers/rotup.pdf. Citado nas páginas 39 e 40.

BERTAGLIA, T. F. C. **Normalização textual de conteúdo gerado por usuário**. 137 p. Dissertação (Mestrado) — Universidade de São Paulo, 2017. Citado na página 123.

CARBERRY, S.; ELZER, S.; GREEN, N.; MCCOY, K.; CHESTER, D. Extending document summarization to information graphics. In: **Text Summarization Branches Out**. [s.n.], 2004. Disponível em: www.aclweb.org/anthology/W04-1002. Citado na página 35.

CARDOSO, P. C.; MAZIERO, E. G.; JORGE, M. L.; SENO, E. M.; FELIPPO, A. D.; RINO, L. H.; NUNES, M. G.; PARDO, T. A. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: **Proceedings of the 3rd RST Brazilian Meeting**. [S.l.: s.n.], 2011. p. 88–105. Citado na página 24.

CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. **Expert Systems with Applications**, Elsevier, v. 78, p. 124–134, 2017. Citado nas páginas 100, 101 e 117.

DEMIR, S.; CARBERRY, S.; MCCOY, K. Generating textual summaries of bar charts. In: **Proceedings of the Fifth International Natural Language Generation Conference**. Association for Computational Linguistics, 2008. p. 7–15. Disponível em: www.aclweb.org/anthology/W08-1103. Citado na página 35.

EISTERHOLD, J.; ATTARDO, S.; BOXER, D. Reactions to irony in discourse: evidence for the least disruption principle. **Journal of Pragmatics**, v. 38, n. 8, p. 1239 – 1256, 2006. ISSN 0378-2166. Focus-on Issue: Discourse and Conversation. Disponível em: http://www.sciencedirect.com/science/article/pii/S0378216604002656. Citado na página 127.

FELDMAN, R. Techniques and applications for sentiment analysis. **Commun. ACM**, ACM, New York, NY, USA, v. 56, n. 4, p. 82–89, abr. 2013. ISSN 0001-0782. Disponível em: doi.acm.org/10.1145/2436256.2436274. Citado nas páginas 38 e 46.

FELLBAUM, C. (Ed.). **WordNet: An Electronic Lexical Database**. Cambridge, MA: MIT Press, 1998. (Language, Speech, and Communication). ISBN 978-0-262-06197-1. Citado na página 50.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. **Artificial Intelligence Review**, v. 47, n. 1, p. 1–66, Jan 2017. ISSN 1573-7462. Disponível em: doi.org/10.1007/s10462-016-9475-9. Citado na página 23.

GANESAN, K. **Opinion Driven Decision Support System**. 132 p. Tese (Doutorado) — Universidade de Illinois, 2013. Citado na página 34.

GENEST, P.-E.; LAPALME, G. Framework for abstractive summarization using text-to-text generation. In: **Proceedings of the Workshop on Monolingual Text-To-Text Generation**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (MTTG '11), p. 64–73. ISBN 9781937284053. Disponível em: dl.acm.org/citation.cfm?id=2107679.2107687. Citado na página 35.

GUO, J.; LU, Y.; MORI, T.; BLAKE, C. Expert-guided contrastive opinion summarization for controversial issues. In: **Proceedings of the 24th International Conference on World Wide Web**. New York, NY, USA: ACM, 2015. (WWW '15 Companion), p. 1105–1110. ISBN 978-1-4503-3473-0. Disponível em: http://doi.acm.org/10.1145/2740908.2743038. Citado nas páginas 28, 47 e 100.

HAHN, U.; MANI, I. The challenges of automatic summarization. **Computer**, v. 33, n. 11, p. 29–36, Nov 2000. ISSN 0018-9162. Citado na página 23.

HOQUE, E.; CARENINI, G. Multiconvis: A visual text analytics system for exploring a collection of online conversations. In: **Proceedings of the 21st International Conference on Intelligent User Interfaces**. New York, NY, USA: ACM, 2016. (IUI '16), p. 96–107. ISBN 978-1-4503-4137-0. Disponível em: doi.acm.org/10.1145/2856767.2856782. Citado na página 22.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: **Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2004. (KDD '04), p. 168–177. ISBN 1-58113-888-1. Disponível em: doi.acm.org/10.1145/1014052.1014073. Citado nas páginas 26, 49, 50, 51 e 57.

_____. Mining opinion features in customer reviews. In: **Proceedings of the 19th National Conference on Artifical Intelligence**. AAAI Press, 2004. (AAAI'04), p. 755–760. ISBN 0-262-51183-5. Disponível em: dl.acm.org/citation.cfm?id=1597148.1597269. Citado nas páginas 50 e 57.

IBEKE, E.; LIN, C.; WYNER, A.; BARAWI, M. Extracting and understanding contrastive opinion through topic relevant sentences. In: _____. **Proceedings of the The 8th International Joint Conference on Natural Language Processing**. [S.l.]: ACL Anthology, 2017. v. 2, p. 395–400. ISBN 978-1-948087-01-8. Citado nas páginas 42 e 47.

JIN, J.; JI, P.; GU, R. Identifying comparative customer requirements from product online reviews for competitor analysis. **Eng. Appl. Artif. Intell.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 49, n. C, p. 61–73, mar. 2016. ISSN 0952-1976. Disponível em: dx.doi.org/10.1016/j.engappai.2015.12.005. Citado nas páginas 28, 30, 40, 41, 42, 43, 44, 47, 49, 59, 61, 62, 64, 67, 85, 88, 99, 100, 108, 109 e 137.

JINDAL, N.; LIU, B. Mining comparative sentences and relations. In: **Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2**. AAAI Press, 2006. (AAAI'06), p. 1331–1336. ISBN 978-1-57735-281-5. Disponível em: dl.acm.org/citation.cfm?id=1597348.1597400. Citado na página 46.

KAN, M.; KLAVANS, J. L.; MCKEOWN, K. Using the annotated bibliography as a resource for indicative summarization. **CoRR**, cs.CL/0206007, 2002. Disponível em: arxiv.org/abs/cs.CL/0206007. Citado na página 34.

KIM, H. D.; ZHAI, C. Generating comparative summaries of contradictory opinions in text. In: **Proceedings of the 18th ACM Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2009. (CIKM '09), p. 385–394. ISBN 978-1-60558-512-3. Disponível em: doi.acm.org/10.1145/1645953.1646004. Citado nas páginas 28, 41, 42, 43, 44, 45, 47, 48, 55, 57, 58, 59, 62, 67, 76, 77, 78, 79, 81, 82, 83, 84, 99, 100, 107, 109, 126 e 137.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **The Annals of Mathematical Statistics**, The Institute of Mathematical Statistics, v. 22, n. 1, p. 79–86, 03 1951. Disponível em: doi.org/10.1214/aoms/1177729694. Citado na página 71.

LERMAN, K.; BLAIR-GOLDENSOHN, S.; MCDONALD, R. Sentiment summarization: Evaluating and learning user preferences. In: **Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EACL '09), p. 514–522. Disponível em: dl.acm.org/citation.cfm?id=1609067.1609124. Citado nas páginas 39, 40, 52, 53, 55, 67, 68, 69, 71, 74 e 107.

LERMAN, K.; MCDONALD, R. Contrastive summarization: An experiment with consumer reviews. In: **Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (NAACL-Short '09), p. 113–116. Disponível em: dl.acm.org/citation.cfm?id=1620853.1620886. Citado nas páginas 21, 28, 41, 42, 43, 47, 48, 52, 53, 59, 62, 64, 67, 69, 71, 72, 99, 100, 107, 108, 126 e 137.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: https://www.aclweb.org/anthology/W04-1013. Citado na página 101.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan and Claypool Publishers, 2012. Disponível em: www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. Citado nas páginas 20, 21, 23, 24, 25, 28, 36, 37, 38, 39, 40, 41, 46, 137 e 138.

LIU, B.; HU, M.; CHENG, J. Opinion observer: Analyzing and comparing opinions on the web. In: **Proceedings of the 14th International Conference on World Wide Web**. New York, NY, USA: ACM, 2005. (WWW '05), p. 342–351. ISBN 1-59593-046-9. Disponível em: doi.acm.org/10.1145/1060745.1060797. Citado nas páginas 28, 30, 38, 41, 42, 43, 47, 48, 49, 50, 64, 99, 100 e 137.

LLORET, E.; PALOMAR, M. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. **Informatica**, v. 34, n. 1, 2010. Citado na página 33.

López Condori, R. E. **Sumarização automática de opiniões baseada em aspectos**. 134 p. Dissertação (Mestrado) — Universidade de São Paulo, 2015. Citado nas páginas 25 e 26.

MANI, I. **Advances in Automatic Text Summarization**. Cambridge, MA, USA: MIT Press, 1999. ISBN 0262133598. Citado na página 33.

_____. **Automatic Summarization**. John Benjamins Publishing, 2001. Disponível em: www.jbe-platform.com/content/books/9789027299109. Citado nas páginas 20, 33, 34 e 117.

_____. Summarization evaluation: An overview. In: **Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization**. [S.l.]: Citeseer, 2001. Citado na página 35.

MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. Wordnet: An on-line lexical database. **International Journal of Lexicography**, v. 3, p. 235–244, 1990. Citado na página 50.

ÖZSOY, M. G.; ÇAKICI, R. Contrastive max-sum opinion summarization. In: JAAFAR, A.; ALI, N. M.; NOAH, S. A. M.; SMEATON, A. F.; BRUZA, P.; BAKAR, Z. A.; JAMIL, N.; SEMBOK, T. M. T. (Ed.). **Information Retrieval Technology**. Cham: Springer International Publishing, 2014. p. 256–267. Citado nas páginas 28 e 47.

PARK, S.; LEE, K. S.; SONG, J. Contrasting opposing views of news articles on contentious issues. p. 340–349, 01 2011. Citado nas páginas 47 e 100.

PAUL, M. J.; ZHAI, C.; GIRJU, R. Summarizing contrastive viewpoints in opinionated text. In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (EMNLP '10), p. 66–76. Disponível em: http://dl.acm.org/citation.cfm?id=1870658.1870665. Citado na página 47.

PEDERSEN, T.; PATWARDHAN, S.; MICHELIZZI, J. Wordnet: Similarity - measuring the relatedness of concepts. In: **Proceedings of the 19th National Conference on Artifical Intelligence**. AAAI Press, 2004. (AAAI'04), p. 1024–1025. ISBN 0-262-51183-5. Disponível em: dl.acm.org/citation.cfm?id=1597148.1597310. Citado na página 78.

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. **Knowledge-Based Systems**, v. 89, p. 14 – 46, 2015. ISSN 0950-7051. Disponível em: www.sciencedirect.com/science/article/pii/S0950705115002336. Citado na página 40.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. In: CASELI, H.; VILLAVICENCIO, A.; TEIXEIRA, A.; PERDIGÃO, F. (Ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 260–271. Citado nas páginas 23 e 24.

RIBALDO, R.; PARDO, T.; RINO, L. Sumarização automática multidocumento com mapas de relacionamento. In: **Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology**. [s.n.], 2011. Disponível em: www.nilc.icmc.usp.br/til/stil2011_English/tilic/artigos/5249bd47_vf.pdf. Citado na página 23.

SANCHAN, N.; BONTCHEVA, K.; AKER, A. Understanding man Preferences for Summary Designs in Online Debates Domain. **Polibits**, scielomx, p. 79 – 85, 12 2016. ISSN 1870-9044. Disponível em: www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-90442016000200079&nrm=iso. Citado nas páginas 43 e 45.

SILVA, R. R.; PARDO, T. A. S. Córpus 4p: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de sumarização contrastiva de opinião. In: **Symposium in Information and Human Language Technology - STIL**. [S.l.]: SBC, 2019. Citado nas páginas 102 e 137.

SOUSA, O. A. F. **Sumarização contrastiva de opinião: uma abordagem com otimização**. 2018. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP). Citado nas páginas 49 e 83.

SOUSA, R. F. **Abordagem TOP (X) para inferir os comentários mais importantes sobre produtos e serviços**. 79 p. Dissertação (Mestrado) — Universidade Federal do Piauí, 2015. Citado na página 123.

SUN, J.-T.; WANG, X.; SHEN, D.; ZENG, H.-J.; CHEN, Z. Cws: a comparative web search system. In: **WWW '06: Proceedings of the 15th international conference on World Wide Web**. New York, NY, USA: ACM, 2006. p. 467–476. ISBN 1-59593-323-9. Disponível em: www.microsoft.com/en-us/research/publication/cws-a-comparative-web-search-system/. Citado na página 46.

TADANO, R.; SHIMADA, K.; ENDO, T. Multi-aspects review summarization based on identification of important opinions and their similarity. In: **Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation**. [S.l.: s.n.], 2010. Citado na página 101.

THONET, T. **Modèles thématiques pour la découverte non supervisée de points de vue sur le Web**. 147 p. Tese (Doutorado) — Université Toulouse 3–Paul Sabatier, 2017. Citado na página 28.

VARGAS, F. A.; PARDO, T. A. S. Aspect clustering methods for sentiment analysis. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 365–374. Citado nas páginas 101, 137, 138 e 142.

WANG, D.; ZHU, S.; LI, T. Sumview: A web-based engine for summarizing product reviews and customer opinions. **Expert Syst. Appl.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 1, p. 27–33, jan. 2013. ISSN 0957-4174. Disponível em: dx.doi.org/10.1016/j.eswa.2012.05.070. Citado na página 40.

WANG, D.; ZHU, S.; LI, T.; GONG, Y. Comparative document summarization via discriminative sentence selection. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 6, n. 3, p. 12:1–12:18, out. 2012. ISSN 1556-4681. Disponível em: doi.acm.org/10.1145/2362383.2362386. Citado nas páginas 43 e 100.

XU, X.; MENG, T.; CHENG, X. Aspect-based extractive summarization of online reviews. In: **Proceedings of the 2011 ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2011. (SAC '11), p. 968–975. ISBN 978-1-4503-0113-8. Disponível em: http://doi.acm.org/10.1145/1982185.1982396. Citado na página 40.

APÊNDICE

A

CONSTRUÇÃO DO CONJUNTO DE DADOS

Este apêndice descreve a construção do **Córpus 4P**, um córpus de opiniões em português brasileiro sobre telefones celulares e câmeras digitais. O córpus inclui anotação manual de aspectos e polaridades das opiniões. As 1437 sentenças do córpus foram coletadas de 542 comentários publicados por compradores no site Buscapé e se referem a quatro produtos diferentes. Esse córpus deve subsidiar pesquisas na área de Análise de Sentimentos, mais especificamente, em Sumarização Contrastiva de Opinião. O córpus também foi descrito em Silva e Pardo (2019).

A criação do Córpus 4P se deu para possibilitar a investigação de métodos de sumarização contrastiva para o português. Nessa aplicação, é necessário um conjunto de textos de tamanho suficiente, com anotação confiável e que ofereça casos de teste diversificados para se avaliar os métodos. Na linha do que fizeram outros trabalhos (LIU; HU; CHENG, 2005; LERMAN; MCDONALD, 2009; KIM; ZHAI, 2009; JIN; JI; GU, 2016), optou-se por coletar textos opinativos sobre 4 produtos eletrônicos. Essa escolha deve-se principalmente à facilidade de coletar esse tipo de texto (dada sua abundância em algumas páginas da web) e também por se acreditar que a identificação de aspectos opinativos nesse domínio é uma tarefa mais bem definida do que em outros casos (por exemplo, avaliações de serviços, resenhas de livros e discussões políticas), como sugerido em (VARGAS; PARDO, 2018).

Aspectos costumam ser o foco de muitas tarefas de Análise de Sentimentos (LIU, 2012). Por exemplo, se uma pessoa opina sobre um telefone móvel, ela pode falar sobre a tela, o peso, o tamanho, etc. Essas características e partes da entidade avaliada são o que se convencionou chamar de **aspectos** (ou aspectos opinativos, para diferenciar de outros possíveis usos). Uma **entidade** é o objeto alvo da opinião; pode ser um produto, serviço, organização, indivíduo ou evento, entre outros (LIU, 2012).

No córpus anotado neste trabalho, cada sentença foi manualmente rotulada quanto aos **aspectos** avaliados e quanto à **polaridade** da opinião, que indica se a opinião é positiva ou negativa. Além dessas anotações, foram identificadas sentenças que não contêm opiniões (por exemplo, *'comprei esse celular semana passada'*) e sentenças que não pertencem ao escopo de avaliação do produto (por exemplo, *'demorou muito para chegar'*).

O córpus foi composto a partir de 542 comentários extraídos do site Buscapé, que é um site brasileiro que oferece serviço gratuito de busca de produtos e comparação de preços em lojas virtuais, permitindo que seus usuários publiquem avaliações sobre os produtos. Os comentários coletados contêm 1437 sentenças, nas quais foram manualmente identificadas

1682 opiniões sobre os produtos de interesse[1]. Considera-se que uma **opinião** é um par formado por um aspecto e sua polaridade. Foram identificadas no córpus 1415 opiniões positivas e 226 negativas[2], além de 164 outras passagens que não contêm opinião sobre os produtos de interesse, que foram separadas em várias categorias para melhor caracterizar o córpus. A Tabela A.1 mostra a quantidade de texto coletado.

Tabela A.1 – Informações quantitativas dos dados coletados para o conjunto de dados.

entidade	comentários	sentenças	opiniões
Motorola Moto G5 Plus	229	592	795
Galaxy S7	221	610	813
Canon EOS Rebel T5	47	129	161
Canon PowerShot SX520 HS	45	106	133
total	**542**	**1437**	**1902**

Visando subsidiar o teste de métodos de sumarização em contextos variados de uso, diferentes arranjos do córpus foram produzidos. Além da configuração inicial, foram construídos subconjuntos do córpus para permitir avaliar os métodos em situações com variedade de dados. Foram criados 12 arranjos dos dados a partir da seleção de sentenças sobre as entidades avaliadas. A seleção foi feita de modo que os arranjos tivessem características diversificadas entre si, como quanto à quantidade de aspectos citados, proporção entre opiniões positivas e negativas, etc.

Após escolhidas as entidades e coletadas as opiniões sobre elas, a primeira etapa da criação do córpus foi definir como ele deveria ser anotado. Foram feitas leituras dos textos coletados e anotações prévias para entender melhor o problema e definir as regras de anotação. Então, o conjunto passou por uma etapa de anotação automática para que os anotadores precisassem apenas conferir a anotação e não ter que inserir todas as informações eles mesmos. Após a revisão manual, o córpus passou por processos de extensão, limpeza e simplificação. Esses passos são descritos a seguir.

A.1 Anotação

A próxima seção relata o processo prático da anotação, e a atual descreve as regras usadas para identificar opiniões. As regras foram definidas empiricamente por meio da inspeção manual dos textos coletados, tendo como base ideias publicadas em (LIU, 2012) e (VARGAS; PARDO, 2018).

A.1.1 Identificação de aspectos

Um aspecto é o assunto principal de que trata uma opinião. Os aspectos foram identificados seguindo uma listagem de 16 aspectos para celulares (Quadro A.1) e 18 aspectos para câmeras Quadro A.2). Além desses, foi atribuído um aspecto especial caso a opinião se referisse ao produto de maneira geral (e não a um aspecto específico), como na sentença *'esse celular é supimpa'*; opiniões assim são ditas **genéricas**.

[1] Algumas sentenças contêm mais de uma opinião, como *'A câmera é boa mas o aparelho trava muito'*, onde se identificam uma opinião positiva sobre a câmera e uma opinião negativa sobre o aparelho.

[2] As outras 41 não são claramente positivas ou negativas.

Quadro A.1 – Aspectos definidos para entidades do tipo 'Celular'.

aspecto	descrição	exemplos
acessório	Itens extras que acompanham o produto ou que podem ser comprados separadamente.	fone de ouvido, alça, cartão de memória, capa de proteção, bolsa, manual de uso
armazenamento	Sistema de armazenamento de arquivos.	espaço de armazenamento interno, expansão com cartão de memória
bateria	Sistema de energia.	duração da bateria, tempo de carregamento
câmera	Captação de imagens.	fotos, vídeos, foco, zoom
desempenho	Performance do processamento.	processador, memória RAM, aquecimento
design	Características relacionadas à aparência.	acabamento, cor, formato
outro	Temas específicos que não estão inclusos em outras categorias.	rádio, televisão, sensor biométrico, NFC, Bluetooth, microfone
peso	Peso do equipamento.	
preço	Avaliações sobre o preço pago e a relação custo-benefício.	
produto	Avaliações genéricas sobre o produto.	
resistência	Capacidade do produto de resistir a situações adversas e ao tempo.	
sistema	Características do software.	sistema operacional
som	Som reproduzido pelo aparelho.	volume, equalização
tamanho	Tamanho do equipamento.	
tela	Parte do equipamento que exibe a interface visual.	visor, cores, resolução, sensores táteis
usabilidade	Facilidade de uso.	facilidade de operação, personalização, interface com usuário

Quadro A.2 – Aspectos definidos para entidades do tipo 'Câmera'.

aspecto	descrição	exemplos
acessório	Itens extras que acompanham o produto ou que podem ser comprados separadamente.	fone de ouvido, alça, cartão de memória, capa de proteção, bolsa, manual de uso
armazenamento	Sistema de armazenamento de arquivos.	espaço de armazenamento interno, expansão com cartão de memória
áudio	Áudio gravado pelo aparelho.	
bateria	Sistema de energia.	duração da bateria, tempo de carregamento
design	Características relacionadas à aparência.	acabamento, cor, formato
foco	Sistema de controle de foco.	
foto	Fotografias obtidas com a câmera.	
funcionalidade	Recursos e opções do produto que adicionam possibilidades de uso.	reconhecimento de face, modos pré-programados
imagem	Características gerais de imagens obtidas pela câmera, sejam fotos ou vídeos	resolução, brilho, granulação, HDR, estabilização, abertura da lente
peso	Peso do equipamento.	
preço	Avaliações sobre o preço pago e a relação custo-benefício.	
produto	Avaliações genéricas sobre o produto.	
resistência	Capacidade do produto de resistir a situações adversas e ao tempo.	
tamanho	Tamanho do equipamento.	
tela	Parte do equipamento que exibe a interface visual.	visor, cores, resolução, sensores táteis
usabilidade	Facilidade de uso.	facilidade de operação, personalização, interface com usuário
vídeo	Vídeos obtidos com a câmera.	
zoom	Recursos de aproximação de imagem.	

A.1.2 Identificação de polaridades

Cada aspecto identificado foi associado a uma polaridade que refletisse o sentimento expresso em relação a ele. A identificação de polaridade foi feita de acordo com as definições da Quadro A.3; os anotadores usaram os símbolos da primeira coluna da tabela para indicar as polaridades. Observam-se quatro categorias de polaridades, indicadas por cores diferentes na tabel: positivas, negativas, neutras e experiência.

Quadro A.3 – Definições de polaridades usadas na anotação.

polaridade		definição	exemplo
+	**positivo**	bom, desejável	*Celular muito rápido.*
+.	**positivo fraco**	condicionalmente bom ou parcialmente bom	*É um valor caro, mas vale a pena.*
++	**positivo forte**	excepcionalmente bom	*Perfeito, sem qualquer reclamação.*
−	**negativo**	ruim, indesejável	*O preço é alto demais.*
−.	**negativo fraco**	condicionalmente ruim ou parcialmente ruim	*Para profissionais não compensa.*
−−	**negativo forte**	excepcionalmente ruim	*Foi o pior aparelho que já comprei.*
.	**mediano**	no meio da escala entre desejável e indesejável	*Aparelho razoável.*
*	**relativo**	sem conceito claro de desejável ou indesejável	*Design discreto.*
..	**dual**	simultaneamente bom e ruim	*Aceita SD mas não expande a memória.*
#	**irresoluto**	indecisão ou falta de opinião	*Não sei se gostei.*
!	**conselho**	informação que ajuda a usar melhor o produto	*Recomendo que seja comprada uma capinha.*
&	**caso de uso**	experiência de uso do produto	*Uso para falar e web.*

A.1.3 Trechos indesejados

Algumas sentenças coletadas não são úteis para avaliar o produto em questão. Essas sentenças foram separadas em três classes:

- **serviço**: trechos que falam não sobre o produto, mas sobre alguma entidade relacionada a ele ou à experiência de compra, como fabricante, vendedor, transportadora, etc.
- **contextualização**: trechos que contêm informação adicional que pode agregar valor ao comentário da pessoa, mas não ajuda a avaliar o produto se lido isoladamente.
- **irrelevante**: trechos que não se relacionam ao produto em questão e sequer agregam valor a comentários sobre o produto.

Também foram identificados trechos **duplicados** (quando a mesma pessoa publica duas vezes o mesmo comentário ou repete algo no mesmo comentário) e trechos **defectivos**, como trechos ininteligíveis (*'xvcxcvc'*) e quebrados (*'pouca, recém comprada'*).

A Tabela A.2 mostra a frequência dos trechos indesejados. Um total de 11,6% dos trechos não são interessantes para a avaliação do produto-alvo.

Tabela A.2 – Frequência de trechos indesejados.

tipo	frequência		exemplo
serviço	**3,4%**	64	*A entrega foi rápida.*
irrelevante	**1,0%**	18	*Não conheço esse produto.*
contextualização	**1,5%**	29	*A promessa foi que o produto seria entregue até dia 9.*
defectiva	**2,8%**	53	*ASHJAKHAKJYADSFHJK*
duplicada	**2,9%**	56	
total	**11,6%**	220	

A.1.4 Segmentação de texto

Os colaboradores foram instruídos a identificar todas as opiniões contidas no texto. Sentenças que contivessem mais de uma opinião foram divididas em trechos de forma que cada trecho contivesse apenas uma opinião. As informações sobre a divisão de sentença foram registradas para que também seja possível usar o córpus com as sentenças completas.

Os colaboradores foram orientados a reescrever trechos para que cada trecho fizesse sentido se lido isoladamente. Por exemplo, a sentença *'o produto é bom mas caro pra dedéu'* poderia ser quebrado nos trechos *'o produto é bom'* e *'o produto é caro pra dedéu'*. Isso seria útil para uma eventual tarefa onde convém considerar trechos contendo uma única opinião. Além disso, essa divisão traz mais segurança à anotação, pois permite identificar exatamente onde cada opinião foi encontrada.

Se achassem necessário, os anotadores também poderiam unir duas sentenças em casos que essa união formasse uma opinião que não se formaria com as sentenças isoladas, como no trecho *'O que eu mais busco num celular é a qualidade e possibilidade de edição nas configurações da câmera. E nesse quesito o S7 não deixa a desejar.'*

A.2 Ferramentas

Após coletados os dados, os textos foram automaticamente divididos em sentenças e a classificação de opiniões começou com uma etapa de identificação automática com um método[3] que identifica aspectos por meio de palavras-chave (como estudaram (VARGAS; PARDO, 2018)) e polaridades por meio de meta-informações sobre a avaliação[4]. Depois da fase automática, duas pessoas trabalharam revisando a anotação. Toda a anotação foi feita em ferramentas de edição de texto.

Os anotadores receberam arquivos de texto puro, um para cada produto, estruturados como no Quadro A.4. Dentro de cada arquivo, os comentários aparecem ordenados pela data de publicação. Cada sentença se inicia com um identificador do tipo (006.015) que indica o comentário de onde a sentença foi extraída e a posição da sentença no arquivo. Depois, existem dois pares de colchetes: o primeiro é preenchido com a polaridade e o segundo contém os aspectos identificados na etapa automática. Depois, há um separador (formado por dois dois-pontos) seguido da sentença.

[3] A ferramenta usada está disponível em github.com/raphsilva/naive-opinion-miner.

[4] Quando uma pessoa publica uma avaliação sobre um produto no Buscapé, ela deve também indicar se recomenda ou não o produto. Considera-se, no método, que se uma pessoa não recomenda o produto, todas as sentenças que ela escreveu são negativas; análogo para positivas.

Quadro A.4 – Formato de dados recebido pelos anotadores.

```
(006.015) [+][PRODUTO TAMANHO DESEMPENHO] :: Sempre quis um aparelho da linha S e o S7 tem o
tamanho perfeito para nao chamar muita atencao e desempenho fantástico.

(007.017) [+][PRODUTO] :: Qualidade.

(007.018) [+][PRODUTO] :: Produto deixa a desejar, bordas metalicas riscam com facilidade,
botão home então nem se fala.

(007.019) [+][TELA] :: Complicado de encontrar películas compatíveis com a tela toda.
```

Quadro A.5 – Revisão manual feita no exemplo do Quadro A.4.

```
 (006.015) [+.][PRODUTO]    :: Sempre quis um aparelho da linha S.

 (006.015)  [+][TAMANHO]    :: O S7 tem o tamanho perfeito para não chamar muita atenção.

 (006.015)  [+][DESEMPENHO] :: Desempenho fantástico.

b(007.017)  [.][PRODUTO]    :: Qualidade.

 (007.018)  [-][PRODUTO]    :: Produto deixa a desejar.

 (007.018)  [-][DESIGN]     :: Bordas metalicas riscam com facilidade.

 (007.018)  [-][OUTRO]      :: Botão home ruim.

 (007.019)  [-][ACESSÓRIO]  :: Complicado de encontrar películas compatíveis com a tela toda.
```

Com os arquivos de texto como no Quadro A.4, os anotadores deveriam corrigir as opiniões que foram identificadas automaticamente e dividir a sentença quando necessário. Foram definidas combinações de caracteres para especificar as polaridades (mostradas no Quadro A.3). Se uma sentença não contivesse informação útil para avaliar o produto, os anotadores identificaram isso no começo da linha com caracteres predefinidos: por exemplo, uma letra 'b' no começo de uma linha indica que o texto contido ali está quebrado; uma letra 'd' indica trecho duplicado; etc. O Quadro A.5 mostra um exemplo de trabalho dos anotadores.

A.3 Extensão do córpus

Os dados coletados são opiniões sobre quatro produtos extraídas do site Buscapé. São dois tipos de produtos: celulares e câmeras. Os comentários sobre os celulares formam o subconjunto rotulado **D1**, e os sobre as câmeras formam o conjunto **D2**. Para diversificar os testes, foram criados artificialmente outros casos de teste a partir de D1 e D2.

O conjunto **D3** é um subconjunto de D1 do qual foram excluídas algumas sentenças de modo a equilibrar a quantidade de opiniões positivas e negativas para cada entidade. Ambos os conjuntos D1 e D2 têm muito mais opiniões positivas do que negativas (ver Tabela A.3 a

seguir). Esse novo conjunto pode ser usado para simular um cenário em que exista forte controvérsia entre as opiniões sobre as entidades.

O conjunto **D4** é um subconjunto de D1 de onde foram excluídas aleatoriamente algumas sentenças para que uma das entidades ficasse com uma quantidade de texto muito menor do que a outra. Com ele, pode-se avaliar casos em que uma das entidades tem mais destaque.

O conjunto **D5** contém os comentários mais recentes de D2, o conjunto **D6** contém os comentários mais antigos de D2 e o conjunto **D8** contém comentários aleatórios de D1. Eles permitem simular situações com conjuntos pequenos.

O conjunto **D7** é um subconjunto de D2 que contém apenas quatro aspectos. Foram escolhidos quatro aspectos (os quatro mais frequentes de D2) e descartadas todas as sentenças que não os citam. Com ele, simula-se situação em que uma entidade tem poucos aspectos.

A.4 Limpeza e simplificação do córpus

Para uso prático em trabalhos de Análise de Sentimentos, foi projetada uma versão limpa do córpus. Além de removidas as sentenças que não são úteis para avaliar o produto a que deveriam se referir (sentenças que foram marcadas pelos anotadores como serviço, irrelevantes, ininteligíveis, etc), foram removidas sentenças genéricas com menos de quatro palavras por se considerar que elas normalmente não ajudam a avaliar as especificidades de um produto (por exemplo, *'não gostei'*).

Para reproduzir melhor o que um sistema automático de anotação faria, algumas informações foram simplificadas. Foram descartadas as divisões de sentenças em trechos, que é uma tarefa difícil de ser executada no contexto deste córpus. Os três níveis de polaridade positiva foram unidos em um só; o mesmo ocorreu para opiniões negativas. Trechos subjetivos que não são positivos ou negativos foram identificados como neutros.

A limpeza foi feita para evitar ruídos e propagação de erros em sistemas que usem o córpus, e a simplificação foi feita para deixá-lo em um estado em que possa simular um conjunto de dados anotado automaticamente, o que seria interessante para projetos de Processamento de Linguagem Natural. A versão original do córpus, antes da limpeza e simplificação, também foi disponibilizada, pois pode ser útil em pesquisas futuras.

A.5 Visão geral do córpus

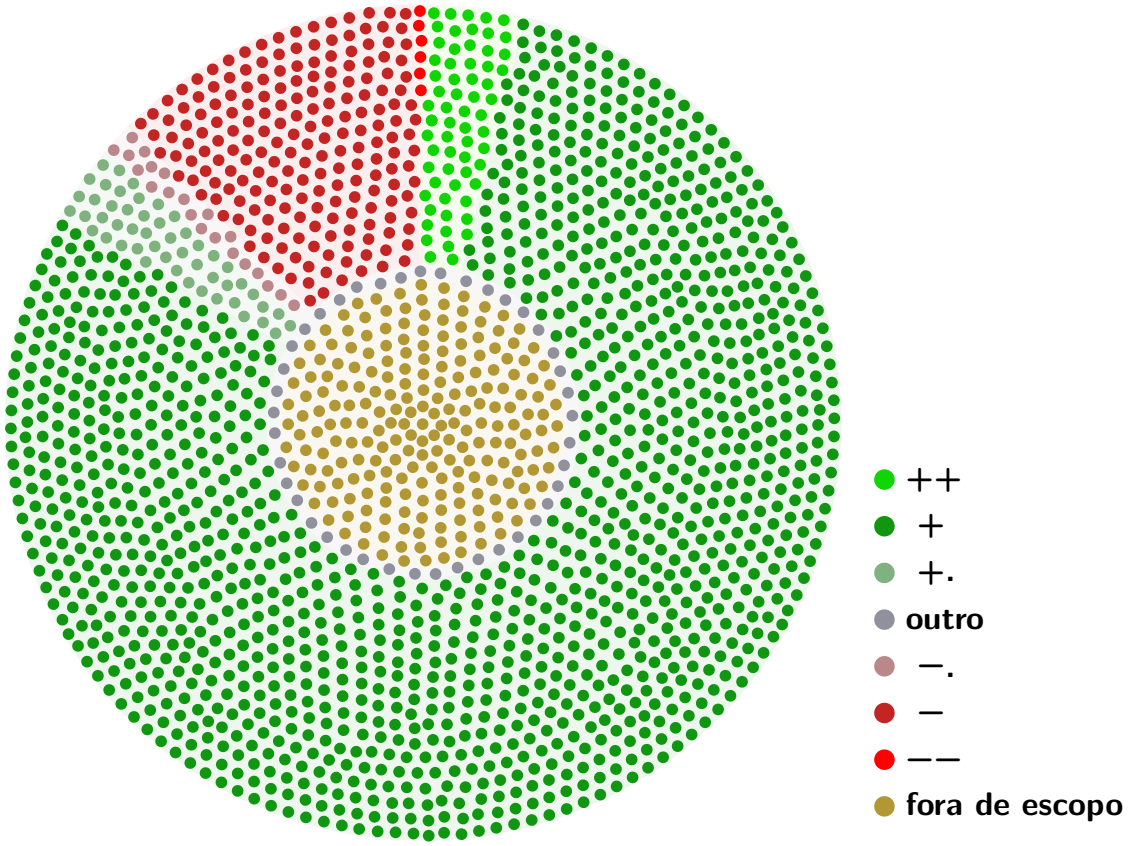
A Tabela A.3 mostra a contagem de aspectos identificados, sentenças e opiniões para cada entidade após a extensão, limpeza e simplificação do córpus.

Dos 1902 trechos de texto anotados, 88% foram marcados como úteis para fins de avaliação do produto em questão. Desses, 84% são opiniões positivas e 13% são negativas; as opiniões marcadas como fortes ou fracas contribuem com 8% das opiniões. O Gráfico A.1 mostra uma visualização de todas as opiniões contidas no conjunto de dados; cada ponto representa um trecho de texto identificado pelos anotadores, sendo mostrados os 1902 trechos.

Tabela A.3 – Quantitativo do conjunto de dados simplificado e estendido.

tipo	nome	entidade	aspectos	sentenças	positivas	negativas	visualização
celular	**D1**	D1a Motorola Moto G5 Plus	15	269	**346**	**101**	
		D1b Galaxy S7	14	253	**342**	**91**	
câmera	**D2**	D2a Canon EOS Rebel T5	13	68	**77**	**11**	
		D2b Canon PowerShot SX520 HS	15	52	**68**	**8**	
celular	**D3**	D3a *(subconjunto de D1a)*	11	150	**143**	**65**	
		D3b *(subconjunto de D1b)*	10	109	**85**	**65**	
celular	**D4**	D4a *(subconjunto de D1a)*	13	43	**56**	**13**	
		D4b *(cópia de D1b)*	14	253	**342**	**91**	
câmera	**D5**	D5a *(subconjunto de D2a)*	12	39	**40**	**10**	
		D5b *(subconjunto de D2b)*	10	30	**37**	**3**	
câmera	**D6**	D6a *(subconjunto de D2a)*	8	29	**37**	**1**	
		D6b *(subconjunto de D2b)*	11	22	**31**	**5**	
câmera	**D7**	D7a *(subconjunto de D2a)*	4	31	**33**	**6**	
		D7b *(subconjunto de D2b)*	4	25	**22**	**4**	
celular	**D8**	D8a *(subconjunto de D1a)*	12	39	**62**	**10**	
		D8b *(subconjunto de D1b)*	12	32	**36**	**15**	

Gráfico A.1 – Visualização de todas as opiniões contidas no conjunto de dados.



Os tipos de opinião mais raros no conjunto foram as relativas e as duais[5], com apenas 2 ocorrências cada. Houve 11 ocorrências de trechos irresolutos, 10 de opiniões medianas, 10 de experiência de uso e 6 de conselhos. A Tabela A.4 mostra a frequência das polaridades dos trechos opinativos.

Tabela A.4 – Frequência das polaridades.

Frequência de polaridades			
polaridade		frequência	
++	positivo forte	3,8%	64
+	positivo	77,4%	1301
+.	positivo fraco	3,0%	50
−.	negativo fraco	1,6%	21
−	negativo	11,8%	199
−−	negativo forte	0,4%	6
.	mediano	0,6%	10
*	relativo	0,1%	2
..	dual	0,1%	2
#	irresoluto	0,7%	11
!	conselho	0,4%	6
&	experiência	0,6%	10
	total	100%	1682

Dos 217 trechos não avaliativos ou sem utilidade, 25% são duplicados, 29% são sobre o serviço, 21% são textos irrelevantes ou de contextualização e 24% apresentam problemas de texto (quebrados ou ininteligíveis).

A.6 Disponibilização

O córpus pode ser acessado pela página do Projeto Opinando (sites.google.com/icmc.usp.br/opinando). Neste trabalho de mestrado, foi usada a versão 1.0.0 do córpus, disponível em github.com/raphsilva/corpus-4p/releases/tag/1.0.0. Duas variações do córpus estão disponíveis:

- Uma versão limpa e estendida, em formato JSON, recomendada para uso direto em processamento de problemas do mesmo tipo que a sumarização contrastiva.
- Uma versão da anotação manual das quatro entidades, que inclui os comentários não pertinentes devidamente marcados como tal e todas as informações da anotação, bem como o script usado para converter a anotação para o formato final do córpus. Essa versão é recomendada para análise manual e para eventuais trabalhos derivados.

[5] Observaram-se alguns casos onde havia opiniões positivas e negativas sobre um mesmo aspecto e os anotadores preferiram separá-las, já que haviam sido orientados a fazer isso sempre que possível. Por exemplo, para a sentença *'a tela não tem o melhor contraste de cores, mas a nitidez é imbatível'*, os anotadores identificaram uma opinião positiva sobre a tela e uma negativa sobre a tela.

A.7 Uso do córpus

Além do uso prático para avaliar sistemas de Análise de Sentimentos, a construção do córpus proporcionou uma experiência que permitiu sistematizar melhor a identificação de opiniões. Essa tarefa, se feita automaticamente, costuma classificar as opiniões em apenas três classes: positiva, negativa e neutra. Com a análise manual, foi possível refinar melhor essa classificação e obter informações que podem ser úteis em futuros trabalhos.

Os números descobertos com a análise do córpus mostram que quase 90% dos trechos encontrados na fonte são úteis para fins de avaliação de produtos, e 70% de todos os trechos foram marcados com polaridade positiva ou negativa pelos anotadores, número que sobe para 86% se se considerarem as positivas e negativas fortes e fracas. Apenas 2% das opiniões são classificadas como outro tipo, o que faz ponderar se realmente vale a pena se preocupar em formas eficientes e sistematizadas de classificar esse tipo de texto em aplicações reais.

Este projeto permitiu obter um córpus que simule um conjunto de dados automaticamente processado com a vantagem de ele ser livre de ruídos. Espera-se que o córpus construído contribua com outros projetos e que as ideias apresentadas aqui somem valor à pesquisa em Análise de Sentimentos e Processamento de Linguagem Natural em língua portuguesa.

APÊNDICE

B

INSTRUÇÕES PARA A AVALIAÇÃO HUMANA

Este apêndice mostra a folha de instruções fornecida aos voluntários para avaliação do resumo, como relatado na Seção 5.2.4. As duas páginas contendo instruções foram impressas em uma única folha (frente e verso) de tamanho A4 e entregue aos voluntários junto com os resumos, também impressos, cada um em uma folha de tamanho A5.

Avaliação da utilidade de resumos contrastivos de opinião – FOLHA DE INSTRUÇÕES

Raphael Rocha da Silva
Universidade de São Paulo
São Carlos, março de 2019

Leia esta folha por completo (frente e verso) antes de começar.

Nesta tarefa, você vai avaliar resumos contrastivos de opinião. Esse tipo de resumo tem como objetivo apresentar de maneira sucinta informações encontradas em textos opinativos sobre dois produtos de forma a facilitar a comparação entre eles.

Você está recebendo um ou mais blocos com seis resumos cada um. Cada bloco contém resumos sobre um par de produtos. Esses produtos podem ser celulares ou câmeras. Dentro de cada bloco, cada resumo foi obtido a partir do mesmo conjunto de textos opinativos porém com uso de métodos diferentes.

A primeira linha de cada quadro contém o código identificador do resumo (pode ser ignorado). A linha seguinte identifica o tipo de produto do qual os textos opinativos foram obtidos. Abaixo dela, figura o resumo propriamente dito.

As frases contidas no resumo foram extraídas do conjunto de textos de maneira independente. Por isso, elas são desconexas, isto é, uma frase não necessariamente tem relação com a frase anterior ou posterior. Isso é comum nesse tipo de resumo.

Os textos foram escritos por várias pessoas. Por isso, é esperado haver opiniões discordantes.

Avaliação

Esta avaliação pretende identificar quão útil cada resumo é. Você vai atribuir uma nota em uma escala de −3 a +3 para cada resumo. Para atribuir a nota, responda à seguinte pergunta, e escolha a nota de acordo com as alternativas listadas abaixo:

> **Se você deseja comprar um dos dois produtos dos quais trata este resumo, quanto este resumo te ajuda a tomar uma decisão de compra?**
> **−3**: o resumo atrapalha muito a tomada de decisão
> **−2**: o resumo atrapalha a tomada de decisão
> **−1**: o resumo atrapalha um pouco a tomada de decisão
> **0**: o resumo nem ajuda nem atrapalha a tomada de decisão
> **1**: o resumo ajuda um pouco a tomada de decisão
> **2**: o resumo ajuda a tomada de decisão
> **3**: o resumo ajuda muito a tomada de decisão

Ao avaliar um resumo, considere apenas o conteúdo das informações contidas nele. Despreze quaisquer outros fatores, como: formatação, tamanho e erros gramaticais.

Para facilitar o trabalho, fique à vontade para rabiscar os resumos marcando pontos que te ajudam a julgá-lo.

É esperado que você compare os resumos entre si, e não apenas avalie cada resumo isoladamente; assim, pretende-se que resumos melhores (na sua opinião) recebam pontuações melhores.

Procedimento

Para avaliar os resumos, você seguirá o procedimento abaixo. Se houver qualquer problema, comunique o condutor do experimento antes de continuar.

1. Escolha um bloco de resumos.
2. Leia todos os resumos do bloco. Marque pontos relevantes em cada um deles. Ainda não atribua notas.
3. Escolha um resumo para avaliar.
4. Leia o resumo. Identifique pontos que contribuem com a utilidade do resumo.
5. Atribua uma nota de −3 a +3 à utilidade do resumo (como especificado na página anterior); escreva-a no lugar indicado. Para atribuir a nota, considere também os outros resumos do mesmo bloco: resumos melhores devem ter notas melhores.
6. Se você tiver uma justificativa para a nota que atribuiu, escreva-a. Você também pode escrever sugestões dizendo como o resumo poderia ser melhor e comentários comparando o resumo a outros do mesmo bloco.
7. Volte ao passo **5** até completar todos os resumos do bloco.
8. Volte ao passo **2** até completar todos os blocos.

Depois da avaliação, responda ao questionário que lhe foi entregue.

Após terminar, você devolverá ao condutor todas as folhas que recebeu.

Guia da escala de notas:

Se você deseja comprar um dos dois produtos dos quais trata este resumo, quanto este resumo te ajuda a tomar uma decisão de compra?

−3: o resumo atrapalha muito a tomada de decisão
−2: o resumo atrapalha a tomada de decisão
−1: o resumo atrapalha um pouco a tomada de decisão
0: o resumo nem ajuda nem atrapalha a tomada de decisão
1: o resumo ajuda um pouco a tomada de decisão
2: o resumo ajuda a tomada de decisão
3: o resumo ajuda muito a tomada de decisão


ICMC
USP
SÃO CARLOS