UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Agrupamento semântico de aspectos para mineração de opinião

Francielle Alves Vargas

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

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Francielle Alves Vargas

Agrupamento semântico de aspectos para mineração de opinião

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestra em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA.*

Área de Concentração: Ciências de Computação e Matemática Computacional.

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo.

USP – São Carlos
Novembro de 2017

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

AL864sa

Alves Vargas, Francielle
Agrupamento semântico de aspectos para mineração de opinião / Francielle Alves Vargas; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2017.
126 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, 2017.

1. Processamento de Linguagem Natural. 2. Mineração de Opinião. I. Salgueiro Pardo, Thiago Alexandre, orient. II. Título.

Francielle Alves Vargas

Semantic clustering of aspects for opinion mining

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC- USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY.*

Concentration Area: Computer Science and Computational Mathematics.
Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo.

USP – São Carlos
November 2017

Uma "coisa", no sentido tradicionalmente amplo, entende-se como algo que de algum modo é. Nessa acepção, um deus é também uma coisa. Somente quando se encontra a palavra para a coisa, a coisa torna-se algo. Nenhuma coisa é, se não for pela palavra. É a palavra que confere ser às coisas. Mas como pode uma simples palavra fazer isso, conferir ser a alguma coisa?

—Heidegger

A única fonte de conhecimento é a experiência.

—Albert Einstein

Agradecimentos

Gostaria de agradecer primeiramente ao meu orientador, Prof.Thiago Pardo, pela confiança, generosidade e a orientação ao longo deste projeto, as professoras do NILC, Graça Nunes, Sandra Aluiso e Ariani Di Felippo por representarem com força, competência e elegância o PLN no Brasil, e aos professores do ICMC, Gustavo Batista e Dilvan Moreira pelas aulas não menos que excelentes.

Aos familiares pelo apoio e conforto durante esses dois anos de trabalho e estudo, especialmente ao vovô Perillo (em memória).

Aos meus queridos amigos do ICMC, Isabelle Carvalho, Guilherme Ponteciano, Danillo Reis, Yuri Magagnatto e Ronnie Shida por terem compartilhado comigo amizade e conhecimentos.

À Dona Ana e a Aurinha pelos sorrisos sempre acolhedores e o cafezinho oferecido com carinho e gentileza nas tardes no instituto, e a equipe da secretaria de pós graduação e os demais funcionários do ICMC pelo suporte prestado.

Por fim, agradeço à CAPES pelo apoio financeiro.

RESUMO

VARGAS, F. A. **Agrupamento semântico de aspectos para mineração de opinião.** 2017. 126p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Com o rápido crescimento do volume de informações opinativas na web, extrair e sintetizar conteúdo subjetivo e relevante da rede é uma tarefa prioritária e que perpassa vários domínios da sociedade: político, social, econômico, etc. A organização semântica desse tipo de conteúdo, é uma tarefa importante no contexto atual, pois possibilita um melhor aproveitamento desses dados, além de benefícios diretos tanto para consumidores quanto para organizações privadas e governamentais. A área responsável pela extração, processamento e apresentação de conteúdo subjetivo é a mineração de opinião, também chamada de análise de sentimentos. A mineração de opinião é dividida em níveis de granularidade de análise: o nível do documento, o nível da sentença e o nível de aspectos. Neste trabalho, atuou-se no nível mais fino de granularidade, a mineração de opinião baseada em aspectos, que consiste de três principais tarefas: o reconhecimento e agrupamento de aspectos, a extração de polaridade e a sumarização. Aspectos são propriedades do objeto avaliado e podem ser implícitos e explícitos. Reconhecer e agrupar aspectos são tarefas críticas para mineração de opinião, no entanto, também são desafiadoras. Por exemplo, em textos opinativos, usuários utilizam termos distintos para se referir a uma mesma propriedade do objeto. Portanto, neste trabalho, focamos no problema de agrupamento de aspectos para mineração de opinião. Para resolução deste problema, optamos por uma abordagem linguística. Investigou-se os principais fenômenos intrínsecos e extrínsecos em textos opinativos a fim de encontrar padrões linguísticos e insumos acionáveis para proposição de métodos automáticos de agrupamento de aspectos correlatos para mineração de opinião. Nós propomos, implementamos e comparamos seis métodos automáticos baseados em conhecimento linguístico para a tarefa de agrupamento de aspectos explícitos e implícitos. Um método inédito foi proposto para essa tarefa que superou os demais métodos implementados, especialmente o método baseado em léxico de sinônimos (baseline) e o modelo estatístico com base em \textit{word embeddings}. O método proposto também não é dependente de uma língua ou de um domínio, no entanto, focou-se no português do brasil e no domínio de produtos da web.

Palavras-chave: Mineração de opinião baseada em aspectos; Agrupamento de aspectos; Processamento de linguagem natural.

ABSTRACT

VARGAS, F. A. **Semantic clustering of aspects for opinion mining**. 2017. 126p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

With the growing volume of opinion information on the web, extracting and synthesizing subjective and relevant content from the web has to be shown a priority task that passes through different society domains, such as political, social, economical, etc. The semantic organization of this type of content is very important nowadays since it allows a better use of those data, as well as it benefits customers and both private and governmental organizations. The area responsible for extracting, processing and presenting the subjective content is opinion mining, also known as sentiment analysis. Opinion mining is divided into granularity levels: document, sentence and aspect levels. In this research, the deepest level of granularity was studied, the opinion mining based on aspects, which consists of three main tasks: aspect recognition and clustering, polarity extracting, and summarization. Aspects are the properties and parts of the evaluated object and it may be implicit or explicit. Recognizing and clustering aspects are critical tasks for opinion mining; nonetheless, they are also challenging. For example, in reviews, users use distinct terms to refer to the same object property. Therefore, in this work, the aspect clustering task was the focus. To solve this problem, a linguistic approach was chosen. The main intrinsic and extrinsic phenomena in reviews were investigated in order to find linguistic standards and actionable inputs, so it was possible to propose automatic methods of aspect clustering for opinion mining. In addition, six automatic linguistic-based methods for explicit and implicit aspect clustering were proposed, implemented and compared. Besides that, a new method was suggested for this task, which surpassed the other implemented methods, specially the synonym lexicon-based method (baseline) and a word embeddings approach. This suggested method is also language and domain independent and, in this work, was tailored for Brazilian Portuguese and products domain.

Keywords: Aspect-based opinion mining; Aspect clustering; Natural language processing.

Sumário

Lista de Abreviaturas

AM	Aprendizagem de Máquina.
AO	Aprendizagem de Ontologia.
EI	Extração de Informação.
EM	*Expectation Maximization.*
FR	Frequência Relativa.
IA	Inteligência Artificial.
LDA	*Latent Dirichlet Allocation.*
PLN	Processamento de Línguas Naturais.
RI	Recuperação de Informação.
SBC	Sistemas Baseados em Conhecimento.
SN	Sintagma Nominal.
SVM	*Support Vector Machine.*

Lista de Algoritmos

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

1.1 Contextualização

Com o crescimento do volume de informações opinativas na web, extrair conteúdo subjetivo e relevante da rede é uma tarefa prioritária e que perpassa vários domínios da sociedade: político, social, econômico, etc. Por exemplo, no âmbito político, minerar conteúdo subjetivo a partir de comentários de usuários sobre as eleições poderia auxiliar na tomada de decisão de melhores estratégias de campanha, ou até mesmo predizer candidatos políticos mais propensos à eleição. No âmbito econômico, de acordo com Yu *et al.* (2011), consumidores geralmente procuram informações de qualidade em revisões de usuários antes da tomada de decisão de compra de um produto, enquanto que algumas empresas usam revisões de usuários como um importante recurso no desenvolvimento e melhoramento de seus produtos, marketing e gestão de relacionamento com clientes. Entretanto, realizar manualmente análise de revisões de consumidores sobre um determinado produto pode levar muito tempo que certamente impactaria de forma negativa o processo de decisão de compra. Portanto, prover de forma organizada conteúdo subjetivo relevante de revisões, dentre esses vários domínios, é uma tarefa importante no contexto atual, pois provê um melhor aproveitamento desses dados, tanto para consumidores quanto para organizações privadas e governamentais.

A área responsável pela extração de conteúdo subjetivo de textos é a mineração de opinião, também chamada de análise de sentimentos. De acordo com Pang *et al.* (2002), mineração de opinião é a tarefa de analisar e classificar as informações subjetivas e os sentimentos associados a um alvo específico. Os termos "mineração de opinião" e "análise de sentimentos" são usados na literatura de forma intercambiada e ambos designam a função de extrair com precisão, de forma automática, características e conteúdo subjetivo de textos. No entanto, para Munezero *et al.* (2014), é necessário que haja maior prudência na utilização desses termos, para que sejam

aplicados corretamente em cada um de seus contextos. Uma das distinções consite na compreensão dos termos *opinião* e *sentimento*. Os autores defendem que opiniões são interpretações pessoais de informações formadas na mente e não necessariamente contêm expectativas sociais. Os sentimentos são, no entanto, socialmente construídos. Além disso, os termos *análise* e *mineração* também merecem atenção quanto as suas peculiariedades. O termo "mineração" está associado a área de mineração de dados, que, de acordo com Fayyad *et al.* (1996), historicamente refere-se à noção de encontrar padrões úteis em dados. Além do termo "mineração de dados", também são empregados os termos "extração de conhecimento", "descoberta de informações", entre outros. A mineração de dados também ganhou popularidade em domínios como banco de dados e na inteligência artificial com a aprendizagem de máquina. A frase "descoberta de conhecimento em bancos de dados" foi cunhado no primeiro *KDD* [1] *workshop* em 1989, para enfatizar que o conhecimento é o produto final de uma descoberta baseada em dados. Portanto, a mineração de opinião herda da mineração de dados a conceitualização inerente à descoberta de conhecimento em bases de dados. No entanto, diferentemente da mineração de dados, que pode abarcar várias fontes de dados distintas (imagem, vídeo, texto, etc), a mineração de opinião atua majoritariamente com dados do tipo texto e com o principal objetivo de reconhecer, tratar, extrair e sumarizar conteúdo subjetivo deste tipo de dado. O termo "análise", por outro lado, é genérico e pode ou não se referir à busca por padrões e descoberta de conhecimento a partir de uma base de dados. Neste trabalho, optamos pelo termo "mineração de opinião" por compreendermos que representa com maior precisão nossa proposta de trabalho.

Mineração de opinião é o campo de pesquisa responsável por propor métodos de análise, processamento, sumarização e classificação de grandes volumes de dados, majoritariamente do tipo texto, para a extração de conteúdo subjetivo. Segundo Liu (2012), existem níveis de granularidade de análise para a mineração de opinião. São eles: (i) nível do documento, (ii) nível da sentença e (iii) nível de aspectos. No nível do documento, contabiliza-se a somatória das opiniões expressas no documento. Por exemplo, um documento composto por conteúdo opinativo é classificado em positivo, negativo ou neutro, de acordo com a contabilização do conteúdo relevante que expressa sentimento. De acordo com Liu (2012), estes conteúdos são expressos geralmente através de adjetivos. Na Figura 1.1, apresentamos um conjunto de revisões de usuários sobre um smartphone.

Note que foram emitidas 1.414 revisões para o produto smartphone. Cada uma das 1.414 revisões refere-se a um documento para sistemas de mineração de opinião. Portanto, no nível de análise do documento, um *score* positivo, negativo ou neutro é emitido para cada documento. Observe também que, neste nível de análise, não é possível saber com precisão o que o usuário gostou ou não. No nível da sentença, o objetivo é determinar a opinião expressa em cada uma das sentenças do documento. Portanto, um conjunto de documentos é segmentado em sentenças e, em seguida, um *score* é emitido para cada uma dessas sentenças. Por exemplo, em um documento composto por x sentenças, para cada uma haverá uma classificação positiva, negativa ou neutra. Vejamos novamente a Figura 1.1. Na terceira revisão, o usuário emite a

[1] KDD é uma conferência internacional sobre descoberta de conhecimento em banco de dados e mineração de dados.

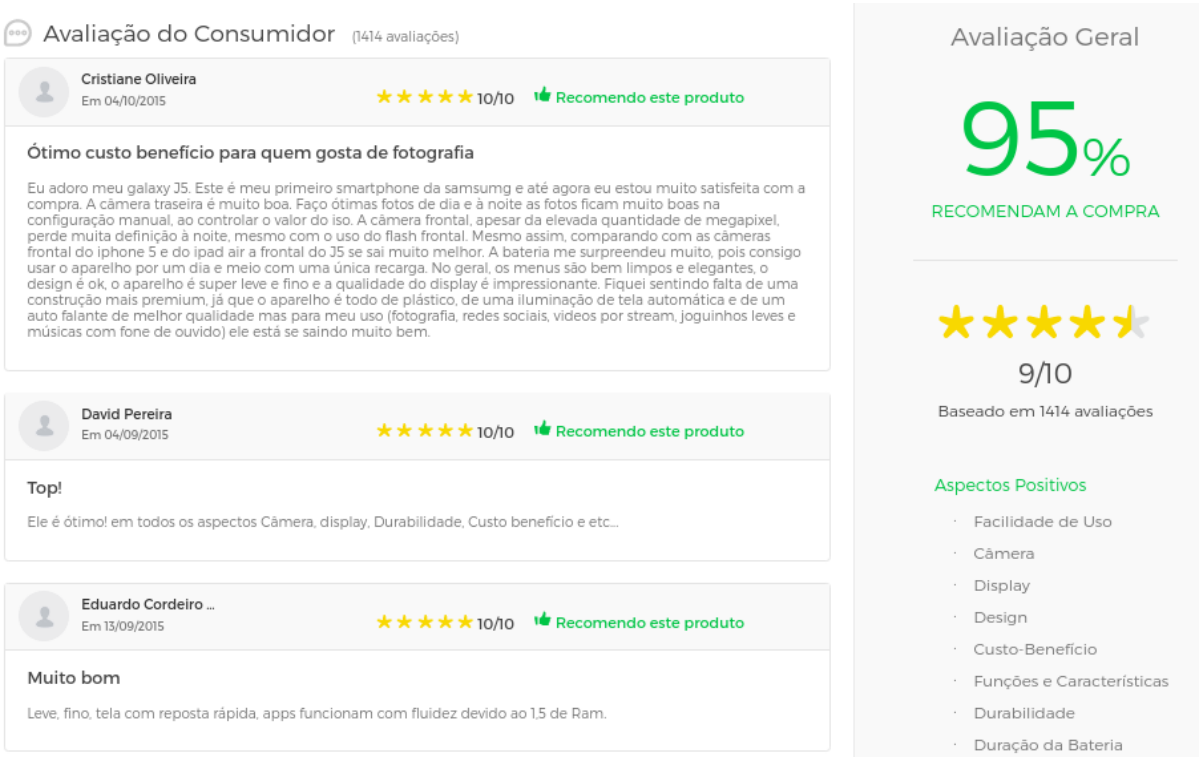

Figura 1.1: Conjunto de revisões de usuários sobre um smartphone e extraído de *Buscape.com*.

seguinte avaliação “Leve, fino, tela com resposta rápida, *apps* funcionam com fluidez devido ao 15 de Ram”. Note que neste nível de análise ainda não é possível saber com precisão as características do produto avaliada pelo usuário. Para solucionar esse problema, Liu *et al.* (2005) argumenta a necessidade de um nível mais fino de análise: a mineração de opinião baseada em aspectos.

Em mineração de opinião baseada em aspectos, os aspectos podem ser encontrados de modo explícito e/ou implícito. Aspectos explícitos são realizações explícitas inerentes à avaliação de uma ou mais propriedades do objeto/alvo da opinião. Por exemplo, vejamos um exemplo de aspectos explícitos e implícitos nas revisões exibidas nas Figura 1.2 e 1.3.

Na revisão da Figura 1.2, os termos “história” e “romancezinho” são usados de forma intercambiada para avaliar a propriedade “história” da entidade livro. Observe que se trata de aspectos explícitos, diferentemente dos termos “menininha” e “cara”, que são termos indicativos de aspectos implícitos. Esses termos podem ser classificados como “termos pistas”, pois indicam a ocorrência de um aspecto implícito. Na revisão exibida na Figura 1.3, os aspectos explícitos são “preço”, “câmera digital”, “vídeo”, “espaço de memória” e “som”. Observe que o aspecto “sinal” também é avaliado pelo usuário nesta revisão, no entanto, foi avaliado de forma implícita. Por exemplo, o usuário, ao utilizar a expressão “recebi chamadas até na beira do Rio Paraná”, está avaliando a propriedade “sinal” do smartphone.

Além dos desafios de reconhecimento de conteúdo implícito e explícito, revisões de usuários são, muitas vezes, numerosas e de difícil compreensão. De acordo com Yu *et al.* (2011), é impraticável para o usuário compreender a visão geral das opiniões de outros usuários sobre todos os aspectos de um produto, por exemplo, em função do grande número de revisões

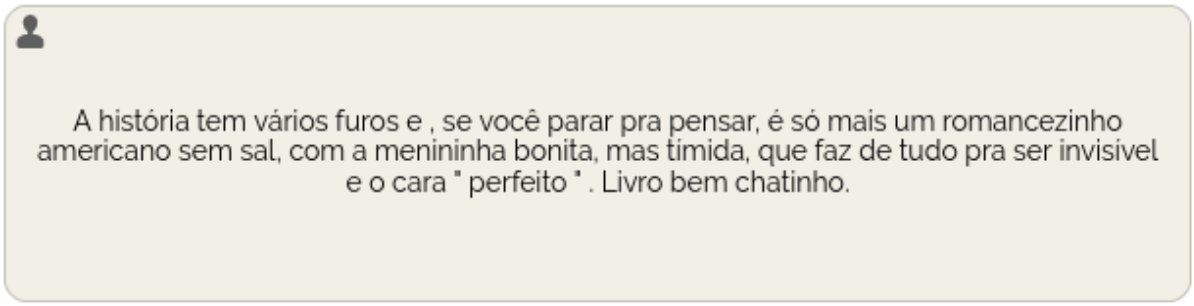

Figura 1.2: Revisão de usuário sobre o livro *Crepúsculo*.

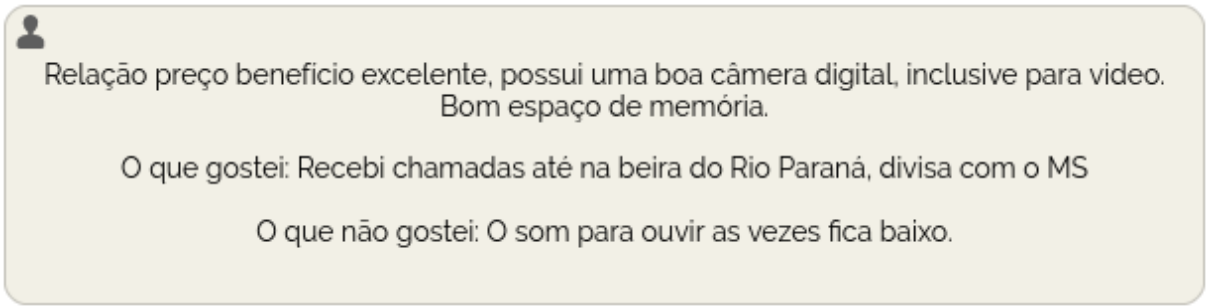

Figura 1.3: Revisão de usuário sobre um smartphone.

emitidas para cada produto. Além disso, ainda segundo o autor, revisões de usuários são desorganizadas e conduzem à dificuldade de navegação de informações e aquisição de conhecimento. Portanto, a organização semântica do conteúdo em avaliações de consumidores sobre os vários aspectos de um produto e as suas respectivas opiniões pode possibilitar ao usuário compreender facilmente a visão geral de opiniões de outros consumidores sobre um produto.

Outro desafio ao trabalhar com textos opinativos é a abundância lexical empregada pelos usuários para se referirem a uma mesma propriedade do objeto ou alvo da opinião. Alvo da opinião é a entidade ou objeto principal avaliado. Por exemplo, na revisão “Achei o preço da câmera caro”, o usuário emprega o termo “preço” para avaliar uma propridade da câmera. No entanto, para avaliar essa mesma propriedade da câmera, os usuários também podem utilizar os termos “custo”, “valor”, “investimento”, “acessível” e “custo-benefício”. Além disso, os usuários podem usar indicativos de aspectos implícitos para se referirem a uma mesma propriedade do objeto. Por exemplo, as expressões “Recebi chamadas até na beira do rio são francisco” e “funciona em qualquer lugar” foram empregadas para avaliar a propriedade “sinal” do aparelho smartphone. Outro exemplo é o termo indicativo de aspecto implícito “compatibilidade”, que foi usado para avaliar o sistema operacional de um smartphone. Concomitante com esse termo, também foram usados os termos “programa”, “sistema” e “aplicação”. Além disso, há uma porção significativa de nomes prórios usados pelos usuários para se referirem a uma mesma propriedade do objeto avaliado. Por exemplo, os nomes “edward”, “edward cullen”, “noelle page”, “larry” e “bella” são usados para avaliar o aspecto “personagem” ou “protagonista” no domínio de livro. E os nomes “josé saramago” e “thalita rebouças” são usados para avaliar a propriedade “autor” do livro. No domínio de câmera, os termos “sony”, “nikon”, “fuji” e “benq” são usados para avaliar a propriedade “marca” de uma câmera digital.

Portanto, em sistemas de mineração de opinião, é fundamental o agrupamento de aspectos correlatos, ou seja, aspectos usados para se referir a uma mesma propriedade do objeto avaliado, pois a ausência desse tipo de agrupamento pode representar falhas na apresentação dos resultados sobre as reais propriedades avaliadas pelos usuários.

Neste trabalho, portanto, focamos no problema de agrupamento de aspectos para mineração de opinião. Em decorrência da complexidade inerente ao domínio de opinião, realizamos um estudo empírico e de aprofundamento linguístico sobre os principais fenômenos em revisões de usuários a fim de encontrar padrões e insumos linguísticos acionáveis em textos opinativos. Além disso, propomos e implementamos métodos automáticos enrriquecidos linguisticamente para a tarefa de agrupamento de aspectos explícitos e termos indicativos de aspectos implícitos. Propomos também um algoritmo inédito de agrupamento de aspectos que superou os demais métodos implementados, especialmente o método baseado em léxico de sinônimos (baseline da literatura) e o método estatístico baseado em *word embeddings*. Esse método também não é dependente de uma lígua, no entanto, nosso focamos no português do brasil.

1.2 Lacunas, hipóteses e objetivos

De acordo com Yu *et al.* (2011), a organização semântica do conteúdo em avaliações de consumidores sobre os vários aspectos de um produto e as suas respectivas opiniões permite que o usuário facilmente compreenda a visão geral de opiniões de outros consumidores sobre um produto. Para a área de mineração de opinião, a investigação de métodos automáticos de organização semântica de aspectos pode provê vários benefícios. Várias pesquisas da área (Lu & Zhai (2008), Cadilhac *et al.* (2010), Yu *et al.* (2011), Freitas & Vieira (2013)), demonstraram que a organização semântica de aspectos melhora a acurácia da mineração de opinião. No entanto, não há, para o português, trabalhos que tratam o problema de organização semântica de aspectos a partir de textos opinativos. Portanto, essa é uma lacuna crítica, no contexto atual, em que muita informação opinativa é produzida na web. Além disso, a exploração de métodos automáticos motivados linguisticamente ainda é pouco representativa na literatura, em detrimento de métodos estatísticos. Por exemplo, não encontramos nenhum trabalho aprofundado sobre os principais fenômenos intrísecos e extrísecos da língua em textos opinativos. Esse tipo de exploração linguística é importante, pois pode prover insumos acionáveis para proposição de métodos automáticos e mais adptáveis ao contexto e à situação social.

A partir das lacunas identificadas na área de mineração de opinião, levantamos as três principais hipóteses que regem este trabalho de mestrado. São elas:

Hipótese 1: A principal hipótese que motivou este trabalho é que a partir de um estudo aprofundado dos principais fenômenos que acometem textos opinativos é possível propor métodos melhorados para mineração de opinião;

Hipótese 2: A segunda hipótese deste trabalho consiste na descoberta de conhecimento em textos opinativos. Acreditamos que, a partir da exploração de córpus, é possível descobrir

e extrair o conhecimento necessário para o mapeamento semântico desse domínio;

Hipótese 3: A terceira hipótese consiste nas especificidades de cada domínio. Acreditamos que haja um conjunto representativo de características semânticas que exprimam as especificidades de um domínio, e que essas especificidades sofrem influências de fatores intrínsecos e extrínsecos da língua.

A partir do mapeamento de lacunas e o levantamento de hipóteses, apresentamos, portanto, os dois principais objetivos desta proposta de mestrado:

- Realizar um estudo exploratório e mapear fenômenos relevantes estatisticamente em textos opinativos a fim de encontrar padrões e insumos linguísticos acionáveis para proposição e implementação de métodos automáticos para sistemas de mineração de opinião baseado em aspectos.
- Propor e implementar métodos de agrupamento de aspectos explícitos e implícitos, a partir de textos opinativos, fortemente motividados linguisticamente.

1.3 Metodologia de trabalho

A metodologia adotada neste trabalho consistiu de quatro macro processos: (i) investigação linguística; (ii) proposição de métodos; (iii) implementação; (iv) avaliação; conforme exibido na Figura 1.4.

Investigação linguística

Nesta etapa, um estudo empírico a partir de córpus e um estudo linguístico aprofundado foi realizado. Nosso objetivo com essa etapa foi compreender os principais fenômenos linguísticos estatisticamente relevantes que acometem textos opinativos. A partir deste estudo, foi possível encontrar insumos linguísticos acionáveis para proposição de métodos automáticos para a mineração de opinião.

Proposição de métodos

A partir dos resultados obtidos com os estudos empírico e de aprofundamento linguístico, foi possível compreender os principais fenômenos em textos opinativos e, a partir da identificação e modelagem desses fenômenos, nós propomos e implementamos seis métodos automáticos para a tarefa de agrupamento de aspectos, fortemente motivados linguisticamente, sendo que um desses métodos é inédito e oriundo das investigações realizadas neste trabalho de mestrado.

Implementação

A partir da proposição dos métodos, selecionamos uma linguagem de programação para implementação e, em seguida, realizamos alguns testes sobre os métodos implementados. Os

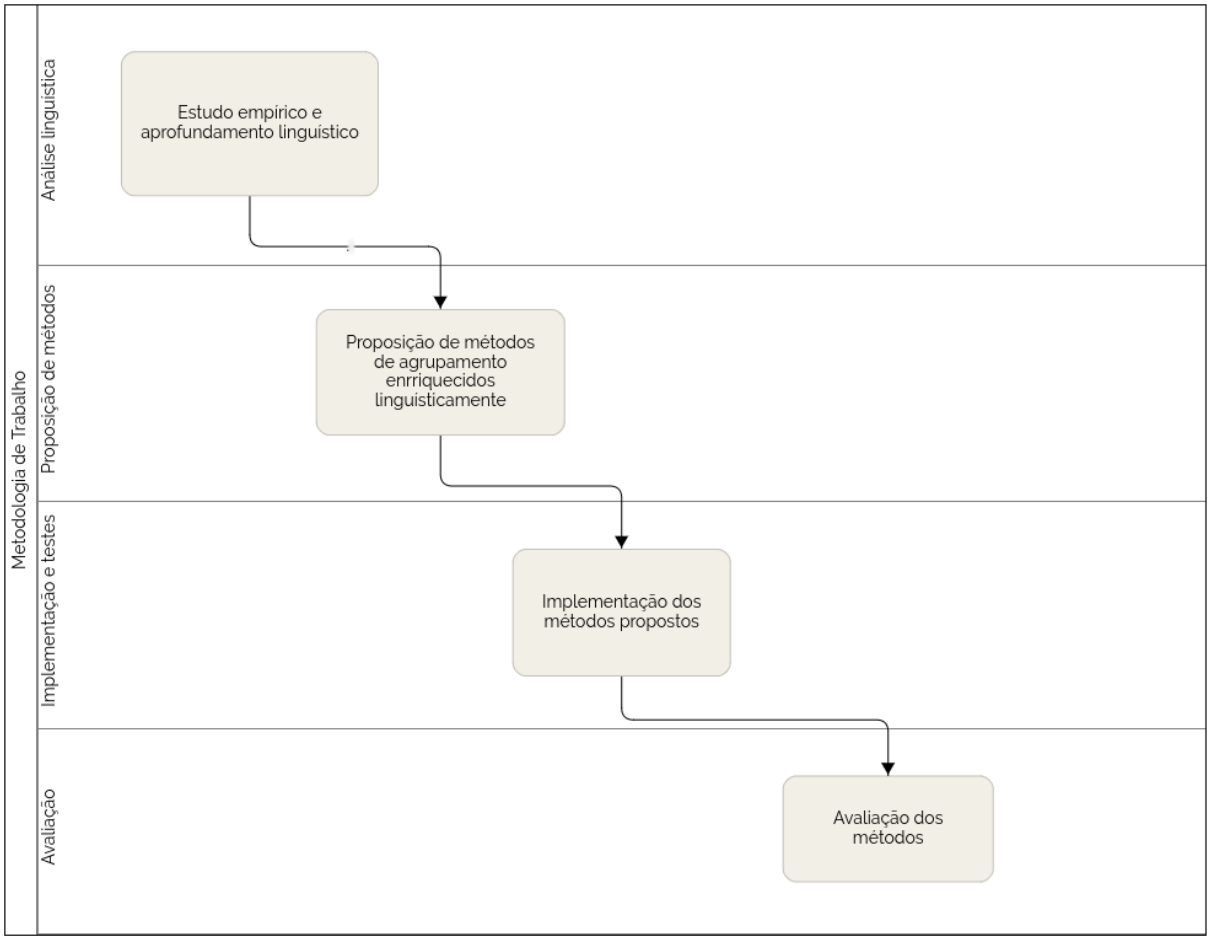

Figura 1.4: Metodologia.

testes serviram para avaliação de cada processo inerente ao método. Por exemplo, para testar o método que extrai relações de sinonímia em revisões de usuários, a partir de um recurso lexical, utilizamos a versão online deste recurso para validação de uma amostra dos resultados obtidos com o método. Ou seja, se a entrada do método implementado é o aspecto "livro", verificamos se os *sinônimos* retonados pelo método eram compatíveis com os *sinônimos* retornados a partir desta mesma entrada, porém na versão online do recurso lexical.

Avaliação

Finalizada a implementação e os testes dos seis métodos, nós utilizamos métricas de avaliação da literatura, como *precisão, cobertura, medida-f e medida-f global*, para avaliação dos resultados obtidos com cada um dos seis método implementados.

1.4 Contribuições

1.4.1 Contribuições teóricas

A principal contribuição teórica deste trabalho é a investigação linguística aprofundada e de viés empírico sobre os principais fenômenos que acometem textos opinativos, especialmente revisões de usuários sobre produtos da web. Caracterizamos e modelamos textos opinativos dos domínios de smartphone, câmera e livro a fim de encontrar padrões e insumos linguísticos acionáveis para proposição de métodos automáticos e fortemente motivados linguisticamente para mineração de opinião. Além disso, a partir da investigação deste trabalho de mestrado, foram publicados, até o momento, 2 relatórios técnicos e 2 *papers* em conferências da área.

1.4.2 Constribuições práticas

A partir das investigações realizadas neste trabalho de mestrado foram desenvolvidos vários recursos computacionais práticos. Alguns deles são: o algoritmo OpCluster-PT, hierarquias de aspectos em formato OWL sobre os domínios de smartphone, câmera e livro, 180 revisões de usuários anotadas quanto ao aspectos explícitos, indicativos de aspectos implícitos e seus grupos, entre outros recursos.

1.5 Estruturação do documento

Estruturamos este documento da seguinte forma: no Capítulo 2, apresentamos a fundamentação teórica e os recursos e ferramentas linguístico-computacionais aplicados nos experimentos; no Capítulo 3, faremos uma discussão sobre a tarefa de agrupamento de aspectos, descrevemos as principais abordagens da literatura e os trabalhos da relacionados; no Capítulo 4, apresentaremos um estudo de córpus e um estudo de aprofundamento linguístico a partir de textos opinativos; no Capítulo 5, serão descritos e discutidos os experimentos; no Capítulo 6,

enfim, apresentaremos os resultados obtidos com os experimentos; e, no Capítulo 7, discorremos sobre as considerações finais, além das limitações e possibilidades de trabalhos futuros.

Capítulo

2

Fundamentação teórica, ferramentas e recursos linguístico-computacionais

Neste capítulo, apresentaremos o principal eixo teórico desta proposta de trabalho: Mineração de Opinião. Discutiremos também a conceitualização de ontologias, pois a proposta deste trabalho de mestrado agrupa aspectos de opinião com base em conhecimento lexical intrínseco em textos opinativos. Este tipo de conhecimento é estrutural e organiza-se, na maioria das vezes, através de relações semânticas hierárquicas. Portanto, nossa proposta de agrupamento semântico de aspectos, a partir de revisões de usuários, flerta com a conceitualização de ontologia. Por fim, serão apresentados as ferramentas e recursos linguístico-computacionais utilizadas nesta proposta de mestrado.

2.1 Mineração de Opinião

2.1.1 Conceitualização

Mineração de opinião, também chamada de análise de sentimentos, é um campo de intersecção entre a linguística e a ciência da computação, cujo objetivo é extrair automaticamente conteúdo subjetivo de textos (Taboada, 2016). De acordo com Liu (2012), os termos "mineração de opinião" e "análise de sentimentos" são geralmente usados como sinônimos e representam o campo de pesquisa de análise de opiniões, sentimentos, avaliações, apreciação, atitudes e emoções direcionados a entidades, tais como produtos, serviços, organizações, indivíduos, assuntos, eventos, tópicos e seus atributos. Para Zhao & Li (2009), o objetivo da mineração de opinião é descobrir opiniões em declarações textuais de forma automática e, portanto, é diferente da mineração de textos tradicional. A mineração de textos, segundo o autor, é baseada em temas objetivos e não em percepções subjetivas, além de centrar-se em temas específicos (por exem-

plo, negócios, viagem), bem como mudanças de tópicos em textos, enquanto que a mineração de opinião abarca problemas mais complexos. Na mineração de textos, os temas são expressos explicitamente através de palavras-chave ou tópicos, enquanto que, na mineração de opinião, as opiniões são, muitas vezes, expressas de forma sútil e implícita, além de possuírem marcas expressivas de subjetividade que são inerentes a este tipo de conteúdo.

Alguns conceitos são centrais na mineração de opinião. O primeiro deles diz respeito a própria definição de opinião. Segundo Liu (2012), uma opinião é definida por uma quíntupla: *entidade, aspecto da entidade, sentimento do aspecto, autor* e *tempo da opinião*. Para exemplificar esta quíntupla, usaremos a revisão exibida na Figura 2.1, em que o usuário descreve sua experiência com o produto câmera digital. Nesta revisão, a *entidade* ou alvo da opinião é uma câmera digital. Os *aspectos* são as propriedades da entidade avaliadas pelo usuário. Por exemplo, na revisão da Figura 2.1, o usuário avaliou os aspectos "câmera", "recursos" e "tela". O *sentimento do aspecto* consiste de palavras que representam a experiência do usuário com as propriedades do objeto avaliado. Por exemplo, na revisão da Figura 2.1, o usuário avaliou positivamente os aspectos usando as expressões "ótima", "intuítivo" e "facilita as operações". O *autor* e *tempo* da opinião são informações do usuário emissor da revisão e da data em que a revisão foi emitida. Na revisão da Figura 2.1, o autor da revisão é "Elcio Seidhy Kakuta" e a data de emissão desta revisão é "15/07/2001".

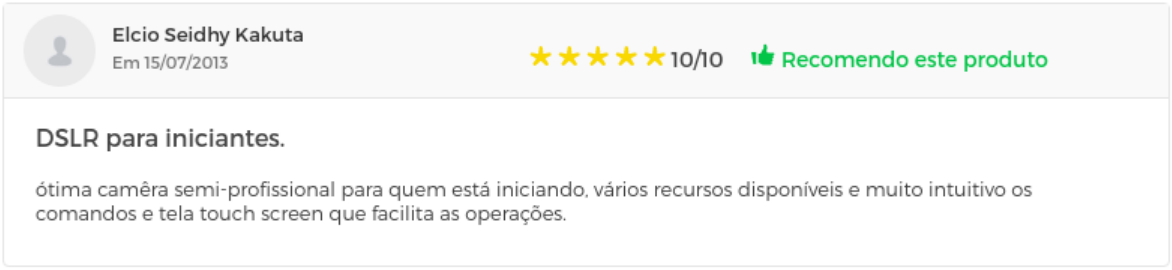

Figura 2.1: Revisão sobre uma câmera digital extraída de *Buscape.com*.

De acordo com Liu (2012), opiniões também podem ser classificadas em: opiniões regulares, comparativas, explícitas e implícitas. A seguir, descreveremos cada uma delas:

- **Opiniões regulares**: de acordo com Liu *et al.* (2005), em opiniões regulares, o autor expressa um sentimento, atitude, emoção ou percepção sobre um determinado alvo, por exemplo "Este filme é muito bom". Opiniões regulares podem ser classificadas em dois grupos. São eles: **opiniões diretas**, em que um objeto é avaliado diretamente pelo usuário, por exemplo, "A qualidade da imagem é alta"; e **opiniões indiretas**, que consiste de opiniões expressas indiretamente sobre uma entidade ou um aspecto de uma entidade. Por exemplo, na revisão "Após a injeção da droga, senti minhas articulações piorarem", a entidade "droga" é avaliada indiretamente pelo usuário emissor da revisão.

- **Opiniões comparativas**: são opiniões representadas por expressões que imprimem relação de semelhança ou diferença entre duas ou mais entidades. Por exemplo, na revisão

"Gosto mais de coca-cola do que pepsi", o usuário emite uma avaliação comparativa entre as entidades "coca-cola" e "pepsi".

- **Opiniões explícitas**: consistem de avaliações realizadas através de uma afirmação subjetiva, que pode ser uma opinião regular ou comparativa. Por exemplo, nas revisões "Eu gosto muito de coca-cola" e "Eu gosto mais de coca-cola do que de pepsi", as entidades avaliadas são "pepsi" e "coca-cola" e foram explicitadas nas revisões.

- **Opiniões implícitas**: tratam-se de indicações subjetivas que implicam uma opinião regular ou comparativa. No exemplo "A câmera é cara", o aspecto avaliado é "preço", no entanto, esse aspecto não foi explícitado, portanto é um aspecto implícito.

É intereressante ressaltar que, de acordo com Liu (2012), grande parte das investigações em mineração de opinião estão voltadas para opiniões explícitas. De acordo com o autor, opiniões explícitas são mais facilmente identificadas em comparação com opiniões implícitas.

Para Bhuiyan *et al.* (2009), pesquisas em mineração de opinião podem ser divididas em dois principais eixos: a classificação do sentimento e a mineração de opinião no nível de características ou aspectos. A classificação do sentimento consiste em reconhecer o sentimento geral presente em um documento ou sentença. Normalmente, essa tarefa é simplificada, classificando um documento ou uma sentença em 3 classes: positivo, negativo ou neutro (Avanço & Nunes, 2014). Mineração de opinião no nível de aspectos ou baseada em aspectos concentra-se, geralmente nas tarefas de: identificação de aspectos de opinião, identificação de polaridade e exibição dessas informações sumarizadas baseada na extração dos aspectos (Liu, 2012). Na Figura 2.2, exibimos estas três tarefas e, a seguir, descreveremos cada uma delas.

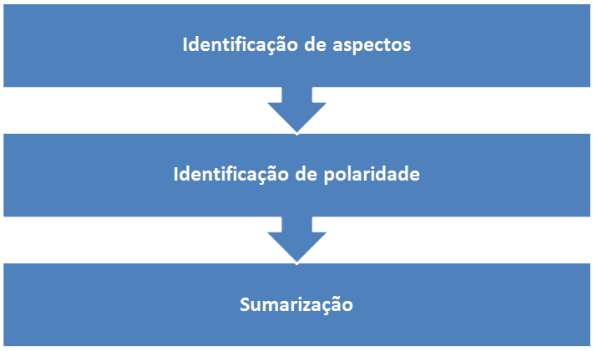

Figura 2.2: Principais tarefas da mineração de opinião baseada em aspectos (Liu *et al.*, 2005).

Na fase de identificação de aspectos, são extraídas características avaliadas pelos usuários sobre o alvo da opinião. Por exemplo, na revisão "A tela do Iphone 6 é ótima", a entidade avaliada é "Iphone". Para extração de aspectos, de acordo com Liu (2012), os principais métodos normalmente utilizados são: métodos baseado em frequência de substantivos e sintagmas nominas, métodos baseados em aprendizagem de máquina supervisionado, métodos baseados em aprendizagem de máquina semi-supervisionado e métodos baseado em modelo de tópicos. Abordagens recentes têm se apropriado de informações de ontologias e técnicas de Extração de

Informação (EI) para a tarefa de extração de aspectos. No trabalho de Freitas & Vieira (2013), uma ontologia de domínio é aplicada para a extração de aspectos de opinião para o domínio de filmes.

Na fase de identificação de polaridade, são extraídos os sentimentos associados aos aspectos. Por exemplo, na revisão "A bateria da câmera é péssima", o sentimento emitido em relação ao aspecto "bateria" é negativo, portanto a polaridade desta revisão é negativa. Para identificação de polaridade, grande parte dos trabalhos utilizam um léxico composto por uma lista de palavras de sentimentos associados com uma polaridade, geralmente sendo positiva, negativa ou neutra (Taboada, 2016).

Na fase de sumarização, o conteúdo mais relevante é exibido através de sumários, geralmente do tipo *extrativo*, que exibem o conteúdo sumarizado através do agrupamento e raqueamento de sentenças ou do tipo *abstrativo*, que não apenas selecionam as sentenças mais relevantes dos textos-fonte, mas analisam o documento e automaticamente geram novas sentenças. Esta abordagem tenta produzir novos textos a partir dos fragmentos originais identificados como relevantes. Não iremos nos aprofundar no detalhamento desses conceitos, pois não é o foco deste trabalho de mestrado. Sobre sumarização de opinião baseada em aspecto para a língua portuguesa, ver o trabalho de Condori (2014).

Além das tarefas de identificação de aspectos, identificação de polaridade e sumarização, de acordo com Taboada (2016), outra tarefa de responsabilidade da mineração de opinião é determinar se um texto, ou parte dele, é subjetivo ou não. De acordo com a autora, conteúdo textual pode conter informação objetiva (fatos, ações) ou informação subjetiva (percepções, opiniões, sentimentos). Além disso, textos subjetivos expressam uma visão positiva ou negativa e essa direção da opinião - se positiva ou negativa - é algo também conhecido como orientação semântica.

Muitas pesquisas utilizam córpus de revisões de usuários dos domínios de filmes, livros e produtos eletrônicos (Hu & Liu, 2004), porque esses domínios possuem relevância tanto a fabricantes quanto a consumidores. Para os fabricantes, é importante avaliar sua reputação, a aceitabilidade e a avaliação de seus produtos. Para os consumidores, a sumarização de revisões de outros usuários facilita na tomada de decisão na hora da compra.

2.1.2 Desafios

Antes do advento da internet, consumidores, para tomarem uma decisão de compra, pediam opiniões de amigos, familiares e organizações quando precisavam encontrar informação do público em geral sobre seus produtos e serviços. No entanto, com a rápida expansão dos serviços de e-commerce, usuários e empresas recorreram a revisões de usuários da web para tomada de decisão de compra. Entranto, processar esse grande volume de conteúdo opinativo, classificá-lo, sumarizá-lo e apresentar ao usuário de forma organizada apenas o conteúdo mais relevante destas revisões é, sem dúvidas, um grande desafio.

Os desafios iniciam com as definições básicas da área. Por exemplo, não há um consenso sobre a distinção entre os termos mineração de opinião e análise de sentimentos. Os dois ter-

mos são usados como sinônimos pela maioria dos pesquisadores da área, porém, de acordo com Liu (2012) e Munezero *et al.* (2014), mineração de opinião e análise de sentimentos não são conceitos equivalentes. Além disso, a distinção entre os conceitos de *aspecto* e *atributo* não são defendidos com clareza. Na maioria dos trabalhos da literatura, esses conceitos também são usados como sinônimos. Liu (2012) apresenta uma distinção rasteira sobre os conceitos de aspecto e atributo. Para o autor, atributos são características dos aspectos. Logo, aspectos são características do alvo da opinião ou entidade, formando, portanto, um nível hierárquico de características de opinião, entretanto essa conceitualização não é usada pela literatura. Além dos problemas de definição, de acordo com Yu *et al.* (2011), conteúdo opinativo geralmente apresenta-se de forma desorganizada, levando a dificuldade na navegação de informações e aquisição de conhecimento. Esse fenômeno é influenciado também pelas características distintas de perfis de usuários que geram conteúdo opinativo na web e que possuem competências linguísticas distintas (Vargas & Pardo, 2017).

Além dos desafios de definição e desorganização do conteúdo de revisões de usuários, para sistemas de mineração de opinião baseados em aspectos, além do reconhecimento de aspectos, é fundamental agrupar aspectos correlatos, no entanto, a tarefa de agrupamento de aspectos explícitos e implícitos não é simples. Por exemplo, agrupar aspectos que expressam características das especificidades do domíno é um grande desafio. Por exemplo, o usuário ao avaliar a propriedade "sinal" de um aparelho smartphone, também pode utilizar o estrangeirismo "*quadriband*" ou o termo "recepção" para avaliar a mesma propriedade do smartphone. Portanto, os termos "sinal", "quadriband" e "recepção" são correlatos e compõem um mesmo campo semântico neste domínio, por isso devem ser agrupados. Note que os três termos carregam marcas das especificidades deste domínio. Outro exemplo é o aspecto "display", que também é usado concomitantemente com os termos "visor", "tela" e "*touchscreen*" para avaliar uma mesma propriedade do smartphone. Outro desafio dessa tarefa de agrupamento é o número expressivo de construções lexicais em desacordo com a variante padrão da língua, além de marcas de oralidade que são comuns em textos produzidos no ambiente web. Por exemplo, em nossas análises, encontramos nove formas lexicais distintas para designar o aspecto "design" (designer, designe, design, desing, etc). As expressões "alto-falante" e "wi-fi" também foram encontradas de várias formas (auto-falante, auto falante, autofalante, alto falante, altofalante) e (wi-fi, wifi, wi fi). Além das formações lexicais em desacordo com a variante padrão da língua, encontramos termos como "cara", usado para avaliar o "autor" de um livro, e o termo "tatá", referindo-se a escritora brasileira "thalita rebouças". Todos esses fenômenos reforçam a predominância de marcas de informalidade e oralidade em discursos opinativos extraídos da web, além das especificidades desses domínios e idiossincrasias[1], que são características relacionadas com os perfis dos usuários emissores das revisões.

Tratando-se de processamento automático de conteúdo opinativo da web, os problemas parecem se maximizar. O ambiente comunicacional da web é predominantemente informal. Além disso, de acordo com Sales *et al.* (2015), a web tem se mostrado um ambiente com identidade

[1]Idiossincrasias são características de comportamento peculiar de um indivíduo ou de um determinado grupo.

linguística própria, carregada de especificidades constitutivas da comunicação textual em um ambiente discursivo virtual. Por exemplo, na revisão "O Iphone 6 é muitooooooo bom". O termo "muitooooooo" é recorrentemente utilizado na web para explicitar alta satisfação. Outros exemplos são "A luz da câmera é uma blz" e "A bateria é jóia". Nestes dois exemplos, a redução "blz" e a utilização do substantivo "jóia" com função adjetiva mostram o quão desafiador é minerar opinião na web. Além disso, em detrimento da efemeridade e fluidez da comunicação no ambiente web, grande parte do conteúdo produzido neste ambiente possui códigos implícitos. Por exemplo, na revisão "Eu adorei a sociedade do big brother" ou em "a câmera é fácil de usar". Na primeira revisão, o usuário avalia, de forma implícita, o livro 1984 de George Orwell. O termo "sociedade do big brother" é o termo indicativo de aspecto implícito para se referir a essa propriedade do objeto. Na seguda revisão, a expressão "fácil de usar" é usada pelo usuário para avaliar a propriedade "usabilidade" de uma câmera digital. Portanto, a tarefa de agrupamento de aspectos explícitos e implicitos é desafiadora e requer métodos que abarquem suas especificidades.

2.1.3 Métodos

Métodos para mineração de opinião utilizam, principalmente duas abordagens: a primeira é baseada em léxico, e um exemplo é usado no trabalho de Taboada (2016); a segunda é baseada em AM, conforme implementado nos trabalhos de Turney (2002) e Pang *et al.* (2002). As abordagens que utilizam léxico classificam a polaridade da opinião utilizando um dicionário de palavras de sentimento. Nos métodos baseados em AM, algumas características são utilizadas para treinamento de classificadores de sentimentos. Recentemente, alguns trabalhos propuseram a utilização de informação semântica para melhorar o desempenho da mineração de opinião. Nestes trabalhos, hierarquias são propostas para identificação de aspectos de opinião, por exemplo, os trabalhos de Freitas & Vieira (2013) e Cadilhac *et al.* (2010), e para identificação de polaridade e palavras de sentimentos, os trabalhos de Zhao & Li (2009) e Freitas & Vieira (2013).

2.1.4 Aplicações

Sistemas de mineração de opinião possuem aplicações em quase todos os negócios e na área social (Liu, 2012). Na área de negócios, no trabalho de Ghose *et al.* (2007), os autores observaram que revisões de usuários em sistemas online influenciam o comportamento dos leitores na hora da compra. Portanto, os autores propuseram um sistema para mensurar a reputação do comércio eletrônico da *Amazon.com*. No trabalho de Chaves *et al.* (2012), os autores relatam a experiência de classificação de sentimentos em revisões de usuários no domínio de hotelaria. Na âmbito social, no trabalho de Van Hee *et al.* (2015), os autores utilizam uma abordagem baseada em aspectos para classificação de eventos de *cyberbullying*. No trabalho de Vaassen (2014), o autor apresenta uma proposta baseada em aspectos para a predição de notas de suicídio em redes sociais.

Para o Português, em relação ao Inglês, poucos trabalhos foram desenvolvidos. No trabalho de Freitas & Vieira (2013), os autores apresentam uma proposta de mineração de opinião baseada em aspectos para o domínio de filmes. Neste trabalho, utilizam-se ontologias para o reconhecimento de aspectos e polaridades. No âmbito da sumarização de opinião, Condori (2014) apresenta uma abordagem baseada em aspecto para geração automática de sumários de opinião. Para identificação de polaridade, o trabalho de Avanço & Nunes (2014), propõe uma abordagem baseada em léxico para a classificação de orientação semântica em revisões de usuários no domínio de produtos. Para a tarefa de identificação de aspectos, o trabalho de Balage Filho & Pardo (2014) apresenta uma proposta baseada em AM.

2.2 Ontologias

De acordo com Gruber (1993), ontologias são especificações explícitas de uma conceitualização. Ontologias, no domínio da computação, foram desenvolvidas em Inteligência Artificial (IA) com o objetivo de modelar o conhecimento humano e facilitar o compartilhamento e reutilização de conhecimento entre humanos e máquinas (Fensel, 2003). A promessa de compartilhamento de informação entre pessoas e máquinas, a compreensão comum de um domínio e o reuso de informação entre agentes de software são justificativas para o uso de ontologias em uma gama de aplicações e por diversas comunidades acadêmicas. No entanto, segundo Aitchison (2003), o entrosamento de palavras e conceitos é um área de estudo que é um "pântano" de complexidade e ignorância. A complexa conexão entre a etiqueta que as pessoas usam e suas concepções das coisas rotuladas ainda é pobremente entendida. Nas próximas seções, discutiremos as várias vertentes conceituais de ontologias, na tentativa de enriquecer um pouco mais essa compreensão.

2.2.1 Definições

Uma gama de aplicações computacionais utiliza ontologias, entretanto não há um consenso conceitual na literatura sobre a definição de ontologia. A seguir, apresentaremos três principais conceitualizações da literatura.

Gruber (1993) define ontologias como artefatos computacionais responsáveis pela representação do conhecimento formal e estruturado de uma conceitualização compartilhada. Conceitualização, para o autor, é uma visão abstrata e simplificada do mundo que se deseja representar. Os componentes básicos de uma ontologia, segundo ele, são: (i) classes (organizadas em uma taxonomia); (ii) relações (representam o tipo de interação entre os conceitos de um domínio); (iii) axiomas (usados para modelar sentenças sempre verdadeiras); e (iv) instâncias (utilizadas para representar elementos específicos, ou seja, os próprios dados).

Guarino (1998) defende uma diferença sutil em relação à proposta de conceitualização de Gruber. Para Guarino, uma ontologia é uma teoria lógica para representar o significado pretendido de um vocabulário formal, ou seja, o seu compromisso ontológico para uma conceitualização particular do mundo. Para o autor, cada base de conhecimento, sistema baseado em

conhecimento, ou agente de nível de conhecimento está empenhado em alguma conceitualização explícita ou implicitamente. Ou seja, o compromisso ontológico reflete indiretamente uma conceitualização subjacente, pela aproximação dos modelos pretendidos. As relações ontológicas, de acordo com o autor, podem ser taxonômicas (hierárquicas) e não-taxonômicas (não hierárquicas). Relações taxonômicas ou "estruturantes" contribuem na estruturação de um domínio e na classificação de conceitos. Relações não-taxonômicas ou "não estruturantes" não estão relacionadas à hierarquia. Essas relações acrescentam informações aos conceitos já encontrados, identificando relacionamentos entre eles. Segundo Guarino, a espinha dorsal de uma ontologia consiste da generalização/especialização hierárquica de conceitos por uma taxonomia. Para o autor, ontologias são meios de modelar formalmente a estrutura de um sistema, ou seja, as entidades e relações que emergem de sua observação, conforme demonstrado na Figura 2.3. Nesta ontologia, proposta para a representação do domínio de recursos humanos de uma empresa de software, os conceitos são representados pelos recursos humanos da empresa. Note que os conceitos "programador" e "gerente" possuem uma relação do tipo *é-um* com o conceito "pessoa", enquanto que o conceito "programador" também possui uma relação do tipo *reportaA* com o conceito "gerente". O conceito "programador" também possui relação do tipo *trabalhaCom* com outro recurso "programador".

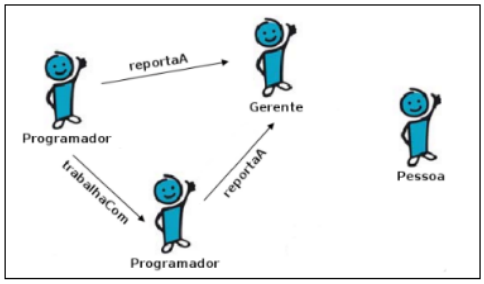

Figura 2.3: Ontologia de recursos humanos em uma empresa de software.

Vossen (2011), no entanto, compreende ontologias como um inventário de objetos e processos de um domínio, bem como a especificação de algumas relações que se mantêm entre eles. O autor especializa esse conceito atrelando-o a tradição ontológica. Por exemplo, para o autor, as tradições que lidam com a estrutura do conhecimento em ontologias podem ser posicionadas baseadas em suas diferentes perspectivas e diferentes propostas, tais como: (i) tradição filosófica, que categoriza as entidades como classes lógicas e tipos; (ii) a tradição cognitiva, cuja categorização das entidades é como o processamento de informação de humanos e inferência; (iii) a tradição da IA que classifica as entidades como funções de máquinas para fazer processamento de informação e inferência; (iv) a tradição semântica lexical, cuja categorização de palavras em um léxico faz parte de uma teoria linguística; (v) a tradição lexicográfica, cuja definição de palavras é extraída a partir de dicionário para usuários humanos; E por fim, (iv) a tradição da ciência da informação, que consiste na categorização da informação a partir de tópicos em ordem recuperável, por exemplo, uma biblioteca.

2.2.2 Tipologia

De acordo com Vossen (2011), os tipos de ontologias são definidos a partir dos artefatos gerados pelo domínio das tradições ontológicas e, esses domínios, se distinguem através das definições de conceitos e relações. Por exemplo, ontologias formais são geradas no âmbito da tradição filosófica. Nestas, os conceitos são entidades abstratas do mundo e as relações são do tipo *subsunção* ou *subtipo*. Na literatura, esse tipo de relação também é chamada de *é-um*. Na tradição cognitiva e na tradição da IA, geram-se modelos de domínio ou ontologias de domínio. Nesses modelos, a hierarquia tenta espelhar a estrutura cognitiva de conhecimento humano. As associações entre os "nós" geralmente são do tipo *parte-todo* em combinação com as relações do tipo *é-um*. Na tradição linguística, geram-se artefatos como léxicos, dicionários e tesauros, cujas relações são, predominantemente, do tipo não-taxonômicas. Nessa última abordagem, diferentemente da IA, o foco é sobre o conhecimento inerente às unidades lexicais.

Guarino (1998) apresenta outra classificação tipológica para ontologias. Para o autor, ontologias podem ser de quatro tipos: (i) alto-nível ou genéricas; (ii) domínio; (iii) tarefa; e (iv) aplicação. A Figura 2.4 ilustra essa tipologia em que ontologias de alto nível explicitam conceitos genéricos como elementos do espaço, da natureza e do tempo, livres de um domínio ou problema específico. Um exemplo é a Wordnet (Miller *et al.*, 1990). Ontologias de domínio representam vocabulários relacionados a domínios específicos. Este tipo é mais utilizado pela literatura por descrever visões de mundo. Um exemplo de ontologia de domínio para o Português é a Ontologia de Nanociência e Nanotecnologia proposta por Kasama (2009). Uma Ontologia de tarefa provê um vocabulário de termos usados para resolver problemas associados a uma tarefa específica, que pode ou não ser realizada em um mesmo domínio. Uma ontologia de tarefa foi proposta por Martins (2011) para reuso de conhecimento no âmbito da engenharia de requisitos. Por fim, ontologias de aplicação são dependentes de domínio e de tarefas específicas.

Figura 2.4: Tipos de ontologia proposto por Guarino (1998).

Neste trabalho de mestrado, nós assumimos as definições de ontologia propostas por Gua-

rino (1998).

2.2.3 Aprendizado de ontologias a partir de textos

Aprendizagem de ontologias é uma área essencialmente multidisciplinar e envovle áreas mais amplas da IA e de PLN. O termo Aprendizagem de Ontologias (AO) foi originalmente usado por Maedche & Staab (2001) e definido como a aquisição de um modelo de domínio a partir de uma fonte de dados. Segundo Maedche & Staab (2004), as abordagens de AO podem ser classificadas de acordo com os tipos de entradas. Essas entradas geralmente são: (i) bancos de dados relacional ou não relacional; (ii) esquemas semi-estruturados; (iii) bases de conhecimento; (iv) dicionários; e (v) textos. Nesta proposta de mestrado, nós utilizamos a fonte de dados do tipo texto, especificamente textos opinativos. Entradas do tipo texto partem de um córpus e geralmente utilizam técnicas de PLN e AM para aprender termos e relacionamentos entre esses termos. Neste sentido, dado um arquivo de entrada (texto), um arquivo de saída (esquema) é gerado correspondente a estruturação dos termos baseados em suas relações (taxonômicas e/ou não taxonômicas). Buitelaar & Magnini (2005) argumentam que a tarefa de aprendizagem de ontologias a partir de textos abrange basicamente os processos exibidos na Figura 2.5. As etapas também podem ser chamadas de camadas, sendo que as camadas superiores exigem técnicas mais complexas de aprendizagem. A seguir, discorreremos sobre cada uma dessas camadas.

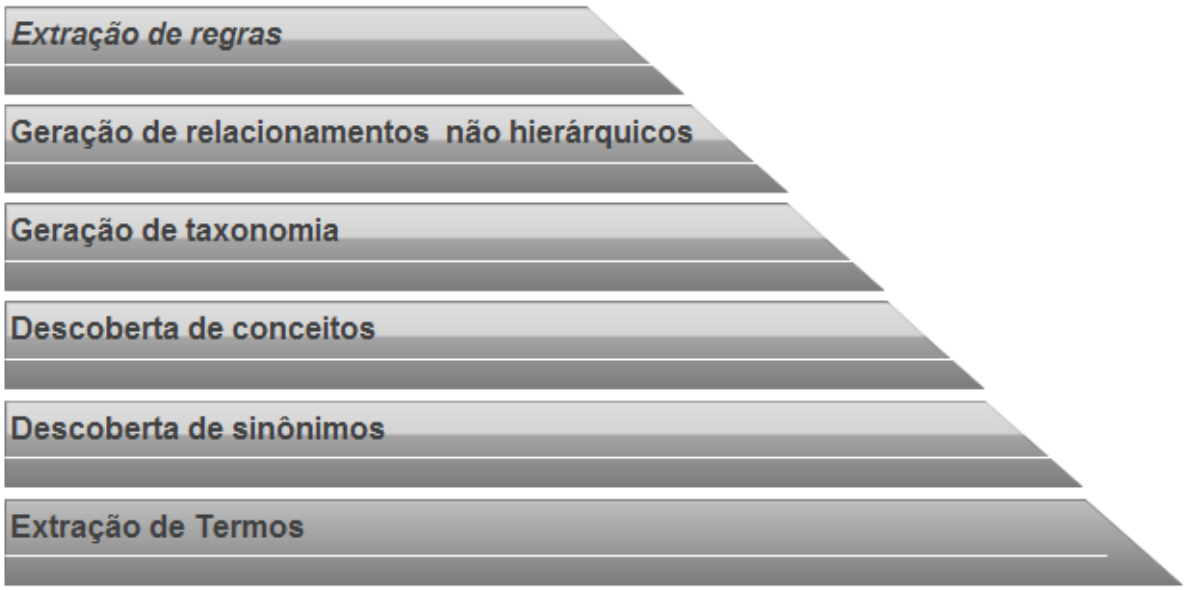

Figura 2.5: Etapas de aprendizagem de ontologia a partir de textos.

- **Extração de termos**: de acordo com a literatura, são aplicadas técnicas de EI e de PLN para extração de termos. Algumas dessas técnicas utilizadas são: (i) tokenização, que consiste do processo de extração de unidades mínimas de um documento textual; (ii) etiquetagem morfossintática (em inglês, *pos-tagging*), que é o processo pelo qual o conteúdo do córpus recebe etiquetas de categorias gramaticais (por exemplo, substantivo, adjetivo,

verbo); (iii) Tf-idf, que é uma medida baseada na frequência do termo *t*, em um documento *d*, em função do número de vezes que *t* ocorre em *d*; (iv) lematização, que consiste no processo de agrupar todas as formas morfológicas relacionadas de uma unidade lexical sob uma única entrada (lexema) e a tarefa correlata de separar os homônimos (Biderman, 2001); (v) Frequência Relativa (FR), que trata-se do resultado obtido da divisão entre a frequência absoluta - o valor que é observado na população - e a quantidade de elementos da amostragem. Para a tarefa de extração de termos, encontramos para o português, a aplicação Exato-LP (Lopes *et al.*, 2009), que recebe como entrada um córpus linguístico anotado pelo *parser* PALAVRAS (Bick, 2000) e extrai automaticamente todos os Sintagmas Nominais (SN) desse córpus, organizando-os segundo o número de *tokens* que os compõem.

- **Descoberta de sinônimos e conceitos**: muitos trabalhos usam léxicos disponíveis como Wordnet (Miller *et al.*, 1990) para reconhecer conceitos e sinônimos. Além disso, outros trabalhos também utilizam redes semânticas como a ConceptNet (Mukherjee & Joshi, 2013) para a descoberta de sinônimos e conceitos a partir de um córpus. Trabalhos recentes têm explorado o conhecimento da web como base léxico-semântico para descoberto de sinônimos e conceitos, por exemplo, no trabalho de Xavier & Lima (2010), as autoras propõem um método semi-automático de aprendizagem de conceitos e sinônimos, para o português, usando o Wikipedia.

- **Geração de taxonomia**: Para a geração de taxonomias geralmente são aplicados os métodos: (i) baseados em padrões léxico-sintáticos (Hearst, 1992) para extração de relações de hiponímia, ou seja, relações do tipo *é-um* entre termos; (ii) baseados na extração de relações de meronímia, ou seja, relações do tipo *parte-todo* entre termos, como por exemplo, usado nos trabalhos de Roberts (2005) e Ittoo *et al.* (2010); (iii) baseados em características semânticas entre pares de unidades lexicais. Essa abordagem utiliza geralmente um recurso lexical como Wordnet (Miller *et al.*, 1990) para extração de informação; (iv) baseados em termos complexos, ou seja, n-grama > 1, que analisam se um termo está lexicalmente contido em outro. Por tanto, se essa condição for verdadeira, esse termo é considerado seu hipônimo. Tais métodos foram utilizados nos trabalhos de Buitelaar & Magnini (2005), Baségio (2006), Ryu & Choi (2006) e Ribeiro Junior (2008); (v) baseados em agrupamento hierárquico de termos. No trabalho de Faure & Nédellec (1998), uma técnica de agrupamento foi aplicada a partir de termos que ocorriam com o mesmo verbo e após a mesma preposição. No trabalho de Yu *et al.* (2011), uma métrica de distância semântica baseada em um conjunto de caracteristicas linguísticas foi aplicada para agrupamento hierárquico de aspectos de opinião. O método usou uma hierarquia inicial de aspectos baseada em conhecimento da web em conjunto com métricas de distância semântica para agrupar aspectos de opinião mais próximos na hierarquia.

- **Geração de relacionamentos não hierárquicos**: vários trabalhos da literatura combinam técnicas de mineração de textos, análises estatísticas e linguísticas para geração de

relacionamentos não hierárquicos. No trabalho de Ciaramita *et al.* (2005), estruturas sintáticas foram exploradas para o reconhecimento de relacionamentos entre os conceitos da ontologia.

- **Extração de regras**: nessa etapa, regras são elaboradas para o reconhecimento, principalmente de conceitos disjuntos. Conjuntos são ditos disjuntos se não tiverem nenhum elemento em comum. Em uma ontologia, conceitos disjuntos são conceitos que não admitem instâncias em comum. Alguns trabalhos da literatura utilizam técnicas para classificação de grupo de conceitos disjuntos, tal como é feito no trabalho de Haase & Völker (2008).

2.2.4 Métodos de avaliação

De acordo com Brank *et al.* (2005), a tarefa de avaliação de ontologias é o problema de avaliar uma determinada ontologia do ponto de vista de um determinado critério de aplicação. O objetivo é determinar qual ontologia melhor se adequa a uma finalidade específica. Muitas ontologias diferentes conceituam o mesmo corpo de conhecimento, portanto é necessário ser capaz de medir qual modelo ontológico melhor se adapta a algum critério pré-definido. Técnicas de aprendizagem de ontologias necessitam igualmente de medidas de avaliação eficazes, que possam ser usadas para selecionar a "melhor" ontologia de muitas candidatas Algumas abordagens para a avaliação de ontologias têm sido consideradas pela literatura. De acordo com Brank *et al.* (2005), a maioria das abordagens de avaliação se enquadram em uma das seguintes categorias: (i) baseada no "padão-ouro", que avalia a ontologia gerada em função de uma ontologia de referência, chamada de "padrão-ouro"; (ii) baseada na aplicação em que a ontologia gerada é avaliada em relação a performance de uma aplicação; (iii) baseada em dados, que outras fontes de dados como coleções de documentos (por exemplo, a web) são usados sobre o domínio de cobertura da ontologia; (iv) baseado na avaliação humana, em que humanos se propõem a medir o quão bem a ontologia reúne um conjunto de critérios pré-definidos, normas e requisitos de um domínio. Segundo Brank *et al.* (2005), além das categorias acima de avaliação, pode-se agrupar as abordagens de avaliação de ontologias baseadas no nível de avaliação, tal como é descrito a seguir:

- **Vocabulário ou lexical**: a avaliação, neste nível, tende a envolver comparações com várias fontes de dados relacionadas ao domínio do problema (por exemplo, córpus de textos de domínios específicos), bem como medidas de similaridade entre termos (Maedche & Staab, 2002). Além de medidas de similaridade entre termos, o conteúdo lexical de uma ontologia pode ser avaliado utilizando os conceitos de *precisão* e *cobertura*, como é conhecido em sistemas de Recuperação de Informação (RI). Neste contexto, a precisão é medida a partir do percentual de entradas lexicais da ontologia gerada (itens lexicais usados como identificadores de conceito), que também aparecem na ontologia "padrão-ouro", relativo ao número total de conceitos da ontologia gerada; *cobertura* é o percentual

de entradas lexicais da ontologia "padrão-ouro" que também aparecem como identificadores de conceitos na ontologia gerada, em relação ao número total de entradas lexicais na ontologia "padrão-ouro".

- **Hierarquia/taxonomia e outras relações semânticas**: neste nível, avaliam-se as relações taxonômicas, ou seja, responsáveis por compor a estrutura da ontologia. De acordo com Biemann (2005), uma taxonomia é composta por relações do tipo *é-um* e/ou do tipo *parte-todo*. Métricas de avaliação de taxonomias foram propostas nos trabalhos de Guarino & Welty (2002) e Brewster *et al.* (2004). Neste último, os autores propõem uma abordagem baseada em aprendizagem de máquina para avaliar o grau de ajuste estrutural entre uma ontologia e um córpus de documentos. A partir do modelo probabilístico gerado, realiza-se a verificação se cada conceito da ontologia se encaixa, pelo menos, em algum tópico. Outra abordagem é proposta no trabalho de Guarino & Welty (2002). Neste trabalho, os autores utilizam uma abordagem baseada em "noções filosóficas" (essencialidade, rigidez, unidade, etc.), usadas para compreender melhor a natureza de vários tipos de relações semânticas nas ontologias. Segundo os autores, relações do tipo *é-um*, muitas vezes, são usadas para expressar características no nível de alguma classe ou são usadas em substituição das relações *parte-todo*, ou ainda utilizadas para indicar que um conceito pode ter significados múltiplos.

- **Contexto ou níveis de aplicação**: esse nível é aplicado a ontologias que geralmente fazem referência a outras fontes de ontologias (por exemplo, uma ontologia pode usar uma classe ou conceito declarado em uma ontologia externa).

- **Nível sintático**: neste nível, a ontologia é geralmente descrita em uma língua formal particular e deve corresponder aos requisitos sintáticos dessa língua. De acordo com Gómez-Pérez (1995), esse nível de avaliação é normalmente aplicado a ontologias construídas manualmente.

- **Estrutura, arquitetura ou desenho**: de acordo com Gómez-Pérez (1995), neste nível, são avaliados critérios que atendam certos princípios de *design* e preocupações estruturais pré-definidas. Este tipo de avaliação também é direcionado para ontologias criadas manualmente.

- **Multíplos critérios:** é possível utilizar uma combinação de critérios de avaliação de uma ontologia. Para cada critério usado, é atribuído uma pontuação numérica. Uma pontuação global para a ontologia é calculada como a soma ponderada dos *scores* dos critérios. Os trabalhos de Burton-Jones *et al.* (2005) e Fox *et al.* (1997) propõem uma abordagem deste tipo. No trabalho de Burton-Jones *et al.* (2005), um conjunto de dez critérios são usados. São eles: legalidade (frequência de erros sintáticos); riqueza (como os recursos sintáticos de uma língua formal são realmente aplicados na ontologia); interpretabilidade (coerência e coesão dos conceitos); consistência (grau de incosistência de conceitos); clareza (conceitos descritos de forma clara); abrangência (número de conceitos da ontologia,

em relação à média de todo o domínio da ontologia); precisão (percentual de conceitos da ontologia que não fazem parte do domínio); relevância (o quão relevante é um conceito para o domínio da ontologia gerada); autoridade (número de outras ontologias que usam conceitos da ontologia gerada); história (quantas vezes a ontologia gerada foi acessada em relação a outras ontologias).

A Tabela 2.1 resume as abordagens geralmente usadas para avaliação de ontologias e a relação com os níves de análise descritos acima.

Tabela 2.1: Uma visão geral das abordagens de avaliação de ontologias (Brank *et al.*, 2005).

Nível	**Baseada no "padrão ouro"**	**Baseada na aplicação**	**Baseada em dados**	**Baseada na avaliação de humanos**
Lexical, vocabulário e conceitos	x	x	x	x
Hierarquia ou taxonomia	x	x	x	x
Outras relações semânticas	x	x	x	x
Aplicações de contexto		x		x
Sintaxe	x			x
Estrutura, arquitetura e desenho				x

2.2.5 Domínios de aplicação

De acordo com Vossen (2011), em PLN, ontologias são aplicadas principalmente às seguintes tarefas: recuperação e extração de informação, sumarização de texto, similaridade semântica e desambiguação de sentido. A seguir, discutiremos três aplicações que utilizam ontologias para as tarefas de EI, sumarização de textos e aprendizagem semi-automática de ontologias a partir de textos. Selecionamos estas aplicações, pois em todas as abordagens uma hierarquia conceitual é proposta para extração e organização de conhecimento semântico do domínio, que utilizam como fonte um conjunto de textos. Portanto, esse tipo de proposta dialoga diretamente com a nossa proposta de mestrado. Organizamos a exibição das aplicações em ordem cronológica.

Extração de Informação

Gaizauskas & Humphreys (1997) propõem uma aplicação que usa uma ontologia de domínio para a tarefa de EI. Uma tarefa típica de EI, por exemplo, pode envolver o processamento de textos de jornais de negócios que contenham diversos anúncios e extrair-lhe os nomes e nacionalidades das empresas participantes, a atividade do empreendimento, a data de início deste empreendimento, sua capitalização, etc. A aplicação de EI proposta no trabalho dos autores foi implementada para processar grandes volumes de textos oriundos do jornal *Wall Street* [2]. O sistema é chamado pelos autores de LaSIE e processa textos utilizando uma ontologia, chamada pelos autores de "modelo de mundo". A ontologia foi especificada em uma rede semântica e

[2] *Wall Street Journal* é um jornal publicado na cidade de Nova Iorque, nos Estados Unidos

implementada em um grafo acíclico direcionado com um único *nó* no topo. Os *nós* no grafo são classes ou instâncias com *nós* de instância ocorrendo apenas como *nós* folha. Qualquer *nó* não-folha pode ser subclasse através *n* dimensões. Cada uma dessas árvores divide-se em ramos mutuamente exclusivos. Por exemplo, o conceito representado pela etiqueta "os vinhos" pode ser classificado pela cor e pela nacionalidade, de modo que, um determinado vinho possa ser branco e francês (dominado pelos *nós* branco e francês), mas não pode ser tanto vermelho e branco.

O sistema LaSIE consiste de três principais fases de processamento: pré-processamento lexical, análise e interpretação semântica e análise do discurso. Na fase de pré-processamento lexical, o sistema lê e tokeniza o texto bruto de entrada, realiza a etiquetagem morfossintática e faz uma correspondência sentencial com base em listas de nomes próprios. Na fase de análise e interpretação semântica, é construído uma representação predicado-argumento das sentenças. Por fim, na fase de análise do discurso, as representações das sentenças em predicado-argumento são adicionadas em uma rede semântica estruturada hierarquicamente, e compõem o "modelo de mundo" do sistema LaSIE. A Figura 2.6 ilustra um recorte da ontologia do sistema LaSIE. Nesta ontologia, o *nó* raiz representa a entidade do domínio, os *nós* filhos, do segundo nível, descrevem os objetos, eventos e atributos do domínio jornalístico. Por exemplo, o *nó* "*organisation*" é um objeto do domínio jornalístico que possui duas instâncias, que são "*company*" e "*government*". O *nó* "*person*" possui o atributo "*animate*" com valor igual a "*yes*". O interessante dessa proposta é a representação dos atributos juntamente com as classes de objetos e eventos na taxonomia. Por exemplo, o atributo "*single-valued*" possui instâncias igual a "*animate*" e "*count*".

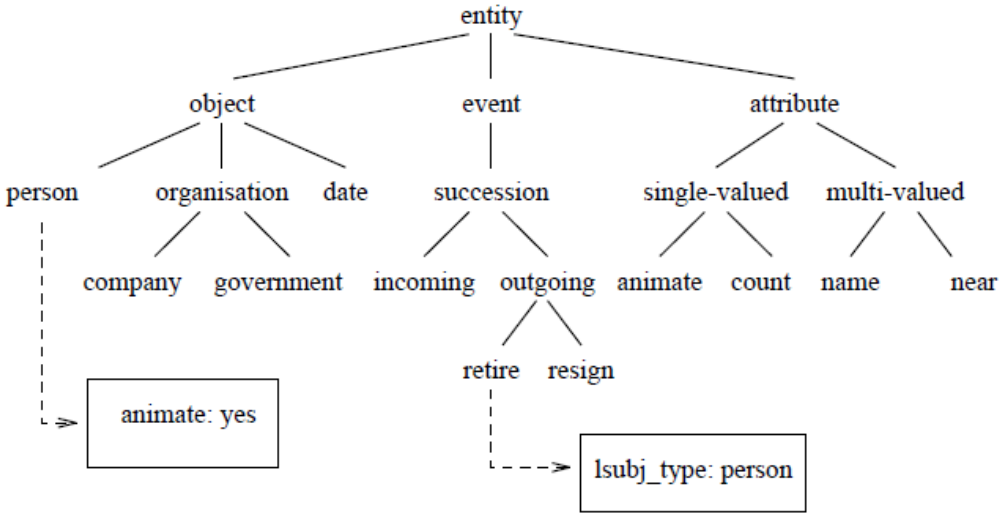

Figura 2.6: Recorte da ontologia usada pelo sistema LaSIE (Gaizauskas & Humphreys, 1997).

Sumarização

Wu & Liu (2003) apresentam uma proposta baseada em ontologia de domínio para a tarefa de sumarização de texto. Nesta proposta, uma análise contrastiva entre a abordagem baseada

em ontologia e outra abordagem baseada em estatística é avaliada para a tarefa de sumarização automática monodocumento. Foram coletados 51 artigos no domínio de negócios, especificamente do domínio *Sony*, incluindo produtos e informação financeira, publicados originalmente nos jornais *New York Time* e *Wall Street*. No total, foram extraídos 882 trechos destes jornais. Cada trecho é composto, normalmente, por uma ou duas sentenças. O método de sumarização basicamente classifica cada trecho com um *score* baseado em sua relevância no texto. Trechos com *scores* maiores são candidatos ao sumário. A seguir, descreveremos o método baseado em frequência e o método baseado em ontologia.

No método baseado em frequência, foram implementados quatro técnicas já citadas em trabalhos da literatura. São elss: (i) baseada na contagem de *tokens* mais frequentes; (ii) baseada no tamanho da sentença, em que apenas sentenças com um comprimento mínimo são selecionadas; (iii) baseada na seleção de "palavras bônus", que é verificado se uma sentença no trecho possui uma "palavra bônus" ou dada como "específica" do domínio para seleção de sentenças relevantes; (iv) baseada em nomes próprios, que o significado de um trecho está relacionado com o número de ocorrências de nomes próprios.

Em seguida, após a execução das técnicas descritas acima, uma pontuação de cada trecho é gerada usando a Equação 2.1, onde G é o conjunto de j trechos, F é o valor de i características de j, W é o peso de i características, e L é *1* se o trecho tem número suficiente de palavras, caso contrário, *0*.

$$G_{(}j) = L\,(w_{(}1)\,f_{(}j1) + w_{(}2)\,f_{(}j2) + \ldots + w_{(}n)\,f_{(}jn)) \tag{2.1}$$

No método baseado em conhecimento, uma ontologia é usada para determinar quais tópicos são relevantes para extração dos trechos. Foram selecionados os sinônimos e termos relacionados ao domínio *Sony*. Em seguida, esses itens foram organizados em uma estrutura de árvore, que pode ser vista na Figura 2.7. O *nó* raiz é representado pelo domínio e os *nós* filhos representam os conceitos e instâncias. O tipo de relação entre conceitos e as intâncias da ontologia não foram explicitados pelos autores. Os autores também não deixam claro se a ontologia foi criada manualmente. Portanto, a partir da construção da ontologia, o método baseado em conhecimento compara as palavras nos trechos jornalísticos com os conceitos da ontologia. Se a palavra não existe na ontologia, ela é ignorada; caso contrário, registra-se na ontologia o número de vezes que a palavra apareceu no trecho jornalístico. Palavras com frequência menor que 10 foram ignoradas.

Os autores usam as medidas de *precisão, cobertura* e *medida-f* para avaliar os resultados e comparar as abordagens baseadas em conhecimento e estatística. De acordo com os resultados apresentados pelos autores, ambos os métodos oferecem equivalência de precisão para seleção e contagem de subtópicos ou conceitos. Entretanto, de acordo com os autores, o método baseado em ontologia encontra subtópicos mesmo que eles não estejam nos itens lexicais mais frequen-

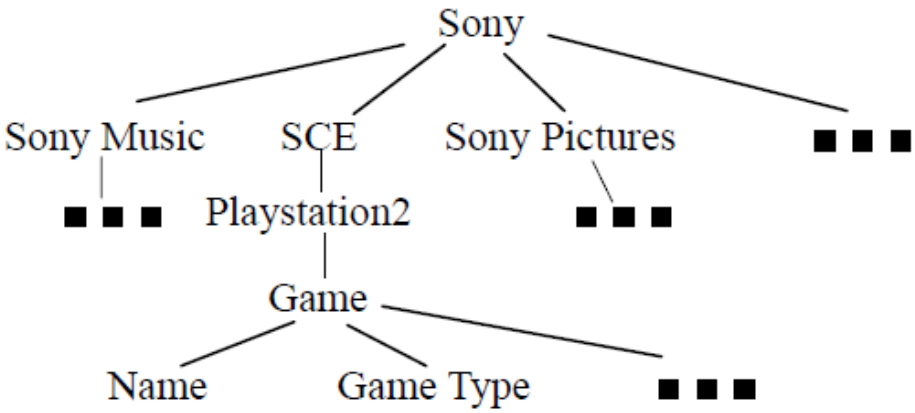

Figura 2.7: Ontologia de domínio usada para sumarização monodocumento (Wu & Liu, 2003).

tes do texto. Ou seja, o método baseado em ontologia encontra conceitos no texto sem depender do critério de frequência. Eles concluiram também que os resultados experimentais demonstraram o valor do método baseado em ontologias para sumarização de textos, porém, de acordo com os autores, projetar, construir e manter uma ontologia, mesmo de um domínio específico, ainda é caro. Os autores relatam que uma das fragilidades do método baseado em ontologia é que não incluía um dicionário de sinônimos. Segundo os autores, a inclusão desse dicionário melhoraria a performace do sistema. Por fim, os autores relatam que, em trabalhos futuros, pretede-se testar a implementação de uma abordagem híbrida, ou seja, baseada em frequência e usando uma ontologia de domínio.

Aprendizagem de ontologias a partir de textos

Ribeiro Junior (2008) apresenta um recurso de aprendizagem semi-automático de ontologias para o Português: a aplicação Onto-LP. De acordo com o autores, as abordagens da literatura para aprendizagem de ontologias a partir de textos baseiam-se fortemente no uso de informacões linguísticas, característica que torna essas abordagens dependentes do idioma. Neste sentido, comparando com outras línguas, para a língua portuguesa, poucos recursos de aprendizagem de ontologias foram desenvolvidos até o momento.

O Onto-LP é um plugin para o ambiente de construção de ontologias Protégé [3] que extrai, de forma semi-automática, ontologias a partir de recursos textuais. Esse recurso lê um córpus previamente anotado com informacões linguísticas pelo *parser* PALAVRAS (Bick, 2000), que provê informacões morfológicas, sintáticas e semânticas, representadas no formato xces/pln-br. [4]. Nessa proposta, a extração de termos é realizada a partir de uma abordagem híbrida: com base em técnicas linguísticas e estatísticas. Para o cálculo de relevância de termos, utilizaram três métricas estatísticas: *FR, Tf-idf, C-value* e *Nc-value*. Para organização hierárquica dos termos são empregados os seguintes métodos: baseado em padrões léxico-sintáticos proposto por Hearst (1992) e padrões de Morin/Jacquemim, adaptado por Baségio (2006) para o Português. Todos esses métodos são discutidos e exemplificados na Seção 2.2.3.

[3] http://protege.stanford.edu/
[4] Padrão de codificação de córpus

A arquitetura do Onto-LP é ilustrada na Figura 2.8. Note que o modelo recebe como entrada um córpus etiquetado pelo *parser* PALAVRAS (Bick, 2000), em seguida extrai os termos deste córpus, organiza hierarquicamente os termos e retorna como saída uma taxonomia.

Os autores relatam que os córpus aplicados durante a fase de testes do Onto-LP foram extraídos dos domínio da NanoCiência e de Pediatria.

Figura 2.8: Arquitetura do modelo Onto-LP (Ribeiro Junior, 2008).

2.3 Ferramentas e recursos linguístico-computacionais

Neste trabalho de mestrado, implementamos seis métodos para agrupamento de aspectos de opinião e, para isso, utilizamos alguns recursos linguístico-computacionais. Os recursos a seguir foram empregrados com o objetivo de reconhecer e extrair relações entre aspectos explícitos e indicativos de aspectos implícitos em textos opinativos. Uma síntese desses recursos é exibida na Tabela 2.2 e serão descritas nas seções seguintes.

Tabela 2.2: Sintése dos recursos linguístico-computacionais

N.	Recursos linguístico-computacionais
1	Onto-PT
2	CORP e CorrefVisual
3	*Word embeddings* do NILC
4	Dicionário de estrangeirismos do iLteC
5	Dicionário de nomes deverbais do iLteC
6	Lista de diminutivos/aumentativos
7	Lematizador do NILC

2.3.1 Onto-PT

Utilizamos, neste trabalho de mestrado, a ontologia lexical do português Onto-PT (Oliveira, 2014). Essa ontologia consiste de um recurso léxico-computacional criado a partir de recursos textuais etiquetados do português e contém no total 109.000 *synsets* envolvidos por pelo menos 105.000 sub-triplas. Quase metade do conteúdo dessa ontologia é composto por relações de *hiperonímia/hiponímia*, com $\approx$ 80.300 itens. Em seguida, essa ontologia possui relações do tipo *propriedade-de*, com $\approx$ 25.100, entre adjetivos e verbos. Além dessas relações, encontramos neste recurso, $\approx$ 9.700 relações entre adjetivos e nomes. As relações remanescentes são do tipo *propósito-de*, com $\approx$ 15.300 itens , e as relações *causativas*, com $\approx$ 9.800 itens, relações do tipo *parte-todo*, com $\approx$ 8.500 itens, relação do tipo *membro-de*, com $\approx$ 7.000 itens, e a relação de *antonímia* com $\approx$ 6.000 itens, entre outras. A Onto-PT encontra-se disponível para download [5] no formato OWL.

2.3.2 CORP

Utilizamos também o sistema de resolução de correferências para a língua portuguesa, o CORP (Fonseca *et al.*, 2016). Trata-se de um sistema desenvolvido em Java e constituído também por aplicações de código aberto como o Cogroo (de desenvolvimento CoGrOO, 2012) e o OpenNLP [6]. O *toolkit* OpenNLP fornece a etiquetagem morfossintática e o reconhecimento de entidades nomeadas, enquanto que o Cogroo fornece enxerto de SN e uma estrutura sintática superficial. Além disso, na versão atual do sistema de resolução de correferências, foram

[5] http://ontopt.dei.uc.pt/
[6] https://github.com/apache/opennlp

adicionados dois módulos semânticos com relações de hiponímia e sinonímia basedas nas relações da Onto-PT (Oliveira, 2014). O CORP (Fonseca *et al.*, 2016) encontra-se disponível nas versões web [7] e, para fins acadêmicos, na versão desktop. O principal objetivo desta ferramenta é mapear as cadeias de referentes ou correferentes de um texto e possui *F-score* médio igual a 46,70%. Utilizamos também uma versão do sistema CORP (Fonseca *et al.*, 2016), o CorrefVisual (Fonseca, 2014) [8].

2.3.3 *Word embeddings*

De acordo com Jurafsky & Martin (2000), esse é um método em que o significado de uma palavra é definido pela frequência com que ocorre perto de outras palavras. Ainda de acordo com o autor, métodos como esse são frequentemente referenciados por semântica vetorial. Neste trabalho de mestrado, nós utilizamos o algoritmo *word2vec* proposto por Mikolov *et al.* (2013) com acurácia média de 53,3% e os modelos pré-treinados propostos por Hartmann *et al.* (2017) e disponíveis no repositório de *word embeddings* do NILC [9]. Esse modelo pré-treinado utiliza textos da língua geral oriundos de diversas fontes (sites de produtos, wikipedia, entre outros).

2.3.4 Dicionário de estrangeirismos

Nós utilizamos também o dicionário de estrangeirismos do iLteC [10] (Ferreira & Janssen, 2017). Esse dicionário consiste de uma lista de estrangeirismos composta por 2.210 unidades lexicais que se encontram na base MorDebe [11] - isto é,trata-se de um dicionário atestado em dicionários de referência ou frequentes em jornais portugueses - e que violam regras da ortografia ou da morfologia. O dicionário de estrangeirismos inclui palavras provenientes de várias línguas, sendo as origens mais frequentes o inglês (“hobby”), o francês (“croissant”), o italiano (“paparazzi”), o alemão (“blitzkrieg”) e o japonês (“karaoke”). Para algumas palavras, na sua maioria internacionalismos como nomes de medidas ou de moedas, não foi indicada qualquer língua de origem.

[7] http://ontolp.inf.pucrs.br/corref/

[8] Nós utilizamos o CorrefVisual para checar apenas as revisões em que ocorriam relações entre aspectos mais complexas (mais específicas) para identificação automática. Essas relações ocorreram principalmente entre aspectos oriundos de gírias ou vocabulário informal, por exemplo, os aspectos “cara”, “tatá” e “porcaria”, e entre termos indicativos de aspectos implícitos com *n-grama* > 1, por exemplo “sociedade do big brother”, “canon rebel T3i” e “o outro lado da meia noite”. Identificamos esses casos e analisamos manualmente apenas as revisões em que os aspectos ocorriam. No total, foram verificados 11,66% do total de revisões do córpus, sendo necessário corrigir manualmente, 10 relações entre aspectos no domínio de smartphone, 12 relações entre aspectos no domínio de câmera e 20 relações entre aspectos no domínio de livro

[9] http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

[10] http://www.portaldalinguaportuguesa.org/

[11] A MorDebe é uma base de dados que contém palavras do português, apresentando mais de 135 000 lemas e cerca de 1,5 milhões de formas flexionadas

2.3.5 Dicionário de nomes deverbais

Nós selecionamos também o dicionário de nomes deverbais do português disponibilizado pelo iLteC (Janssen & Ferreira, 2007). Nomes deverbais são nomes derivados de um verbo, que expressam seu sentido de forma abstrata. Por exemplo, "pensamento" é um nome deverbal derivado do verbo "pensar". Este dicionário consiste de uma base composta por uma média de 6.000 unidades lexicais divididas entre verbos e suas respectivas construções deverbais.

2.3.6 Lista de diminutivos e aumentativos

Para o propósito deste trabalho de mestrado, nós construímos uma lista de unidades lexicais composta por algumas construções de diminutivos/aumentativos do português e recorrentes em revisões de usuários no domínio de produtos. Por exemplo, os pares ("leve", "levinho") ou ("livro", "livrinho") são unidades que populam essa lista [12].

2.3.7 Lematizador

Para lematização do córpus, nós optamos pela utilização do lematizador para o português desenvolvido pelo NILC e disponível no repositório Sucinto [13].

[12] Nós soubemos de uma iniciativa da Faculdade de Linguística da UFSCAR sobre o desenvolvimento de um léxico de diminutivos para o português. No entanto, tentamos o contato para disponbilização deste recurso e não obtivemos resposta.

[13] http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/resources.html

Capítulo

3

Trabalhos relacionados

Neste capítulo, discutiremos os principais trabalhos relacionados. Observamos que a tarefa de agrupamento de aspectos tem sido superficialmente explorada pela literatura. Na realidade, a definição de "grupos de aspectos" tem sido pobremente compreendida. No entanto, sistemas de mineração de opinião que não compreendam e tratem as especificidades do problema de agrupamento de aspectos incorrem no risco de apresentar resultados em desacordo com a realidade semântica do domínio. Portanto, além dos trabalhos relacionados, nós apresentaremos e discutiremos a tarefa de agrupamento de aspectos para mineração de opinião.

3.1 A tarefa de agrupamento de aspectos para mineração de opinião

A tarefa de agrupamento de aspectos explícitos e implícitos em textos opinativos possui grande relevância para sistemas de mineração de opinião, no entanto, não é uma tarefa trivial. O fenômeno linguístico de "abundância lexical" que se materializa em textos de uma língua natural através de unidades lexicais correlatas semanticamente, ou seja, unidades lexicais que possuem correspondência interpretativa em um dado domínio, ocorre principalmente em decorrência da subjetividade dos falantes que incidem visões de mundo distintas sobre um mesmo objeto e, essas visões de mundo, são impressas no vocabulário através de novas acepções lexicais e de significação. Para exemplicar como esse fenômeno acomete textos opinativos, vejamos o diagrama exibido na Figura 3.1. Nesta figura, apresentamos um recorte dos grupos de aspectos identificados no domínio de smartphone.

Começemos pelo grupo G1. Neste grupo, a propriedade do objeto avaliada é "internet". Note que os usuários utilizam os termos "3g", "wifi" e "wireless", que são tipos de conexão de internet, para avaliar essa mesma propriedade do aparelho. Perceba também que esses ter-

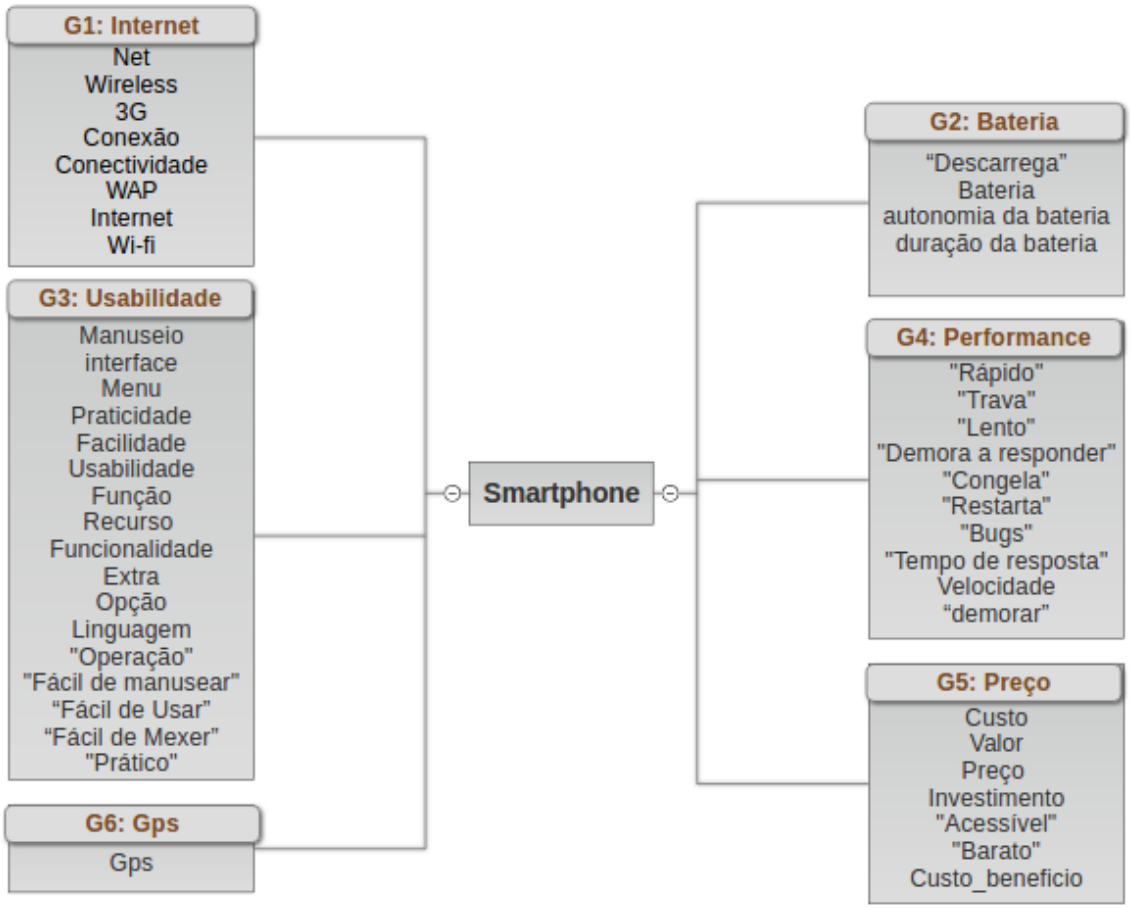

Figura 3.1: Recorte de grupos de aspectos do domínio de smartphone.

mos, por representarem as especicidades do domínio, nem sempre são encontrados em recursos linguístico-computacionais da língua como wordnets [1]. Além desses termos, os usuários também utilizam os termos "net", "internet", "conexão" e "conectividade" para avaliar a mesma propriedade do aparelho celular. No grupo G2, encontramos um fenômeno recorrente no córpus de revisões de usuários, os atributos de aspectos. De acordo com Liu (2012), aspectos possuem atributos que apresentam propriedades de aspectos. Por exemplo, as expressões "autonomia da bateria" e "duração da bateria" são propriedades do aspecto "bateria". Neste caso, há uma relação intrínseca entre as unidades lexicais. É também uma relação de *substring* [2]. O grupo G3 consiste de aspectos utilizados para avaliar a propriedade "usabilidade" do smartphone. Veja como o processo de agrupamento destes termos não é simples, pois muitos deles são termos que denotam vageza ("opção", "função", "recurso", "extra"). De acordo com Zipf (1970), a maioria das palavras tem múltiplas definições, no entanto, palavras mais frequentes tendem a ser mais ambíguas. Ainda sobre os itens do grupo G3, podemos observar expressões indicativas de aspectos implícitos. Por exemplo, as expressões "fácil de mexer", "fácil de manusear" e "fácil de usar", além dos termos "operação" e "prático", que são usadas para designar o aspecto "usabilidade" do smartphone. Perceba a dificuldade de agrupar itens de natureza tão distintas (verbos, sinônimos, adjetivos) em um mesmo grupo. No grupo G4, os usuários avaliaram a propriedade "performance" do smartphone. O termo "*bugs*", oriundo de estrangeirismo e a relação de deverbalidade entre a expressão "demora a responder" e o verbo "demorar" são indicativos de aspectos implícitos e usados pelos usuários para avaliar a propriedade "performance" do celular. No grupo G5, é interessante observar dois fenômenos em especial. O primeiro fenômeno consiste dos termos "acessível" e "barato", que são *termos pistas*, ou seja, usados como indicativos de aspectos implícitos. Veja que os termos"acessível" e "barato" são termos com alta ambiguidade, sendo necessário o mecanismo de inferência no domínio para correta correspondência interpretativa desses itens. O segundo fenômeno é representado pelo termo "investimento". Observamos o fenômeno de neologismo semântico que atribui a essa unidadade lexical o valor agregado de "custo" ou "preço" do aparelho. Por fim, o grupo G6 representa a ocorrência de grupos unitários no córpus de revisões de usuários. Os grupos unitários representam unidades únicas sem nenhuma correspondência semântica localizável no plano do conteúdo em revisões de usuários em que ocorrem. Por exemplo, não encontramos no córpus nenhum termo correlato ao aspecto "gps" do aparelho smartphone, portanto esse aspecto forma um grupo unitário. Portanto, a tarefa de agrupamento de aspectos pode ser definida pelo reconhecimento de aspectos correlatos semanticamente, ou seja, aspectos que possuem correspondência interpretativa em um determinado domínio.

A seguir, discutiremos as duas principais abordagens utilizadas para identificação de aspectos de opinião e os trabalhos relacionados.

[1] Wordnets são grandes banco de dados lexicais de uma língua em que, substantivos, verbos, adjetivos e advérbios, por exemplo, são agrupados em conjuntos de sinônimos *synsets*, cada um expressando um conceito distinto (Miller *et al.*, 1990).

[2] Uma *string* que aparece dentro de palavras no texto. Por exemplo, a *string* "ando" é uma *substring* de "caminhando"

3.2 Abordagens

De acordo com Zhai *et al.* (2011), são duas principais abordagens utilizadas para resolução do problema de agrupamento de aspectos. A primeira é baseada em conhecimento e a segunda é baseada em estatística.

Abordagens que implicam recursos baseados em conhecimentos pré-existentes utilizam predominantemente recursos lexicais como ontologias, tesauros, redes semânticas e léxicos. Essa abordagem foi proposta nos trabalhos de Alvarez & Lim (2007) e Hughes & Ramage (2007), cujo principal objetivo consiste em extrair medidas de similaridade linguística entre termos que podem ser relações lexicais de sinonímia, hiperônimia, hiponímia, merônimia entre outras. Essas relações linguísticas são usadas para mensurar similaridade entre dois termos.

Abordagens baseadas em estatística, basicamente utilizam a distribuição de palavras no córpus. Exemplos desta abordagem podem ser vistos nos trabalhos de Bollegala *et al.* (2007), Lin (1998) e Pereira *et al.* (1993). A abordagem estatística parte do pressuposto central de que palavras com significados semelhantes tendem a aparecer em contextos semelhantes (Harris, 1968).

A seguir, apresentaremos e discutiremos os trabalhos relacionados. Agrupamos esses trabalhos de acordo com a abordagem e os organizamos em ordem cronológica.

3.2.1 Abordagens baseadas em estatística

3.2.1.1 Zhai *et al.* (2011)

Neste trabalho, os autores afirmam que métodos estatísticos não supervisionados, usados para solução do problema de agrupamento de aspectos, não apresentam resultados satisfatórios. Essa afirmação é baseada nos resultados obtidos pelos trabalhos da literatura. Portanto, o problema de agrupamento de aspectos foi modelado como um problema de aprendizado semi-supervisionado. No entanto, vale ressaltar que esse tipo de método é custoso, pois requer um conjunto de exemplos etiquetados. Para etiquetagem do *dataset*, os autores exploraram características lexicais superficiais do problema para identificar automaticamente alguns exemplos rotulados. O método dos autores apresentou melhores resultados em relação a alguns outros métodos estatísticos clássicos de aprendizado não supervisionado e os métodos baseados em conhecimento que utilizam apenas similaridade lexical a partir de relações de wordnets.

Córpus

O córpus é composto por revisões de usuários sobre cinco domínios: *home theater*, seguro, colchão, carro e o aparelho de limpeza *vacuum*. A língua é o inglês. O córpus foi obtido de uma empresa comercial que fornece serviços de análise de sentimentos. Todas as expressões de aspectos foram etiquetadas quanto ao aspecto e o grupo, e também foram obtidas dessa mesma empresa comercial. Os detalhes do córpus são exibidos na Tabela 3.1.

Tabela 3.1: Córpus e base de referência (Zhai *et al.*, 2011).

	Home Theater	**Seguro**	**Colchão**	**Carro**	***Vacuum***
Sentenças	6355	12446	12107	9731	8785
Revisões	587	2802	933	1486	551
Aspectos	237	148	333	317	266
Grupos	15	8	15	16	28

É interessante observar um número de grupos de aspectos identificados pouco representativo do domínio. Além disso, se comprarado com os grupos de aspectos identificados para os domínios de smartphone, câmera e livro desta proposta de mestrado, esse número de grupos proposto pelos autores é expressivamente menor. Por exemplo, neste trabalho de mestrado, foram identificadas 180 expressões de aspectos para o domínio de smartphone e 48 grupos neste domínio; 132 expressões de aspectos para câmera e 36 grupos neste domínio; por fim, 103 expressões de aspectos para livro e 21 grupos neste domínio (ver uma descrição completa destes dados no Capítulo 4).

Descrição do método

Uma vez que os autores relatam que métodos não supervisionados não obtem bons resultados aplicados à tarefa de agrupamento de aspecto, o problema foi reformulado pelos autores como um problema de aprendizagem semi-supervisionado. No entanto, na aprendizagem semi-supervisionada são necessários alguns exemplos rotulados. Portanto, o método proposto pelos autores primeiramente rotula automaticamente alguns desses dados e, em seguida, classifica os grupos. Para classificação dos grupos, utilizou-se um melhoramento do algoritmo EM (Dempster *et al.*, 1977)(em inglês, *expectation–maximization*). Para geração de dados etiquetados, foram aplicadas três etapas principais, que serão descritas a seguir.

- **Etapa 1:** relações de *substring* entre expressões de aspectos. Relações de *substring* são relações entre expressões de aspectos contidas em outras expressões de aspectos. Por exemplo, os aspectos "serviço" e "serviço ao cliente" são aspectos que atendem esse critério. Em seguida, esses pares de expressões de aspectos foram adicionados em um grafo *G* como um conjunto de vértices que gerou um novo grafo *Gsc*. O grafo *Gsc* é exibido na Figura 3.1. Note que apenas o critério de *substring* parece insuficiente para o agrupamento de aspectos, pois como demonstrado no grafo da Figura 3.2, apenas alguns vértices foram conectados.

- **Etapa 2:** nesta etapa, as relações de sinonímia entre aspectos foram extraídas utilizando a Wordnet (Miller *et al.*, 1990). Por exemplo, os aspectos "imagem" e "foto" possuem relação de sinonímia, portanto nessa etapa, serão agrupados.

- **Etapa 3:** nesta última etapa, os autores selecionam um critério baseado em "componentes líderes" a partir de dados rotulados *L*. No entanto, os autores não deixam claro como

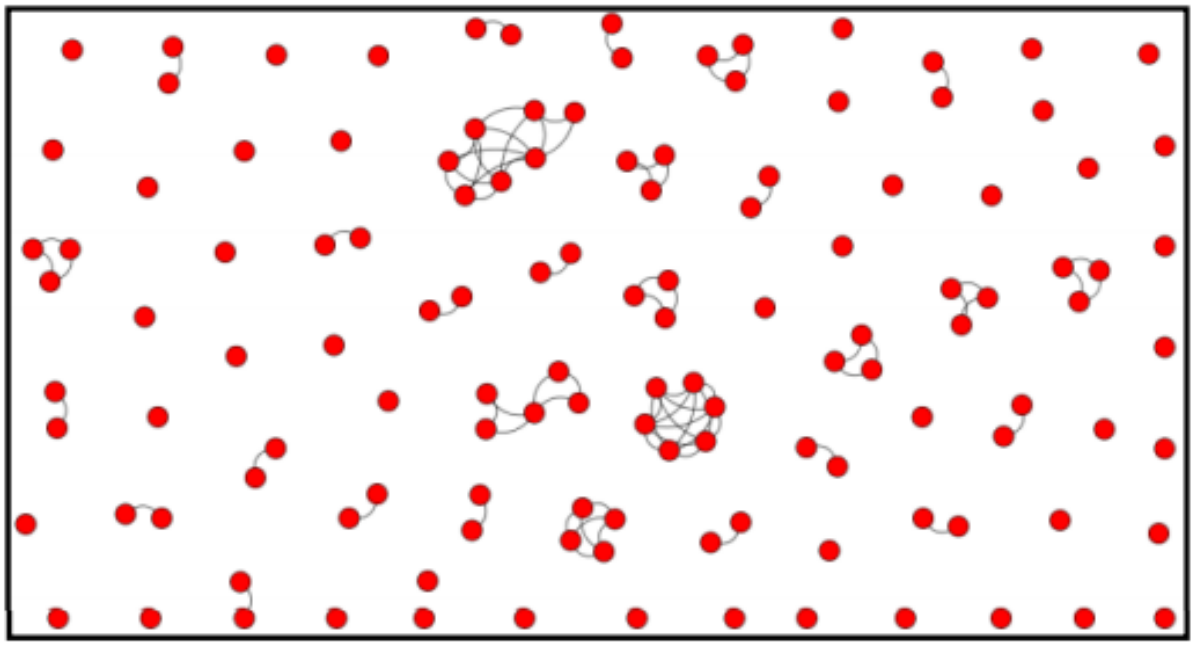

Figura 3.2: Grafo *Gsc* (Zhai *et al.*, 2011).

foi realizada essa seleção e não definem com clareza o que são esses "componentes líderes". Os autores apenas afirmam que, de acordo com o grafo *Gsc* e exibido na Figura 3.2, os grupos estão altamente desconectados, o que implica um número limitado de agrupamentos ou fusões nessa etapa. Portanto, foram selecionados os melhores componentes ou "componentes líderes" para formação dos dados rotulados com k grupos ou *clusters*.

Resultados

Os autores comparam o método proposto por eles com outros métodos da literatura. O primeiro grupo de métodos não utiliza nenhum conhecimento pré-existente. São eles: o algoritmo *K-means* (MacQueen, 1967) e o modelo LDA (Blei *et al.*, 2003). Em seguida, foram implementados métodos baseados em similaridade lexical que usam a Wordnet (Miller *et al.*, 1990) para extração de relações lexicais formais entre aspectos. Foram implementados dois métodos usando relações da Wordnet, que os autores chamaram de *CHC* e *SHC*. No entanto, não é clara quais as relações exatas usadas em cada um desses métodos. Por fim, o algoritmo EM (Dempster *et al.*, 1977) sem modificação também foi implementado a fim de compará-lo com o mesmo método ampliado proposto pelos autores. Esse método é chamado pelos autores de *L-Rand2* e o método ampliado é chamado pelos autores de *L-EM*. Os autores discutem os resultados e relatam que métodos baseados em similaridade lexical e os métodos baseados em aprendizado não-supervisionado trouxeram resultados inferiores em comparação aos métodos semi-supervisionados. Além disso, o método dos autores superou os demais métodos implementados.

O trabalho dos autores tem alguns pontos fortes, um deles é a comparação de vários métodos usadas pela literatura para a resolução do problema de agrupamento de aspectos. No entanto, um ponto fraco desta proposta é o não tratamento de aspectos implícitos. Para o domíno de

opinião, uma parcela significativa de aspectos são implícitos, portanto é fundamental que esse conhecimento também seja incorporado. Além disso, métodos como o proposto pelos autores são custosos, pois exigem um conjunto de dados etiquetados, além de pouco adaptável a outros domínios.

3.2.1.2 Zhang *et al.* (2011)

Neste trabalho, o problema de agrupamento de aspectos é modelado pelos autores como um problema não supervisionado e o algoritmo *K-Means* (MacQueen, 1967) é utilizado para a tarefa. Foram utilizados revisões de usuários nos domínios de câmera, smartphone, automóvel e notebook. A língua é o chinês.

Córpus

Dois anotadores etiquetaram 22.000 revisões quanto ao aspecto e seus grupos. Os grupos anotados e a quantidade de aspectos para cada grupo são exibidos na Tabela 3.2. Note que o grupo de etiqueta "aparência" possui o maior número de aspectos [3]. Entretanto, os autores não discorrem sobre os grupos formados nos outros domínios analisados. Além disso, não foram descritas as instâncias de aspectos para cada um dos grupos formados.

Tabela 3.2: Grupos anotados em revisões do domíno de câmera (Zhang *et al.*, 2011).

Categoria	N. de aspectos de câmera
Lente	56
Visor	62
Aparência	110
Bateria	18
Fotografia	76
Total	322

Descrição do método

Inicialmente, os autores identificam dois principais tipos de relações entre aspectos. São elas: a relação intra baseada em morfemas e a relação inter baseada em palavras de opinião. A relação intra baseada em morfema, de acordo com os autores, são relações entre aspectos com compatibilidade morfêmica, por exemplo, os aspectos "poder" e"bateria", que na língua chinesa, compartilham um mesmo morfema. Além disso, expressões de aspectos que possuem uma mesma palavra em comum, tal como os aspectos "bateria" e "resistência da bateria", também são enquadrados neste critério. Os autores argumentam que, em chinês, os morfemas são principalmente monossilábicos, embora haja alguns morfemas polissilábicos excepcionalmente integrados à língua por empréstimo linguístico. A relação inter baseado em palavras de opinião, de acordo com os autores, são relacionamentos entre aspectos que compartilham palavras

[3]Esse fenômeno também ocorreu no córpus analisado nessa proposta de mestrado, para o domínio de câmera.

de sentimentos em comum. Os autores relatam que palavras de sentimentos possuem informação semântica e essa relação reflete a interrelação entre aspectos similares. Por exemplo, os aspectos "forma" e "aparência", de acordo com os autores, recebem recorrentemente as mesmas palavras de sentimento, bem como, os aspectos "bonito" e "*fashion*".

Resultados

Os resultados são exibidos na Tabela 3.3. Observe que a melhor performance obtida com o método ocorreu no domínio de notebook, com Medida-F igual a 0,3830.

Tabela 3.3: Resultados (Zhang *et al.*, 2011).

Domínio	**Precisão**	**Cobertura**	**Medida-F**
Automóvel	0,2435	0,3326	0,2811
Câmera	0,3512	0,3563	0,3537
Telefone	0,3920	0,3539	0,3720
Notebook	0,3782	0,3880	0,3830

Nessa abordagem, é interessante observar a tentativa dos autores de explorar características linguísticas para a tarefa de agrupamento de aspectos. No entanto, apenas a extração de relações de *substring* e características morfêmicas entre aspectos é superficial, do ponto de vista linguístico, e, do ponto de vista computacional, também parece ineficiente, de acordo com os resultados obtidos pelos autores e exibidos na Tabela 3.3. Além disso, note que o conjunto de grupos de aspectos é definido de forma arbitária, além de compor um conjunto genérico e pouco representativo do domínio. Por exemplo, características como "conectividade" e "preço" são recorretemente avaliadas por usuários do domínio de câmera digital, no entanto, nenhum desses aspectos poderiam ser adequadamente agrupados nos grupos propostos nesta proposta, ou seja, não há compatibilidade semântica óbvia entre esses aspectos com os grupos arbitrariamente selecionados. Por fim, assim como nos outros trabalhos analisados, os autores não reportam sobre o reconhecimento e agrupamento de aspectos implícitos e os desafios com o agrupamento de aspectos específicos do domínio.

3.2.1.3 Abu-Jbara *et al.* (2013)

Nesta proposta, os autores aplicam um método robusto para mineração de opinião no domínio político. De acordo com os autores, no domínio de discussões políticas, quando dois usuários compartilham uma mesma opinião, esses usuários tendem a se concentrar em aspectos correlatos sobre o alvo da discussão e, portanto, acabam por enfatizar aspectos similares e que reforçam essas opiniões.

Neste trabalho, um método não supervisionado é proposto para a resolução do problema de agrupamento de aspectos de opinião a partir de textos opiniativos do domínio político e em língua árabe.

Córpus

Foram extraídos revisões sobre 36 debates políticos que compõem um total de 711 revisões escritas por 326 usuários. O número médio de revisões por discussão é de 19,75 e o número médio de participantes por discussão é de 13,08. Os autores propõem métodos de extração de conteúdo subjetivo, extração de aspectos e polaridade, além da tarefa de agrupamento de aspectos. No entanto, iremos abordar apenas o método de agrupamento de aspectos, pois é o foco desta proposta de mestrado.

Descrição do método

Para a tarefa de agrupamento de aspectos, os autores utilizam o modelo LDA (Blei *et al.*, 2003) (do inglês, *Latent Dirichlet Allocation*). O LDA é um modelo probabilístico gerativo aplicado a córpus. A idéia básica deste modelo consiste na representação de documentos como "mixes" aleatórios sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre a "palavra" no documento. O LDA assume alguns processos gerativos para cada documento w em um córpus D. Basicamente, o processo generativo proposto neste trabalho seleciona cada palavra N, no documento w, e escolhe aleatoriamente um tópico da distribuição sobre tópicos e, por fim, escolhe aleatoriamente uma palavra do tópico correspondente (distribuição ao longo do vocabulário).

O termo *targets* é usado pelos autores para se referir aos aspectos de opinião. Os *targets* de opinião são extraídos pelos autores a partir das sentenças que continham alguma palavra de opinião, reconhecido através do SAMAR (Abdul-Mageed *et al.*, 2012), um sistema de análise de subjetividade e sentimentos para o gênero de mídia social da língua árabe. Das sentenças selecionadas, foram extraídos sintagmas nominais que ocorreram pelo menos em duas revisões escritas por dois participantes distintos. Em seguida, foram selecionados apenas os sintagmas nominais associados a alguma expressão de opinião mais próxima do alvo.

Resultados

Os resultados obtidos com a aplicação deste método são exibidos na Tabela 3.4.

Tabela 3.4: Resultados (Abu-Jbara *et al.*, 2013).

Agrupamento	**Medida-F**
Apenas *targets* de opinião	0,65

Nesta proposta, os autores também não relatam o agrupamento de aspectos implícitos, no entanto, essa abordagem é interessante, pois avança ao propor um domínio distinto ao domínio de produtos e serviços. Domínios de produtos e serviços são predominantemente investigados pela literatura atual. O domínio político certamente apresenta outros desafios, além de outras especificidades, porém as especificidades e desafios desse domínio foram fracamente exploradas nesta proposta.

3.2.1.4 Zhou *et al.* (2015)

Nesta proposta, os autores afirmam que o passo crucial para alcançar um melhor desempenho na tarefa de agrupamento de aspectos para sistemas de mineração de opinião é a implementação de técnicas oriundas da engenharia de conhecimento. No entanto, esse tipo de abordagem consome muito esforço humano e, de acordo com os autores, pode ser instável quando o domínio do produto muda. Portanto, os autores optaram pela modelagem do problema de agrupamento de aspectos como um problema supervisionado. Vale ressaltar que abordagens supervisionadas também são custosas e altamente dependentes de domínio.

Córpus

Foram utilizadas revisões de usuários em inglês e do domínio de restaurante. A seguir, na Tabela 3.5, apresentamos os dados do córpus. Os autores utilizaram um conjunto de dados de referência lançado pelo SemEval-2014 [4]. A coluna de *treinamento* consiste no número de revisões do conjunto de dados e a coluna de *teste* o número de revisões utilizadas pelo autores. Os grupos de aspectos identificados para esse domínio foram: comida, serviços, ambiente e um conjunto genérico chamado "outros".

Tabela 3.5: Informações do córpus.

Categoria	Treinamento	Teste
Comida	1232	418
Serviços	597	83
Ambiente	431	172
Outros	1132	118

Descrição do método

Os autores propõem um conjunto de métodos, semi-supervisionados e supervisionados, para categorização de aspectos de opinião no domínio de restaurante. Foram utilizados os algoritmos *word2vec* (Mikolov *et al.*, 2013) para seleção de grupos de palavras similares e, em seguida, uma rede neural supervisionada foi aplicada para a captação de características mais profundas e, por fim, um classificador de regressão logística foi treinado com recursos híbridos para prever as categoria dos aspectos.

Resultados

Os autores implementaram vários métodos para resolução do problema de agrupamento de aspectos a fim de compará-los com o método proposto por eles. Na Tabela 3.6, apresentamos os resultados obtidos pelos autores com a implementação de cada um desses métodos. Foram implementados os métodos supervisionados *Naive Bayes (NB)* (Zhang, 2004), Regressão Logística

[4]SemEval (Avaliação Semântica) consiste de uma série contínua de avaliações de sistemas de análise semântica computacional, organizada pelo SIGLEX, o Grupo de Interesse Especial sobre o Léxico da Associação de Linguística Computacional

(LR) (Fisher, 1936) e Máquinas de Vetores de Suporte (SVM) (Steinwart & Christmann, 2008), usando atributos como *unigramas e bigramas*. Além disso, foram usados algoritmos baseados no modelo *word embeddings*, tais como *word2vec* (Mikolov *et al.*, 2013), C&W (Collobert & Weston, 2008), HLBL (Mnih & Hinton, 2009) e *GloVe* (Pennington *et al.*, 2014). Foram usados modelos pré-treinados publicamente disponíveis na web. Os autores também usaram o resultado médio de desemepenho dos sistemas no SemEval-2014 para comparação dos métodos implementados.

Tabela 3.6: Resultados (Zhou *et al.*, 2015).

	Método	**Medida-F**
1	KNN	63,89
2	LR	66,01
3	NB	66,70
4	SVM	80,81
5	SVM-DS	70,97
6	SemEval-Avg	73,79
7	NRC-Lexicon	84,08
8	NRC (Melhor sistema classificado no SemEval)	88,57
9	HLBL	69,69
10	C&W	72,55
11	GloVe	81,12
12	GloVe-re	84,55
13	word2vec	83,31
14	word2vec-re	87,67
15	Método proposto pelos autores	90,10

Observamos que o método proposto pelos autores superou os demais métodos implementados. Portanto, um ponto forte desta proposta, além do resultado expressivo para a tarefa de categorização de aspectos, é o uso conjunto de vários métodos (redes neurais, SVM, *word embeddings*). No entanto, vale ressaltar que métodos supervisionados são dispendiosos e pouco adaptáveis a outros domínios. Os autores também não relatam sobre a categorização dos aspectos implícitos e aspectos específicos do domínio, o que torna a tarefa menos complexa, do ponto de vista de classificação automática. Por fim, constatamos também que o número de grupos de aspectos categorizados para o domínio analisado é muito genérico e pouco representativo. Por exemplo, pode ser ineficiente apresentar um sumário de avaliações cujas categorias do domínio são superficiais e não representam o que realmente foi avaliado pelo usuário. Um usuário, ao requerer informações mais específicas de um domínio, como, por exemplo, os aspectos "especialidades", "cárdapio", "formas de pagamento", "estacionamento" e "conectividade", não as encontrarão nos grupos genéricos definidos pelos autores.

3.2.1.5 Chen *et al.* (2016)

Neste trabalho, os autores modelam o problema de agrupamento de aspectos como um problema de aprendizado não supervisionado. Os autores utilizam um córpus de revisões de usuários da lingua chinesa nos domínios de câmera digital e telefone celular.

Córpus

A visão geral do córpus é exibida na Tabela 3.7. O córpus do domínio da câmera possui 138 comentários, dos quais 4.039 aspectos foram identificados manualmente e anotados antes da remoção de aspectos duplicados e 1.189 aspectos permaneceram após a remoção de duplicados. Aspectos duplicados são análogos aos *tokens* e aspectos não duplicados são análogos aos *types*. O córpus do domínio do telefone celular contém 123 revisões, das quais 1.490 aspectos foram identificados manualmente e anotados antes da remoção de duplicados, e 757 aspectos permaneceram após a remoção de duplicados. A Tabela 3.5 indica que cada aspecto ocorreu, em média, 3,4 vezes no córpus do domínio da câmera e 2,0 vezes no domínio do telefone celular.

Tabela 3.7: Informações do Córpus.

Descrição	**Câmera**	**Telefone Móvel**
Reviews	138	123
Aspectos (antes da remoção de duplicados)	4.039	1.490
Aspectos (depois da remoção de duplicados)	1.189	757
Aspectos únicos	867	574
Aspectos múltiplos	322	183
Média de aspectos	3,4	2,0

Descrição do método

Foram propostos dois algoritmos. O primeiro extrai relações entre aspectos que os autores denominam de *relações relevantes e irrelevantes*. Em seguida, os dados são rotulados e utilizados como entrada pelo segundo algoritmo. No Algoritmo 2, exibimos este primeiro algoritmo responsável pela etiquetagem. Os autores compreendem relações relevantes entre aspectos como relações entre expressões de aspectos em que um aspecto está contido em outro aspecto, ou seja, esse tipo de relação também pode ser chamada de relação de *substring*. Por exemplo, os aspectos “lente” e “lente retrato”, ou os aspectos “imagem” e “qualidade de imagem”, são exemplos desse tipo de relação. As relações irrelevantes são caracterizadas pelos autores como o resultado da subtração entre o conjunto total de aspectos e o conjunto de relações relevantes. Portanto, os aspectos que não forem classificados no conjunto de relações relevantes farão parte do conjunto de relações irrelevantes. De acordo com os autores, os aspectos que aparecem na mesma frase podem ser considerados aspectos distintos se não estiverem relacionados pela relação de *substring* (Chen *et al.*, 2016). Por exemplo, os aspectos “foto” e “resolução” não estão contidos lexicalmente um no outro e aparecem juntos na mesma sentença, portanto, de acordo com a definição dos autores, esse exemplo caracteriza uma relação irrelevante. Os autores apresentam o algoritmo exibido a seguir para a identificação de relações relevantes e irrelevantes entre aspectos.

O algoritmo recebe como entrada um conjunto de aspectos e um córpus composto por revisões de usuários do domínio de câmera ou telefone celular. Em seguida, verificam-se as sentenças que contêm aspectos. No segundo laço, é verificado se o aspecto está contido em outro, ou

Algoritmo 1: Algoritmo de aquisição de conjuntos de aspectos relevantes e irrelevantes (Chen *et al.*, 2016)

Entrada: um aspecto relevante a, um aspecto a_i e um córpus C
Saída: um conjunto de aspectos relevantes R e um conjunto de aspectos irrelevantes IR

```
1  início
2      repita
3          se sentença s contém a então
4              repita
5                  se Há uma relação de inclusão entre a e a_i então
6                      R = R U {a_i}
7                  fim
8                  senão
9                      IR = IR U {a_i}
10                 fim
11             até Para cada aspecto a_i em s;
12         fim
13     até Para cada sentença s em C;
14 fim
```

seja, se há uma relação de *substring* entre os aspectos. Se atender essa condição, classifica-se essa relação como "relevante", senão é classificada como uma relação "irrelevante".

No segundo algoritmo, utiliza-se um cálculo estatístico partindo de dois eixos: (i) verificação se dois aspectos são sinônimos de acordo com suas definições de dicionário; e (ii) o cálculo literal da semelhança entre dois aspectos. Em outras palavras, neste último cálculo, cada aspecto é tratado como um vetor de caracteres. Utilizou-se a medida de similaridade de cosseno e alguns cálculos de similaridade de conjuntos relevantes e irrelevantes especificamente definidos pelos autores.

A seguir, no Algoritmo 2, exibimos o segundo algoritmo proposto pelos autores. Esse método toma como entrada os dados etiquetados pelo primeiro algoritmo (Algoritmo 1) e retorna como saída grupos de aspectos $\boldsymbol{A}$ no córpus $\boldsymbol{C}$.

Resultados

Um ponto interessante desta proposta é o víes híbrido em que conhecimento lexical e estatística são explorados em conjunto para resolução do problema de agrupamento de aspectos de opinião. Muitas abordagens da literatura têm apresentado excelentes resultados a partir de abordagens híbridas. A respeito da extração de relações entre aspectos, proposta pelos autores, podemos afirmar que a relação de *substring* entre aspectos é estatisticamente relevante em córpus de textos opinativos. No entanto, esse tipo de relação representa, em média, apenas 25% das relações entre aspectos de um domínio, de acordo com um estudo empírico realizado neste trabalho de mestrado (iremos apresentar os resultados do estudo de córpus no Capítulo 4). Portanto, há outras relações entre aspectos que devem ser exploradas. Nesta proposta, os autores também não reportam a extração e o agrupamento de aspectos implícitos e os desafios com o reconhecimento e agrupamento de aspectos específicos dos domínios analisados.

Algoritmo 2: Algoritmo de agrupamento hierárquico baseado em novas medidas de similaridade (Chen *et al.*, 2016)

Entrada: conjunto de aspectos **A**, **A** = {a_1, a_2,...a_n}; cada aspecto é descrito pela relação **R** e **IR**.
Saída: Clusters de aspectos **{A,C}**.

início
  1. Defina cada aspecto como um *cluster*, determinado por **C** = {c, c_2...c_n};
  2. Calcule a semelhança entre cada par de *clusters*;
  repita
    se *a similaridade entre* $\mathbf{c_i}$ *e* $\mathbf{c_j}$ *é máximo e maior que zero* **então**
      agrupe $\mathbf{c_i}$ **e** $\mathbf{c_j}$ em um novo cluster;
    fim
    3. Repita 2 até o número de *clusters* não mudar;
    4. Os *clusters* finais são **{A,C}**;
  até *Para cada sentença s em* ***C***;
fim

3.2.2 Abordagens baseadas em conhecimento

3.2.2.1 Patra *et al.* (2014)

Neste trabalho, os autores propõem um método robusto para extração e categorização de aspectos e identificação da polaridade. As sentenças foram coletadas de revisões sobre clientes de restaurantes e notebooks. A língua é o inglês. Os autores propõem um método baseado em conhecimento linguístico usando duas técnicas: etiquetagem morfossintática e extração de relações de hiperonímia do Wordnet (Miller *et al.*, 1990).

Córpus

As informações do córpus são descritas na Tabela 3.8.

Tabela 3.8: Córpus (Patra *et al.*, 2014).

	Domínio	**Sentenças**
1	Restaurante	3041
2	Notebook	3045

As revisões sobre restaurante consistem de 3041 sentenças em inglês com anotações quanto a expressão de aspecto, categorias de aspectos e suas respectivas polaridades. As revisões de notebook contém 3045 sentenças, anotadas também quanto a expressão de aspecto e categoria de aspecto, juntamente com a sua polaridade.

Descrição do método

Para a tarefa de agrupamento de aspectos, os autores elegeram arbitrariamente quatro categorias: serviço, preço, comida e ambiente. Em cada uma das categorias, os aspectos são

agrupados usando relações de hiperonímia da Wordnet (Miller *et al.*, 1990). Esse método, baseado em conhecimento, utiliza até o segundo grau de hiperônimos para agrupar os aspectos nas quatro categorias definidas pelos autores.

Resultados (Patra *et al.*, 2014)

Os resultados obtidos com a tarefa de agrupamento de aspectos nas categorias serviço, preço, comida e ambiente são exibidos na Tabela 3.9.

Tabela 3.9: Resultados (Patra *et al.*, 2014).

Precisão	Cobertura	Medida-F
0,7307	0,6802	0,7046

Um ponto forte da abordagem dos autores é a proposta de extração de conhecimento intrísico da língua e, neste método, o conhecimento explorado é a similaridade lexical entre unidades da língua. Além disso, os resultados obtidos pelo método dos autores são superiores em comparação com trabalhos que utilizam apenas conhecimento estatístico sem supervisão. No entanto, ontologias lexicais da língua, como wordnets, na maioria das vezes, não comportam aspectos específicos do domínio. Além disso, os autores relatam que o principal problema enfrentado nesta tarefa foi atribuir alguns aspectos apenas nas categoria eleitas. Segundo os autores, existem muitos casos em que os aspectos ocorreram em outras categorias e, nesses casos, o sistema falhou.

3.2.2.2 García *et al.* (2014)

Neste trabalho, os autores descrevem um método basedo em ontologias para o agrupamento de aspectos de opinião. Os autores utilizam a Wordnet (Miller *et al.*, 1990) e o Wikipedia [5] para a tarefa de agrupamento de aspectos no domínio de restaurante e notebook. A língua é o inglês.

Córpus

A seguir, na Tabela 3.10, exibimos as informações do córpus.

Tabela 3.10: Córpus (García *et al.*, 2014).

Domínio	N. de Sentenças	Língua
Restaurante	3000	Inglês
Notebook	3000	Inglês

Descrição do método

Nesta proposta, os autores utilizam a wordnet e o wikipedia para agrupamento de aspectos. Os autores categorizam somente aspectos "multipalavras". Por exemplo, o aspecto "autonomia

[5] http://wiki.dbpedia.org/

da bateria" é um aspecto "multipalavra". Eles também utilizam um conjunto de regras para extração desses termos. Por exemplo, *se a palavra N e a palavra N + 1 são substantivos, e essa combinação é uma entrada no WordNet (ou no Wikipedia), selecione esse termo. Por exemplo, tem-se os aspectos "duração da bateria"*. No entanto, não é clara e precisa a forma como o processo de agrupamento de aspectos ocorreu. Por exemplo, os aspectos "qualidade da tela" e "qualidade do serviço" seriam agrupados em uma mesma categoria baseado nesse critério? Esse processo não é descrito com clareza no trabalho dos autores.

Resultados

Os resultados obtidos com o método proposto pelos autores foi comparado com o melhor resultado obtido no SemEval-2014. Os resultados são exibidos na Tabela 3.11.

Tabela 3.11: Resultados (García *et al.*, 2014).

Método	**Precisão**	**Cobertura**	**Medida-F**
SemEval *Baseline*	0,671	0,602	0,638
Método proposto pelos autores	0,638	0,569	0,602

É interessante observar que os resultados obtidos com o método dos autores aproximou-se dos resultados obtidos pelo melhor sistema do SemEval-2014. Além disso, um ponto forte desta proposta é a incorporação de conhecimentos do Wikipedia. Conhecimentos oriundos de dados abertos parece-nos interessante, pois esse tipo de dado pode ser usado enquanto arcabouço semântico-lexical com unidades lexicais mais "reais" e próximas da língua em uso, além de ser um dado computacional "barato" e disponível em grande escala. Porém, o método dos autores, do ponto de vista linguístico, não cobre a complexidade inerente aos textos opinativos. Os autores agrupam apenas expressões de aspectos em relação de *substring*, por exemplo, os aspectos "bateria" e "autonomia da bateria". No entanto, em revisões de usuários, ocorrem várias outras relações entre aspectos que não implicam necessariamente apenas as relações de *substring*, como por exemplo, os aspectos "design" e"modelo" ou os aspectos "áudio", "alto-falante" e "som". Os fenômenos que acometem textos opinativos são complexos, o que implica a adequada compreensão de características linguísticas e extra-linguísticas profundas. Além disso, o método dos autores não se propõe a agrupar aspectos implícitos.

3.3 Considerações finais

A tarefa de agrupamento de aspectos consiste em identificar grupos de unidades lexicais correspondentes em um domínio. No entanto, esse tipo de agrupamento é complexo, do ponto de vista de processamento automático, pois implica o agrupamento de termos que, muitas vezes, são de naturezas distintas. Portanto, é fundamental a compreensão adequada deste fenômeno. Grupos de aspectos são formados por aspectos de opinião correlatos em um domínio, usados pelos usuários de forma concomitante para se referir a uma mesma propriedade do objeto avaliado.

Além disso, nós observamos que não há para a língua portuguesa nenhum trabalho de agrupamento de aspectos de opinião. Encontramos trabalhos em que a identificação de sinônimos é proposta usando os mesmos métodos discutidos nesse capítulo. No entanto, para sistemas de mineração de opinião, nenhum método foi proposto. Dentre os trabalhos da literatura, mesmo os trabalhos em outras línguas (como o inglês), nenhum deles aborda e propõe o agrupamento de indicativos de aspectos implícitos. Todos os trabalhos encontrados tratam apenas do agrupamento de aspectos explícitos. Até mesmo os trabalhos que se propõem a agrupar aspectos sinônimos em outros domínios. Também não encontramos dados quantitativos e qualitativos, em nenhum dos trabalhos sobre agrupamento de aspectos de opinião, sobre a indicidência de aspectos específicos do domínio. Esse dado é muito relevante, pois a indicidência desse tipo aspecto varia entre domínios, além de potencializar o grau de dificuldade da tarefa de agrupamento de aspectos.

Capítulo

4

Estudo de córpus e aprofundamento linguístico

Neste capítulo, apresentaremos um estudo linguístico aprofundado, cujo objetivo foi compreender e mapear os principais fenômenos linguísticos e estatisticamente relevantes em revisões de usuários, sobretudo no domínio de produtos na web. Organizamos este capítulo da seguinte forma: na Seção 4.1, apresentamos os dados, a metodologia e os resultados do estudo de córpus e, na Seção 4.2, apresentamos um estudo linguístico teórico aprofundado baseado na observação dos dados linguísticos, na tentativa de compreender o fenômeno de "abundância lexical", em que várias unidades lexicais distintas são usadas pelos falantes de uma língua para se referir a um único objeto/entidade no mundo, portanto, de cardinalidade 1:N, ou seja, para uma única propriedade semântica do objeto avaliado são atribuídas pelo menos uma ou várias unidades lexicais distintas.

4.1 Estudo de córpus

Nossa principal motivação com esse estudo de córpus é compreender as características e desafios no processo de reconhecimento de grupos de aspectos e a organização semântica desses grupos. Nosso objetivo é propor soluções automáticas fortemente motivadas linguisticamente para sistemas de mineração de opinião. Nós selecionamos três domínios distintos: smartphone, câmera e livro a fim de compreender as convergências e divergências de comportamento entre os domínios. A análise empírica foi realizada manualmente e serviu como base de referência, além de recurso para a pesquisa. Neste estudo, apresentamos diversos dados quantitativos e qualitativos sobre a tarefa de agrupamento, além de algumas evidências empíricas que o comportamento linguístico varia entre os domínios e que essas variações tem fortes ligações com as

especificidades de conhecimento do domínio e com os perfis do escritor/usuário que produz o conteúdo.

4.1.1 Descrição dos dados

Neste trabalho, foram selecionados 60 revisões dos produtos smartphone, câmera digital e livro. Optamos por selecionar apenas 60 revisões para cada domínio, principalmente por tratar-se de um estudo empírico, realizado manualmente e de cunho qualitativo. Nossa principal hipótese é que, para cada domínio, há comportamentos linguísticos e fenômenos distintos. Uma síntese do córpus analisado é exibido na Tabela 4.1.

Tabela 4.1: Visão geral dos dados

Domínio	**Nº de Revisões**	***Tokens***	***Types***
Livro	60	35.771	1.577
Smartphone	60	6.077	1.496
Câmera	60	3.887	1.060

No domínio de livro houve um salto significativo no número de *tokens*, quando comparado aos domínios de smartphone e câmera. Neste domínio, caracterizamos um número expressivo de conteúdo irrelevante. Essas questões serão discutidas na Seção 4.1.4.

4.1.2 Metodologia

Nesta seção, apresentaremos a metodologia utilizada para realização do estudo de córpus. Para cada revisão, primeiramente foram identificados manualmente todos os aspectos, inclusive aspectos implícitos. Para a identificação e quantificação de aspectos implícitos, anotamos o termo indicativo de aspecto implícito e o chamamos de *termo pista*. Estes termos pistas foram anotados e diferenciados utilizando aspas duplas, portanto, desta forma, foi possível mensurar a ocorrência de aspectos implícitos em cada domínio. O reconhecimento de grupos de aspectos foi realizado revisão por revisão. A progressão na identificação dos grupos ocorria a cada revisão analisada. Anotou-se e quantificou-se cada novo grupo de aspectos na ordem em que surgiam. Esse processo se repetiu até a finalização das 180 revisões, sendo sessenta revisões para cada um dos três domínios. Nessa tarefa, optamos por agrupar também os atributos dos aspectos no grupo do respectivo aspecto. Como já apresentado, aspectos representam propriedades ou partes das entidades que são avaliadas pelos usuários, em textos opinativos, como em comentários em sites e blogs na web (Liu, 2012). Contudo, não é clara na literatura a distinção entre atributos e aspectos de uma entidade. Na maioria das vezes, são usados como sinônimos. Com isso, por exemplo, o atributo “qualidade do som” inerente ao aspecto “som” foi incorporado ao grupo de aspectos “som”. Outro exemplo é o aspecto “qualidade da imagem”, que foi incorporado ao grupo de aspectos “imagem”. Optamos também por selecionar o lexema (unidade abstrata do léxico (Biderman, 2001)) relativo ao aspecto. Por exemplo, os usuários, ao se referirem ao aspecto “foto”, podem utilizar esse termo no singular e no plural. Ou, ao se

referirem ao aspecto “autor” do livro, os usuários podem utilizar as flexões de gênero. Portanto, lematizamos as revisões e exibimos nos grupos apenas o lexema correspondente a cada unidade lexical.

Na Figura 4.1, apresentamos o processo de reconhecimento e agrupamento de aspectos explícitos e aspectos implícitos. O primeiro processo é o reconhecimento de aspectos implícitos. Dada a entrada (revisão), é verificado se há algum aspecto implícito; se sim, anota-se o *termo pista*, ou seja, o indicativo do aspecto implícito naquela revisão; senão, anota-se o aspecto explícito, se houver. Em seguida, é verificado se o aspecto anotado é uma ocorrência nova. Se não for, o aspecto é agrupado no seu grupo semântico respectivo; se for um aspecto novo e possuir correspondência semântica naquele domínio com outro aspecto já analisado, agrupa-se esse aspecto no grupo semântico respectivo, caso contrário, um novo grupo é criado.

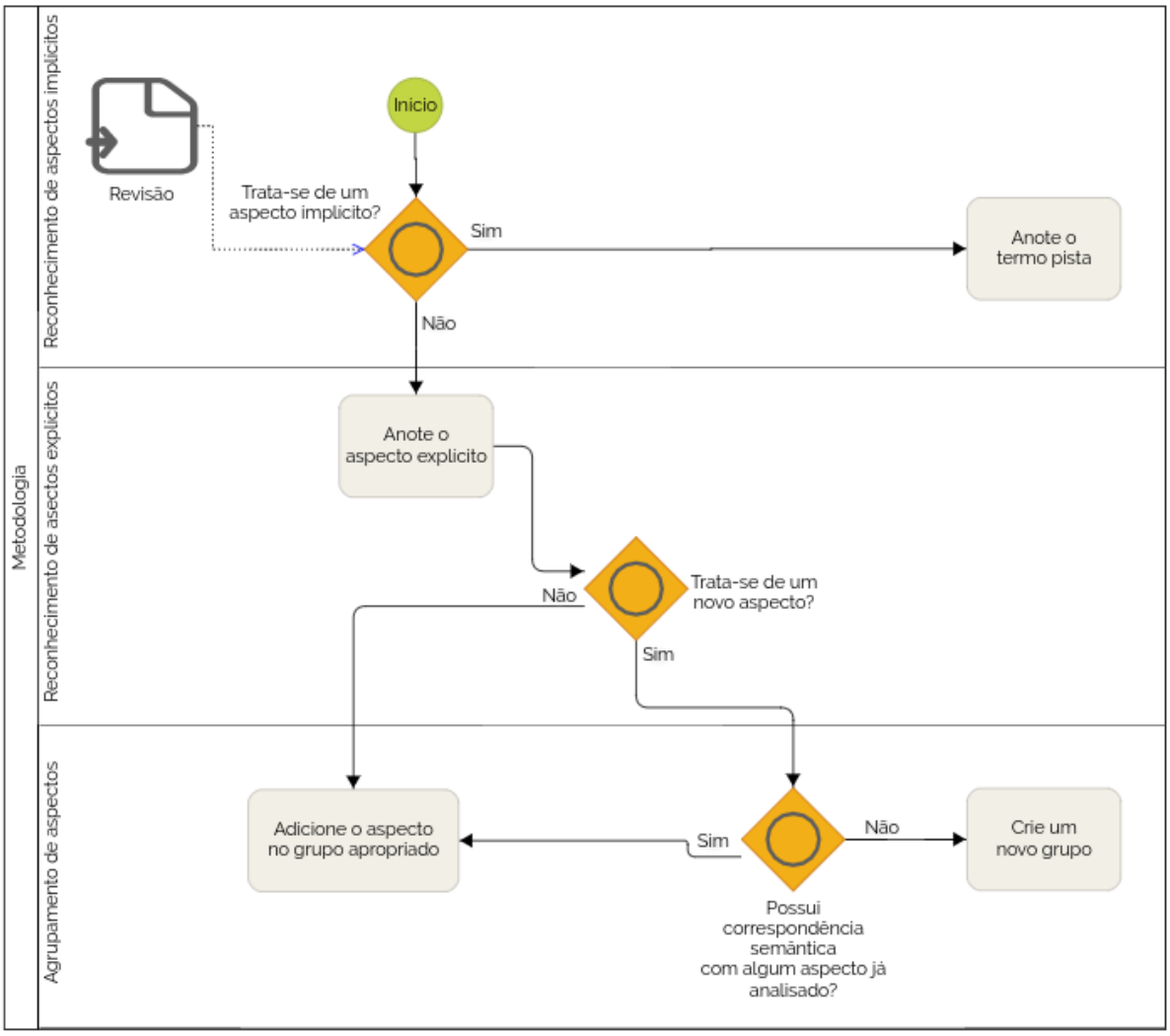

Figura 4.1: Reconhecimento e agrupamento de aspectos explícitos e aspectos implícitos.

4.1.3 Resultados

Nesta seção, apresentaremos os resultados obtidos com o trabalho de análise de córpus. Abordaremos os resultados do processo de reconhecimento dos grupos de aspectos e suas di-

versas implicações.

4.1.3.1 Visão geral

Foram identificados 48 grupos de aspectos para o domínio smartphone, 36 para o domínio de câmera digital e 21 para o domínio de livro. Houve uma diferença significativa no número de grupos de aspectos dos domínios de smartphone e câmera em relação ao domínio de livro. Os produtos smartphone e câmera digital são produtos tecnológicos populares e de aspectos facilmente reconhecidos por usuários que se envolvem mais com as características desses produtos e, portanto, tornam-se "mais especialistas" no assunto, diferentemente do domínio de livro, em que os usuários geralmente são apenas leitores e não especialistas em literatura ou críticos literários, além de não se interessarem por avaliar aspectos técnicos de livros (como "tamanho" e "tipo de papel", por exemplo). Esses usuários, portanto, conseguem avaliar um número superficial de aspectos do produto, geralmente *aspectos prototípicos* [1] do objeto ou aspectos superficiais do produto. Esses dados podem ser visualizados na Tabela 4.2.

Tabela 4.2: Classificação Geral

	Smartphone	**Câmera**	**Livro**	**Média**
Número total de aspectos (c/repetição)	459	342	323	374,66
Aspectos explícitos	392	289	274	318,33
Aspectos implícitos	67	53	45	55,00
Aspectos únicos (sem repetição)	180	132	103	138,33
Aspectos explícitos	142	109	91	114,00
Aspectos implícitos	38	23	12	24,33
Grupos de aspectos	48	36	21	35,00
Aspectos específicos do domínio	67	34	30	43,66

Note que o número total de aspectos por domínio e o número médio de aspectos por revisão parece-nos uma relevante evidência empírica para a relação entre perfil de usuário e o nível de informatividade [2] em textos opinativos. Revisões de usuários *experts* ou especialistas possuem maior grau de informatividade, ou seja, esses usuários possuem mais conhecimento sobre o domínio, o que os possibilita avaliar um número maior de aspectos da entidade avaliada. Além disso, constamos que 21,11% do total de aspectos (únicos) que ocorreram no domínio de smartphone são implícitos e 37,22% são específicos do domínio [3]. No domíno de câmera, 17,24% do total de aspectos (únicos) que ocorreram no domínio são implícitos e 16,66% são específicos do domínio. Por fim, no domínio de livro, 11,65% do total de aspectos (únicos) que ocorreram no domínio são implícitos e 29,12% são específicos do domínio. Quanto aos percentuais médios dos três domínios, obtivemos 84,96% de aspectos explícitos e 15,03% de

[1] As categorias linguísticas apresentam uma estrutura prototípica (baseada em protótipos). Mais precisamente, a linguística cognitiva afirma que os vários membros ou propriedades de um objeto possuem, geralmente, diferentes graus de saliência (uns são prototípicos e outros periféricos).

[2] De acordo com Koch (2004), a informatividade de um texto está associada a sua capacidade de apresentar informações novas.

[3] Aspectos específicos representam o conhecimento relativo específico do domínio em que são empregrados.

aspectos implícitos (com repetição); 84,41% de aspectos explícitos e 17,58% (sem repetição); Sendo que, em média, 28,67% são aspectos específicos do domínio.

4.1.3.2 Conteúdo relevante e irrelevante em revisões de usuários

De acordo com Bronckart (1997), uma língua natural baseia-se em um código ou sistema que não pode ser considerado estável - como já afirmava Saussure (2002) - e só pode ser apreendida por meio de produções verbais efetivas/empíricas, de caráter diversificado, sobretudo por serem articuladas em situações muito diferentes. A essas formas de realizações empíricas, o autor denomina de texto. Ainda, de acordo com o autor, os textos são produtos da atividade de linguagem em funcionamento permanente nas formações sociais. Em função de seus objetivos, interesses e questões especifícas, essas formações elaboram diferentes espécies de textos, que apresentam características relativamente estáveis (justificando que sejam chamadas de gêneros de texto) e que ficam disponíveis no intertexto como modelos indexados, para os contemporâneos e para as gerações posteriores. Portanto, toda forma de regularidade ocorre na forma de semiotização do discurso [4] e se vincula aos tipos de discurso, que podem ser da ordem do "narrar" ou do "expor", por exemplo. De acordo com Bronckart (1997), os tipos de discurso se relacionam com as representações dos mundos discursivos, que implicam unidades estruturais com combinações de diversas proposições organizadas e, além disso, constituem o produto da (re)organização dos conhecimentos disponíveis na memória do falante, que dividem-se em *narrativa, descritiva, explicativa, argumentativa, dialogal e injuntiva*. Ainda segundo o autor, os tipos de discurso apresentam um conjunto de fases que definem as peculiariedades das formações textuais, orientadas por dois eixos, o eixo do "narrar" e o eixo do "expor". Na Tabela 4.3, exibimos os tipos de discursos proposto pelo autor.

Tabela 4.3: Tipos de discurso por Bronckart (1997)

Tipo	**Peculiariedade**	**Fase**
Narrativo	Configuração de um processo de intriga	(i) fase de situação inicial: apresentação do "estado inicial das coisas"; (ii) fase de complicação : introdução do movimento de transformação ao previsto na ação discursiva e cria uma tensão; (iii) fase de resolução: introdução de acontecimentos que amenizam a tensão; (iv) fase de situação ao final: explicitação do novo equilíbrio obtido por essa resolução.

[4] O processo de semiotização do discurso, ou a operação de discursivização na língua, possibilita a passagem de uma referenciação externa à língua para o real construído pelo discurso, o que corresponde a um conjunto de operações estratégicas que permitem fazer a passagem do significado (sentido de língua) para a significação (sentido de discurso). Vocábulos, quando atualizados discursivamente. Assim, no enunciado "O homem é mortal" (Sócrates), o termo refere-se ao ser humano, mas colocado numa placa, em uma porta ao fundo de um bar, por exemplo, "homem" ganha significação de banheiro masculino e "mulher", de banheiro feminino.

Descritivo	Composição por fases que não se organizam em ordem linear obrigatoriamente, mas que se combinam e se encaixam em ordem hierárquica ou vertical	(i) fase de ancoragem: apresentação do tema-título que inicia a descrição (é ancoragem porque esse tema-título pode ser retomado ao longo de todo o processo descritivo); (ii) fase de aspectualização: enumeração de aspectos ligados ao tema-título; (iii) fase de relacionamento: assimilação dos elementos descritos a outros, por meio de operações de caráter comparativo ou metafórico.
Argumentativo	Existência de uma tese discutível	(i) fase de premissas: exposição de uma constatação de partida; (ii) fase de apresentação de argumentos: exposição de elementos que orientam para uma conclusao provável; (iii) fase de apresentação de contra-argumentos: restrição à orientação argumentativa; (iv) fase de conclusão: integração dos efeitos de argumentos e contra-argumentos apresentados.
Explicativo	Constatação de um fenômeno incontestável	(i) fase de constatação inicial: introdução de um fenômeno não contestável (objeto, situação, fato, etc); (ii) fase de problematização: explicitação de uma questão da ordem do porque ou do como, associada a um enunciado de contradição aparente; (iii) fase de resolução: introdução de informações suplementares capazes de responder a questões delineadas na fase de problematização; (iv) fase de conclusão-avaliação: reformulação e complementariedade da contestação inicial.

Dialogal	Realizações concretas somente nos segmentos de discursos interativos dialogados	Ocorre em três níveis. 1o nível - fase de abertura: exposição de caráter fático, na qual os interactantes estabelecem um contato com base nas convenções sociais; fase transacional: construção do conteúdo temático da interação (relação de interdependência dos tópicos e subtópicos conversacionais); fase de encerramento: exposição, também de caráter fático, na qual se póe fim à interação. 2o nível - fase dialogal ou de troca: caracterização de cada uma das fases gerais da interação, nas quais ocorrem diálogos entre os interactantes; 3o nível - fase de intervenção: decomposição da interação em atos discursios, ou seja, enunciados que realizam um ato de fata determinado (pedido, afirmação, injunção).
Injutivo	Orientação que visa a um fazer agir direcionado a um destinatário em uma determinada direção	fase 1 - descritiva: na qual há a exposição de elementos, conforme o objetivo a que se destina o texto; fase 2 - de procedimentos: tambem é uma etapa descritiva, porém apresenta um detalhamento da ação a ser realizada. Como o objetivo desta sequência é fazer agir, destacam como condições para sua constituição: o uso de formas verbais no infinitivo ou no imperativo e a ausência de estruturação espacial ou hierárquica.

Interessa-nos, neste trabalho, o *modelo de discurso descritivo*. Neste modelo, segundo o autor, na fase de aspectualização, o detentor de discurso enumera "aspectos" relacionados ao tema-título. Note a representatividade do discurso opinativo, neste conceito. O autor também caracteriza esse tipo de discurso por fases que não se organizam em uma ordem linear obrigatória, mas que se combinam e se encaixam em uma ordem hierárquica ou vertical. Essa característica representa com precisão revisões de usuários, em que os aspectos avaliados podem estar relacionados por relações semânticas hierárquicas.

As fases do discurso descritivo, segundo o autor, são, inicialmente, a fase de ancoragemgem, em que se apresenta o tema-título e a fase de aspectualização, que representa a fase de enumeração de aspectos ligados ao tema-título. Essas duas fases representam o modelo textual de revisões de usuários, em que o detentor do discurso apresenta o alvo/entidade avaliado ou tema-título e, em seguida, discorre sobre os aspectos desse alvo. Na última fase realiza-se a assimilação dos elementos descritos. Portanto, o texto descritivo por excelência consiste em uma percepção sensorial no intuito de relatar as impressões capturadas, de modo a propiciar a criação de uma imagem do objeto descrito na mente do leitor. Alem disso, essa descrição pode ser retratada apoiando-se sobre dois eixos: *o objetivo e o subjetivo*. Na descrição objetiva, o foco é

relatar as características do objeto de maneira precisa, próximo ao factual. A subjetiva perfaz-se de uma linguagem mais pessoal, na qual são permitidas opiniões, expressões de sentimentos e emoções, além do emprego de construções livres que revelem a identidade e a individualidade do autor do discurso.

Como já discutido anteriormente, a tarefa da mineração de opinião preocupa-se principalmente com o reconhecimento e extração de conteúdo subjetivo em textos. Ao analisar os três domínios distintos de revisões de usuários - smartphone, câmera e livro - constatamos que é possível encontrar tanto conteúdo objetivo quanto conteúdo subjetivo, e que o grau de ocorrência desses conteúdos em revisões tem fortes ligações com o domínio. Portanto, a heterogeneidade composicional textual e discursiva permite que caracterizemos revisões de usuários sobretudo no modelo *discursivo descritivo objetivo-subjetivo*.

Com o objetivo de mensurar as proporções de conteúdo descritivo objetivo e subjetivo em revisões de usuários, dividimos em duas classes a identificação dos aspectos: (i) aspectos sem nenhuma avaliação associada; (ii) aspectos com avaliação do usuário associada (geralmente sendo positiva, negativa ou neutra). Portanto, foram contabilizados todos os aspectos que possuíam associação com alguma opinião/sentimento e os aspectos que não possuíam nenhuma opinião/sentimento associado. Vejamos os seguintes exemplos exibidos nas Figuras 4.2, 4.3 e 4.4.

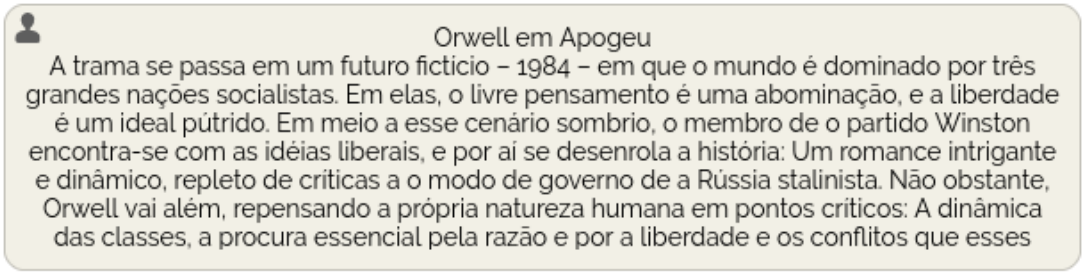
Orwell em Apogeu
A trama se passa em um futuro ficticio – 1984 – em que o mundo é dominado por três grandes nações socialistas. Em elas, o livre pensamento é uma abominação, e a liberdade é um ideal pútrido. Em meio a esse cenário sombrio, o membro de o partido Winston encontra-se com as idéias liberais, e por aí se desenrola a história: Um romance intrigante e dinâmico, repleto de críticas a o modo de governo de a Rússia stalinista. Não obstante, Orwell vai além, repensando a própria natureza humana em pontos críticos: A dinâmica das classes, a procura essencial pela razão e por a liberdade e os conflitos que esses

Figura 4.2: Revisão do domínio de livro (Freitas *et al.*, 2012).

Na revisão da Figura 4.2, alguns dos aspectos são “trama”, “romance”, “cenário” e “autor”. Veja que, nesta revisão, apesar do usuário citar esses aspectos inerentes à entidade “livro”, nao há nenhuma avaliação associada aos aspectos, tratando-se, portanto, de uma descrição objetiva dessas características. Os aspectos somente estão presentes na revisão para compor a descrição das características da entidade. Alem disso, é interessante observar que, apesar de alguns aspectos estarem acompanhados por adjetivos qualificadores, por exemplo, “cenários sombrios”, consideramos que esses adjetivos não remetem à avalição da entidade, e sim, a uma composição da descrição feita pelo usuário. Na revisão da Figura 4.3, observamos o mesmo fenômeno.

Na revisão da Figura 4.3, os aspectos são “livro/obra” e “leitura”. Veja que também não há nenhum tipo de avaliação associada aos aspectos, tratando-se também de uma descrição das características do objeto. Entretanto, o fenômeno muda na revisão da Figura 4.4. Nesta revisão, alguns dos aspectos são: “livro”, “autor”, “protagonista/personagens” e “história”. Note que há avaliações associadas aos aspectos, o que implica que o usuário explícitou sua experiência com o produto, avaliando subjetivamente as partes e propriedades deste produto.

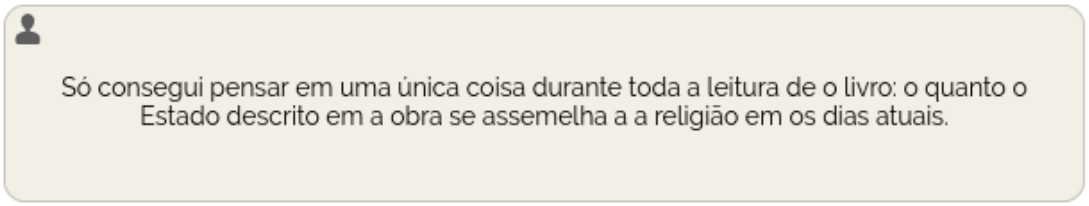

Figura 4.3: Mais uma revisão do domínio de livro (Freitas *et al.*, 2012).

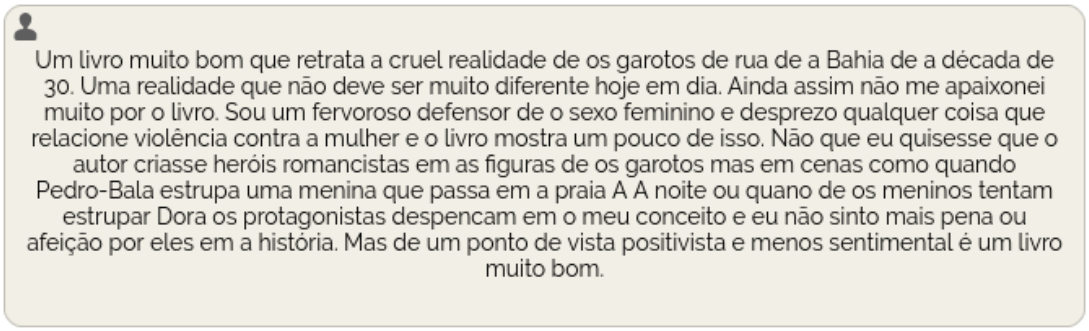

Figura 4.4: Mais uma revisão do domínio de livro (Freitas *et al.*, 2012).

Após a análise e identificação dos aspectos, de acordo com o conteúdo descritivo-objetivo e descritivo-subjetivo, obtivemos os resultados demonstrados na Tabela 4.4. Para o domínio de livro, de todos os aspectos presentes nas revisões de usuários, apenas 52,01% desses aspectos eram para avaliar o produto; o restante, cerca de 47,98%, se referiam apenas a uma descrição objetiva dos usuários em relação às propriedades do produto. Esses resultados demonstram o quão complexas são as tarefas de mineração de opinião, especialmente a mineração baseada em aspectos. Um sistema automático que não considere como critério de processamento os fenômenos linguísticos e/ou as especificidades do domínio, por exemplo, incorre no risco de classificar aspectos que não foram avaliados pelo usuário, portanto retornarão um resultado em desacordo com a realidade apresentada na revisão. Além disso, notamos que conteúdo descritivo-subjetivo, ou seja, que possuía opinião/sentimento explicitamente, estava acompanhado majoritariamente de verbos psicológicos, como ocorre, por exemplo, em "Achei a história meio parada", "Amei o livro" e "Embora eu não tenha gostado da história", sem necessariamente apresentar adjetivos. Nos domínios de smartphone e câmera, o conteúdo objetivo não foi estatisticamente relevante.

Tabela 4.4: Panorama de conteúdo descritivo objetivo e subjetivo no domínio de livro

Domínio	**Conteúdo Objetivo**	**Conteúdo Subjetivo**
Livro	47,98%	52,01%

4.1.3.3 Especificidades do domínio

Constatamos o quão complexo é identificar grupos de aspectos em diferentes domínios. Cada domínio exige um conhecimento relativo específico para que seja possível identificar e distinguir bem os grupos. Há muitas especificidades de domínio importantes que exigem certo

conhecimento de *background* para identificação. Por exemplo, o produto câmera digital possui o aspecto "lente", que também é conhecido pelo público especializado por "objetiva". Outro exemplo é o aspecto "*presets*", que trata-se de uma propriedade de pré-definições de ajustes de fotos, sendo usada recorrentemente por usuários especializados. Outro exemplo interessante está relacionado ao aspecto "resolução". Essa propriedade tambem é usada pelos usuários de forma intercambiada com o termo "megapixels". Portanto, os usuários consumidores do produto câmera digital, ao avaliarem os "megapixels" de uma câmera, estão avaliando a "resolução" dela. Para o domínio de smartphone, um exemplo interessante é a propriedade "*quadri-band*". Essa propriedade diz respeito ao tipo de sinal de comunicação do aparelho, ou seja, o usuário está avaliando o aspecto "sinal" do smartphone.

4.1.3.4 Ambiguidade

Durante o processo de agrupamento de aspectos, um dos desafios encontrados foi tratar a ambiguidade que é inerente às línguas naturais. Por exemplo, para o domínio de smartphone, os usuários utilizam os termos "recursos" e "funções" simultaneamente ora para falar de todos os aspectos do smarthpone, ora para designar algum aspecto, função, recurso ou aplicação específica, como "tv", "radio", etc. Para o domínio de livro, os usuários ora utilizam o termo "situações", referindo-se às passagens e/ou acontecimentos do livro, ora referindo-se ao assunto da história. Ainda no domínio de livro, os usuários também utilizam os termos "narrativa" ora para se referirem a "história", ora para se referirem ao "tipo de história". Esse comportamento também é recorrente com os termos "romance" e "trama". O termo "trama" ora é usado para se referir ao "romance do livro", ora para se referir ao "assunto da história". O termo "leitura" também é utilizado de forma ambígua. Em alguns casos, utiliza-se para se referir ao "tipo de leitura", por exemplo, em "É uma leitura pesada" ou em "A leitura do livro é instigante"; é também usado para se referir à entidade "livro", por exemplo, em "Eu recomendo a leitura do livro". Alem disso, aspectos que denotam vagueza são usados recorrentemente pelos usuários. Por exemplo, os aspectos "função" e "aplicativo" são muitas vezes usados pelos usuários para se referirem ao mesmo aplicativo do smartphone. Esse comportamento é intensificado quando se considera a informalidade desse tipo de texto produzido por usuários no ambiente web.

4.1.3.5 Aspectos implícitos

Nós reconhecemos e agrupamos aspectos implícitos e explícitos no domínio de smartphone, câmera e livro. Mensuramos os aspectos implícitos dos domínios a partir dos *termos pistas*. Um termo pista é um termo indicativo de aspecto implícito. Um panorama deste cenário é exibido na Tabela 4.5. Um dado interessante sobre esse estudo é a proximidade de comportamento entre os domínios de smartphone e câmera digital e o leve distanciamento dos resultados desses dois domínios em comparação com o domínio de livro. Note que há, em média, 0,95 de aspecto implícito em cada revisão para os três domínios. Além disso, foram reconhecidos até 10 aspectos implícitos em um mesmo grupo de aspectos, no domínio de câmera, 9 no domínio de livro e 8 no domínio de smartphone.

Tabela 4.5: Aspectos implícitos

	Smartphone	**Câmera**	**Livro**	**Média**
Total de aspectos implícitos	67	53	45	55
Total de aspectos implícitos (únicos)	38	23	12	24,33
Número médio por revisão	1,11	0,91	0,85	0,95
Numero máximo por revisão	4	5	3	4
Numero máximo por grupo	8	10	9	9

Realizamos também um mapeamento das classes gramaticais dos termos indicativos de aspectos implícitos. A seguir, na Tabela 4.6, apresentamos os resultados deste mapeamento. Dividimos os termos indicativos de aspectos em duas classes, *nominais* e *verbais*, a fim de mensurar a proporção de cada uma dessas classes nos domínios analisados. Na classe de nominais, enquadramos itens lexicais não-verbais, ou seja, substantivos, adjetivos, advérbios, etc. Na classe de verbais, foram enquadrados itens lexicas verbais, ou seja, verbos.

Tabela 4.6: Classificação dos termos indicativos de aspectos implícitos

Domínio	**Nominais**	**Verbais**
Smartphone	73,68%	26,31%
Câmera	69,56 %	30,43%
Livro	50,00%	50,00%
Média	**41,08**	**35,58**

Observe que a média dos domínios de ocorrência de indicativos de aspectos implícitos do tipo nominal é próxima à classe de verbais. Esse tipo de dado é importante, pois direciona esforços para exploração de conhecimentos sobre a classe de verbos de uma língua para aplicações de mineração de opinião. Além disso, é interessante observar um salto de termos indicativos de aspectos implícitos da classe dos nominais para o domínio de smartphone, com 73,68% de indicativos nominais.

4.1.3.6 Aspectos fora do domínio

Em nossas análises, observamos também que revisões de usuários podem conter aspectos que não implicam propriedades da entidade do domínio. Nos domínios câmera e smartphone, aspectos como “entrega”, “atendimento ao consumidor”, “sac” e “assistência técnica” foram avaliados pelos usuários, mas essas características não condizem com propriedades das entidades nos domínios analisados. Tratam-se, na verdade, de características relacionadas à empresa que vendeu o produto ou a marca do produto.

4.1.3.7 Relações entre aspectos

Os grupos completos, obtidos com a tarefa de reconhecimento e agrupamento de aspectos, encontram-se no Apêndice. As relações entre aspectos identificadas para os três domínios foram, principalmente de hiperonímia/hiponímia, meronímia/holonímia e sinonímia, construções deverbais e correferências. Descreveremos a seguir cada uma dessas relações:

- **Relação de sinonímia** é um tipo de relação semântica vertical entre uma ou mais unidades lexicais sinônimos. Por exemplo, os termos "tela" e "visor";

- **Relação de hiperonímia/hiponímia** é um tipo de relação semântica hierárquica entre duas unidades lexicais. É também chamado pela literatura como relação do tipo *i-sa* ou *é-um* (Vossen, 1997). Por exemplo, o termo "câmera" é um *hipônimo* do termo "máquina", logo "câmera" possui relação do tipo *é-um* com a unidade "máquina".

- **Relação de meronímia/holonímia** consiste em um tipo de relação semântica hierárquica entre duas unidades lexicais, uma denotando a parte (merônimo), que implica referência a um todo (holônimo), relativo a essa parte. Essa relação também é classificada pela literatura por *parte-todo* (Vossen, 1997). Por exemplo, a unidade lexical "tecla" é *merônimo* da unidade lexical "teclado", logo "tecla" é *parte-de* "teclado".

- **Construções deverbais** são relações entre uma unidade lexical não-verbal oriunda pela interação social de uma unidade verbal. Por exemplo, o aspecto "manuseio" é um deverbal da unidade lexical "manusear".

- **Relações entre referentes também chamados de correferência**, de acordo com Wasow (1967), é a possibilidade cognitiva de se estabelecer relação entre dois elementos *A* e *B*, quando *B*, tecnicamente chamado de elemento anafórico, recebe o conteúdo semântico total ou parcial de *A*, que é antecedente de *B*. Por exemplo, nomes próprios são utilizados de forma intercambiada com algum objeto. Veja as seguintes revisões, "Jorge amado é ruim" e "Eu não gostei do autor". Note que, nas duas revisões, o usuário avalia o mesmo aspecto, a propriedade "escritor".

Para o domínio de livro, houve 46,60% de relações de hiperonímia/hiponímia; para o domínio de smartphone, obtivemos 45,00%; e, no domínio de câmera digital, também obtivemos um número considerável de relações de hiperonímia/hiponímia, sendo igual a 37,12%. Identificamos também grupos unitários, que consistem de grupos em que não foram reconhecidos nenhum outro aspecto com correspondência semântica naquele domínio, portanto são grupos formados com apenas uma unidade lexical. Note a importância da tarefa de agrupamento de aspectos para mineração de opinião. Apenas 10% dos aspectos que ocorrem no domínio de smartphone não possuem correspondência semântica com outros aspectos, ou seja, são grupos unitários; o restante dos aspectos, ou seja 90%, estão relacionados com outros aspectos dentro do domínio. Note que, sem a execução da tarefa de agrupamento de aspectos, sistemas de mineração de opinião incorrem no risco de apresentar aspectos que se referem a uma mesma propriedades do objeto como sendo propriedades distintas.

4.1.3.8 Grupos prototípicos do domínio

Observamos a ocorrência de grupos de aspectos que são usados mais frequentemente pelos usuários em detrimento de outros grupos. Esses grupos de aspectos formam normalmente os

Tabela 4.7: Principais relações entre aspectos

	Smartphone	**Câmera**	**Livro**	**Média**
Hiperonímia/hiponímia	45,00%	37,12%	46,60%	42,90%
Sinonímia	23,88%	18,93%	26,21%	23,00%
Meronímia/holonímia	8,91%	15,18%	7,76%	10,61%
Construção deverbal	5,55%	6,81%	9,70%	7,35%
Correferências	6,66%	8,33%	0,00%	4,99%
Grupos unitários	10,00%	13,63%	9,73%	11,12%

grupos de aspectos prototípicos do domínio. Por exemplo, para o domínio de smartphone, alguns grupos de aspectos prototípicos são: "smartphone", "usabilidade", "design", "valor", "bateria', "marca", etc. Plotamos os resultados da identificação desses grupos prototípicos nas Figuras 4.5, 4.6, 4.7. Nesses gráficos, relacionamos o número de avaliações (eixo "número de avaliações") para cada grupo de aspectos (eixo "grupo de aspectos") identificados para os domínios smartphone, câmera e livro. Os grupos de aspectos marcados com as cores mais escuras representam os grupos de aspectos prototípicos (mais frequentemente avaliados pelo usuário) do domínio.

4.1.3.9 Curvas de aprendizagem

Com o objetivo de mensurar o comportamento do agrupamento de aspectos de opinião e identificar o ponto de estabilização para identificação de novos grupos em um domínio, descrevemos o que denominamos de *curvas de aprendizagem*, resultantes do processo de identificação de novos grupos de aspectos para os dominios de smartphone, câmera digital e livro. As curvas são exibidas nas Figuras 4.8, 4.9 e 4.10. O eixo X das curvas de aprendizagem representa a quantidade de revisões analisadas e o eixo Y a quantidade de novos grupos de aspectos identificados. Por exemplo, após análise da revisão número 1, exibida pela Figura 4.8, no eixo X, houve o reconhecimento de 8 grupos de aspectos, como mostra o eixo Y. Após a análise das dez primeiras revisões, houve o reconhecimento de 33 grupos de aspectos, e assim sucessivamente. Nós observamos que, para os domínios de smartphone, câmera digital e livro são necessários, em média, 40 revisões de usuários para o aprendizado de grupos de aspectos representativos do domínio.

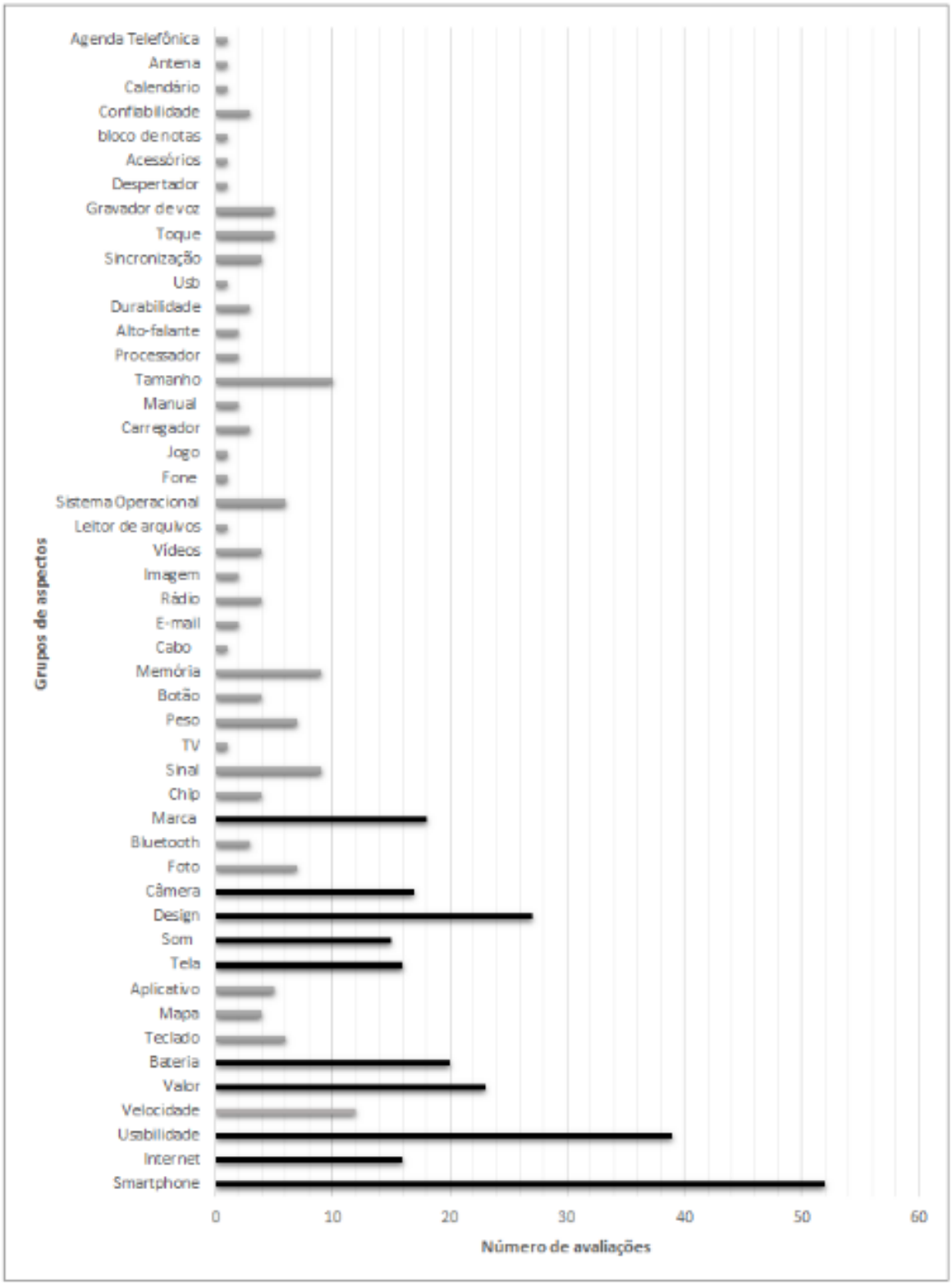

Figura 4.5: Número de avaliações para os grupos de aspectos do domínio de smartphone.

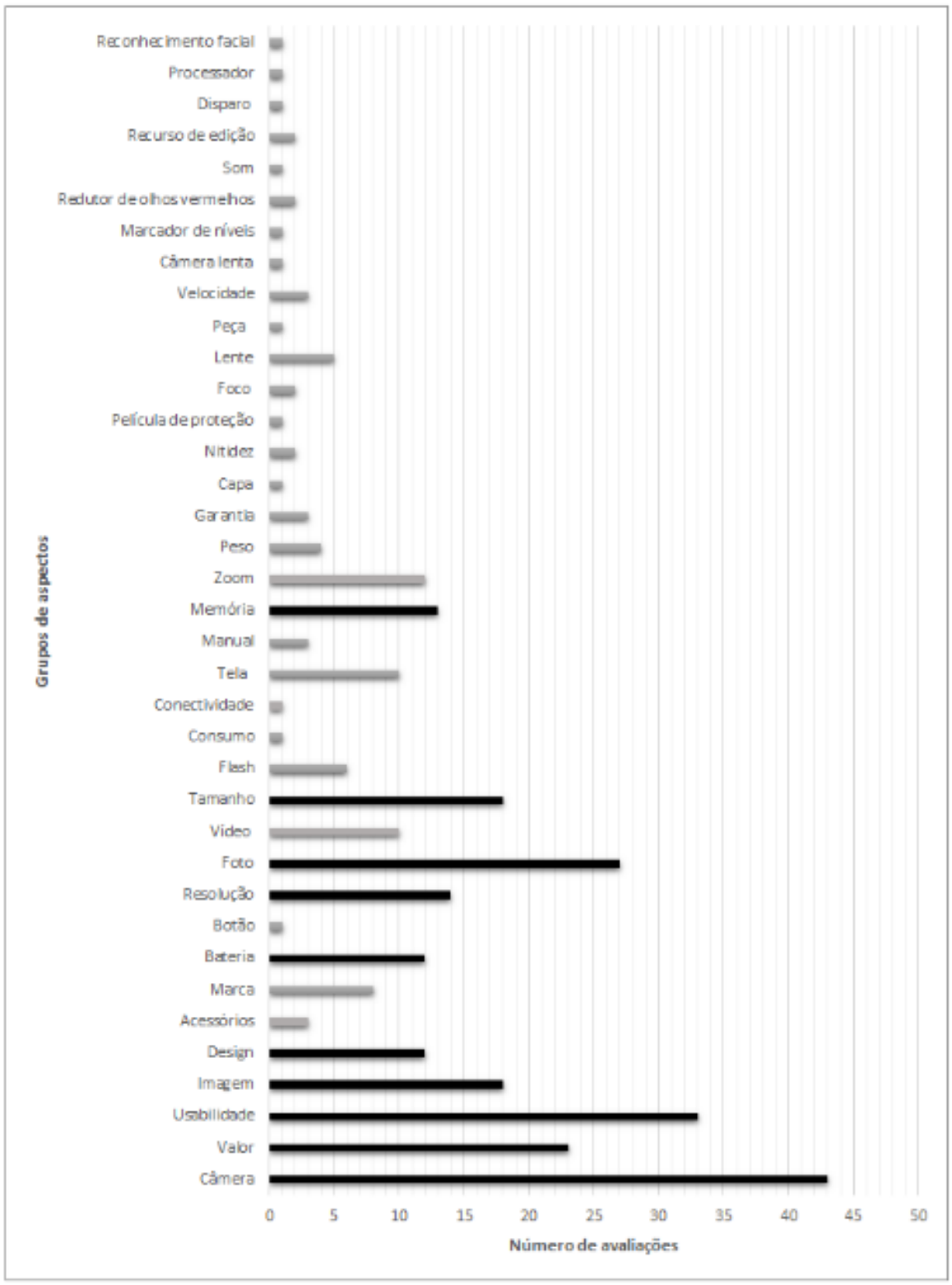

Figura 4.6: Número de avaliações para os grupos de aspectos do domínio de câmera digital.

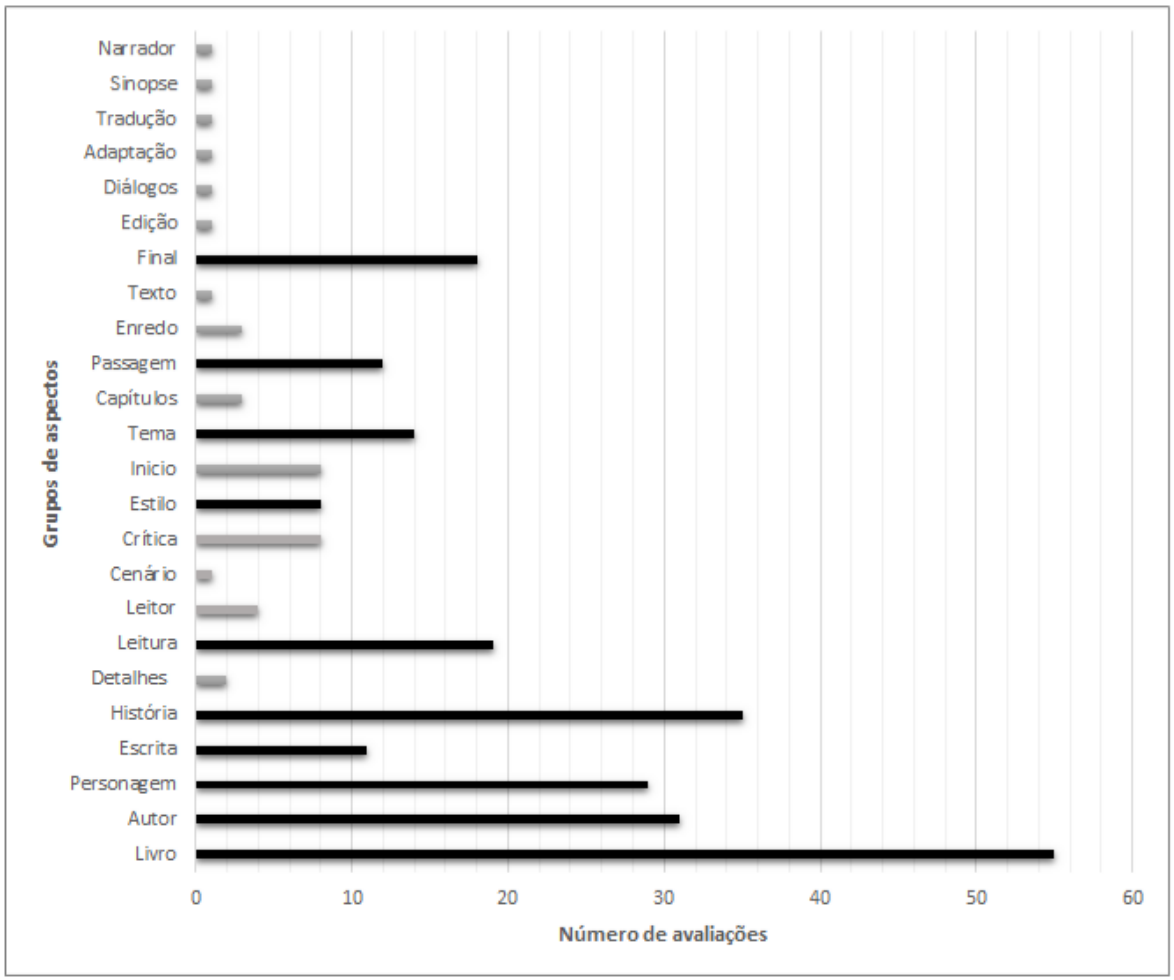

Figura 4.7: Número de avaliações para os grupos de aspectos do domínio de livro.

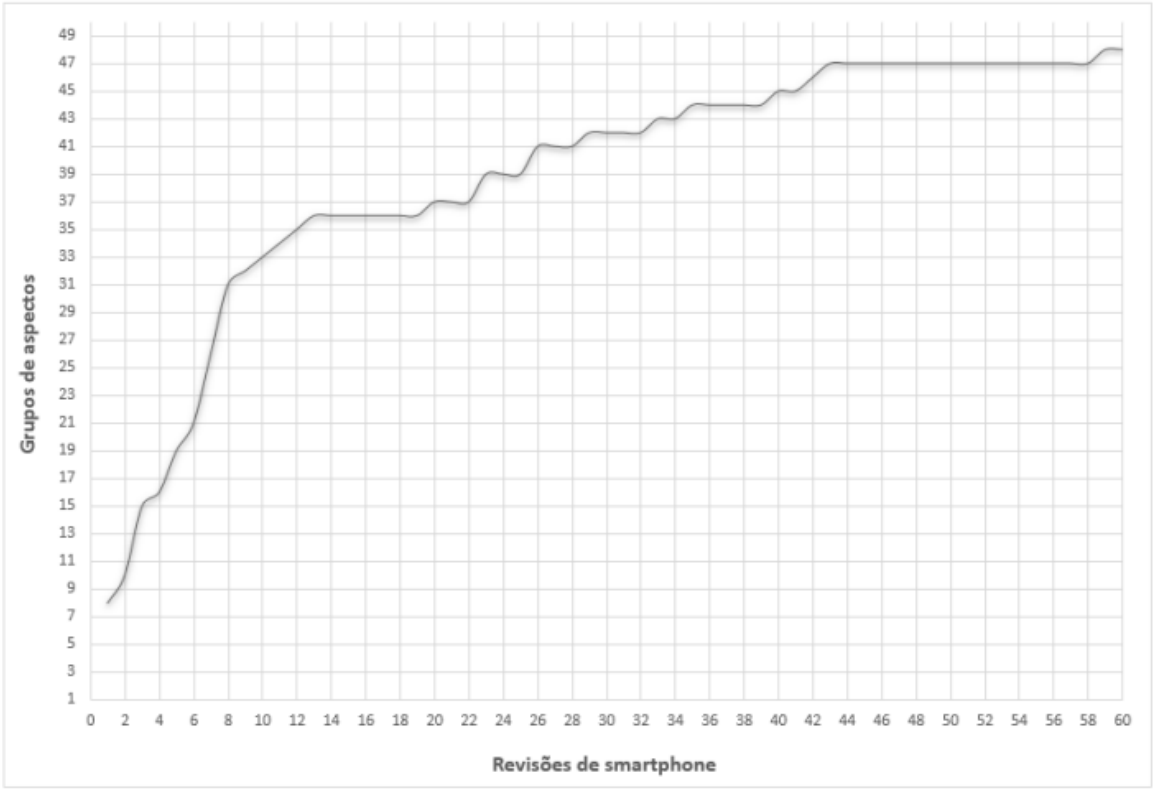

Figura 4.8: Curva de grupos aprendidos no domíno de smartphone.

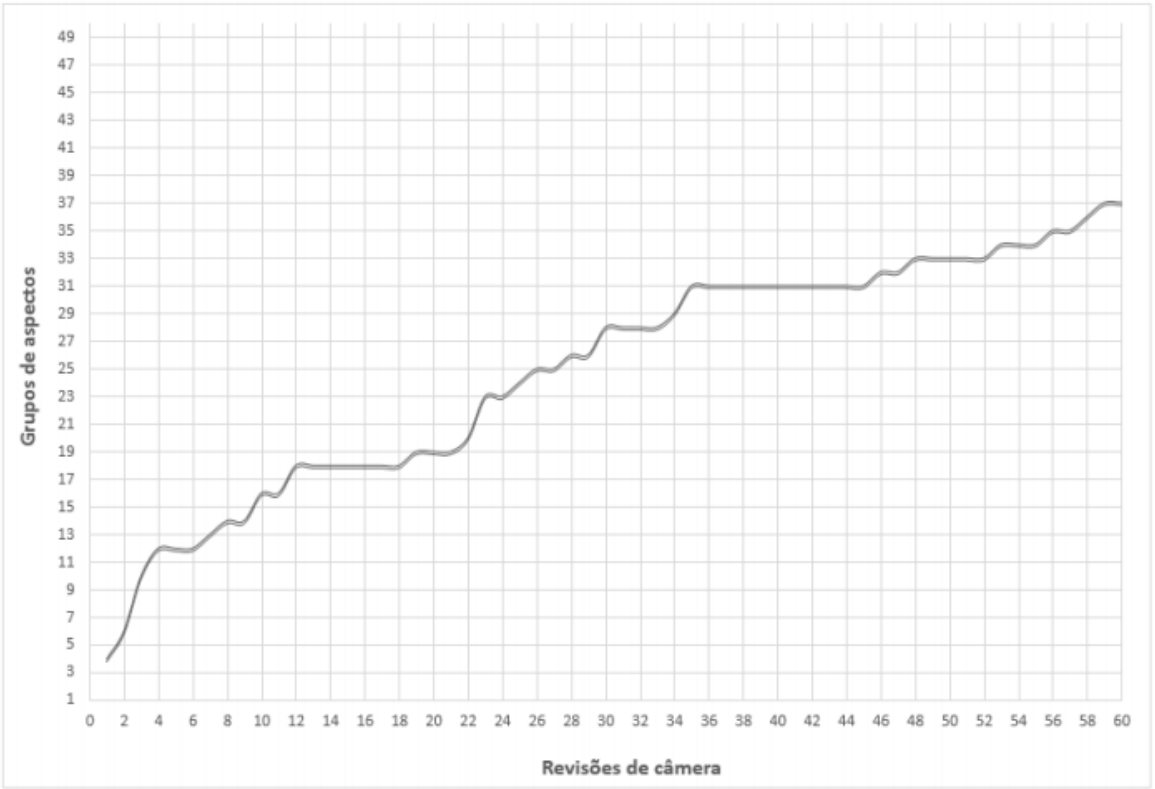

Figura 4.9: Curva de grupos aprendidos no domíno de câmera digital.

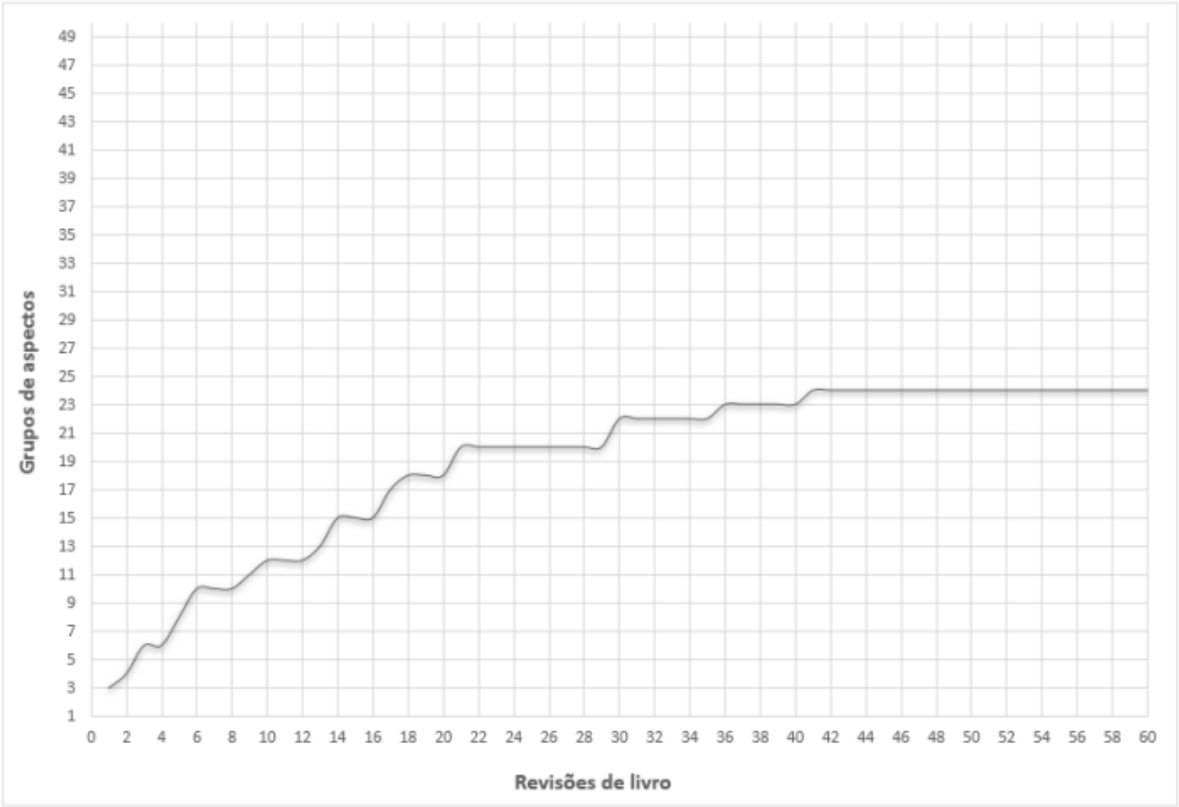

Figura 4.10: Curva de grupos aprendidos no domíno de livro.

4.2 Aprofundamento linguístico

De acordo com Lopes (1995), a linguística é uma ciência interdisciplinar, pois solicita empréstimo à sua instrumentação metalinguística dos dados elaborados pela estatística, pela teoria da informação, pela lógica matemática, etc., e, por outro lado, ela empresta os métodos e conceitos que elaborou à psicanálise, à musiologia, à antropologia, à teoria e crítica literária, à sociologia, etc. Ela também se dá como linguística aplicada ao ensino das línguas e à tradução mecânica. Ainda, de acordo com Lopes (1995), a aprendizagem, conservação, transformação e transmissão da cultura realizam-se através de uma grande variedade de práticas sociais. As práticas sociais organizam-se para expressar a cultura das comunidades humanas, assumindo a condição de sistemas de signos para transmitir essa cultura de um indivíduo para outro, de uma geração para a geração seguinte. Além disso, as línguas naturais ocupam a posição hierárquica predominante entre todos os sistemas semióticos, porque as línguas constituem a única realidade imediata para o pensamento de cada um dos indivíduos. Portanto, uma língua natural é um domínio complexo envolvido por fenômenos semióticos, cognitivos, empíricos e sociais e que influenciam na formação da significação. Sobre o eixo da significação, Todorov (1966) afirma que o fenômeno da significação é estrutural, pois os significados constituem estruturas dentro das línguas naturais e que a lexicologia forneceu a prova disso: a introdução de uma nova palavra ou um novo significado no léxico da língua não altera a estrutura semântica global da língua, pois cada nova unidade léxical é absorvida no interior de um campo ou grupo semântico

afim.

A imersão de uma nova unidade lexical e/ou de significação na língua é caracterizada pela neologia linguística como um processo que constitui a principal forma de inovação lexical de uma língua e que consiste a produção de formas e significados inéditos no léxico de uma língua (Lopes, 1995). Essa inovação lexical pode ser classificada pelas tipologias: neologia formal, neologia semântica e neologia de empréstimos. A neologia formal é aquela que acontece por meio de processos internos ao sistema linguístico. Esses processos acontecem tanto em nível morfológico, quanto sintático e fonológico. Um exemplo de neologia formal são as construções de diminutivos e construções deverbais do português, tais como "livro" e "livrinho" (diminutivo) e "pensar" e "pensamento" (construção deverbal). A neologia de empréstimos se dá pela importação de elementos de outros sistemas linguísticos, havendo ou não adaptação da forma importada. Exemplos clássicos são os estrangeirismos, tanto do inglês quanto do francês, por exemplo, os termos "deletar", "smartphone", "shampoo" e "abajur". A neologia semântica é aquela que opera no nível do significado e acontece mediante a atribuição de novos significados às formas lexicais existentes na língua, resultado de processos de expansão de sentido como metáfora, metonímia e polissemia (Ferraz, 2008). Um exemplo é o termo "engarrafamento", originalmente introduzido na língua com o significado de *ato ou processo industrial de engarrafar* [5], no entanto, também é usado para designar *acúmulo de veículos em determinado ponto da via pública, congestionamento* [6].

Analisando as relações estatisticamente relevantes entre aspectos, no domínio de produtos, foi possível observar um tipo de padrão relacional não óbvio entre essas unidades lexicais. Esse tipo de relação é resultado do fenômeno de *isotopia linguística* [7]. Relações isotópicas podem ser classificadas deste modo, pois representam relacionamento entre unidades lexicais correlatas ou termos com correspondência eclipsa [8], ou seja, são unidades lexicais que apesar de distintas quanto à forma, são equivalentes no campo da ideologia/significação. No domínio de opinião, especialmente em revisões de usuários sobre produtos, identificamos as relações lexicais formais de sinonímia, hiperonímia/hiponímia, meronímia/holonímia e os fenômenos de estrangeirismos, diminutivos/aumentativos e correferências. Nós obseramos também que os fenômenos de estrangeirismos e os diminutivos estavam presentes imersas nas relações lexicais formais identificadas. Por exemplo, os aspectos "livro" e "bestseller" são hiperônimo/hipônimo, no entanto, também trata-se de um fenômeno de estrangeirismo. Outro exemplo, "romance" e "romancezinho" são hiperônimo/hipônimo, no entanto, também é uma construção de diminutivo. A fim de quantificar a ocorrência de estrangeirismos e diminutivos no córpus, contabilizamos os estrangeirimos e diminutivos em cada um dos domínios e apresentamos na Tabela 4.8. Note que ocorreu, em média, 14,33 estrangeirismos e 3,33 diminutivos nos domínios analisados.

[5]Extraído de http://dicionariocriativo.com.br/

[6]Extraído de http://dicionariocriativo.com.br/

[7]De acordo com Lopes (1995), poder-se-ia, por exemplo, utilizar o nome de isomorfia para a correspondência localizável no sistema (ou na estrutura) de dois códigos, reservando-se o nome isotopia para a correspondência interpretativa, quer dizer, localizável no plano do conteúdo. Diríamos, então, que há isotopia (= correspondência de sentido).

[8]O fato de cada termo expressar uma visão de mundo diferente, pois supõe que sejam plenamente equivalentes.

Tabela 4.8: Estrangeirismos e diminutivos

	Smartphone	**Câmera**	**Livro**	**Média**
Estrangeirismos	26	16	1	14,33
Diminutivos	2	3	5	3,33

Além das relações e fenômenos descritos acima, nós observamos outro fenômeno muito interessante. Algumas *relações causativas* parecem também relacionar aspectos correlatos em textos opinativos. Por exemplo, identificamos os aspectos "fim" e o indicativo de aspecto implícito "terminar" no domínio de livro. Esses dois itens são usados pelos usuários para avaliar a propriedade "desfecho" da história do livro. Por exemplo, nas revisões "Eu amei o fim da história entre a bella e o vampiro" ou em "A história termina não menos que incrivelmente". Note que nas duas revisões o usuário avalia a mesma propriedade do livro.

Observamos que as *relações causativas* não ocorreram de forma expressiva no córpus analisado, no entanto, vale salientar que as condições linguísticas em que textos opinativos são produzidos, potencializam a probabilidade de ocorrência deste fenômeno, que pode sofrer variações em decorrência das especificidades do domínio e dos perfis dos usuários.

Por fim nós constatamos que aspectos correlatos não são encontrados em textos opinativos apenas pela relação lexical de sinonímia. Os insumos acionáveis para compreensão desse fenômeno são oriundos, principalmente da neologia e da correferência linguística, ou seja, a partir do reconhecimento e tratamento desses fenômenos de uma língua natural é possível propor métodos automáticos de agrupamento de aspectos, pois são esses os principais fenômenos linguísticos que desencadeiam o problema.

Capítulo

5

Experimentos

O principal objetivo desta proposta de pesquisa de mestrado foi compreender os fenômenos entre aspectos em revisões de usuários, mapeá-los e propor métodos automáticos baseados em conhecimento e especialmente motivados linguisticamente para resolução do problema de agrupamento de aspectos para sistemas de mineração de opinião. Portanto, a partir de um estudo linguístico aprofundado e empírico baseado em córpus, nós propusemos, implementamos e comparamos 6 métodos de reconhecimento de grupos de aspectos a partir de revisões sobre produtos. O córpus utilizado consiste de textos opinativos sobre smartphones, câmeras e livros. Esses dados linguísticos foram anotados quanto ao aspecto explícito, o termo indicativo de aspecto implícito e o grupo de aspectos (todo o processo de anotação foi apresentado na Capítulo 4). Um conjunto de referência (humano) foi criado, a partir do processo de anotação desses dados. Esse conjunto de referência foi usado para avaliação dos métodos implementados e, além disso, servirá como recurso para pesquisas futuras. As informações do córpus e do conjunto de referência são exibidas na Tabela 5.1.

Tabela 5.1: Informações do córpus e do conjunto de referência (humano)

N.	**Domíno**	**N. de revisões**	**Total de aspectos (únicos)**	**Aspectos explícitos**	**Aspectos implícitos**	**Grupos**
1	Smartphone	60	180	452	38	48
2	Câmera	60	132	109	23	36
3	Livro	60	103	91	12	21

Nós observamos uma diferença significativa entre o número de aspectos e os grupos de aspectos entre os domínios de smartphone e câmera em relação ao domínio de livro. Nossa hipotése para esse comportamento consiste, principalmente, sobre as especificidades de domínio e no perfil dos usuários. Por exemplo, nos domínios de smartphone e câmera, os usuários

possuem maior conhecimento sobre os aspectos mais específicos dos objetos por se tratarem de produtos "populares". No entanto, no domínio de livro, os usuários, geralmente não são críticos literários ou profissionais de literatura, ou seja, são usuários leitores e não especialistas no domínio e, por isso, são capazes de avaliar apenas as características superficiais do objeto.

A partir de um estudo de córpus sobre os fenômenos linguísticos e estatisticamente relevantes em textos opinativos, nós propusemos e implementamos 6 métodos de agrupamento de aspectos. Os 6 métodos agrupam aspectos explícitos e indicativos de aspectos implícitos, sendo que os 3 primeiros métodos implementados são baseados em similaridade lexical, 1 método baseado em similaridade lexical e correlações linguísticas, 1 método estatístico baseado em semântica vetorial, e, por fim, propusemos um método novo, resultado do estudo linguístico aprofundado e do refinamento das análises sobre os resultados obtidos com os demais métodos implementados.

O primeiro método do experimento reconhece relações lexicais de sinonímia entre aspectos usando a Onto-PT (Oliveira, 2014). Esse método tem sido usado recorremente pela literatura como *baseline* (Zhai *et al.* (2011); Zhang *et al.* (2011)). O segundo método implementado reconhece relações de sinonímia incrementado por relações de hiperonímia/hiponímia. Também utilizamos a Onto-PT (Oliveira, 2014) para extração das relações. O terceiro método utiliza relações de sinonímia e hiperonímia/hiponínia, incrementado por relações de meronínia/holonímia, e a Onto-PT (Oliveira, 2014) foi novamente usada. O quarto método extrai, além das relações lexicais entre aspectos descritas no método anterior, as cadeias de referentes ou correferências. Foram utilizados a Onto-PT (Oliveira, 2014) e o sistema de resolução de correferência CORP (Fonseca *et al.*, 2016) e sua versão CorrefVisual (Fonseca, 2014). Para implementação do quinto método, optamos por um modelo estatístico baseado na proposta de semântica vetorial. Utilizamos *word embeddings* e o algoritmo *word2vec* (Mikolov *et al.*, 2013) e optamos pela arquitetura *skip-gram* de 300 dimensões. Por fim, nós propomos e implementamos um método novo. O algoritmo OpCluster-PT é resultado do refinamento dos métodos implementados e do estudo linguístico aprofundado sobre textos opinativos. Neste método, nós utilizamos a Onto-PT (Oliveira, 2014) para extração de relações lexicais entre aspectos de sinonímia, meronímia/holonímia e as relações causativas *resultadoDaAção* e *serveParaAccao*. Para o reconhecimento de correferências, nós utilizamos o sistema de resolução de correferências CORP (Fonseca *et al.*, 2016) e o CorrefVisual (Fonseca, 2014). Também utilizamos o dicionário de estrangeirismos (Ferreira & Janssen, 2017) e o dicionário de nomes deverbais (Janssen & Ferreira, 2007) do português desenvolvido pelo iLteC. Para identificação de diminutivos, nós criamos uma lista de diminutivos/aumentativos, pois não encontramos esse recurso acessível para o português. A descrição deste e dos demais recursos linguístico-computacionais utilizados nesta proposta de mestrado encontra-se na Seção 2.3.

Os métodos foram implementados usando a linguagem de programação Python, versão 2.7. Utilizamos, principalmente, as bibliotecas *RdfLib* [1] para construção de *querys* de navegação no modelo de dados da ontologia lexical, e a biblioteca *BeautifulSoup* [2] para as buscas nos arquivos

[1] https://github.com/RDFLib

[2] https://pypi.python.org/pypi/beautifulsoup4

XML gerado pelo CORP (Fonseca *et al.*, 2016). A seguir, na Tabela 5.2, exibimos uma síntese dos experimentos e dos recursos linguístico-computacionais aplicados.

Tabela 5.2: Sintése dos experimentos

N.	Método	Recursos linguístico-computacionais
1	Sinônimos	Onto-PT.
2	Sinônimos + hiperônimos/hipônimos	Onto-PT.
3	Sinônimos + hiperônimos/hipônimos + merônimos/holônimos	Onto-PT.
4	Sinônimos + hiperônimos/hipônimos + merônimos/holônimos + correferências	Onto-PT, CORP e CorrefVisual.
5	*Word embeddings*	Repositório de word embeddings do NILC.
6	OpCluster-PT	Onto-PT, CORP e CorrefVisual, dicionários de estrangeirismos e nomes deverbais do iLteC e uma lista de diminutivos/aumentativos.

5.1 Métodos baseados em similaridade lexical

Nós implementamos 3 métodos baseados em similaridade lexical para a tarefa de agrupamento de aspectos. A Onto-PT (Oliveira, 2014) foi usada para extração automática de relações de sinonímia, hiperonímia/hiponímia e meronímia/holonímia. Os métodos foram implementados de forma incremental a fim de avaliar os resultados obtidos em cada nível de incremento. Por exemplo, o primeiro método implementado, baseado em similaridade lexical, foi o método de extração automática de relações de sinonímia. O segundo método consiste da extração automática de relações de sinonímia incrementado pela extração de relações de hiperonímia e hiponímia. No terceiro método, além das relações de sinonímia e hiperonímia/hiponímia, incrementamos com a extração de relações de meronímia/holonímia. A seguir, faremos uma descrição detalhada da implementação de cada método e apresentaremos um indicativo inicial do desempenho desses métodos.

5.1.1 Relações de sinonímia

O primeiro método implementado consiste da extração automática de relações de sinonímia entre aspectos, em revisões de usuários. O algoritmo é exibido a seguir (ver Algoritmo 3).

O Algoritmo 3 recebe como entrada uma lista de aspectos A, ordenados de forma decrescente com base na frequência em que ocorrem no córpus. O item a_i de A é lido no laço de repetição. Se a_i possuir sinônimos na Onto-PT (Oliveira, 2014), os sinônimos encontrados são armazenados em b_{sin} de B. Em seguida, é verificado se B possui itens duplicados e, se houver, eles são excluídos. Em seguida, o grupo G_{i} é formado com os itens da intersecção *(A,B)*. O

Algoritmo 3: Algoritmo de agrupamento com base em relações de sinonímia

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência;

Saída: Grupos de aspectos $G = \{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

```
início
    Declare B = {b_sin}, tal que B contém o resultado da busca por aspectos em relação de sinonímia;
    Declare contador = 0;
    repita
        se a_i de A possuir sinônimos na base do Onto.PT então
            Adiciona em b_sin os sinônimos encontrados;
        fim
        Exclua itens duplicados de B = {b_sin}, se houver;
        Crie grupo G_i e adicione em G_i os aspectos da intersecção (A,B);
        Incremente contador;
        Remova de A os aspectos da intersecção;
        Esvazie B;
    até A esvaziar;
fim
```

contador é incrementado. Por fim, removem-se de *A* os itens da intersecção *(A,B)* e esvazia-se *B*.

Na Tabela 5.3, apresentamos o número de grupos gerados por esse método automático em relação ao número de grupos da referência (humano). Note uma diferença expressiva no número de grupos obtidos pelo método automático em relação ao número de grupos da referência (humano), portanto esse método automático agrupou um número pequeno de aspectos. Foram obtidos automaticamente 162 grupos e, desses grupos, 145 eram grupos unitários para o domínio de smartphone. No domínio de câmera, obtivemos 126 grupos automaticamente e, destes grupos, 117 eram grupos unitários. Para o domínio de livro, foram obtidos 84 grupos automaticamente, dos quais, 69 eram grupos unitários. Esses números são um indicativo inicial da ineficiência deste método para a tarefa de agrupamento de aspectos nos domínios analisados. Por exemplo, este método retornou 89,50% de grupos unitários para o domínio de smartphone, sendo que a referência (humano), para o mesmo domínio, possui 24,13% de grupos unitários. No Capítulo 6, apresentaremos também a avaliação deste método e de todos os outros métodos implementados usando medidas de avaliação comumente usadas pela literatura. Os dados apresentados na Tabela 5.3 são indicativos iniciais do desempenho deste método.

Tabela 5.3: Grupos gerados pelo Algoritmo 3.

Domínio	**Smartphone**	**Câmera**	**Livro**
Referência (humano)	48	36	21
Método automático	162	126	84

5.1.2 Relações de sinonímia e hiperonímia/holonímia

O segundo método implementado consiste da extração automática de relações de sinonímia e hiperonímia/holonímia entre aspectos. O algoritmo é exibido a seguir (ver Algoritmo 4).

Algoritmo 4: Algoritmo de agrupamento com base em relações de sinonímia e hiperonímia/hiponímia

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência;
Saída: Grupos de aspectos $G = \{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

```
1  início
2      Declare B = {b_sin, b_hipe e b_hipo }, tal que B contém o resultado da busca por aspectos
         em relação de sinonímia e hiperonímia/hiponímia;
3      Declare contador = 0;
4      repita
5          se a_i de A possuir sinônimos na base do Onto.PT então
6              Adiciona em b_sin os sinônimos encontrados;
7          fim
8          se a_i de A possuir hiperônimos/hipônimos imediatos na base do Onto.PT então
9              Adiciona em b_hipe os hiperônimos encontrados e em b_hipo os hipônimos
                 encontrados;
10         fim
11         Exclua itens duplicados de B = {b_sin, b_hipe, b_hipo}, se houver;
12         Crie grupo G_i e adicione em G_i os aspectos da intersecção (A,B);
13         Incremente contador;
14         Remova de A os aspectos da intersecção;
15         Esvazie B;
16     até A esvaziar;
17 fim
```

O Algoritmo 4 recebe como entrada uma lista de aspectos A, ordenados de forma decrescente com base na frequência em que ocorrem no córpus. O item a_i de A é lido no laço de repetição. Se a_i possuir sinônimos na Onto-PT (Oliveira, 2014), os sinônimos encontrados são armazenados em b_{sin}. Em seguida, é verificado se a_i possui hiperônimos/hipônimos na Onto-PT (Oliveira, 2014) e, se houver, os hiperônimos encontrados são adicionados em b_{hipe} e os hipônimos encontrados são adicionados em b_{hipo}. Verifica-se, em seguida, se B possui itens duplicados e, se houver, eles são excluídos. Em seguida, o grupo G_i é formado com os itens da intersecção *(A,B)*. O *contador* é incrementado. Por fim, removem-se de A os itens da intersecção *(A,B)* e esvazia-se B.

Na Tabela 5.4, apresentamos o resultado do número de grupos obtidos automaticamente em relação aos grupos da referência (humano). Por exemplo, esse método automático retornou, para o domínio de livro, 76% de grupos unitários, sendo que os grupos unitários do conjunto de referência para o mesmo domínio é igual a 28,57%. Observamos também um número alto de grupos gerados automaticamente se comparados com a referência (humano), no entanto, esse número ainda é menor comparado ao número de grupos obtidos pelo método que extrai

apenas relações de sinonímia. No Capítulo 6, iremos apresentar e discutir a precisão deste método e veremos que, apesar desse método ter agrupado mais aspectos em relação ao método de sinônimos, a precisão dele é menor, ou seja, muitos aspectos foram agrupados indevidamente.

Tabela 5.4: Grupos gerados pelo método Algoritmo 4.

Domínio	**Smartphone**	**Câmera**	**Livro**
Referência (humano)	48	36	21
Método automático	146	116	68

5.1.3 Relações de sinonímia, hiperonímia/hiponímia e meronímia/holonímia

O terceiro método implementado consiste da extração automática de relações lexicais de sinonímia e hiperonímia/hiponímia, incrementado pelas relações de meronínima/holonímia. O algoritmo é exibido a seguir (ver Algoritmo 5).

Algoritmo 5: Algoritmo de agrupamento com base em relações de sinonímia, hiperonímia/hiponímia e meronímia/holonímia

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência;
Saída: Grupos de aspectos G = $\{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

início
  Declare **B** = $\{b_{sin}, b_{hipe}, b_{hipo}, b_{mero}$ e $b_{holo}\}$, tal que **B**, contém o resultado da busca por aspectos em relação de sinonímia, hiperonímia/hiponímia e meronímia/holonímia;
  Declare *contador* = 0;
  repita
    se a_i *de* ***A*** *possuir sinônimos na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{sin}}$ os sinônimos encontrados;
    fim
    se a_i *de* ***A*** *possuir hiperônimos/hipônimos imediatos na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{hipe}}$ os hiperônimos encontrados e em $\mathbf{b_{hipo}}$ os hipônimos encontrados;
    fim
    se a_i *de* ***A*** *possuir meronímias/holonímias imediatos na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{mero}}$ os merônimos encontrados e em $\mathbf{b_{holo}}$ os holônimos encontrados;
    fim
    Exclua itens duplicados de **B** = $\{b_{sin}, b_{hipe}, b_{hipo}, b_{mero}, b_{holo}\}$, se houver;
    Crie grupo $\mathbf{G_i}$, tal que adicione em $\mathbf{G_i}$ os aspectos da intersecção **(A,B)**;
    Incremente *contador*;
    Remova de **A** os aspectos da intersecção;
    Esvazie **B**;
  até ***A*** *esvaziar*;
fim

O Algoritmo 5 recebe como entrada uma lista de aspectos, ordenados por frequência de ocorrência no córpus em ordem decrescente. O item a_i de A é lido no laço de repetição. Se a_i possuir sinônimos na Onto-PT (Oliveira, 2014), os sinônimos encontrados são armazenados em b_{sin}. Em seguida, é verificado se a_i possui hiperônimos/hipônimos imediatos na Onto-PT (Oliveira, 2014) e, se houver, os hiperônimos encontrados são adicionados em b_{hipe} e os hipônimos encontrados são adicionados em b_{hipo}. Conseguinte, é verificado se a_i possui merônimos/holônimos na Onto-PT (Oliveira, 2014) e, se houver, os merônimos encontrados são adicionados em b_{mero} e os holônimos encontrados adicionados em b_{holo}. Verifica-se, em seguida, se B possui itens duplicados e, se houver, eles são excluídos. Em seguida, o grupo G_{i} é formado com itens da intersecção *(A,B)*. O *contador* é incrementado. Por fim, removem-se de A os itens da intersecção *(A,B)* e esvazia-se B.

Neste método, nós observamos resultados muito próximos do método anterior (Algoritmo 4). Uma hipotése para esse comportamente consiste da composição de relações lexicais da Onto-PT (Oliveira, 2014). Por exemplo, essa ontologia possui um número inferior de relações de meronímia/holonímia em detrimento das relações de sinonímia e as relações de hiperonímia/holonímia. (ver descrição completa desses dados na Seção 2.3). Além disso, constatamos que em textos opinativos, as relações de meronímia/holonímia ocorrem predominantemente como *substring* e/ou através de aspectos específicos do domínio. Por exemplo, os aspectos "câmera" e "imagem da câmera" ou os aspectos "lente" e "foco", que são aspectos em relação de *substring* (primeiro par de aspectos) e com relação entre aspectos específicos do domínio (segundo par de aspectos). Na tabela 5.5, apresentamos o número de grupos formados por esse método em relação ao número de grupos da referência (humano).

Tabela 5.5: Grupos gerados pelo Algoritmo 5.

Domínio	**Smartphone**	**Câmera**	**Livro**
Referência (humano)	48	36	21
Método automático	147	116	68

5.2 Método baseado em similaridade lexical e correlações

Propusemos e implemetamos um método baseado em similaridade lexical que extrai automaticamente relações lexicais de sinonímia, hiperonímia/hiponímia, meronímia/holonímia a partir da Onto-PT (Oliveira, 2014) e, além disso, extrai correlações entre aspectos referentes ou correferentes usando o sistema de resolução de correferências para o português CORP (Fonseca *et al.*, 2016). A descrição detalhada deste método será realizada a seguir. É interessante ressaltar que esta abordagem é exclusiva deste trabalho de mestrado, pois não encontramos na literatura da área nenhum método que explorasse as correferências para agrupamento de termos correlatos da língua portuguesa.

5.2.1 Relações de sinonímia, hiperonímia/hiponímia, meronímia/holonímia e correferências

Para implementação deste método, extraímos automaticamente, além das relações lexicais de sinonímia, hiperonímia/hiponímia e meronímia/holonímia, as correferências em revisões de usuários. O algoritmo é exibido a seguir (ver Algoritmo 6).

Algoritmo 6: Algoritmo de agrupamento com base em relações de sinonímia, hiperonímia/hiponímima, meronímia/holonímia e correferências

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência; Revisões processadas pelo Corp $R = \{r_1, r_2, ...r_n\}$, em que os aspectos de A ocorrem;

Saída: Grupos de aspectos $G = \{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

```
1  início
2      Declare B = {b_sin, b_hipe, b_hipo, b_mero, b_holo e b_corref }, tal que B contém o resultado da
         busca por aspectos em relação de sinonínimia, hiperonímia/hiponímia,
         meronímia/holonímia e correferências;
3      Declare contador = 0;
4      repita
5          se a_i de A possuir sinônimos na base do Onto.PT então
6              Adiciona em b_sin os sinônimos encontrados;
7          fim
8          se a_i de A possuir hiperônimos/hipônimos imediatos na base do Onto.PT então
9              Adiciona em b_hipe os hiperônimos encontrados e em b_hipo os hipônimos
                 encontrados;
10         fim
11         se a_i de A possuir merônimos/holônimos imediatos na base do Onto.PT então
12             Adiciona em b_mero os merônimos encontrados e em b_holo os holônimos
                 encontrados;
13         fim
14         se a_i de A, nas revisões em que ocorre, possuir correferências classificadas pelo
             CORP então
15             Adiciona em b_corref as cadeias de correferentes encontrados;
16         fim
17         Exclua itens duplicados de B = {b_sin, b_hipe, b_hipo, b_mero, b_holo, b_corref}, se houver;
18         Crie grupo G_i, tal que adicione em G_i os aspectos da intersecção (A,B);
19         Incremente contador;
20         Remova de A os aspectos da intersecção;
21         Esvazie B;
22     até A esvaziar;
23 fim
```

O Algoritmo 6 recebe como entrada uma lista de aspectos A, ordenados por frequência de ocorrência no córpus em ordem decrescente. O item a_i de A é lido no laço de repetição. Se a_i possuir sinônimos na Onto-PT (Oliveira, 2014), os sinônimos encontrados são armazenados em b_{sin}. Em seguida, é verificado se a_i possui hiperônimos/hipônimos diretos na Onto-PT (Oliveira, 2014) e, se houver, os hiperônimos encontrados são adicionados em b_{hipe} e os hipônimos encon-

trados são adicionados em b_{hipo}. Conseguinte, é verificado se a_i possui merônimos/holônimos na Onto-PT (Oliveira, 2014) e, se houver, os merônimos encontrados são adicionados em b_{mero} e os holônimos encontrados em b_{holo}. Na última condição, é verificado se a_i possui cadeias de correferentes nas revisões em que ocorrem, classificadas pelo CORP (Fonseca *et al.*, 2016) e, se houver, as correferências encontradas são adicionadas em b_{coref}. Verifica-se, em seguida, se *B* possui itens duplicados e, se houver, eles são excluídos. Em seguida, o grupo G_i é formado com os itens da intersecção *(A,B)*. O *contador* é incrementado. Por fim, removem-se de *A* os itens da intersecção *(A,B)* e esvazia-se *B*.

Na Tabela 5.6, apresentamos o número de grupos obtidos por esse método automático em relação aos grupos de referência (humano). É interessante observar o recuo no número de grupos obtidos neste método em relação aos métodos que utilizam apenas similaridade lexical. Portanto, o número de grupos gerados automaticamente mais próximo do número de grupos da referência é um bom indicativo inicial da potencialidade deste método para a tarefa de identificação de grupos de aspectos para mineração de opinião. Nós constatamos também que a utilização de correferências captura um número maior de relações entre aspectos que representam as especificidades de um domínio. Por exemplo, a relação entre o aspecto "autor" e o aspecto "jorge amado" é uma relação do tipo hiperonímia/hiponímia, em que o aspecto "autor" é *hiperônimo* do aspecto "jorge amado". No entanto, essa relação não foi identificada através do método que utiliza apenas a ontologia lexical para extração de relações de hiperonímia/hiponímia, pois o aspecto "jorge amado" é um aspecto específico do domínio "livro" e unidades lexicais específicas de um domínio, geralmente são difíceis de serem encontradas em ontologias lexicais da língua.

Tabela 5.6: Grupos gerados pelo Algoritmo 6.

Domínio	**Smartphone**	**Câmera**	**Livro**
Referência (humano)	48	36	21
Método automático	71	100	54

5.3 Semântica Vetorial

Neste trabalho, utilizamos o conceito de *word embeddings* para a tarefa de identificação de grupos de aspectos de opinião. De acordo com Jurafsky & Martin (2000), *word embeddings* consistem em uma técnica em que o significado de uma palavra é definido pela frequência com que ocorre perto de outras palavras. De acordo com o autor, métodos como esse são frequentemente referenciados por semântica vetorial. A seguir, faremos uma descrição detalhada do desempenho deste método para a tarefa de agrupamento de aspectos, além de suas implicações.

5.3.1 *Word Embeddings*

Em métodos baseados em semântica vetorial, parte da compreensão do fenômeno da significação de uma unidade lexical consiste das unidades lexicais vizinhas. De acordo com Firth

(1957), para se conhecer o significado de uma palavra, basta olhar as companhias que ela mantém. Assim também propõe Matoré (1973), através do conceito de *palavras-testemunho*, que consiste de unidades lexicais "vizinhas" de uma unidade lexical nominal (não-verbal) alvo que, de acordo com o autor, carregam traços semânticos/ideológicos desta unidade lexical alvo.

Nesta proposta de mestrado, optamos pela utilização do algoritmo *word2vec* [3] e o modelo linear *skip-gram* de 300 dimensões proposto por Mikolov *et al.* (2013). O *word2vec* consiste de dois modelos lineares para computação de *word embeddings*: (i) o modelo CBOW, (do inglês, *continuous bag-of-words*), que prediz a palavra atual com base nas palavras do contexto [4]; e o (ii) modelo *skip-gram*, que prediz palavras em torno da palavra atual. A decisão pelo algoritmo, modelo e dimensões foi tomada a partir da nossa compreensão de que os modelos definidos seriam adequados à tarefa dessa proposta de trabalho. Utilizamos, especificamente para essa tarefa, um modelo pré-treinado proposto por Hartmann *et al.* (2017) e disponível no Repositório de *Word Embeddings* do NILC [5]. O modelo foi treinado a partir textos de domínios distintos da língua geral e extraídos a partir de websites em português do Brasil.

Em revisões de usuários, observamos que os textos geralmente são curtos e caracterizados por discursos contendo certo grau de conteúdo implícito, portanto, para compreensão semântica deste domínio, fazem-se necessários mecanismos mentais de inferência para a adequada compreensão e interpretação do conteúdo. Por exemplo, nas revisões "A recepção é muito ruim, porque o *slot* do sim não suporta minichip" e "Recebi chamadas até na beira do Rio Paraná, divisa com o MS", as expressões "recepção", "sim" e "recebi chamadas" são usadas para avaliar a propriedade "sinal" do aparelho smartphone, porém essa informação é esplícitada na revisão. Captamos essa infomação por inferência no contexto. Outro exemplo é a revisão "A câmera é feia". Note que o aspecto avaliado não é "câmera". Neste exemplo, o usuário avalia a "aparência" ou "design" da câmera. Portanto, o campo de significação desses itens lexicais não pode ser reduzido apenas às unidades lexicais vizinhas, e sim ao mecanismo complexo de inferência no contexto e na situação social.

A proposta do modelo com *word embeddings* parece inicialmente subsidiada linguisticamente pelos estudos de viés mentalista, proposto por Harris (1968). Para o autor, se *A* e *B* ocorrem em ambientes idênticos, eles podem ser sinônimos. No entanto, Saussure (2002) já havia introduzido essa percepção a partir das denominações de *parole* e *langue*. A *parole* se desenvolve sintagmaticamente, ao longo de um eixo virtual de sucessões, onde cada elemento discreto ("palavra") ocupa uma posição significativa. Portanto, o significado desse elemento não provém de sua natureza, mas sim, por um lado, da posição que ele ocupa por referência aos outros elementos coocorrentes em seu contexto e, por outro lado, ele depende dos elementos ausentes desse mesmo contexto, mas por ele evocados, na memória implícita da *langue*. Trier (1931) também observou que as unidades léxicas de uma língua se deixam reunir em grupos estruturados de tal modo que cada unidade fica ali definida pelo lugar empírico que ocupa

[3] https://code.google.com/archive/p/word2vec/

[4] Note que "contexto", nesta aplicação, refere-se às palavras que acompanham a palavra alvo, ou seja, suas vizinhas.

[5] http://www.nilc.icmc.usp.br/nilc/index.php

respectivamente à posição das demais (Lopes, 1995).

Em revisões de usuários, encontramos marcas expressivas de subjetividade no discurso, portanto, trata-se de um tipo de conteúdo complexo, pois acarreta variáveis linguísticas e extralinguísticas, fato este que eleva o nível de complexidade deste tipo de dado na perspectiva do processamento automático. Além disso, revisões de usuários possuem baixa adequação à variante padrão da língua, ou seja, muitos termos com erros ou inadequações ortográficas e morfológicas, marcas de oralidade e inadequações sentenciais, do ponto de vista sintático e semântico. Além disso, trata-se de um conteúdo composto por códigos implícitos e de mensagens curtas, sendo necessários mecanismos de inferência para decodificação semântica adequada dessas mensagens. Portanto, para mineração de opinião, acreditamos que sejam necessários métodos que comportem as especificidades do domínio e que abarquem as complexidades intrínsecas e extrínsecas da língua.

Assim como despendemos esforços para análises linguísticas na tentativa de compreender o comportamento de aspectos em textos opinativos, também despendemos esforços para analisar o comportamento deste método estatístico baseado no conceito de *word embeddings* para resolução do problema de agrupamento de aspectos para mineração de opinião. Portanto, a seguir, iremos discorrer sobre a análise do desempenho deste modelo referenciando o comportamento do método ao se deparar com os principais fenômenos linguísticos mapeados pelo nosso estudo em revisões nos domínios de smartphone, câmera e livro. A seguir, descreveremos detalhadamente os resultados desta análise.

Verbos

Em revisões de usuários, constatamos que aproximadamente 40% dos termos indicativos de aspectos implícitos compõem a classe de verbos (ver mais na Seção 4.1.3.5). Observamos que, tratando-se de unidades lexicais da classe de verbos, a performance deste método estatístico é superior se comparada aos outros fenômenos que acometem textos opinativos (iremos apresentar nas seções seguintes). A justificativa desse resultado deve-se ao fenômeno de interdependência entre algumas classes de verbos e algumas classes de nomes, em que o nome é o sujeito do verbo, por exemplo: as unidades lexicais *ave e voar* e *peixe e nadar*; entre adjetivos e substantivos, por exemplo, *cabelos e loiros* e *leite e coalhada*; entre verbos e "objetos normais", por exemplo, *guiar e carro*; entre verbos e substantivos ligados por uma relação instrumental, por exemplo, *morder e dentes* e *chutar e pé*; e assim por diante (Lyons, 1970). Portanto, a valência verbal é responsável pela interdependência entre argumentos verbais, de tal modo que os elementos que "coocorrem" com as unidades verbais seguirão "certo padrão". Para exemplificar, aplicamos os aspectos "gostar","refletir" e "demorar" como entrada do modelo treinado utilizado nesta proposta de mestrado que usa o algoritmo *word2vec* (Mikolov *et al.*, 2013). As três unidades lexicais utilizadas como entrada são indicativos de aspectos implícitos nas revisões de usuários sobre smartphone, câmera e livro. Nas Tabelas 5.7, 5.8 e 5.9, exibimos os conjuntos retornados pelo modelo. Na coluna *unidade lexical*, são exibidas as palavras mais similares de acordo com a entrada. Por exemplo, na Tabela 5.7, a entrada foi o termo "gostar"

e os termos “gosta”, “odiar” e “gostava” são algumas palavras de vetores similares, retornados pelo modelo, de acordo com o termo de entrada. A coluna *score* é a pontuação inerente ao cálculo usado pelo método para definir o quão similar a unidade de entrada é em relação a unidade retornada. Esse valor varia entre 0 (menos similar) e 1 (mais similar).

Tabela 5.7: *Word embeddings* do indicativo de aspecto “gostar”.

Unidade lexical	***Score***
gosta	0.6405864953994751
odiar	0.6015201807022095
gostava	0.5822267532348633
gostarem	0.5632543563842773
gostou	0.5526309013366699
desconfiar	0.5460933446884155
gostei	0.5422440767288208
gostam	0.5347926616668701
goste	0.533798336982727
gostando	0.5308782458305359)

Tabela 5.8: *Word embeddings* do indicativo de aspecto “refletir”.

Unidade lexical	***Score***
reflectir	0.711204469203949
questionar	0.6180825233459473
questionar-se	0.5624111294746399
reflectirem	0.5547906160354614
especular	0.5535352230072021
prevalecer	0.5454685688018799
impactar	0.5451191663742065
influir,	0.5408214926719666
teorizar	0.5383339524269104
compreender	0.5361512899398804

Tabela 5.9: *Word embeddings* do indicativo de aspecto “demorar”.

Unidade lexical	***Score***
durar	0.7959119081497192
demorarável	0.6619040966033936
prolongar-se	0.6616053581237793
demore	0.5905470848083496
demoraria	0.5808826684951782
demorou	0.5485048294067383
perdurar	0.5337100028991699
custar	0.5312538743019104
durarável	0.5251177549362183
demoram	0.5208966732025146

Observe que o conjunto da Tabela 5.7, composto por unidades lexicais retornadas a partir da entrada "gostar", é majoritariamente composto por verbos do tipo psicológico [6], por exemplo, os verbos *desconfiar, odiar, etc*. No segundo conjunto, demonstrado na Tabela 5.8, utilizamos como entrada o indicativo de aspecto implícito "refletir". Note que todos os verbos deste conjunto são sinônimos, portanto, possuem alta similaridade lexical e correpondência semântica com o verbo "refletir". No último conjunto, exibido na Tabela 5.9, a entrada utilizada é o verbo "durar". Observe que as unidades lexicais retornadas também são similares e possuem correspondência interpretativa com o verbo "durar". Portanto, é possível observar, nos três conjuntos, alta correspondência semântica em relação às unidades lexicais usadas como entrada, ou seja, dadas as constatações empíricas iniciais apresentadas acima e a fundamentação linguística sobre o fenômeno de valência verbal, o modelo estatístico baseado em *word embeddings*, especificamente para a classe de verbos, parece-nos eficiente para classificação semântica.

Ambiguidade

Um dos fenômenos linguísticos mais complexos de tratamento automático é a ambiguidade inerente às línguas naturais. Por exemplo, o termo "sim" é majoritariamente usado no português do Brasil como advérbio de afirmação. No entanto, no domíno de smartphone, encontramos esse termo na forma de substantivo e sujeito da sentença. Esse termo foi usado pelos usuários em revisões para designar um "cartão" ou "chip" do aparelho celular. Vejamos o comportamento do modelo treinado dado a entrada lexical "sim" (ver Tabela 5.10).

Tabela 5.10: *Word embeddings* do aspecto "sim".

Unidade lexical	***Score***
téo-so	0.5924373865127563
obviamente	0.5268886685371399
evidentemente	0.5187929868698120
infelizmente	0.5107123851776123
penso	0.4490032494068146
acredito	0.4471751451492309
felizmente	0.4400550425052643
não	0.4400264024734497
lamentavelmente	0.43296197056770325

Note que o conjunto de unidades lexicais obtidas a partir desta entrada evidência a significação prototípica (mais frequente) do termo "sim", no português do Brasil. Itens como "obviamente", "evidentemente" e "felizmente" são advérbios de afirmação.

Vejamos outro exemplo deste mesmo fenômeno, porém para as entradas lexicais "bateria" e "disparo". Os conjuntos retornados pelo modelo para essas duas entradas são exibidos nas Tabelas 5.11 e 5.12, respectivamente.

[6] Verbos psicológicos expressam uma experiência de natureza emocional, sendo assim também designados por verbos de emoção, e dividem-se, na literatura, em verbos com *experenciador sujeito*, por exemplo, o verbo amar, e em verbos com *experenciador objeto*, por exemplo, o verbo preocupar.

Tabela 5.11: *Word embeddings* do aspecto “bateria”.

Unidade lexical	***Score***
guitarra	0.719579279422760
percussão	0.646325945854187
teclados	0.6124431490898132
contrabaixo	0.5962961316108704
teclado	0.5848445892333984
sanfona	0.5814942717552185
trompete	0.5804557800292969
violão	0.5762271881103516
fabbio	0.5646268129348755
escaleta	0.5611575841903687

Tabela 5.12: *Word embeddings* do aspecto “disparo”.

Unidade lexical	***Score***
disparos	0.7292324304580688
projéctil	0.6630128622055054
projétil	0.6388976573944092
morteiro	0.6125934720039368
disparado	0.6001366376876831
assaltante	0.5879765748977661
tiro	0.5749138593673706
rojão	0.5593981146812439
disparar	0.5533016324043274
assalto	0.5424944758415222

Na Tabela 5.11, encontramos o conjunto de unidades lexicais para a entrada “bateria” e, na Tabela 5.12, o conjunto de unidades lexicais para a entrada “disparo”, ambos aspectos do domínio de câmera digital. Assim como ocorre em várias unidades lexicais de uma língua natural, “bateria” e “disparo” são polissêmicos, ou seja, para esses itens é atribuída uma matriz semântica dinâmica que muda de acordo com o contexto e a situação social em que os termos são inseridos. No entanto, as unidades mais similares com “bateria”, de acordo com os conjuntos retornados pelo modelo, são os itens do sema *música e/ou instrumentos musicais* (ver Tabela 5.11); e as unidades mais similares com o aspecto “disparo” são unidades lexicais de sema *armas de fogo e/ou confronto armado* (ver Tabela 5.12). Note que a matriz semântica formada para os aspectos “bateria” e “disparo” estão em desacordo com a compreensão semântica desses termos nas revisões em que ocorrem. O aspecto “bateria” é usado pelos usuários nas revisões analisadas para avaliar parte do equipamento do smartphone ou câmera reponsável por fonecer “energia” àquele equipamento eletrônico. O termo “disparo” foi usado nas revisões analisadas para avaliar um recurso da câmera digital por onde passa a luz que será captada no momento do disparo da foto. Esse comportamento é resultado do conjunto de textos usados para treinamento do modelo utilizado. O modelo utilizado neste mestrado foi treinado usando textos da língua geral do português brasileiro.

Polarização semântica

Observamos também que este modelo estatístico classifica como similares unidades lexicais em relação de antonímia, portanto polarizadas semanticamente. Por exemplo, no conjunto de unidades exibidas pela Tabela 5.10, utilizamos como entrada o aspecto “sim”, e um dos itens desse conjunto, é a unidade lexical “não”, além de outros itens em relação semântica de oposição com o dado de entrada, tais como “infelizmente” e “lamentavelmente”. Outro exemplo para retratar esse comportamento é demonstrado na Tabela 5.13. Neste exemplo, a entrada lexical usada foi o aspecto “fim”. Note que os itens deste conjunto são antônimos, o que implica polarização semântica. Por exemplo, as unidades “começo” e “fim”, “início” e “final”, etc.

Tabela 5.13: *Word embeddings* do aspecto “fim”.

Unidade lexical	***Score***
começo	0.6296569108963013
início	0.5840789675712585
fim	0.5724589824676514
ínicio	0.5362707376480103
inicio	0.5212630629539490
finais	0.51150095462799070
anterior	0.46888995170593260
decorrer	0.46558552980422974
ício	0.46451112627983093
final	0.46126979589462280

Portanto, esse tipo de comportamento parece ser uma evidência empírica que o modelo não

é tão eficiente para a classificação de similaridade semântica em textos em que a polaridade é um elemento essencial, especialmente textos opinativos. Na verdade, métodos superficiais dificilmente serão suficientes para classificações semânticas, discursivas e paradigmáticas de uma língua natural, pois as línguas naturais são fruto da cognição e da atividade humana, e a atividade humana é naturalmente ambígua. Além disso, nos níveis semânticos, discursivos e paradgmáticos, além da ambíguidade, fatores extra-linguísticos acometem em maior proporção, que os torna mais complexos do ponto de vista do processamento automático. Portanto, parece-nos mais eficiente a incursão de métodos que comportem conhecimentos de domínio e da situação social, e não apenas estatísticas de coocorrência.

Estrangeirismos, diminutivos e nomes próprios

Alguns fênomenos linguísticos identificados no domínio de revisões de produtos foram: (i) construções de diminutivos (“romancezinho”, “livrinho”, “capinha”, etc.); (ii) estrangeirimos (*presets*, *slow motion*, *bugs*, etc); e (iii) nomes próprios (“crepúsculo”, “fuji”, “sony”, “h09”, etc). O uso de diminutivos na língua sofre influência de fatores linguísticos (utilização do morfema “inho” e“inha”, por exemplo) e fatores extra-linguísticos (é recorrente em ambientes informais e usado com maior frequência por falantes do gênero feminino em detrimento do gênero masculino (Labov, 1994)). O fenômeno de estrangeirismo em uma língua é complexo, no entanto, sofre influência principalmente por relações políticas, econômicas e geo-políticas. Os nomes próprios acometem qualquer língua natural por fenômenos distintos. Vejamos como o método se comporta ao se deparar com cada um desse fênomenos. Na Tabela 5.14, exibimos o conjunto retornado a partir da entrada lexical “romancezinho”, aspecto extraído do domínio de livro. Note que o desempenho foi ruim, pois há ausência de correspondência semântica entre os itens do conjunto em relação ao contexto semântico em que ocorre o aspecto de entrada. Itens como “fiozinho”, “urinol” e “engordurado” não possuem proximidade semântica óbvia com o item “romancezinho”.

Tabela 5.14: *Word embeddings* do aspecto “romancezinho”.

Unidade lexical	***Score***
bigode-de-broxa	0.5420284867286682
início	0.5840789675712585
inarmânico	0.540293276309967
fiozinho	0.5318080186843872
desencapado	0.5224076509475708
belai	0.5121059417724609
smidge	0.5115442276000977
engordurado	0.511095404624939
banalíssimo	0.5075555443763733
floreio	0.5069879293441772

Quanto aos estrangeirimos e os nomes próprios, observamos um resultado mais significativo em relação aos outros fenômenos da mesma classe de nomes (não-verbais). Vejamos os con-

juntos a seguir, exibidos pelas Tabelas 5.15, 5.16, 5.17 e 5.18, cujas entradas foram os aspectos "touchscreen" e "interface" (estrangeirismos) e "canon" e "fuji" (nomes próprios), todos eles do domínio de câmera.

Tabela 5.15: *Word embeddings* do aspecto "touchscreen".

Unidade lexical	***Score***
touchscreen	0.5909452438354492
sensor	0.5897029638290405
projetor	0.579769492149353
display	0.5631867051124573
visores	0.5607553720474243
oled	0.5523111820220947
capacitivo	0.551577091217041
ecrê	0.5467293858528137
monitor	0.5434030890464783
touchpad	0.5423129796981812

Tabela 5.16: *Word embeddings* do aspecto "interface".

Unidade lexical	***Score***
api	0.6434847116470337
inteface	0.5985379219055176
colloquy	0.5883268117904663
plataforma	0.5766987800598145
middleware	0.5713101625442505
cérebro-computador	0.5711454749107361
smidge	0.5115442276000977
desktop	0.5703732967376709
multimá1quinas	0.5639685988426208
multi-plataforma	0.562650740146637

Tabela 5.17: *Word embeddings* do aspecto "canon".

Unidade lexical	***Score***
cameramake	0.6056790351867676
cameramodel	0.5953608751296997
minolta	0.5847787261009216
nikon	0.5757133960723877
fnumber	0.5749174356460571
dslr	0.5720962285995483
kodak	0.5603956580162048
fujifilm	0.5367375612258911
exposuretime	0.5329309701919556
epson	0.5328077673912048

Note que, em todos os quatro conjuntos retornados pelo modelo, há unidades lexicais com alta correspondência semântica. Por exemplo, no conjunto exibido na Tabela 5.15, o método

Tabela 5.18: *Word embeddings* do aspecto “fuji”.

Unidade lexical	***Score***
epson	0.5328077673912048
tokyo	0.5449055433273315
nippon	0.5353261232376099
tbs	0.5245823264122009
caburaed	0.5184643864631653
0kids	0.4917868673801422
at-x	0.4837615489959717
mbs	0.482839971780777
tucuju	0.4823723733425140
wowow	0.4811611473560333

retornou itens correspondentes semanticamente com a entrada “touchscreen”, tais como “display”, “visores”, “ecrê” e “monitor”. No conjunto da Tabela 5.16, em que utilizamos a entrada “interface”, o modelo também retornou um grupo de unidades lexicais com correspondência semântica de acordo com a entrada. São eles: “api”, “plataforma” e “’cérebro-computador”. Quanto aos dois nomes próprios, “canon” e “fuji”, exibidos nas Tabelas 5.17 e 5.18, o método também retornou itens próximos semanticamente, como “kodak”, “nippon”, “tokyo” e “fujifilm”. No entanto, é importante salientar que as classes de estrangeirismos e de nomes próprios da língua natural são classes com baixa ambiguidade.

Especificidades do domíno

O método também não apresentou um bom desempenho ao se deparar com aspectos que representam as especificidades do domínio. Por exemplo, para os aspectos “3G”, “wap”, “wifi”, “gps”, “hit”, “quadriband” e alguns outros exemplos no córpus que imprimem as especificidades do domínio, o método retornou resultado nulo. Por exemplo, ao inserirmos como entrada o aspecto “3g”, o método não encontrou nenhuma correspôndencia semântica para esse item no modelo treinado. Isso ocorre em função da ausência desses termos no conjunto de textos usados para treinamento (conjuntos de textos da língua geral, que não necessariamente abarcam conteúdos específicos de domínios). Esse comportamento ocorreu também nos métodos baseados em similaridade lexical para extração de relações entre aspectos. No entanto, houve um número reduzido de aspectos específicos do domínio para os quais o método estatístico baseado no conceito de *word embeddings* encontrou vetores de palavras similares, porém os itens lexicais do conjunto retornado para essas entradas consistiam de unidades lexicais com baixa correspondência semântica. Portanto, é um desafio reconhecer e agrupar aspectos específicos do domínio.

Expressões de aspectos

Nos domínios analisados, aproximadamente 35% dos aspectos são constituídos de *n-gramas*, ou seja, são expressões de aspectos ou aspectos compostos. Por exemplo, no domínio

de livro, encontramos as expressões de aspectos “sociedade do big brother”, “técnica de escrita”, “crítica social”, “jorge amado”, “capitães de areia”, etc. No domínio de câmera digital, identificamos as expressões de aspectos “cartão de memória”, “câmera amador”, “acionamento de funções”, “manual de instruções”, entre outras. Além disso, identificamos também algumas expressões indicativas de aspectos implícitos, como “acesso aos dados” e “facilidade de uso”, que são fenômenos mais complexos de identificação e agrupamento. Observamos que o modelo aceita como entrada apenas termos do tipo *unigrama*, portanto aspectos em composição de *n-gramas* não foram comportados por este método.

O modelo baseado em semântica vetorial, especificamente o algoritmo *word2vec* (Mikolov *et al.*, 2013) possui recursos notadamente atraentes para solução do problema de identificação de grupos de aspectos, pois esse método parece identificar superficialmente alguns traços semânticos entre unidades lexicais nos domínios analisados. No entanto, uma língua natural é um fenômeno complexo e sofre influência de vários processos (cognitivos, empíricos, sociais), dos quais acreditamos que, quanto mais conhecimento do domínio e da situação social são empregrados, maior a probabilidade de eficiência na análise e no processamento automático desses dados.

Vale resaltar que nós utilizamos um modelo já treinado de *word embeddings* com dados a língua geral, pois a realização de um novo treinamento com base nos textos opinativos usados neste trabalho de mestrado não fazia parte do escopo deste trabalho, aĺém disso, a composição de dados utilizada neste trabalho não é suficiente para o treinamento de *word embeddings*.

5.4 Método proposto - OpCluster-PT

Neste trabalho, propusemos e implementamos um algoritmo inédito para a resolução do problema de agrupamento de aspectos para sistemas de mineração de opinião. Nosso algoritmo agrupa aspectos explícitos e termos indicativos de aspectos implícitos. O algoritmo foi proposto a partir de um estudo linguístico aprofundado e baseado em córpus de textos opinativos, em que exploramos fenômenos linguísticos e relevantes estatisticamente nos domínios de smartphone, câmera e livro. Nós constatamos que aspectos correlatos no domínio de opinião encontram-se em relação de isotopia linguística, ou seja, são unidades lexicais distintas que possuem correspondência interpretativa no domínio em que ocorrem, caracterizado por cadinalidade *1:N*, ou seja, *1* unidade lexical possui *N* unidades lexicais correspondentes. As principais relações lexicais formais identificadas e que compõem esse fenômeno são as relações de sinonímia, hiperonímia/holonímia e meronímia/holonímia, além das relações causativas e dos fenômenos de deverbalidade, correferência, estrangeirismo e diminutivos, principalmente. Portanto, a partir das investigações linguísticas e da observação dos resultados obtidos com a implementação dos demais métodos, nós propomos o algoritmo OpCluster-PT, que consiste de um método novo, baseado em conhecimento linguístico. Nas seções seguintes, apresentaremos a arquitetura do método e o algoritmo OpCluster-PT.

5.4.1 Arquitetura

A arquitetura do nosso método é exibida na Figura 5.1. O método recebe como entrada um conjunto de revisões de domínio, assim como uma lista de aspectos inerentes a este domínio. Em seguida, para cada aspecto, são extraídos sinônimos e, em seguida, merônimo/holôninos, relações causativas, especificamente as relações *resultadoDaAçãoDe e serveParaAção*. Conseguinte, são extraídos construções deverbais, estrangeirismos, correferências, diminutivos e relações de *substring*. A saída do método consiste em um conjunto de grupos de aspectos. Na seção seguinte, iremos detalhar todos esses processos a partir da descrição do algoritmo.

A arquitetura do nosso método é composta por vários recursos linguístico-computacionais, tais como ontologia lexical, sistema de resolução de correferências, dicionários lexicais da língua, e uma lista criada neste trabalho. Para extração de relações de similaridade lexical, utilizamos a ontologia Onto-PT (Oliveira, 2014). A Onto-PT (Oliveira, 2014) é usada para extração de relações lexicais de sinonímia, meronímia/holonímia e as relações *resultadoDaAçãoDe e serveParaAção*. Para identificação de estrangeirimos e deverbais, usamos dicionários do iLteC (Janssen & Ferreira (2007); Ferreira & Janssen (2017)). Para identificação de correferências, utilizamos o sistema CORP (Fonseca *et al.*, 2016) e a versão CorrefVisual (Fonseca, 2014). Por fim, uma lista de unidades lexicais composta por construções de diminutivos/aumentativos foi construída especificamente para essa tarefa. Por exemplo, itens como “leve” e “levinho” ou “livro” e “livrinho” são exemplos que compõem a lista criada. A descrição detalhada de todos os recursos linguístico-computacionais utilizados nesta proposta de mestrado encontra-se na Seção 2.3.

5.4.2 Algoritmo

A seguir, apresentamos o algoritmo OpCluster-PT, proposto por esse trabalho de mestrado para resolução do problema de identificação de grupos de aspectos explícitos e termos indicativos de aspectos implícitos para mineração de opinião (ver Algoritmo 7). O algoritmo recebe como entrada um conjunto de revisões R e um conjunto de expressões de aspectos (explícitos e indicativos de aspectos implícitos) descobertos em R. O algoritmo proposto atribui os aspectos descobertos em R aos grupos G_{n}.

A seguir, apresentaremos um relato detalhado do funcionamento do algoritmo OpCluster-PT.

Entradas

O algoritmo recebe como entrada (i) revisões de usuários e uma (ii) lista de expressões de aspectos explícitos e termos indicativos de aspectos implícitos [7], ordenados de forma decrescente de acordo com o critério de frequência de ocorrência desses aspectos nas revisões em que

[7]Como o escopo deste trabalho não abarca a extração automática de aspectos, nós extraímos manualmente os aspectos das revisões em que ocorriam para compor a lista de aspectos usada como entrada pelo algoritmo.

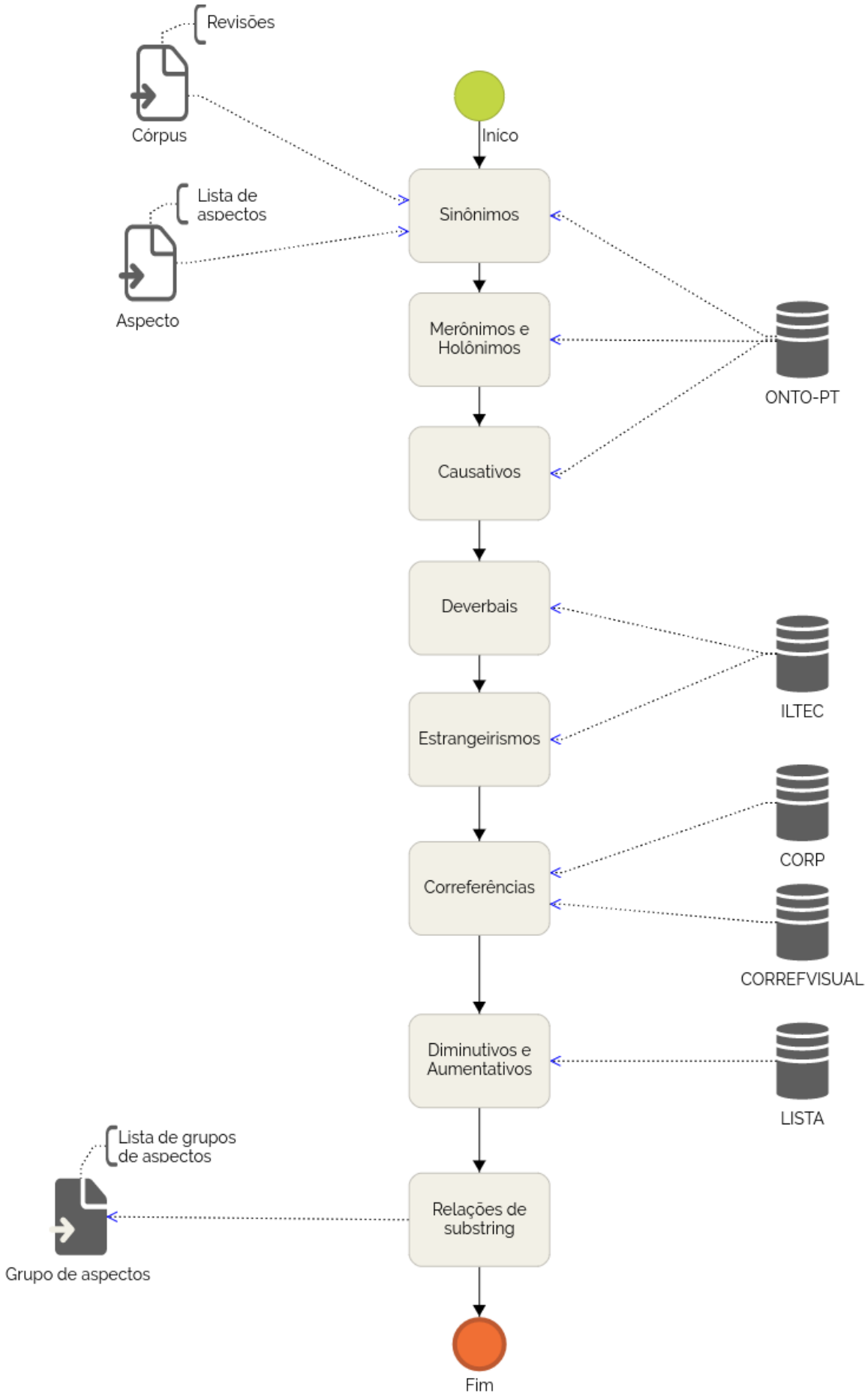

Figura 5.1: Arquitetura do OpCluster-PT.

Algoritmo 7: Algoritmo OpCluster-PT

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência;
Revisões processadas pelo CORP $R = \{r_1, r_2, ...r_n\}$, em que os aspectos de A ocorrem;
Saída: Grupos de aspectos G = $\{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

início
  Declare **B** = $\{b_{sin}, b_{parte\text{-}todo}, b_{causa}, b_{devb}, b_{estrag}, b_{dimin}, b_{corref}, b_{subst}\}$, tal que **B** contém o resultado da busca por aspectos em relação de sinonínimia, meronímia/holonímia, causativa e construções deverbais, correferentes, estrangeirismos, diminutivos (por exemplo, b_{sin} contém os aspectos sinônimos ao aspecto de interesse);
  Declare **U** = $\{u_1\ u_2, ...\ u_n\}$, tal que cada conjunto u_i contém um grupo unitário de **G**;
  Declare *contador* = 0;
  Declare *posicao* = 0;
  repita
    se a_i *de* ***A*** *possuir sinônimos na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{sin}}$ os sinônimos encontrados;
    fim
    se a_i *de* ***A*** *possuir merônimos e/ou holônimos imediatos na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{parte\text{-}todo}}$ os merônimos e/ou holônimos encontrados;
    fim
    se a_i *de* ***A*** *possuir relações causativas do tipo resultadoDaAção e/ou serveParaAccao na base do Onto.PT* **então**
      Adiciona em $\mathbf{b_{causa}}$ os itens em relação *resultadoDaAção* e/ou *serveParaAccao* encontrados;
    fim
    se a_i *de* ***A*** *possuir construções deverbais na base do iLteC* **então**
      Adiciona em $\mathbf{b_{devb}}$ as construções deverbais encontradas;
    fim
    se a_i *de* ***A*** *possuir estrangeirismos na base do iLteC* **então**
      Adiciona em $\mathbf{b_{estrag}}$ os estrangeirismos encontrados;
    fim
    se a_i *de* ***A*** *possuir construções de diminutivos na lista de diminutivos/aumentativos* **então**
      Adiciona em $\mathbf{b_{dimin}}$ os diminutivos encontrados;
    fim
    se a_i *de* ***A*** *possuir relações de substring com outros aspectos de* ***A*** **então**
      Adiciona em $\mathbf{b_{subst}}$ os aspectos em relações de *substring* encontradas;
    fim
    se a_i *de* ***A****, nas revisões em que ocorre, possuir correferentes classificados pelo CORP* **então**
      Adiciona em $\mathbf{b_{corref}}$ as cadeias de correferentes encontradas;
    fim
    Exclua itens duplicados de **B** = $\{b_{sin}, b_{parte\text{-}todo}, b_{causa}, b_{devb}, b_{estrag}, b_{dimin}, b_{corref}, b_{subst}\}$, se houver;
    Incremente *contador*;
    Crie grupo $\mathbf{G_i}$ e adicione em $\mathbf{G_i}$ os aspectos da intersecção **(A,B)**;
    Remova de **A** os aspectos da intersecção;
    Esvazie **B**;
    repita
      se *aspecto de* ***G****, nas revisões em que ocorre, possuir correferentes classificados pela aplicação CORP* **então**
        Adiciona em $\mathbf{b_{corref}}$ as cadeias de correferentes encontradas;
      fim
      se *aspecto de* ***G*** *possuir estrangeirismos na base do iLteC* **então**
        Adiciona em $\mathbf{b_{estrag}}$ os estrangeirismos encontrados;
      fim
      se *aspecto de* ***G*** *possuir construções de diminutivos na lista de diminutivos/aumentativos* **então**
        Adiciona em $\mathbf{b_{dimin}}$ os diminutivos encontrados;
      fim
      Exclua itens duplicados de **B** = { $b_{corref}, b_{estrag}, b_{dimin}$ }, se houver;
      Adicione em $\mathbf{G_i}$ os aspectos da intersecção **(A,B)**;
      Remova de **A** os aspectos da intersecção;
      Esvazie **B**;
      Guarde em **posição** a última posição do elemento adicionado em $\mathbf{G_i}$
    até *a posição dos elementos de* ***G*** *for maior que valor de* ***posição***;
  até ***A*** *esvaziar*;
  repita
    Selecione os grupos unitários e adicione em $\mathbf{U_i}$;
    se $\boldsymbol{U_i}$ *estiver contido em aspectos de* $\boldsymbol{G_i}$ *por relação de substring* **então**
      Adicione em $\mathbf{G_i}$ o aspecto de $\mathbf{U_i}$;
      Remove $\mathbf{U_i}$ de **G**
    fim
  até ***G*** *esvaziar;*;
fim

ocorrem. Portanto, os aspectos mais frequentes nas revisões em que ocorrem serão os primeiros elementos da lista.

Primeiro laço de repetição

O algoritmo lê a_i da lista de aspectos ***A*** e verifica no primeiro laço de repetição as seguintes condições:

- Se o aspecto analisado possui relações de sinonímia, meronímia/holonímia e as relações causativas *resultadodaAçãoDe e serveParaAccao* na Onto-PT (Oliveira, 2014);
- Se o aspecto analisado possui construções deverbais e estrangeirismos nos dicionários do iLteC (Janssen & Ferreira, 2007);
- Se o aspecto analisado possui diminutivos na lista de diminutivos/aumentativos construída para essa tarefa;
- Se o aspecto analisado possui relações de *substring* com outros aspectos da lista de aspectos ***A***;
- Se o aspecto analisado possui relações de correferência anotadas pelo Corp (Fonseca *et al.*, 2016) nas revisões em que ocorrem;

Se as condições forem verdadeiras, o conjunto de resultados obtidos em cada uma das condições é adicionado em ***B***. Por exemplo, se a_i possuir relações de sinonímia na Onto-PT (Oliveira, 2014), os itens sinônimos de a_i serão adicionados em b_{sin}; em seguida, se a_i possuir relações de meronimia/holonímia na Onto-PT (Oliveira, 2014), os itens merônimos/holônimos de a_i serão adicionados em $b_{parte\text{-}todo}$; e assim sucessivamente.

Funções das linhas 31 à 35

Na linha 31, são excluídos itens duplicados de ***B***, se houver. Em seguida, na linha 32, o ***contador*** é incrementado. Na linha 33, um grupo de aspectos $\boldsymbol{G_i}$ é criado com os aspectos da intersecção ***(A,B)***. Na linha 34, são excluídos da lista de aspectos ***A*** os aspectos da intersecção ***(A,B)***. E, na linha 3.5, ***B*** é esvaziado.

Segundo laço de repetição

Neste laço de repetição, o algoritmo lê o primeiro item de ***G*** e verifica as seguintes condições:

- Se o aspecto analisado possui correferências classificadas pelo Corp (Fonseca *et al.*, 2016) nas revisões em que ocorre;
- Se o aspecto analisado possui estrangeirismos e diminutivos nos dicionários do iLteC (Ferreira & Janssen (2017); Janssen & Ferreira (2007));

Se as condições do segundo laço de repetição forem verdadeiras, o conjunto de resultados é adicionado em ***B***.

Funções das linhas 46 à 50

Na linha 46, são excluídos os itens duplicados de ***B***, se houver. Em seguida, na linha 47, adicionam-se em G_i os aspectos da intersecção ***(A,B)***. Na função 8.0, removem-se da lista de aspectos ***A*** os itens da intersecção ***(A,B)***. Na função 9.0, ***B*** é esvaziado e, na função 10, a posição do último elemento adicionado em G_i é capturada e armazenada em ***posição***.

Terceiro laço de repetição

Nesse laço de repetição, realiza-se uma nova verificação sobre os grupos unitários adicionado em ***G***. Por exemplo, caso algum aspecto não tenha sido agrupado e, caso ainda possua relações de *substring* com outro elemento adicionado de ***G***, estes elementos serão capturados por este laço de repetição.

Primeiramente, verifica-se a ocorrência de grupos unitários em ***G***. Se G_i é um grupo unitário, adiciona-se G_i em ***U***. Se U_i possuir relação de *substring* com os elementos de ***G***, o grupo unitário U_i é adicionado em G_i e o elemento de U_i é removido de ***G***.

Saída

O algoritmo retorna como saída grupos G_i de aspectos de ***A***. Por exemplo, G_1, no domínio de livro, pode ser formado pelos aspectos {livro, obra, bestseller, livreto, livrinho, Crepúsculo, 1984, Ensaio sobre a cegueira, etc.}, e G_2 composto pelos elementos {protagonista, bella, Isabella, Vamipiro, Menino, Edward, personagens, herói, etc.}, e assim sucessivamente, formando grupos de aspectos correlatos no domínio em que ocorrem.

A seguir, descreveremos como os recursos linguítico-computacionais são acessados.

Relações de sinonímia, meronímia, holonímia, e as relações causativas *resultadodaAçãoDe* e *serveParaAccao*

Para extração dessas relações, utilizamos a ontologia lexical do português Onto-PT (Oliveira, 2014). Por exemplo, aspectos "valor" e "custo" estão relacionados pela relação de sinonímia; "teclado" e "tecla" relacionados pela relação de meronímia/holonimia; "escrita" e "escrever" pela relação de *resultadodaAçãoDe*. Portanto, esses itens serão extraídos nesta etapa.

Construções deverbais

Para extração de construções deverbais, utilizamos o dicionário de nomes deverbais para o português do iLteC (Janssen & Ferreira, 2007). Relações entre aspectos do tipo "manusear" e "manuseio" são extraídos nessa etapa.

Estrangeirismos

Para a extração de estrangeirismos, utilizamos outro recurso lexical do iLteC: o dicionário de estrangeirismos (Ferreira & Janssen, 2017). Aspectos como "display" e "expositor", que são unidades lexicais inseridas na língua pelo fenômeno de estrangeirismo, são capturados a partir deste recurso. Optamos também por incorporar a este dicionário algumas unidade lexicais mais específicas do domínio e que não foram encontradas no dicionário. O dicionário de estrangeirismos, de acordo com os fundadores, sofrerá uma atualização nos próximos meses e possivelmente alguns dos itens incorporados manualmente neste trabalho estarão presentes na nova versão do dicionário a partir desta atualização. Por exemplo, incorporamos a unidade lexical "presets", que diz respeito a um recurso de pré-edição de uma câmera digital.

Construções de diminutivos e aumentativos

Para esse tipo de construção lexical, optamos pelo desenvolvimento de uma lista, por exemplo, com unidades lexicais do tipo "livro" e seu correspondente diminutivo "livrinho". Essa decisão foi tomanda em função de não encontrarmos nenhum léxico disponível, para o português, composto por variações de grau sintético de substantivos.

Substrings

As relações de *substring* consistem de relações intrínsecas entre unidades lexicais, por exemplo, o aspecto "câmera" e "câmera digital" ou os aspectos "escrita" e "estilo de escrita".

Correlações linguísticas

As cadeias de referentes ou correferentes foram obtidas através da classificação de correferências realizada pelo sistema de resolução de correferência CORP (Fonseca *et al.*, 2016). O CORP recebe como entrada os documentos de revisões de usuários e retorna como saída arquivos no formato XML com marcações dos grupos de correferentes.

Observamos que a utilização do CORP (Fonseca *et al.*, 2016) é mais eficiente na identificação de relações de hiperonímia/holonímia entre aspectos nos domínios analisados em detrimento da utilização da ontologia lexical Onto-PT (Oliveira, 2014). Nós observamos que, especialmente no domínio de smartphone e câmera, em que os usuários possuíam mais conhecimento sobre as propriedades dos produtos, na maioria das vezes, os aspectos específicos desses domínios não foram identificados na ontologia lexical. Por exemplo, os aspectos "canon" e "h70" são aspectos específicos do domínio e não foram identificados na Onto-PT (Oliveira, 2014), no entanto, esses termos foram identificados pelo CORP (Fonseca *et al.*, 2016). Para exemplificar melhor, usemos como referência o aspecto "câmera", que possui relação do tipo i-sa com os aspectos "nikon", "sony" e "benq". Essa relação entre aspectos foi reconhecida e anotada pelo CORP (Fonseca *et al.*, 2016), no entanto, não foi identificada pela Onto-PT (Oliveira, 2014). Além disso, as relações de hiperônia/hipônimia da Onto-PT (Oliveira, 2014) são

genéricas das língua o implica maior cobertura semântica, porém, implica também, menor precisão. Portanto, optamos por não extrair relações de hiperonímia e holonímia usando a Onto-PT (Oliveira, 2014) e, para extração dessas relações, optamos pela utilizamos do CORP (Fonseca *et al.*, 2016).

Capítulo

6

Resultados

Neste capítulo, apresentaremos o desempenho dos métodos de agrupamento de aspectos implementados nesta proposta de trabalho. Utilizamos para avaliação dos métodos as medidas de *precisão, cobertura, medida-f e medida-f global*. Apresentaremos, na Seção 6.1, uma descrição das medidas de avaliação; na Seção 6.2, serão apresentados os resultados obtidos; e, na Seção 6.3, realizaremos a discussão sobre esses resultados.

6.1 Medidas de avaliação

Foram implementados 6 métodos para resolução do problema de agrupamento de aspectos para mineração de opinião. Para avaliação dos métodos automáticos, nós utilizamos o córpus anotado nesta proposta de mestrado, que serviu de referência para aplicação das medidas de *precisão, cobertura, medida-f e medida-f global*, exibidas pelas Equações 6.1, 6.2, 6.3 e 6.4, respectivamente.

A medida de *precisão* indica a proporção de aspectos do grupo de aspectos de referência que está no grupo gerado automaticamente. A *cobertura* indica a proporção de aspectos do grupo de referência que foi coberto pelo grupo gerado automaticamente. Tais medidas são complementares e, por isso, costuma-se calcular a *medida-f*, que representa a média harmônica entre a precisão e a cobertura. A *medida-f global* de cada grupo gerado automaticamente, em relação a todo o conjunto de grupos, se baseia no grupo que melhor descreve cada grupo de referência. Deste modo, o valor da *medida-f global* é dado pela Equação 6.4, onde N é o número total de aspectos a serem agrupados no domínio, K o conjunto de grupos de referência, C o conjunto de grupos e $n_{(ij)}$ o número de aspectos do grupo $k_{(i)} \in K$ que estão presentes no grupo $c_{(j)} \in C$.

$$\textbf{Precisão} = \frac{\textit{Aspectos do grupo gerado automaticamente em comum com aspectos do grupo de referência}}{\textit{Aspectos do grupo gerado automaticamente}} \tag{6.1}$$

$$\textbf{Cobertura} = \frac{\textit{Aspectos do grupo gerado automaticamente em comum com aspectos do grupo de referência}}{\textit{Aspectos do grupo de referência}} \tag{6.2}$$

$$\textbf{Medida-F} = 2 * \frac{\textit{Precisão} * \textit{Cobertura}}{\textit{Precisão} + \textit{Cobertura}} \tag{6.3}$$

$$\textbf{Medida-F Global} = \frac{\sum_{|k_(i)|*\max c_(j) \in C\{F(k_(i)c_(j))\}}}{N} \tag{6.4}$$

6.2 Apresentação dos resultados

Nesta proposta de mestrado, nós implementamos 6 métodos no total, sendo: 3 métodos baseados em similaridade lexical; 1 método baseado em similaridade lexical e correlações linguísticas; 1 método estatístico baseado no conceito de *word embeddings*; e, por fim, um método novo, baseado em conhecimento linguístico, foi proposto por esse trabalho de mestrado. Os resultados obtidos pelos métodos automáticos descrito acima são exibidos nas Tabelas 6.1, 6,2, 6.3 e 6.4.

Tabela 6.1: Precisão

N.	Métodos	Livro	Câmera	Smartphone	Média
1	sinônimos *(baseline)*	**0,974**	**0,987**	**0,973**	**0,978**
2	sin + hipe/hipo	0,916	0,967	0,940	0,949
3	sin + hipe/hipo + mero/holo	0,916	0,967	0,943	0,942
4	sin + hipe/hipo + mero/holo + corref	0,945	0,963	0.953	0,953
5	*word embeddings*	0,953	0,962	0,956	0,957
6	OpCluster-PT	0,925	0,933	0,947	0,935

Na Tabela 6.1, apresentamos a *precisão* de cada um dos seis métodos de agrupamento de aspectos implementados nesta proposta. Note que os resultados de precisão dos métodos está

entre 0,916 e 0,987, o que implica que todos os métodos de agrupamento implementados apresentaram bons resultados de precisão. O método de agrupamento de aspectos baseado em relações de sinonímia (ver Método 1 da Tabela 6.1) apresentou o melhor resultado, dentre os outros métodos, para a tarefa. Neste método, nós observamos um número alto de grupos unitários obtidos em relação ao número de grupos unitários da referência, ou seja, o método agrupou um número reduzido de aspectos em decorrência do alto número de grupos unitários obtidos. Esse comportamento aumenta a precisão do método, pois a precisão de cada grupo unitário é de 100%. No entanto, note que esse resultado não implica que o método seja eficiente para a tarefa de agrupamento de aspecto.

Tabela 6.2: Cobertura

N.	Métodos	Livro	Câmera	Smartphone	Média
1	sinonimos (baseline)	0,231	0,281	0,296	0,269
2	sino + hipe/hipo	0,242	0,287	0,314	0,281
3	sino + hipe/hipo + mero/holo	0,242	0,287	0,310	0,279
4	sino + hipe/hipo + mero/holo + corref	0,321	0,307	0,364	0,330
5	word embeddings	0,231	0,292	0,300	0,274
6	OpCluster-PT	**0,748**	**0,687**	**0,550**	**0,661**

Na Tabela 6.2, apresentamos os resultados obtidos a partir da medida de avaliação de *cobertura*. Observe que, para os métodos 1, 2, 3, 4 e 5, obtivemos resultados entre 0,281 e 0,364. Somente o método OpCluster-PT apresentou resultado superior, com 0,748 para o domínio de livro, 0,687 no domínio de câmera e 0,550 no domínio de smartphone. O método Opcluster-PT também apresentou melhores resultados nas avaliações de *medida-f e medida-f global* (veremos a seguir nas Tabelas 6.3 e 6.4). Para o domínio de livro, obtivemos o melhor resultado, com a aplicação do método OpCluster-PT, em relação aos domínios de câmera e smartphone. Nestes dois últimos domínios, observamos maior ocorrência de aspectos específicos do domínio, fato este que potencializa a complexidade de identificação e agrupamento automático desses aspectos.

Tabela 6.3: Medida-F

N.	Métodos	Livro	Câmera	Smartphone	Média
1	sinonimos (baseline)	0,374	0,438	0,454	0,422
2	sino + hip	0,383	0,442	0,471	0,432
3	sino + hipe/hipo + mero/holo	0,383	0,442	0,466	0,430
4	sino + hipe/hipo + mero/holo + corref	0,480	0,466	0,527	0,491
5	word embeddings	0,372	0,448	0,457	0,425
6	OpCluster-PT	**0,827**	**0,792**	**0,702**	**0,773**

Na Tabela 6.3, apresentamos os resultados obtidos a partir da avaliação da *medida-f* para cada um dos seis métodos implementados. O método com melhor desempenho é o método proposto neste trabalho de mestrado, o algoritmo de agrupamento de aspectos de opinião OpCluster-PT. Note que os demais métodos implementados apresentaram um desempenho

ruim, com média entre 0,422 e 0,491, sendo que o método OpCluster-PT apresentou uma média de 0,773.

Tabela 6.4: Medida-F global

N.	Métodos	Livro	Câmera	Smartphone	Média
1	sinonimos (baseline)	0,300	0,351	0,347	0,332
2	sino + hipe/hipo	0,249	0,319	0,333	0,300
3	sino + hipe/hipo + mero/holo	0,244	0,319	0,333	0,298
4	sino + hipe/hipo + mero/holo + corref	0,399	0,409	0,508	0,438
5	word embeddings	0,280	0,336	0,350	0,322
6	OpCluster-PT	**0,711**	**0,605**	**0,583**	**0,633**

Na Tabela 6.4, apresentamos os resultados obtidos a partir da avaliação da *medida-f global*. O método OpCluster-PT também apresentou os melhores resultados em detrimento dos demais métodos implementados. Em seguida, o método 4 também apresentou resultados melhores em relação aos métodos 1, 2, 3 e 5. Nós observamos que a utilização do recurso de reconhecimento de correferências retornou resultados melhores comparado aos métodos que utilizam apenas relações lexicais extraídas de ontologias lexicais. Portanto, a identificação de correferências parece eficiente para a tarefa de identificação e agrupamento de aspectos para mineração de opinião, porque esse tipo de recurso pode mapear unidades lexicais específicas do domínio, diferentemente de recursos baseados apenas em relações em ontologias lexicais em que, geralmente, não são encontradas unidades lexicais mais específicas do domínio. Na seção seguinte, faremos uma discussão sobre os resultados aprensentados nesta seção, bem como suas implicações.

6.3 Discussão dos resultados

O primeiro método implementado neste trabalho utiliza uma ontologia lexical que foi usada para extração automática de relações de sinonímia. Veja que, assim como os métodos 2 e 3 (ver tabelas da Seção 6.2), em que também são usados relações de similaridade lexical, no geral, todos esses métodos apresentaram um desempenho ruim, assim como o método baseado em *word embeddings*. Os métodos 4 e 6 (ver tabelas da Seção 6.2) foram os métodos que apresentaram melhor desempenho de *cobertura, medida-f e medida-f global* para a tarefa de agrupameto de aspectos. No método 4 (ver tabelas da Seção 6.2), foram extraídas automaticamente relações entre aspectos utilizando uma ontologia lexical e, de modo incremental, foram reconhecidas automaticamente cadeias de referentes ou correferências. Observamos que o reconhecimento de correferências aumentou significativamente a performance do método. Portanto, este tipo de recurso é interessante para a tarefa de agrupamento de aspectos. Nós observamos que correferências podem marcar relações entre aspectos que representam as especificidades do domínio, sendo que apenas a utilização de léxicos ou ontologias lexicais, na maioria das vezes, parece insuficiente para o reconhecimento de termos específicos do domínio.

O método OpCluster-PT é o método proposto por este trabalho de mestrado. O algoritmo Opcluster-PT foi proposto a partir de um estudo linguístico aprofundado sobre os principais fenômenos linguísticos em textos opinativos, e este método superou os demais métodos implementados, exceto quanto à *precisão*, em que o método baseado em relações de sinonímia apresentou melhores resultados em detrimento dos outros métodos. No entanto, discutimos esse comportamento na Seção 6.2. Veja que o algoritmo proposto utiliza vários recursos baseados em conhecimento linguístico (relações de similaridade lexical, correferências, léxicos de deverbais e estrangeirismos da língua, lista de diminutivos/aumentativos, a extração de *substrings*), e a identificação desses elementos só foi possível a partir de um estudo linguístico aprofundado sobre textos opinativos da web.

Observamos também uma pequena diferença no desempenho dos métodos em relação aos domínios analisados. Por exemplo, no domínio de livro, o método OpCluster-PT teve um melhor resultado se comparado aos domínios de smartphone e câmera. Para o domínio de livro, obtivemos um *score* de 0,827 de *medida-f*. Nós observamos também que, nos domínios de câmera e smartphone, houve um número maior de ocorrência de aspectos relacionados às especificidades do domíno. Esse tipo de aspecto representa maior dificuldade tanto quanto ao reconhecimento quanto ao agrupamento automático. Por exemplo, no domíno de livro, apenas 21 grupos foram identificados, contrapondo aos 36 grupos identificados para o domínio de câmera e aos 49 grupos do domínio de smartphone. Notamos que, para o domínio de livros, os usuários emissores de revisões não possuíam perfis "especializados", ou seja, não eram críticos literários ou especialistas em literatura, portanto não possuíam conhecimento suficiente para avaliar aspectos "mais especializados" deste domínio, diferentemente dos domínios de câmera e smartphone. Neste dois últimos domínios, os produtos são populares e de fácil identificação de características mais especializadas. Portanto, assim como evidenciado anteriormente, um dos desafios da tarefa de agrupamento de aspectos é o reconhecimento e agrupamento de aspectos específicos do domínio.

Por fim, iremos apresentar através de alguns exemplos o que seriam "bons" grupos de aspectos e grupos de aspectos "ruins" gerados automaticamente. Por exemplo, um "bom grupo" implica que os aspectos desse grupo estão contídos em maior quantidade no grupo de referência (humano). Vejamos a Figura 6.1. Neste grupo, apresentamos um exemplo de grupos "bem" formados nos domínios de smartphone, câmera e livro. No domíno de smartphone, o grupo gerado automaticamente possui os aspectos "custo", "custo_benefício", "preço", "valor", "investimento" e "barato". Esse grupo foi gerado automaticamente e possui 90% dos aspectos do grupo da referência (humano) (ver Tabela 7.1 do apêndice). Portanto, o método automático obteve uma boa taxa de acerto. No entanto, alguns grupos também foram "mal" formados, ou seja, o método pode ter agrupado indevidamente alguns itens no grupo ou ter deixado de agrupá-los. Vejamos a Figura 6.2.

Observe que o grupo gerado automaticamente do domínio de smartphone é composto pelos aspectos "empresa", "lg", "nokia", "sony", "sony_ericson", "programa", "design", "sistema" e "modelo'. Neste grupo, os aspectos "design", "sistema", "programa" foram agrupados indevi-

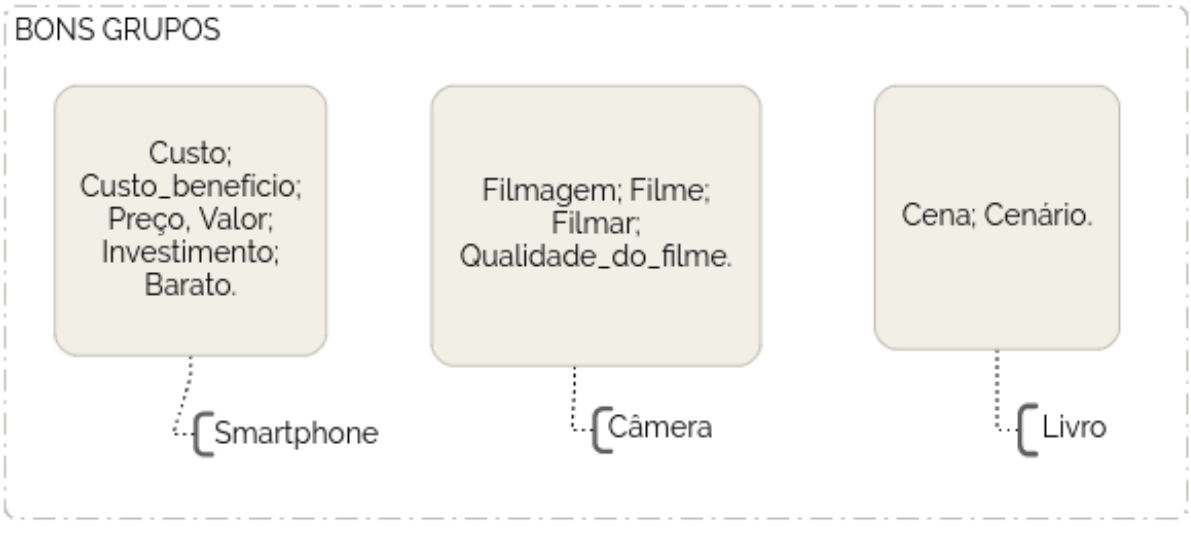

Figura 6.1: Exemplo de “bons” grupos formados automaticamente.

damente (ver Tabela 7.1 do apêndice). No domínio de livro, observe que obtevemos automaticamente um grupo unitário. Esse também é um grupo “ruim”, pois o indicativo de aspecto “explorar” faz parte do grupo de termos usados pelo usuário para avaliar a propriedade “tema” do livro (Ver Tabela 7.3 do apêndice), no entanto, o método não foi capaz de agrupá-lo.

Figura 6.2: Exemplo de grupos “ruins” formados automaticamente.

Capítulo

7

Considerações finais

7.1 Considerações finais

De acordo com Pang *et al.* (2002), a tarefa de mineração de opinião requer compreensão profunda de características das línguas naturais e do contexto textual. Portanto, neste trabalho, optamos por uma abordagem linguística para resolução do problema de agrupamento de aspectos para mineração de opinião. A partir de um estudo empírico e de aprofundamento linguístico sobre os principais fenômenos que acometem textos opinativos, nós propusemos e implementamos métodos automáticos para resolução do problema de identificação de grupos de aspectos correlatos para sistemas de mineração de opinião. Nossa abordagem não é dependente de uma língua, no entanto, apresentará maior adequação ao português do brasil, pois foram usados um conjunto de revisões de usuários desta língua.A partir da identificação dos principais fenômenos linguísticos em textos opinativos que relacionam aspectos correlatos, nós implementamos seis métodos: nos três primeiros métodos, utilizamos uma ontologia lexical para extração automática de relações lexicais entre aspectos com base em similaridade lexical; no quarto método, utilizamos, além da ontologia lexical, o sistema de resolução de correferência para extração automática de relações entre aspectos; no quinto método, utilizamos o modelo estatístico *word embeddings* para extração de unidades lexicais similares no contexto; por fim, nós propusemos e implementamos para essa proposta de mestrado o algoritmo OpCluster-PT. O algoritmo proposto utiliza vários recursos linguístico-computacionais para extração automática de relações entre aspectos em textos opinativos e agrupamento destes aspectos para sistemas de mineração de opinião. O algoritmo Opcluster-PT apresentou resultados superiores de *cobertura, medida-f e medida-f global* em relação aos outros cinco métodos implementados. Além disso, constatamos que o reconhecimento de relações de sinonímia, meronímia/holonímia, hiperonímia/hiponímia e causativas, além dos fenômenos de estrangeirismo, deverbalidade, diminu-

tivos/aumentativos e correferências são fundamentais para o reconhecimento de relações entre aspectos em textos opinativos. Observamos, também que, especialmente a partir do reconhecimento das relações de correferências, é possível captar relações entre aspectos que representam as especificidades de um domínio. Conluímos também que são necessárias, em média, 40 revisões para o reconhecimento de grupos de aspectos de um domínio, no entanto, isso pode variar de acordo com o perfil do usuário emissor da revisão. Além disso, em média, 40% dos termos indicativos de aspectos implícitos de um domíno fazem parte da classe de nomes (substantivos, adjetivos, advérbios, etc.) e 35,58% são verbos. Quanto aos aspectos explícitos, a maior parte é composta por nomes (não-verbais) e uma porção não relevante estatisticamente consite de verbos. Por fim, concluímos que, para as tarefas de mineração de opinião, a compreensão e reconhecimentos dos fenômenos intrínsicos e extrínsecos da língua parece-nos fundamental para a melhor compreensão desse tipo de dado e pode trazer melhores resultados com a proposição de métodos automáticos não-supervisionados e mais facilmente adaptáveis a outros domínios de aplicação. Acreditamos que esse tipo de abordagem de investigação, que trata as "causas" linguísticas para resolução dos problemas de processamento computational de uma língua natural, especialmente no domínio de opinião, além de propiciar conhecimentos teóricos às áreas da linguística e computação, pode prover métodos "mais baratos" e mais adaptáveis ao domínio e à situação social, se comparados a métodos superficiais. Por fim, salientamos que as hipóteses levantadas neste trabalho de mestrado foram confirmadas, ou seja, a partir de um estudo de córpus é possível extrair o conhecimento necessário para compreensão semântica de um domínio. Além disso, os resultados da exploração de métodos baseados em conhecimentos linguísticos para sistemas de mineração de opinião podem trazer melhores resultados para extração das especificidades de um domínio, bem como promover métodos mais "adaptáveis" e mais "baratos". Adaptáveis porque o conhecimento explorado é o conhecimento da língua geral, que pode ser reutilizado em outros domínios e aplicações. Além disso, é um método relativamente "barato" porque recursos da língua (por exemplo, wordnets, lexicos, wikipedia, etc.) são facilmente encontrados, além deste tipo de método não exigir um conjunto de dados etiquetados.

7.2 Limitações

As duas principais limitações enfrentadas neste trabalho de mestrado foram: as variáveis tempo e recurso. Infelizmente, o tempo do trabalho de mestrado (obrigações acadêmicas, investigação, desenvolvimento da pesquisa, artigos, relatórios, implementação, testes e escrita) é curto. Além disso, recursos linguístico-computacionais da língua portuguesa, infelizmente, ainda são escassos. Não encontramos disponível, por exemplo, nenhum dicionário lexical composto por construções de diminutivos e aumentativos para o português. Por fim, também constatamos um número muito reduzido de trabalhos na área de mineração de opinião do português do brasil, em comparação aos trabalhos desenvolvidos em outras línguas, por exemplo, inglês, chinês, etc. Especialmente para a tarefa de agrupamento de aspectos, não encontramos nenhum

trabalho para o português. Além disso, há uma limitação sobre o desempenho de alguns recursos linguístico-computacionais utilizados, como a Onto-PT (Oliveira, 2014) e o CORP (Fonseca *et al.*, 2016), que apresentaram uma performance ruim, em alguns casos. Por exemplo, grande parte dos aspectos específicos dos domínios não foram encontrados na ontologia lexical e alguns conjuntos de correferências não foram anotados adequadamente pelo CORP (Fonseca *et al.*, 2016). Por fim, constatamos também que nosso método ainda possui algumas limitações. Por exemplo, para o agrupamento de aspectos: (ii) *oriundos de gírias*, por exemplo "o cara é um gênio", em que "cara" é usado para avaliar o aspecto "autor" do livro; (iii) *oriundos de nomes próprios*, por exemplo, "a malu é maravilhosa", em que "malu" é usado para avaliar o aspecto "personsagem" do livro; e (iii) *oriundos de conteúdo implícito, especialmente indicativos de aspectos implícitos representados por n-gramas*, por exemplo "recebi chamada até na beira do rio são franciso" e "sociedade do big brother", sendo que, no primeiro exemplo, o usuário avalia o aspecto "sinal" do smartphone e, na segunda revisão, o usuário avalia um "livro" específico do domínio de livro.

7.3 Trabalhos futuros

Para trabalhos futuros, a exploração de tecnologias semânticas baseadas em dados abertos conectados (do inglês, *linked open data* [1]), por exemplo, podem ser explorados para proposição de métodos melhorados para sistemas de mineração de opinião, bem como a investigação desses métodos em conjunto com métodos estatísticos baseados em aprendizagem de máquina modernos (*deep learning*). Além disso, o algoritmo proposto ainda pode ser adaptado para outros domínios e aplicações de PLN. É possível também, explorar o estudo linguístico e os recursos computacionais desenvolvidos neste trabalho de mestrado para proposição de métodos para as tarefas de sintése temporal de preferências de usuários, bem como para a sumarização automática.

[1] O termo *linked open data* refere-se ao conjunto de melhores práticas para publicação e conexão de dados estruturados na Web.

Referências Bibliográficas

Abdul-Mageed, M.; Diab, M.; Kübler, S. (2012). Samar: Subjectivity and sentiment analysis for arabic social media. *Proceedings of the 3th Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, p. 20–37, Jeju, Republic of Korea.

Abu-Jbara, A.; King, B.; Diab, M. T.; Radev, D. R. (2013). Identifying opinion subgroups in arabic online discussions. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, p. 829–835, Sofia, Bulgaria.

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Blackwell Publishing, 3ª edição.

Alvarez, M.; Lim, S. (2007). A graph modeling of semantic similarity between words. *Proceedings of the Conference on Semantic Computing*, p. 355–362, Irvine, United States.

Avanço, L.; Nunes, G. M. V. (2014). Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. *Proceedings of the Brazilian Conference on Intelligent Systems*, p. 277–281, São Carlos, Brazil.

Balage Filho, P. P.; Pardo, T. A. S. (2014). Aspect extraction using semantic labels. *Proceedings of the 8th International Workshop on Semantic Evaluation*, p. 433–436, Dublin, Ireland.

Baségio, T. (2006). Uma abordagem semiautomática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do brasil. Dissertação (Mestrado), Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil.

Bhuiyan, T.; Xu, Y.; Josang, A. (2009). State-of-the-art review on opinion mining from online customers' feedback. *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, p. 385–390, Tokyo, Japan.

Bick, E. (2000). *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Aarhus, 1ª edição.

Biderman, M. T. (2001). *Teoria Linguística: teoria lexical e linguísitica computacional*. Martins Fontes. 1ª.

Biemann, C. (2005). Ontology learning from text: A survey of methods. *LDV Forum*, v. 20, p. 75–93.

Blei, D. M.; Ng, Y. A.; Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022.

Bollegala, D.; Matsuo, Y.; Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. *Proceedings of the 16th International Conference on World Wide Web*, p. 757–766, New York, United States.

Brank, J.; Grobelnik, M.; Mladenic, D. (2005). A survey of ontology evaluation techniques. *Proceedings of the Conference on Data Mining and Data Warehouses*, p. 1–4, Ljubljana, Slovenia.

Brewster, C.; Alani, H.; Dasmahapatra, S.; Wilks, Y. (2004). Data-driven ontology evaluation. *Proceedings of the 4th Language Resources and Evaluation Conference*, p. 164–168, Lisbon, Portugal.

Bronckart, J. P. (1997). *Activité langagière, textes et discours pour un interactionisme socio-discursif*. Lausanne: Delachaux et Niestlé, 1ª edição.

Buitelaar, P.; Magnini, B. (2005). *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, 1ª edição.

Burton-Jones, A.; Storey, V. C.; Sugumaran, V.; Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering*, v. 55, p. 84–102.

Cadilhac, A.; Aussenac-Gilles, N.; Benamara, F. (2010). Ontolexical resources for feature-based opinion mining: a case-study. *Proceedings of the 23th International Conference on Computational Linguistics*, p. 77–86, Pekin, China.

Chaves, M. S.; Freitas, L. A.; Souza, M.; Vieira, R. (2012). PIRPO: an algorithm to deal with polarity in portuguese online reviews from the accommodation sector. *Proceedings of 17th International Conference on Applications of Natural Language to Information Systems*, p. 296–301, Groningen, The Netherlands.

Chen, Y.; Zhao, Y. and, Q. B.; Liu, T. (2016). Product aspect clustering by incorporating background knowledge for opinion mining. *PLOS ONE*, v. 11, p. 1–16.

Ciaramita, M.; Gangemi, A.; Ratsch, E.; Šaric, J.; Rojas, I. (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, p. 659–664, Edinburgh, Scotland.

Collobert, R.; Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, p. 160–167, Helsinki, Finland.

Condori, R. E. L. (2014). Sumarização automática de opiniões baseada em aspectos. Dissertação (Mestrado), Universidade de São Paulo, São Carlos, Brasil.

de desenvolvimento CoGrOO, T. (2012). *CoGrOO: Corretor Gramatical acoplável ao LibreOffice e Apache OpenOffice*. Centro de Competência de Software Livre do Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, Brasil.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, v. 39, p. 1–38.

Faure, D.; Nédellec, C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, p. 5–12, Granada, Spain.

Fayyad, U.; Piatetsky-shapiro, G.; Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37–54.

Fensel, D. (2003). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, 2ª edição.

Ferraz, A. P. (2008). Neologismos semânticos na publicidade imprensa: uma abordagem cognitivista. Isquierdo, A. N.; Finatto, M. J. B., editores, *As ciências do léxico: lexicologia, lexicografia, terminologia*, v. 4, p. 65–80. Campo Grande, Brasil.

Ferreira, J. P.; Janssen, M. (2017). *Dicionário de Formas Não Adaptadas*. Instituto de Linguística Teórica e Computacional, 1ª edição.

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in Linguistic Analysis (Special Volume of the Philological Society)*, v. 1952, p. 1–32.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, v. 7, p. 179–188.

Fonseca, E. B. (2014). Resolução de correferências em língua portuguesa: pessoa, local, organização. Dissertação (Mestrado), Pontifícia Universidade Católica de Minas Gerais, Porto Alegre, Brasil.

Fonseca, E. B.; Vieira, R.; Vanin, A. A. (2016). Corp: Coreference resolution for portuguese. *Proceedings of the 12th International Conference on the Computational Processing of Portuguese*, p. 9–11, Tomar, Portugal.

Fox, M. S.; Barbuceanu, M.; Gruninger, M.; Lin, J. (1997). An organization ontology for enterprise modelling. *Proceedings of the International Conference on Enterprise Integration Modeling Technology*, p. 1–25, Torino, Italy.

Freitas, C.; Motta, E.; Milidiú, R.; Cesar, J. (2012). Vampiro que brilha... rÁ! desafios na anotação de opinião em um córpus de resenhas de livros. *Anais do XI Encontro de Linguística de Corpus*, p. 1–13, São Carlos, Brasil.

Freitas, L. A.; Vieira, R. (2013). Ontology based feature level opinion mining for portuguese reviews. *Proceedings of the 22th International Conference on World Wide Web*, p. 367–370, Rio de Janeiro, Brazil.

Gaizauskas, R.; Humphreys, K. (1997). Using a semantic network for information extraction. *Cambridge University Press*, v. 3, p. 147–169.

García, A.; Cuadros, M.; Rigau, G.; Gaines, S. (2014). V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation)*, p. 833–837, Dublin, Ireland.

Ghose, A.; Ipeirotis, P.; Sundararajan, A. (2007). Opinion mining using econometrics: A case study on reputation systems. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 416–423, Prague, Czech Republic.

Gómez-Pérez, A. (1995). Some ideas and examples to evaluate ontologies. *Proceedings of the 11th Conference on Artificial Intelligence for Applications*, p. 299–305, Washington, United States.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Academic Press Ltd.*, v. 5, p. 199–220.

Guarino, N. (1998). Formal ontology and information systems. *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning*, p. 3–15, Trento, Italy.

Guarino, N.; Welty, C. (2002). Evaluating ontological decisions with ontoclean. *Association for Computing Machinery*, v. 45, p. 61–65.

Haase, P.; Völker, J. (2008). Ontology learning and reasoning - dealing with uncertainty and inconsistency. *Uncertainty Reasoning for the Semantic Web I*, v. 5327, p. 366–384.

Harris, Z. S. (1968). *Mathematical structures of language*. Interscience tracts in pure and applied mathematics, 21ª edição.

Hartmann, N.; Fonseca, E.; Shulby, C.; Treviso, M.; Rodrigues, J.; Aluisio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *Proceedings of the Symposium in Information and Human Language Technology*, p. 122–131, Uberlandia, Brazil.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics*, p. 539–545, Stroudsburg, United States.

Hu, M.; Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, p. 168–177, Seattle, United States.

Hughes, T.; Ramage, D. (2007). Lexical Semantic Relatedness with Random Graph Walks. *Computational Linguistics*, v. 7, p. 581–589.

Ittoo, A.; Bouma, G.; Maruster, L.; Wortmann, H. (2010). Extracting meronymy relationships from domain-specific, textual corporate databases. *Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems*, p. 48–59, Cardiff, United Kingdom.

Janssen, M.; Ferreira, J. P. (2007). *Dicionário de nomes deverbais*. Intituto de Linguística Teórica e Computacional, 1ª edição.

Jurafsky, D.; Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, 1ª edição.

Kasama, A. f. (2009). Estruturação do conhecimento e relações semânticas : uma ontologia para o domínio da naonociência e nanotecnologia. Dissertação (Mestrado), Universidade Federal do Espírito Santo, Vitória, Brasil.

Koch, I. G. V. (2004). *Introdução à Linguística Textual*. Martins Fontes, 1ª edição.

Labov, W. (1994). *Principles of linguistic change: Internal Factors*. Oxford, 1ª edição.

Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, p. 768–774, Stroudsburg, United States.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 1ª edição.

Liu, B.; Hu, M.; Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. *Proceedings of the 14th International Conference on World Wide Web*, p. 342–351, Chiba, Japan.

Lopes, E. (1995). *Fundamentos da Linguística Contemporânea*. Editora Cultrix, 13ª edição.

Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. (2009). ExATO LP - An Automatic Tool for Term Extraction from Portuguese Language Corpora. *Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 427–431, Poznan, Poland.

Lu, Y.; Zhai, C. (2008). Opinion integration through semi-supervised topic modeling. *Proceedings of the 17th International Conference on World Wide Web*, p. 121–130, Beijing, China.

Lyons, J. (1970). *Linguistique générale - Introduction à la linguistique théorique*. Librarie Larousse. 1ª.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 50st Berkeley Symposium on Mathematical Statistics and Probability*, p. 281–297, California, United States.

Maedche, A.; Staab, S. (2001). Ontology for the semantic web. *IEEE Educational Activities Department*, v. 16, p. 72–79.

Maedche, A.; Staab, S. (2002). Measuring similarity between ontologies. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, p. 251–263, London, United Kingdom.

Maedche, A.; Staab, S. (2004). Ontology learning. *Handbook on Ontologies*, v. 1, p. 173–189.

Martins, A. F. (2011). Construção de ontologias de tarefa e sua reutilização na engenharia de requisitos. Dissertação (Mestrado), Universidade Federal do Espírito Santo, Vitória, Brasil.

Matoré, G. (1973). *La méthode en lexicologie: domaine français*. Nouv.éd.:Didier, 1ª edição.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013). Efficient estimation of word representations in vector space. *Computing Research Repository*, v. 1301.3781.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, v. 3, p. 235–244.

Mnih, A.; Hinton, G. E. (2009). A scalable hierarchical distributed language model. Koller, D.; Schuurmans, D.; Bengio, Y.; Bottou, L., editores, *Advances in Neural Information Processing Systems 21*, p. 1081–1088.

Mukherjee, S.; Joshi, S. (2013). Sentiment aggregation using conceptnet ontology. *Proceedings of the 6th International Joint Conference on Natural Language Processing*, p. 570–578, Nagoya, Japan.

Munezero, M.; Montero, C. S.; Sutinen, E.; Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, v. 5, p. 101–111.

Oliveira, H. G. (2014). Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT. *Proceedings of the Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, p. 64–68, São Carlos, Brazil.

Pang, B.; Lee, L.; Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 79–86, Stroudsburg, United States.

Patra, B. G.; Mandal, S.; Das, D.; Bandyopadhyay, S. (2014). Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation*, p. 370–374, Dublin, Ireland.

Pennington, J.; Socher, R.; Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1532–1543, Doha, Qatar.

Pereira, F.; Tishby, N.; Lee, L. (1993). Distributional clustering of english words. *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, p. 183–190, Stroudsburg, United States.

Ribeiro Junior, L. C. (2008). Ontolp: construção semiautomática de ontologias a partir de textos da língua portuguesa. Dissertação (Mestrado), Universidade do Vale do Rio dos Sinos, Porto Alegre, Brasil.

Roberts, A. (2005). Learning meronyms from biomedical text. *Proceedings of the Association for Computational Linguistics Student Research Workshop*, p. 49–54, Michigan, United States.

Ryu, P.; Choi, K. S. (2006). Taxonomy learning using term specificity and similarity. *Proceedings 2th Workshop on Ontology Learning and Population*, p. 41–48, Sydney, Australia.

Sales, S. R.; Ferreira, A. G.; Vargas, F. A. (2015). Juventude emdiálogo: tecnologias digitais na extensão universitária. *Revista Conexão da Universidade Estadual de Ponta Grossa*, v. 11, p. 293–316.

Saussure, F. (2002). *Curso de linguística geral*. Pensamento-Cultrix, 24ª edição.

Steinwart, I.; Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1ª edição.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, v. 2, p. 325–347.

Todorov, T. (1966). Recherches sémantiques. *Langages*, v. 1, p. 5–43.

Trier, J. (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg, C. Winter, 1ª edição.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 417–424, Stroudsburg, United States.

Vaassen, F. (2014). Measuring emotion: Exploring the feasibility of automatically classifying emotional text. Dissertação (Mestrado), University of Antwerp, Antwerp, Belgium.

Van Hee, C.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. *Proceedings of the 10th Recent Advances in Natural Language Processing*, p. 672–680, Hissar, Bulgaria.

Vargas, F. A.; Pardo, T. A. S. (2017). Clustering and hierarchical organization of opinion aspects: a corpus study. *Proceedings of the 14th Meeting of Linguistics of Corpus and 9th Brazilian School of Computational Linguistics*, p. 342–351, São Leopoldo, Brazil.

Vossen, P. (1997). Eurowordnet: a multilingual database for information retrieval. *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, p. 5–7, Zurich, Switzerland.

Vossen, P. (2011). Ontologies. *Mitkov, R. (Org). The Oxford Handbook of Computational Linguistics*, v. 1.

Wasow, T. (1967). *Anaphoric relations in english*. Tese (Doutorado), Massachusetts Institute of Technology: MIT, Massachusetts, United States.

Wu, C.-W.; Liu, C.-L. (2003). Ontology-based text summarization for business news articles. *Proceedings of the 3th International Symposium on Computer Architecture*, p. 389–392, Honolulu, United States.

Xavier, C. C.; Lima, V. L. S. (2010). A semi-automatic method for domain ontology extraction from portuguese language wikipedia's categories. *Proceedings of the 20th Brazilian Symposium on Artificial Intelligence*, p. 11–20, São Bernardo do Campo, Brazil.

Yu, J.; Zha, Z.; Wang, M.; Wang, K.; Chua, T. (2011). Domain-assisted product aspect hierarchy generation: Towards hierarchical organization of unstructured consumer reviews. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 140–150, Edinburgh, United Kingdom.

Zhai, Z.; Liu, B.; Xu, H.; Jia, P. (2011). Clustering product features for opinion mining. *Proceedings of the 4th International Conference on Web Search and Data Mining*, p. 347–354, New York, United States.

Zhang, H. (2004). The optimality of naive bayes. *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, p. 1–6, Florida, United States.

Zhang, S.; Jia, W.; Xia, Y.; Meng, Y.; Yu, H. . (2011). Product features extraction and categorization in chinese reviews. *Proceedings of the 6th International Multi-Conference on Computing in the Global Information Technology*, p. 38–42, Nice, France.

Zhao, L.; Li, C. (2009). Ontology based opinion mining for movie reviews. *Proceedings of the 3th International Conference on Knowledge Science, Engineering and Management*, p. 204–214, Berlin, Germany.

Zhou, X.; Wan, X.; Xiao, J. (2015). Representation learning for aspect category detection in online reviews. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, p. 417–423, Texas, United States.

Zipf, G. (1970). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1ª edição.

Apêndice

Grupos de aspectos

Neste apêndice, apresentamos os grupos de aspectos identificados (referência humana) para os domínios de smartphone (ver Tabela 7.1), câmera digital (ver Tabela 7.2) e livro (ver Tabela 7.3), além da organização hierárquica desses grupos (ver Figuras 7.1, 7.2 e 7.3), respectivamente.

Tabela 7.1: Grupos de aspectos no domínio de smartphone.

Grupo	**Aspectos** [2]
G1	Aparelho, Telefone, Produto, Celular, "Porcaria", Smartphone, Aparelho celular, N70 K700i, XT317.
G2	Net, Wireless, 3G, Conexão, Conectividade, WAP, Internet, Wifi.
G3	Manuseio, Interface, Menu, Praticidade, Facilidade, Usabilidade, Função, Recurso, Funcionalidade, Extra, Opção, Linguagem, "Operação", "Fácil de manusear", "Fácil de Usar", "Fácil de Mexer", "Prático".
G4	"Rápido", "Trava", "Lento", "Demora a responder", "Congela", "Restarta", "Bugs", "Tempo de resposta", Velocidade, "Demorar".
G5	Custo, Valor, Preço, Investimento, "Acessível", "Barato", "Custo-beneficio".
G6	"Descarrega", Bateria, Autonomia da bateria, Duração da bateria, Carregamento.
G7	Tecla, Teclado.
G8	GPS.
G9	Aplicativo.
G10	Tela, Visor, Vidro, Display, "Sensibilidade", Tamanho do visor, Touchscreen, Touch, Touch screen.
G11	Áudio, Som, Volume, Sonorização, Música, Mp3, Mp3 player, Qualidade do áudio, Qualidade do som, Volume do áudio, Qualidade sonora, Alto falante.
G12	Toque, Hits, Toques polifônicos
G13	Google maps
G14	Design, Estético, Estilo, Modelo, Elegância, Beleza, "Robusto", "Lindo", "Moderno", "Arrojado", "Chique", "atual", "Bonito", "Volumoso".
G15	Câmera, Foco da câmera, Resolução da Câmera, Flash da câmera, Luz do flash, Zoom da câmera, "Filmar", Câmera imbutida, Filmadora, Câmera digital, Megapixels.
G16	Fotografia, Qualidade da foto, Foto Panorama, Foto.
G17	Bluetooth
G18	Fabricante, Fábrica, Marca, Empresa, Motorola, LG, Sony, Nokia, Sony ericson, Siemens.
G19	Dual Chip, SIM

[2]Os aspectos com aspas duplas são usados para caracterizar os aspectos implícitos.

G20	“Recebi chamadas até na beira do rio são francisco”, “Funciona em qualquer lugar”, Recepção, Ligação, *Quadriband*, Sinal.
G21	Tv
G22	“Pesado”, “Leve”, “Leveza”, “Levinho”, “Versátil”.
G23	Botão Liga/Desliga, Botão de toque, Botão.
G24	Sd de memória, Memória interna, Cartão de memória, Expansão de memória, Cartão de expansão, Espaço de memória, Memória.
G25	Cabo para TV, Cabo de dados.
G26	Email
G27	Rádio
G28	Imagem, Definição de imagem.
G29	Vídeo.
G30	Leitor de pdf.
G31	Sistema, Software, Programa, “Falta de compatibilidade”.
G32	Sincronização, “Acesso aos dados”.
G33	Fone de ouvido, Fone.
G34	Jogo.
G35	Carregador de carro, Carregador.
G36	Manual.
G37	Tamanho, “Pequeno”, “Compacto”.
G38	Processador.
G39	Durabilidade.
G40	Usb.
G41	Viva voz, Gravador, Gravador de voz.
G42	Despertador.
G43	Acessório.
G44	Bloco de notas.
G45	Confiabilidade, “Confiança na marca”.
G46	Calendário.
G47	Antena.
G48	Agenda Telefônica.

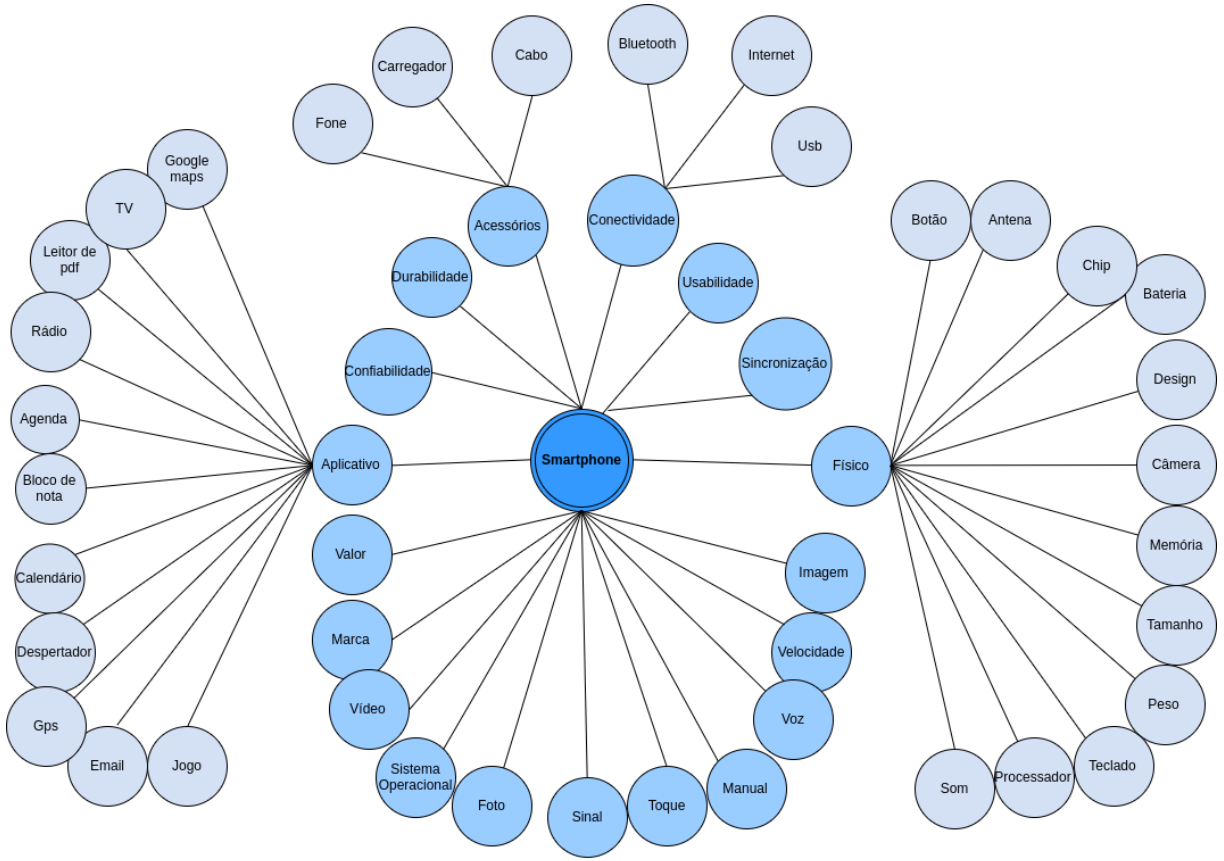

Figura 7.1: Organização hierárquica de aspectos no domínio de smartphone.

Tabela 7.2: Grupos de aspectos no domínio de câmera.

Grupo	**Aspectos** [3]
G1	Câmera, Canon rebel T3i, Câmera amador, Câmera digital, Câmera semiprofissional, Câmerazinha, Equipamento, Máquina, Produto, Qualidade da câmera, WB2000, H10, Máquina digital, Fz35, Canon.
G2	Custo, Preço, Valor, Custo-benefício, "Barato", Investimento.
G3	Modo de imagem, Qualidade da imagem, Imagem, Cor da imagem, Filtro de imagem, Modo noite.
G4	Praticidade, Manuseio, Facilidade, "Fácil de usar", Função, Recurso, Recurso de configuração, Recurso de edição, Acionamento de função, "Fácil de utilizar", "Fácil de manusear", "Fácil de operar", "Prático", "Facilidade de mexer", "Intuitiva", "Auto explicativa", Opção, Funcionalidade, Menu, "Facilidade de uso".
G5	"Beleza", Acabamento, Design, Aparência, Material, "Linda", "Bonita", "Elegante".
G6	Acessório.
G7	Sony, Fuji, Empresa, Nikon, Benq, Marca.
G8	Bateria, Bateria reserva.
G9	Botão.
G10	Resolução da foto, Qualidade de foto, Opção de foto, Cor da foto, Navegação na foto, Foto, Fotografia, Foto dentro d'agua, Foto Noturna, Foto panorâmica, Nigthshot.
G11	Resolução do vídeo, Qualidade de vídeo, Qualidade do filme, Filmagem, Vídeo, Gravação, Filme, "Filmar".
G12	"Fino", "Compacta", "Pequena", "Grande", Tamanho, "Medida", Volume.
G13	Flash.
G14	Consumo de energia, Consumo.
G15	Conectividade.
G16	Visor ocular, Tamanho da tela, Touchscreen, Tela, Display, Visor.
G17	Manual em português, Manual de instrução, Manual.
G18	Sd card, Cartão de memória, Memória interna, Cartão, Cartão SD, Memória, *Memory Stick*.
G20	Zoom, Ultrazoom, Zoom ótico.
G21	Peso, "Leve", "Leveza", "Versátil".
G22	Garantia.
G23	Capinha.
G24	Película de proteção.
G25	Foco.

[3]Os aspectos com aspas duplas são usados para caracterizar os aspectos implícitos.

G26	Lente, Lente auxiliar, Objectiva.
G27	“Demora para responder”, Velocidade, “Rápida”, Demorar.
G28	*Slow Motion*.
G29	Redutor de olhos vermelhos.
G30	Som.
G31	Disparo.
G32	Processador.
G33	Reconhecimento facial
G34	Megapixels, Resolução, Nitidez.
G35	Mostrador de níveis.
G36	Pilha.

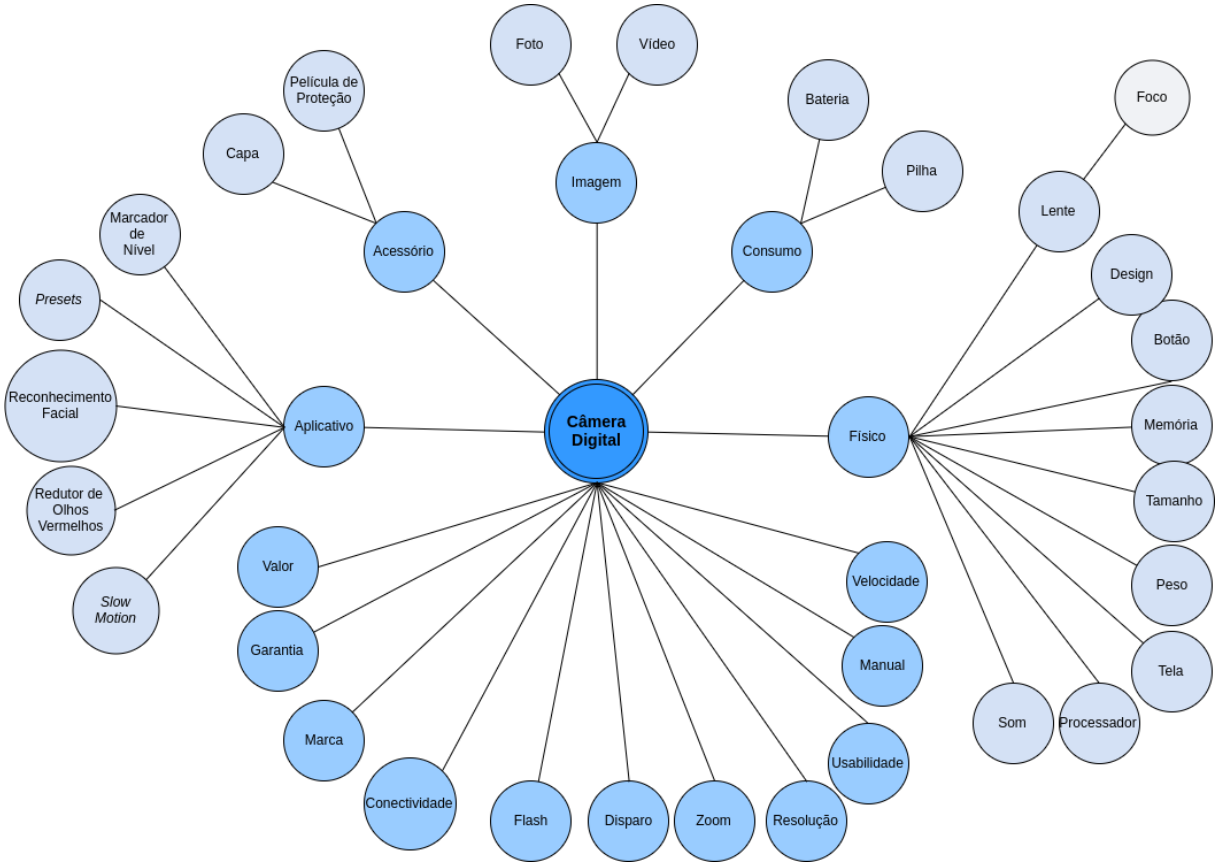

Figura 7.2: Organização hierárquica de aspectos no domínio de câmera digital.

Tabela 7.3: Grupos de aspectos no domínio de livro.

Grupo	**Aspectos** [4]
G1	1984, Bestseller, Capitães de areia, Crepúsculo, Ensaio sobre a cegueira, Fala sério, amiga!, Livrinho, Livro, O grande irmão, O outro lado da meia noite, Obra, "Sociedade do big brother".
G2	leitura, "ler".
G3	Autor, "O cara é um gênio", Escritor, Orwell, George Orwell, Saramago, Sidney, Sidney Sheldon, Stephenie Meyer, Tatá, Thalita Rebouças, Sheldon, José Saramago, Jorge Amado.
G4	Assunto, Clímax, Enredo, Questão, Tema, Trama, Situação, Ponto, Acontecimento, Essência, "Conta coisas...", "É mostrado", "Remetendo", "Retrata", "Explora", Mensagem.
G5	Personagem, Protagonista, Herói, Garoto, Mocinho, Moleque, Menino, Edward Cullen, Edward, Isabella Swanchega, Isabella Swan, Bella, Pedro Bala, Catherine, Malu, Nolle, Noelle Page, Winston, Larry.
G6	Final, Fim, Desfecho, "Termina".
G7	Escrita, Estilo de escrita, Técnica de escrita, "Escreve de forma envolvente".
G8	Romance, Caso, Estória, Ficção, Narrativa, Romancezinho, Aventura, Literatura, Tipo de história, História, Crônica, Suspense.
G9	Crítica, Crítica social, Reflexão, "Refletir", Pensamento, "Pensar".
G10	Estilo.
G11	Início, Começo, "Começar".
G12	Passagem, Página, Capítulo.
G13	Diálogo, Frase, Palavra, Linguagem, Expressão.
G14	Detalhe.
G15	Leitor.
G16	Cenário, Cena.
G17	Adaptação.
G18	Edição.
G19	Sinopse.
G20	Narrador.
G21	Tradução.

[4]Os aspectos com aspas duplas são usados para caracterizar os aspectos implícitos.

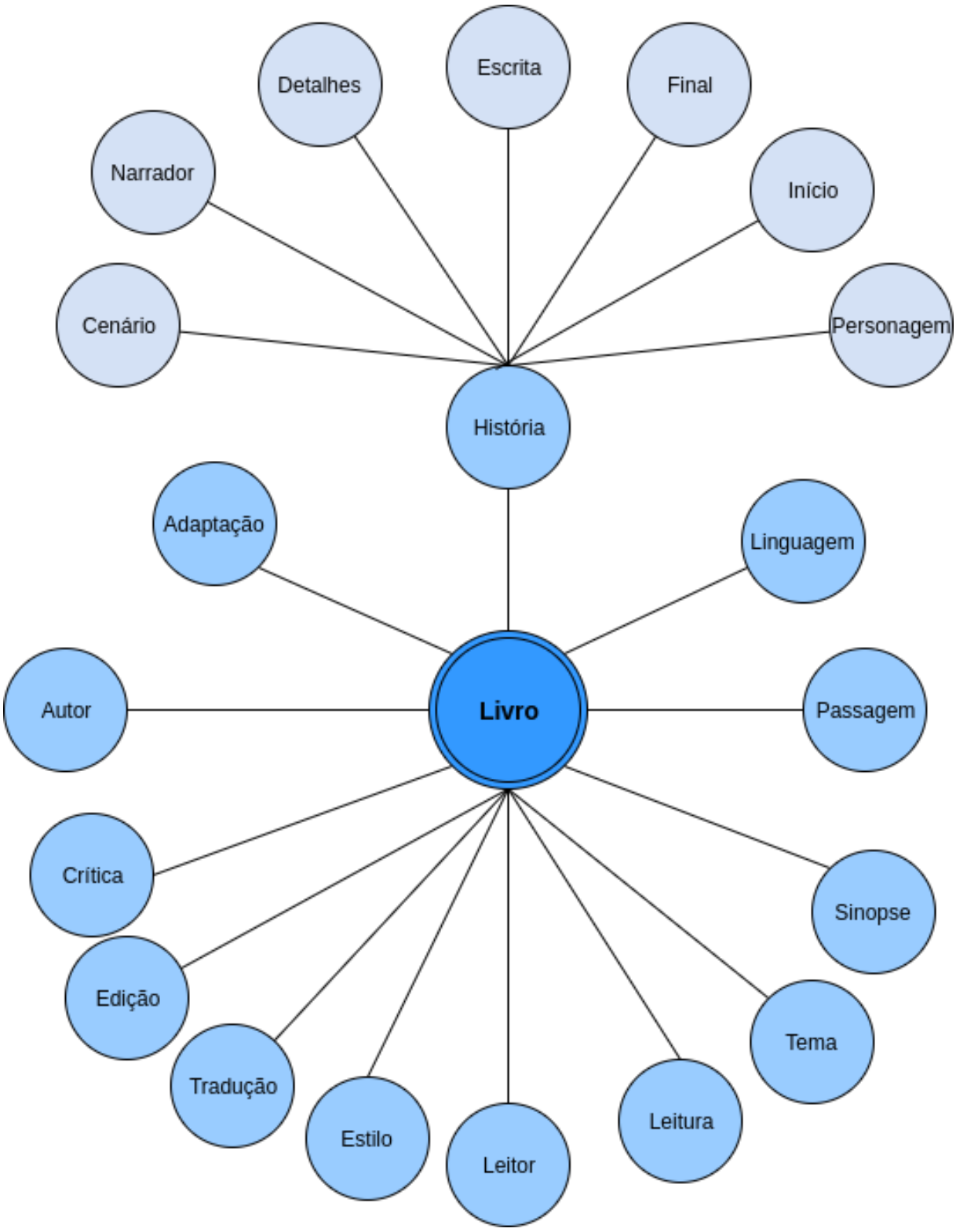

Figura 7.3: Organização hierárquica de aspectos no domínio de livro.


ICMC
USP
SÃO CARLOS