UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

ESTUDO EMPÍRICO SOBRE AGRUPAMENTO E ORGANIZAÇÃO HIERÁRQUICA DE ASPECTOS PARA MINERAÇÃO DE OPINIÃO

FRANCIELLE ALVES VARGAS
THIAGO ALEXANDRE SALGUEIRO PARDO

Nº 418

RELATÓRIOS TÉCNICOS

São Carlos – SP
Mar./2017

Estudo empírico sobre agrupamento e organização hierárquica de aspectos para mineração de opinião

Francielle Alves Vargas, Thiago Alexandre Salgueiro Pardo

[1]Núcleo Interinstitucional de Linguística Computacional
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
francielleavargas@usp.br, taspardo@icmc.usp.br

Resumo. *Este relatório consiste em um estudo empírico, como parte de um projeto de mestrado, para a tarefa de mineração de opinião. Nossa principal motivação é analisar e entender as características e os desafios do processo de identificação de grupos de aspectos de opinião e sua organização hierárquica, com o objetivo de propor soluções automáticas baseadas em conhecimento e motivadas linguisticamente. Selecionamos três domínios de trabalho - câmera digital, smartphone e livro - a fim de identificar convergências e divergências de comportamento. A análise foi realizada manualmente e servirá como referência para pesquisas futuras. Apresentamos diversos dados quantitativos e qualitativos sobre a tarefa e algumas evidências de que o comportamento linguístico varia entre os domínios e que essas variações têm fortes ligações com as especificidades de conhecimentos de domínio e com os perfis de escritor/usuário que produz o conteúdo.*

1. Introdução

Com o 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. De acordo com [Liu 2012], sistemas computacionais para esse fim possuem aplicações diversas. Por exemplo, 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. Nesta trabalho, os autores propõem 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 propõem a classificação de eventos de *cyberbullying* em redes sociais. Para o Português, em relação ao Inglês, relativamente poucos trabalhos foram desenvolvidos. No trabalho de [Freitas and Vieira 2013], os autores apresentam uma proposta de análise de opinião para o domínio de filmes usando uma ontologia para identificação de aspectos da opinião e polaridades. No âmbito da sumarização de opinião, [Condori 2014] apresenta abordagens para geração de sumários de opinião. Para identificação de polaridade de opiniões, o trabalho de [Avanço and 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 idenficação de aspectos abordados em opiniões, o trabalho de [Balage Filho and Pardo 2014] apresenta uma proposta baseada em Aprendizado de Máquina.

Prover de forma organizada conteúdo subjetivo e relevante de revisões, dentre vários domínios, é uma tarefa importante no contexto atual, pois possibilita um melhor aproveitamento dos dados subjetivos produzidos na web que trazem benefícios tanto para consumidores quanto para organizações privadas e governamentais. A área responsável pela extração, processamento e apresentação de conteúdo subjetivo em textos é a mineração de opinião ou 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. Mineração de opinião é, portanto, o campo de estudo que analisa as opiniões, sentimentos, avaliações, atitudes e emoções relacionados a entidades como produtos, serviços, organizações, indivíduos, eventos, tópicos e seus atributos [Liu 2012]. No entanto, de acordo com [Munezero et al. 2014] e [Tsytsarau and Palpanas 2012], é necessário que haja maior prudência na utilização desses termos, para que sejam aplicados corretamente em cada um de seus contextos. Segundo [Tsytsarau and Palpanas 2012], originalmente estudados em diferentes comunidades, os problemas da mineração de opinião e da análise do sentimentos possuem noções ligeiramente diferentes. A mineração de opinião é originária da comunidade de Recuperação de Informação (RI) e visa extrair e processar as opiniões dos usuários sobre produtos, filmes ou outras entidades. A análise de sentimentos, por outro lado, foi inicialmente formulada no Processamento de Línguas Naturais (PLN) com a tarefa de recuperação de sentimentos expressos em textos. Contudo, estes dois problemas são semelhantes em sua essência e recaem sob o escopo da análise de subjetividade. Neste trabalho, optamos pelo termo mineração de opinião por compreendermos que representa com maior clareza nossa proposta de trabalho.

Segundo [Liu 2012], existem níveis de granularidade de análise para a mineração de opinião. São eles: nível do documento, da sentença e do aspecto. No nível do documento, realiza-se a análise do conteúdo relevante que expressa os sentimentos de um documento como um todo, emitindo-se um *score* positivo, negativo ou neutro para cada documento analisado. No nível da sentença, o objetivo é determinar a opinião expressa em cada sentença do documento. Por exemplo, em um documento composto por X sentenças, para cada uma haverá uma classificação positiva, negativa ou neutra. No entanto, nestes dois níveis de análise, não se sabe exatamente do que o usuário gostou ou não gostou [Liu 2012]. Para solucionar esse problema, o autor argumenta que é necessário um nível mais fino de análise, a mineração de opinião baseada em aspectos, ou seja, que analisa como o usuário aborda cada aspecto do item avaliado em uma revisão. Por exemplo, ao avaliar um smartphone, o usuário pode abordar os aspectos "preço", "tamanho" e "resolução da tela". Iremos apresentar as principais tarefas da mineração de opinião baseada em aspectos no Capítulo 2.

De acordo com [Liu 2012], a mineração de opinião representa um delicioso desafio. Antes do advento da internet, consumidores, para tomarem uma decisão de compra, pediam opiniões aos amigos, familiares e organizações quando precisavam encontrar informação relevante do público em geral sobre 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. Entretanto, processar esse grande volume de conteúdo opinativo, classificá-lo, sumarizá-lo e apresentar ao usuário apenas o conteúdo mais relevante é repleto de desafios. A língua é muita rica e permite expressar subjetividade de diversas formas. Nem toda opinião é expressa diretamente e nem todo aspecto

aparece explicitamente. Por exemplo, em "A câmera é cara", o aspecto avaliado é "preço", mas ele é implícito, não estando explicitado na sentença e, portanto, sendo inferido do contexto. Outro exemplo é demonstrado na revisão da Figura 1. Nesta revisão, o usuário avalia aspectos de um smartphone. Os aspectos "resolução da tela", "câmera", "preço" e "design" foram avaliados positivamente pelo consumidor. Outros aspectos também foram avaliados nesta revisão, porém não estão explícitos. Por exemplo, quando o usuário diz "bem rápido", está avaliando o aspecto implícito "velocidade". Por fim, quando o usuário fala "fácil de usar", está avaliando o aspecto implícito "usabilidade".

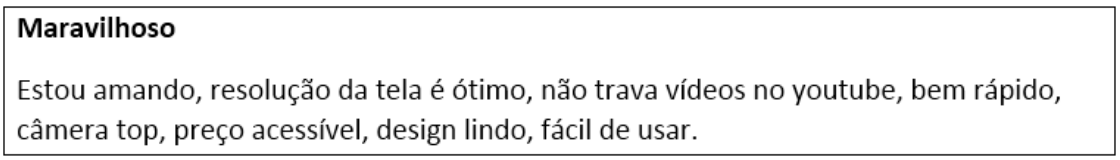

Figura 1. Revisão extraída do *Buscape.com* sobre um smartphone

Outro desafio da mineração de opinião é identificar e agrupar aspectos que se referem a uma mesma característica da entidade. Os usuários recorrentemente se referem a uma característica de um produto ou serviço utilizando termos distintos. Por exemplo, os consumidores podem usar os termos "valor" e "custo" para designar o preço de um determinado produto, ou utilizar os termos "equipamento", "câmera" ou "produto" para qualificar uma câmera fotográfica. Portanto, é útil à mineração de opinião a classificação de aspectos por grupos. A maioria dos trabalhos da literatura limita-se à utilização de léxicos de sinônimos para solucionar esse problema, entretanto esse critério não é suficiente para abarcar todos os grupos. Além de relações de sinonímia, revisões de usuários podem conter aspectos em relação de hiperonímia/hiponímia. Por exemplo, vejamos a revisão da Figura 2. Nesta revisão, o usuário utiliza os termos "câmera", "câmera digital" e "câmera canon" para designar o mesmo aspecto. O termo "câmera" é hiperônimo de "câmera digital" e de "câmera canon". Além das relações de sinonímia e hiperonímia, é possível identificar que, em alguns casos, o usuário se refere ao mesmo aspecto por uso de correferência. Por exemplo, em outros textos, além dos casos tradicionais de anáforas pronominais, os usuários utilizam os termos "modelo" e "produto" para designar um mesmo aspecto[1] ou utilizam os termos "web", "net" e "conexão" para se referirem ao aspecto "conexão de internet" de um smartphone. Se consultarmos uma base de conhecimento semântico como a Wordnet de Princeton[2], por exemplo, verifica-se que os conceitos correspondentes não são sinônimos ou mesmo hiperônimos/hipônimos diretos ou indiretos entre si.

Outro desafio significativo da mineração de opinião, de acordo com [Yu et al. 2011], é que revisões de usuários são numerosas e desorganizadas, conduzindo à dificuldade de navegação de informações e aquisição de conhecimento. Portanto, é impraticável para o usuário compreender a visão geral das opiniões sobre todos os aspectos de um produto, dado o grande número de revisões emitidas para cada produto. Para se ter

[1]É interessante notar que "modelo" poderia ser um aspeto/atributo de "produto", mas, na forma como foi usado no texto do usuário, refere-se ao próprio produto. Isso claramente também se deve à informalidade desse tipo de texto.

[2]http://wordnetweb.princeton.edu/perl/webwn

Sensacional

A três anos fiz uma pesquisa aqui no site Buscape sobre câmeras digitais e fui tirando minhas dúvidas, enfim escolhi a câmera Canon EOS Rebel T5. Depois que comprei esta câmera minhas fotos melhoraram Muito. Agora tenho fotos a nível profissional, com ela já fiz até alguns ensaios e adquiri uma renda extra. Não tenho oque reclamar deste aparelho, simplesmente sensacional.

Figura 2. Revisão extraída do *Buscape.com* sobre uma câmera digital

uma ideia do volume de revisões, o produto *Smartphone Samsung Galaxy J5* possui mais de 800 avaliações de consumidores[3]. A Figura 3 ilustra a proposta de [Yu et al. 2011] para organização hierárquica de aspectos. Nesta proposta, foram analisados 9.245 revisões sobre o produto *iPhone 3G*. Neste trabalho, o autor propõe a organização de avaliações de consumidores sobre os vários aspectos de um produto e as suas respectivas opiniões. A organização hierárquica de aspectos, de acordo com o autor, permite que um consumidor facilmente aprenda a visão geral de opiniões de outros consumidores sobre um determinado produto, serviço ou entidade. Portanto, manter a organização hierárquica do conteúdo de revisões de usuários permite uma maior estruturação desses dados para que se tornem inteligíveis tanto por máquinas quanto por humanos.

Em função dos desafios apresentados, relatamos neste documento o trabalho realizado com base em dois principais eixos: (i) o estudo empírico sobre a identificação de grupos de aspectos para mineração de opinião, além de mapear o comportamento linguístico entre domínios distintos; e (ii) a organização hierárquica dos grupos de aspectos identificados. Nossa proposta com estes dois processos é analisar e entender o fenômeno e seus desafios e criar uma base de referência para o desenvolvimento, aplicação e avaliação de metódos automáticos de identificação de grupos de aspectos e de organização hierárquica destes. Selecionamos três domínios para análise: câmera digital, smartphone e livro. A escolha dos domínios foi feita, principalmente, para permitir analisar convergências e/ou divergências entre domínios diferentes. Esse tipo de estudo é relevante para várias tarefas, especialmente para mineração de opinião.

Estruturamos este documento da seguinte forma: no Capítulo 2, apresentamos os coceitos relacionados à mineração de opinião baseada em aspectos; no Capítulo 3, realizamos a descrição dos dados; no Capítulo 4, apresentamos a metodologia aplicada para a análise; no Capítulo 5, discutimos os resultados obtidos; por fim, as considerações finas são apresentadas no Capítulo 6.

2. Mineração de opinião baseada em aspectos

Pesquisas em mineração de opinião possuem dois principais eixos: a classificação do sentimento e a mineração de opinião no nível de características ou aspectos [Bhuiyan et al. 2009]. Aspectos representam atributos/propriedades ou partes das entidades que são avaliadas pelos usuários, em textos opinativos, como em comentários em sites e blogs na web [Liu 2012]. A mineração de opinião no nível de aspectos ou baseada em aspectos, de acordo com [Liu 2012], concentra-se, geralmente, nas tarefas de (i)

[3]Extraído de Buscape.com

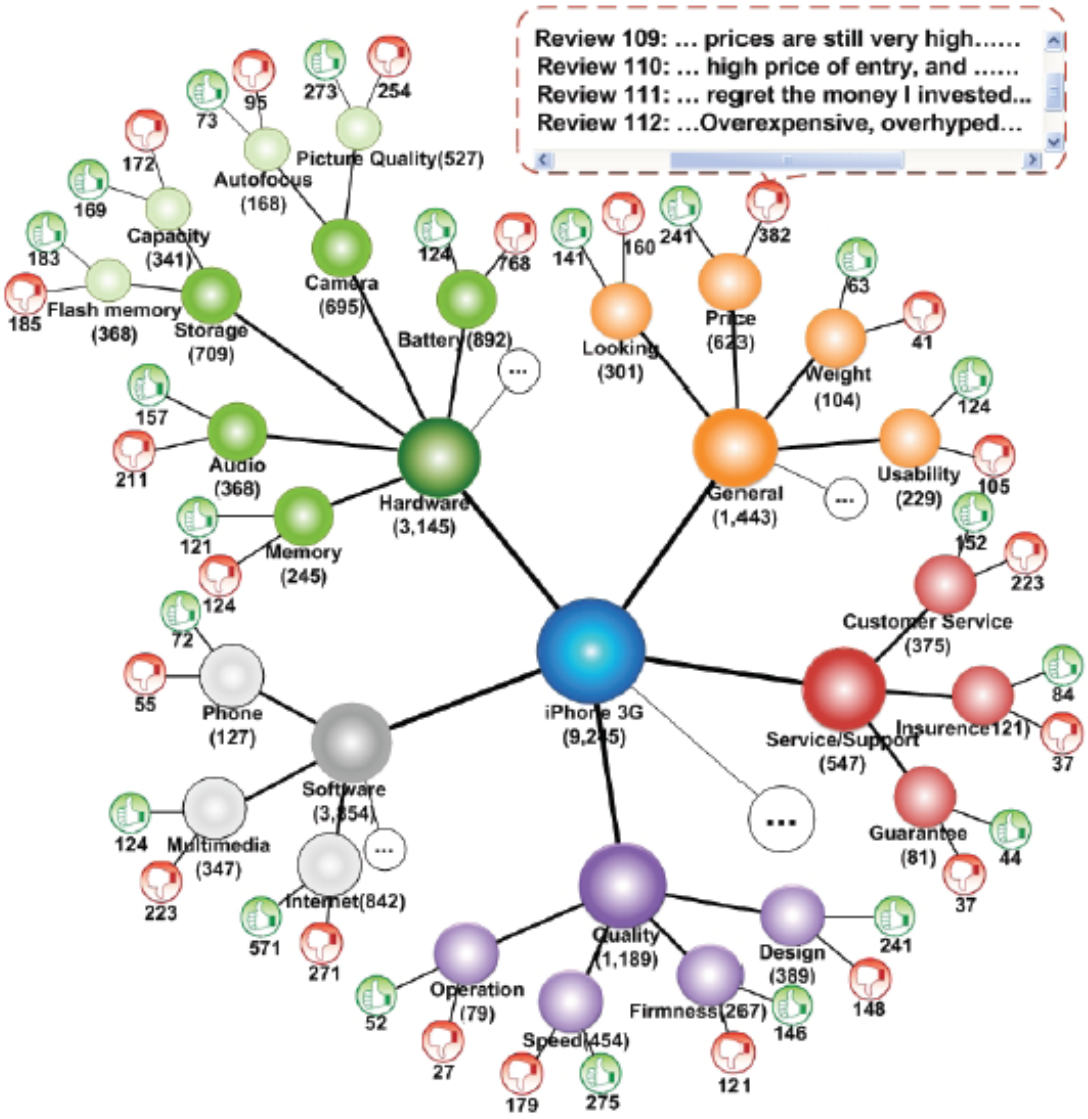

Figura 3. Organização hierárquica de aspectos de opinião do produto iPhone 3G [Yu et al. 2011]

identificação de aspectos de opinião, (ii) identificação de polaridade e (iii) exibição dessas informações sumarizadas, conforme exibido na Figura 4.

Na fase de identificação de aspectos, são extraídas características avaliadas pelos usuários sobre o alvo da opinião. Por exemplo, em "A tela do Iphone 6 é ótima", o alvo da opinião é o "Iphone" e o aspecto avaliado é "tela". Para extração de aspectos, de acordo com [Liu 2012], os principais métodos normalmente utilizados são: (i) método baseado em frequência de substantivos e sintagmas nominais; (ii) método baseado em aprendizado de máquina supervisionado; (iii) método baseado em aprendizado de máquina semissupervisionado; e (iv) método 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. Por exemplo, em [Freitas and Vieira 2013], os autores utilizam uma ontologia de domínio para extração de aspectos de opinião para o domínio

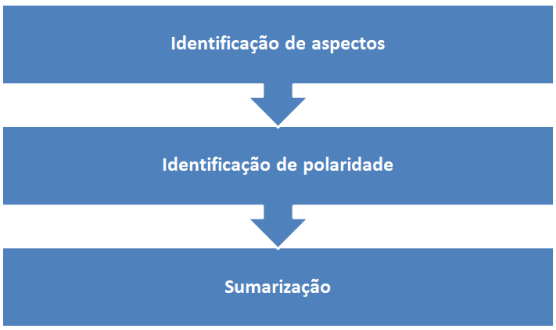

Figura 4. Principais tarefas da mineração de opinião baseada em aspectos

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 associado ao aspecto "bateria" é negativo. Portanto, a polaridade dessa revisão é negativa. Para identificação de polaridade, grande parte dos trabalhos utiliza um léxico composto por uma lista de palavras de sentimentos associadas às respectivas polaridades (veja, por exemplo, [Taboada 2016]).

Na fase de sumarização, o conteúdo mais relevante é exibido por meio de sumários, geralmente do tipo *Extrativo*, que exibe o conteúdo sumarizado através da seleção e justaposição das sentenças mais relevantes dos textos originais, ou *Abstrativo*, que exibe o conteúdo sumarizado através de técnicas de reescrita de trechos do textos originais. Veja, por exemplo, o trabalho de sumarização de opinião baseada em aspectos, voltado para o Português, em [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, o conteúdo textual pode conter informação objetiva (fatos, ações) ou informação subjetiva (percepções, opiniões, sentimentos). Além disso, textos subjetivos normalmente expressam uma visão positiva ou negativa. A direção da opinião - se positiva ou negativa - é algo conhecido como orientação semântica.

Muitas pesquisas têm utilizado córpus de revisões de usuários do domínio de filmes, livros e produtos eletrônicos [Hu and Liu 2004] por esses domínios possuírem relevância tanto para fabricantes quanto para 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 de um item de consumo, por exemplo.

3. Descrição dos dados

Neste trabalho, foram selecionadas sessenta revisões dos seguintes produtos: smartphone, câmera digital e livro. Optamos por selecionar apenas sessenta revisões para cada domínio em função de tratar-se de um estudo empírico, realizado manualmente e de cunho qualitativo. Nossa hipotése é que, para cada domínio, haja comportamentos linguísticos e fenômenos distintos. Nosso objetivo é caracterizar cada domínio a fim de futuramente

propor bons metódos automáticos para as tarefas de mineração de opinião. Uma síntese dos dados é exibida na Tabela 1.

Tabela 1. Visão geral dos dados

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

A escolha por esses três domínios foi motivada pela recorrência da utilização desses itens na literatura. De acordo com [Zhao and Li 2009], a maioria das obras existentes no âmbito da mineração de opinião baseia-se em revisões de produtos e filmes. Com relação ao número de *types*, observamos certo balanceamento entre os três domínios. Entretanto, em relação ao número de *tokens*, no domínio de livro, observamos um salto em relação aos domínios de smartphone e câmera. Para o domíno de livro, constatamos que, diferentemente dos outros domínios, as revisões são muito maiores e houve maior ocorrência de conteúdo irrelevante, com comentários tangenciais ao conteúdo abordado. Além disso, identificamos que, no domínio de livro, os usuários possuíam perfis mais variados e distintos. Esses fenômenos não ocorreram de forma significativa nos domínios de smartphone e câmera digital. Essas questões serão discutidas no Capítulo 5.

Para os domínios selecionados, utilizamos dados do córpus ReLi [Freitas et al. 2012] e do córpus Buscapé [Hartmann et al. 2014]. O córpus ReLi é composto por 1.601 resenhas de 13 livros de 7 autores distintos. São 260 mil palavras e 12 mil sentenças. Para cada livro, há cerca de 200 resenhas. Além disso, este córpus foi anotado em relação à opinião, ao objeto da opinião e a sua polaridade. Optamos pela seleção das resenhas do córpus ReLi de modo aleatório, contemplando pelo menos uma revisão dos treze livros que compõem o córpus. As Figuras 5 e 6 ilustram duas revisões retiradas do córpus ReLi. Na revisão da Figura 5, o usuário avalia o livro "Fala Sério, Amiga!", escrito por Thalita Rebouças, cujo público alvo é formado majoritariamente por adolescentes de classe média e do sexo feminino. Na revisão mostrada na Figura 6, o usuário avalia o livro "1984", escrito por George Orwell. Este livro, diferentemente do anterior, possui conceitos complexos de ciência política e requer um leitor mais politizado, sendo voltado para um público com maior escolaridade. É interessante ressaltar as divergências entre os perfis dos usuários de um mesmo domínio e a explícita diferença entre os vocabulários empregados em cada uma das revisões. Na revisão da Figura 6, há marcas explicítias de uma linguagem mais rebuscada e de alta adequação à variante padrão da língua, além de maior clareza na explanação de ideias. Na Figura 5, é clara a baixa adequação à variante padrão da língua, além da presença de marcas de oralidade, baixa informatividade e vagueza.

O córpus Buscapé é composto por 85.910 revisões de usuários sobre produtos (TVs, celulares, smartphones, câmeras digitais, perfumes, jogos, condicionadores de ar, notebooks, tablets, etc), coletados em setembro de 2013. No total, há 4.097.905 *tokens* e 68.633 *types*. Como exemplos, as Figuras 7 e 8 exibem revisões extraídas deste córpus. Na revisão da Figura 7, o usuário avalia um smartphone e, na Figura 8, trata-se de uma revisão sobre uma câmera digital. As revisões do córpus Buscapé, diferentemente do córpus

Fala sério, Amiga!

Gente, esse livro é ótimooo vale a pena ler. É muito gostoso ler que dá vontade de ler mais de mil vezes. Eu indico pra todo mundo e posso dizer que amei eleee. Afinal, como não amar os livros de Thalita Rebouças? Ela é uma das minhas escritoras favoritas e com todo esse jeito único dela escrever, eu sempre amei os livros delaaa.

Figura 5. Revisão do córpus ReLi sobre o livro *Fala Sério, Amiga!*

Ainda não acabei de ler. Estou nos primeiros capítulos. Mas cada página do livro é um convite a imaginação e a reflexão. Incrivelmente atual, o livro foi escrito em 1949 (salvo engano) remetendo a época o que seria o futuro (1984), mas um futuro sufocado e alienado em que cada passo dos indivíduos é vigiado, seus pensamentos rigorosamente monitorados por um "Estado" altamente burocrático e totalitarista. A forma como a personagem principal descreve em primeira pessoa o mundo em que vive e a sua percepção sobre ele é fantástica. Parece se basear em uma descrição caricaturizada do "socialismo real", embora não se traga importantes reflexões sobre diversas questões enfrentadas nesta era atual em que o argumento de segurança produz contantes incursões sobre o território reservado a liberdade e a a privacidade dos indivíduos. Enfim, qualquer análise seria reducionista, mas a leitura é altamente recomendável!

Figura 6. Revisão do córpus ReLi sobre o livro *1984*

ReLi, são parcialmente estruturadas. Nestas, o conteúdo de cada revisão é topicalizado em *o que gostei* e *o que não gostei*, após uma parte de comentários gerais.

O aparelho é excelente! Vale a pena. Vc tem um ótimo smart phone com um preço acessível.

O que gostei: Bateria tem ótima durabilidade, fácil de usar.. Wi Fi funciona direitinho... muito rápido... ótimo para escrever mensagens de texto.. Teclado qwert facilita muito... tem muitas funções!

O que não gostei: Acho que só pq não tem gps, mesmo assim dá pra usar o guia de mapas que já vem no aparelho.

Figura 7. Revisão do córpus Buscapé sobre um smartphone

4. Metodologia

Nesta seção, apresentaremos a metodologia utilizada para realização deste trabalho. A descrição da proposta metodológica foi dividida em duas subseções: (i) agrupamento de aspectos e (ii) organização hierárquica desses grupos.

4.1. Agrupamento de aspectos

Para cada revisão, primeiramente foram identificados manualmente todos os aspectos, inclusive aspectos implícitos. Para a identificação e quantificação de aspectos implícitos,

O preço é compatível com a qualidade imbatível de uma Sony. Voltaria a comprar com certeza. Eu recomendo.

O que gostei: Altíssima resolução. Bateria com uma duração excelente. Dá para tirar centenas de fotos, filmar a vontade que ela aguenta. E na hora de carregar ela é excelente.

O que não gostei: Achei aquele botão redondo que fica na parte traseira da máquina, onde você busca as funções da máquina, de manuseio um pouco lento. Isso se dá o fato dele ser bem rente ao corpo da máquina, sendo que foi deixado apenas um rebaixo na lateral para facilitar o uso.

Figura 8. Revisão do córpus Buscapé sobre uma câmera digital

anotamos o termo que fazia referência ao aspecto implícito e o chamamos de *termo pista*. Estes aspectos foram anotados e diferenciados utilizando aspas duplas. Desta forma, foi possível mensurar a ocorrência de aspectos implícitos em cada domínio. A identificação dos grupos de aspectos foi realizada revisão por revisão. A progressão na identificação dos grupos ocorria a cada revisão analisada. Anotou-se e quantificou-se cada novo grupo de aspecto na ordem em que surgia. Esse processo se repetiu até a finalização das sessenta revisões para cada um dos três domínios. Uma tabela (em Microsoft Excel) foi progressivamente construída a fim de apoiar esse processo, conforme exibida na Figura 9. A primeira coluna é composta pelas sessenta revisões do domínio. No exemplo, as revisões são do domínio de smartphone. Na coluna seguinte, é identificada a quantidade de aspectos implícitos anotados para aquela revisão e, logo depois, os aspectos explícitos. Na sequência, nós temos a quatidade de grupos de aspectos novos identificados e, em seguida, o número identificador da revisão. Esse número diz respeito à ordem em que as revisões foram organizadas (por exemplo, revisão 1, 2, 3 até 60). Em seguida, mostra-se o resultado acumulado da identificação dos grupos de aspectos. As colunas cujo cabeçalho é azul correspondem aos grupos identificados na ordem de ocorrência. O grupo *G1* indica o primeiro grupo de aspectos identificado. Na linha abaixo de *G1*, encontra-se a etiqueta do grupo, neste caso, "smartphone". Na primeira revisão, foram identificados nove grupos de aspectos, tais como "smartphone"[4], "conexão com internet", "usabilidade", "velocidade", "funções", "valor", "bateria", "teclado" e "mapas", conforme exibido na Figura 9. Nessa tarefa, optamos por agrupar também os atributos dos aspectos no grupo do respectivo aspecto. Aspectos representam propriedades ou partes das entidades que são avaliadas pelos usuários, em textos opinativos, como em comentários em sites e blogs na Internet [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, tal como "qualidade da imagem", foi incorporado ao grupo de aspectos "imagem". Optamos também por selecionar o lexema (forma canônica) relativo ao aspecto. Por exemplo, os usuários, ao se referirem ao aspecto "foto", utilizam os termos "foto" e "fotos". Portanto, lematizamos os itens lexicais e exibimos nos grupos apenas o lexema correspondente. Optamos também por identificar

[4]É interessante notar que, em mineração de opinião com base em aspectos, é comum tratar a entidade avaliada como um aspecto também.

grupos unitários. Por exemplo, os grupos “usb”, “rádio”, “tv” e “manual”, entre outros, são grupos unitários, ou seja, há apenas um item nesses grupos. Essas decisões foram tomadas em função também do nosso segundo objetivo de organizar os grupos de aspectos classificados de forma hierárquica. Desta forma, esse tipo de agrupamento facilitará a organização automática de aspectos e seus atributos na árvore.

Para cada um dos três domínios analisados, foi construída uma tabela similar a exibida pela Figura 9. Anotamos os aspectos implícitos e explícitos e identificamos os grupos de aspectos na medida em que ocorriam. Nós chamamos de *curva de aprendizagem* a progressão da identificação de novos grupos de aspectos para o domínio. As curvas de aprendizagem obtidas são demonstradas nas Figuras 15, 16 e 17. Uma descrição detalhada das curvas de aprendizagem será realizada na Seção 5.1.6 Nosso objetivo com este processo é verificar o padrão de comportamento das curvas de aprendizagem nos domínios analisados, assim como identificar um ponto médio de estabilização para a identificação de novos grupos, de forma que se tenha uma cobertura significativa de grupos de aspectos mais representativos de um domínio.

						G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Revisões	Aspectos Implícito	Aspectos Explícitos	Grupos Novos	Id	Total de grupos	Smartphone	Conexão com internet	Usabilidade	Velocidade	Funções	Valor	Bateria	Teclado	Mapas	Aplicativos
O **aparelho** é excelente! Vale a pena. Vc tem um ótimo **smart phone** com um **preço** acessível. O que gostei: **Bateria** tem ótima durabilidade, fácil de usar.. **Wi Fi** funciona direitinho... muito rápido... ótimo para escrever mensagens de texto.. **Teclado qwert** facilita muito... tem muitas **funções**! O que não gostei: Acho que só pq não tem **gps**, mesmo assim dá pra usar o guia de mapas que já vem no **aparelho**.	2	7	9	1	9	APARELHO , SMARTPHONE	WIFI	"FÁCIL DE USAR"	"RÁPIDO"	FUNÇÕES	PREÇO	BATERIA	TECLADO	GPS	
Com menos de 12 dias de uso ele travou o **touch**, os **aplicativos** caem com frequência na hora do uso, se ele estiver em algum aplicativo com Facebock, a **bateria** dele dura muito pouco tempo, a **internet** dele se não for **wi-fi** é muito ruim, **trava** na hora que estou em uma ligação, parou de funcionar o **touch** no menu superior em 10 dias de uso... ARREPENDIMENTO TOTAAALLLL... Enviei para assistência da motorola e depois de 15 dias eles me devolveram sem conserto pq na nota fiscal não tem o número IMEI, solcitei o número do IMEI junto a empresa que me vendeu o produto e se vão mais 30 dias para eles enviarem a nf com os tais números. Depois mais 30 para a **assistência** me devolver consertado. Só prejuízo, estou super insatisfeita com a compra. O que gostei: Este celular é muito **fácil de mexer**, o touch dele é bem sensível. O que não gostei: Bateria que não dura, touch travado, **aplicativos** e internet que caem a todo momento	2	4	2	2	11		INTERNET, WIFI	"FÁCIL DE MEXER"	"TRAVA"			BATERIA			Aplicativos

Figura 9. Tabela de suporte para identificação de grupos de aspectos

4.2. Organização hierárquica de grupos de aspectos

De acordo com [Biemann 2005], organizações hierárquicas facilitam a compreensão do texto e o processamento automático de recursos textuais. Ainda, segundo [Russell and Norvig 2003], o problema da compreensão da língua humana exige a compreensão do assunto e do contexto, e não apenas a compreensão da estrutura da sentença. Portanto, mesmo que compreendamos a estrutura sentencial de uma revisão de usuário, muitas vezes não é suficiente para extraírmos todo o conhecimento subjacente a ela, sendo necessário lidar com a semântica das sentenças. O mapeamento semântico de um domínio geralmente é feito através da identificação de conceitos e a descoberta de relações entre esses conceitos. Por exemplo, em um conjunto de revisões de usuários, os aspectos de opinião representariam os conceitos e a organização desses aspectos indicaria suas relações.

Vejamos a hierarquia exibida na Figura 10, extraída de [Aciar et al. 2006]. Neste trabalho, os autores utilizam uma hierarquia de aspectos de opinião com a proposta de melhorar a performance de sistemas de recomendação. Essa hierarquia foi implementada utilizando revisões de usuários sobre o produto *Canon PowerShot D630*. Este é o domínio e representa o *nó* raiz da árvore. Os aspectos avaliados pelo usuário foram: "bateria", "tamanho", "lentes", "lcd", "velocidade", "flash", "modo" e "qualidade da imagem". Estes aspectos têm relação de *parte-todo* com a entidade câmera e representam os *nós* filhos da árvore. Nesta hierarquia, os autores também associaram aos aspectos seus sentimentos (expressos pelos usuários). Por exemplo, o aspecto "bateria" foi avaliado negativamente pelo usuário e o aspecto "tamanho" foi avaliado positivamente.

Nesse trabalho, nosso objetivo converge com a hierarquia dos autores [Aciar et al. 2006] em relação à construção da hierarquia a partir dos grupos de aspectos identificados. Entretanto, não pretendemos associar o sentimento aos aspectos na hierarquia (dado que isso foge ao escopo desta pesquisa). Com esse fim, organizamos manualmente, com base nas revisões e em introspecção, os grupos identificados para os domínios de câmera digital, smartphone e livro, e comparamos nosso resultado com outras hierarquias disponíveis na área. As análises serão descritas no Capítulo 5.

5. Resultados

Nesta seção, apresentaremos os principais resultados obtidos neste trabalho. Primeiramente, abordamos os resultados do processo de identificação dos grupos de aspectos e suas diversas questões. Em seguida, abordamos a organização hierárquica destes grupos.

5.1. Agrupamento de aspectos

Foram identificados 48 grupos de aspectos para o domínio de smartphone, 37 para o domínio de câmera digital e 24 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 identificados por usuários, que se envolvem mais com as características dos produtos e são "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,

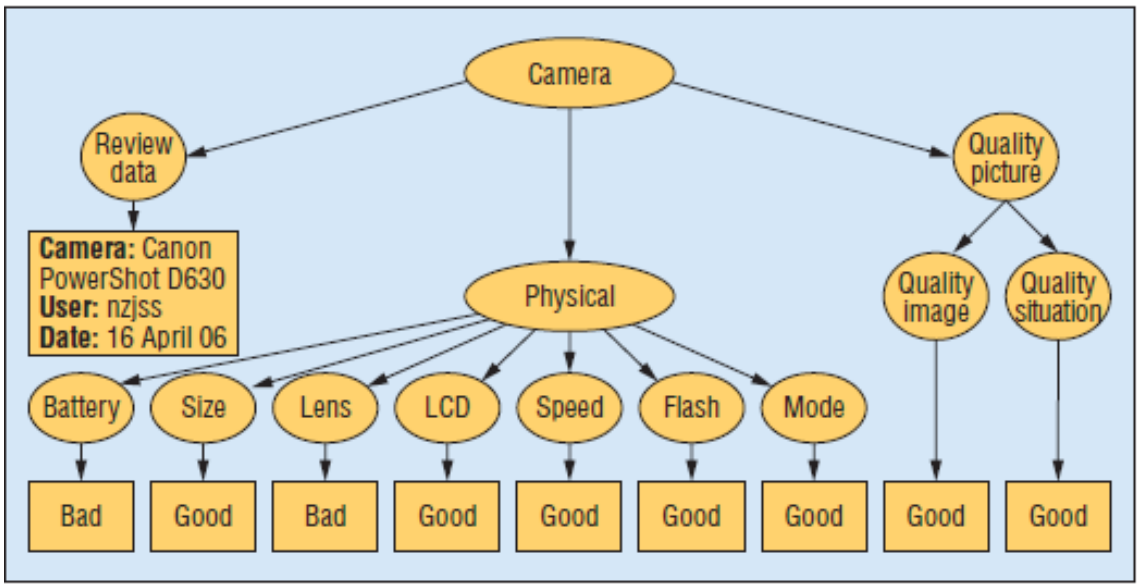

Figura 10. Organização hierárquica de aspectos para o domínio de câmera [Aciar et al. 2006]

portanto, conseguem avaliar um número limitado de aspectos do produto, geralmente *aspectos prototípicos* do objeto ou aspectos superficiais do produto. Esses dados podem ser visualizados na Tabela 2.

Tabela 2. Classificação geral

	smartphone	**câmera**	**livro**	**média**
número total de aspectos	459	342	323	**374,66**
aspectos únicos (sem repetição)	180	132	103	**138,33**
aspectos explícitos	392	289	298	**326,33**
aspectos implícitos	67	53	25	**48,33**
grupos de aspectos	48	37	24	**36,33**

O número total de aspectos por domínio e o número médio de aspectos por revisão parece ser evidência para a relação entre perfil de usuário e informatividade dos textos opinativos. Revisões de usuários "mais" especialistas possuem maior grau de informatividade[5], ou seja, esses usuários possuem maior conhecimento sobre o domínio que os possibilita avaliar um número maior de aspectos da entidade.

5.1.1. 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 [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. Essas

[5]De acordo com [Koch 2009], a informatividade de um texto está associado a sua capacidade de apresentar informações novas e inesperadas.

formas de realização 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 específicas, 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 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 representam unidades estruturais que combinam diversas proposições organizadas e constituem o produto da (re)organização dos conhecimentos disponíveis na memória e dividem-se em (ver Tabela 3): narrativa; descritiva; explicativa; argumentativa; dialogal; injuntiva. Ainda segundo o autor, os tipos de discurso apresentam um conjunto de *fases* que definem as *peculiaridades* das formações textuais, orientadas pela ordem do "narrar" ou do "expor".

Tabela 3: Tipos de discurso [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 previsto na ação discursiva e cria uma tensão; (iii) fase de resolução: introdução de acontecimentos que amenizam a tensão; (iv) fase de situação final: explicitação do novo equilíbrio obtido por essa resolução.
Descritivo	Composição por fases que não se organizam em uma ordem linear obrigatória, mas que se combinam e se encaixam em uma 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 conclusão 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 porquê 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	Realização concreta 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: coconstrução do conteúdo temático da interação (relação de interdependência dos tópicos e subtópicos conversacionais); fase de encerramento: exposição, também de caráter fático, na qual se põe fim à interação. 2º nível: fase dialogal ou de troca: caracterização de cada um das fases gerais da interação, nas quais ocorrem diálogos entre os interactantes; 3º nível: fase de intervenção: decomposição da interação em atos discursivos, ou seja, enunciados que realizam um ato de fala determinado (pedido, afirmação, injunção, etc)
Injuntiva	Orientação que visa a um fazer agir direcionado a um destinatário em uma determinada direção.	1 – fase descritiva: na qual há a exposição de elementos, conforme o objetivo a que se destina o texto; 2 – fase de procedimentos: também é uma etapa descritiva, porém apresenta um detalhamento da ação a ser realizada. Como o objetivo desta seqüência é fazer agir, destacam como condições para sua constituição: 1 – uso de formas verbais no infinitivo ou no imperativo; 2 – ausência de estruturação espacial ou hierárquica.

Interessa-nos, neste trabalho, o modelo de discurso descritivo, em que, segundo o autor, a pecularidade é composta 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 especificidade revisões de usuários, em que os aspectos avaliados podem ter relação hierárquica do tipo *parte-todo* em relação ao alvo da opinião. As fases do discurso descritivo, segundo o autor, são, inicialmente, a (i) fase de ancoragem, em que se apresenta o tema-título, e a (ii) fase de espectualização, que representa a fase de enumeração de aspectos ligados ao tema-título. Essas duas fases dialogam com o modelo textual de revisões de usuários, em que o usuário apresenta o *alvo/entidade* que

ele está avaliando ou *tema-título* e, em seguida, discorre sobre os aspectos desse alvo. Na fase (iii), 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. Além disso, essa descrição pode ser retratada apoiando-se sobre dois eixos: o objetivo e o subjetivo. Na descrição objetiva, o foco principal é relatar as características do objeto de maneira precisa. A subjetiva perfaz-se de uma linguagem mais pessoal, na qual são permitidas opiniões, expressões de sentimentos e emoções e o emprego de construções livres que revelem a identidade e individualidade do leitor.

Como já dito, a tarefa da mineração de opinião preocupa-se principalmente com a extração de conteúdo subjetivo em textos. Na análise dos três domínios diferentes de revisões de usuários, constatamos que é possível encontrar tanto conteúdo descritivo-objetivo quanto conteúdo descritivo-subjetivo, e que o grau desses conteúdos tem fortes ligações com o domínio. 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, do domínio de livro, dividimos a identificação dos aspectos entre as propriedades da entidade descritas na revisão de usuário sem avaliação associada a elas e aquelas avaliadas pelo usuário (geralmente sendo positivas, negativas ou neutras), respectivamente. Contabilizamos todos os casos e verificamos quantos deles possuíam associação com alguma opinião/sentimento e quantos não estavam associados a nenhuma opinião/sentimento. Vejamos os seguintes exemplos exibidos nas Figuras 11, 12 e 13.

A **trama** se passa em um futuro fictício – 1984 – em que o mundo é dominado por três grandes nações socialistas. Nelas, 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 eventos geram.

Figura 11. Revisão sobre o livro 1984 de George Orwell extraída do córpus Reli [Freitas et al. 2012]

Só consegui pensar em uma única coisa durante toda a **leitura** do **livro**: o quanto o Estado descrito na **obra** se assemelha à religião nos dias atuais.

Figura 12. Outra revisão sobre o livro 1984 de George Orwell extraída do córpus Reli [Freitas et al. 2012]

Na revisão da Figura 11, alguns dos aspectos são "trama/romance", "cenário" e "autor". Veja que, nesta revisão, apesar do usuário citar esses aspectos da entidade *livro*, não há nenhuma avaliação associada aos aspectos, tratando-se, portanto, de uma descrição

Um **livro** *muito bom* que retrata a cruel realidade dos garotos de rua da Bahia da década de 30. Uma realidade que não deve ser muito diferente hoje em dia. Ainda assim não me *apaixonei* muito pelo **livro**. Sou um fervoroso defensor do sexo feminino e desprezo qualquer coisa que relacione violência contra a mulher e o **livro** mostra um pouco de isso. Não que eu *quisesse* que o **autor** criasse heróis romancistas em as figuras de os garotos mas em cenas como quando **Pedro-Bala** estrupa uma menina que passa em a praia à noite ou quano dos meninos tentam estrupar **Dora** os ***protagonistas*** *despencam* no meu conceito e eu não sinto mais *pena* ou *afeição* por eles na **história**. Mas de um ponto de vista *positivista* e *menos sentimental* é um **livro** muito bom.

Figura 13. Mais uma revisão sobre o livro 1984 de George Orwell extraída do córpus Reli [Freitas et al. 2012]

objetiva. Os aspectos somente estão presentes na revisão para compor a descrição das características da entidade. Além 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 12, ocorre o mesmo fenômeno. Os aspectos são "livro/obra" e "leitura". Veja que também não há nenhum tipo de avaliação associada aos aspectos, tratando-se também apenas de uma descrição das características do objeto. O fenômeno muda na revisão da Figura 13. Os aspectos desta revisão são: "livro", "autor", "protagonista/personagens" e "história". Nesta revisão, há uma avaliação associada a cada um destes aspectos, o que implica que o usuário descreveu sua experiência com o produto, avaliando subjetivamente as partes e propriedades deste produto.

Após a análise e identificação dos aspectos de acordo com o conteúdo descritivo-objetivo e descritivo-subjetivo, obtivemos os resultados demonstrados na Tabela 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 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 as especificidades do domínio, por exemplo, incorre no risco de classificar aspectos que não foram avaliados pelo usuário e retornará 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 gostei da história*, sem necessariamente apresentar adjetivos.

Tabela 4. Panorama de conteúdo descritivo objetivo e subjetivo no domínio de livro

Domínio	Conteúdo objetivo	Conteúdo subjetivo
Livro	47,98%	52,01%

5.1.2. 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 recorrentemente utilizam os termos "recursos" e "funções", ora para falar de todos os aspectos do smarthpone, ora para designar algum aspecto, função, recurso ou aplicação específica, como "tv", "rádio", etc. Para o domínio de livro, os usuários ora utilizam o termo "situações" referindo-se a passagens e/ou acontecimentos do livro, ora referindo-se ao cenário 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 "leitura" também é utilizado de forma ambígua. Em alguns casos, é utilizado para se referir ao tipo de leitura, por exemplo, em *é uma leitura pesada* ou *a leitura do livro é instigante*, e também é usado para se referir ao livro, por exemplo, em *recomendo a leitura do livro*. Além disso, aspectos vagos ou genérios são usados de forma intercambiável 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 itensificado quando se considera a informalidade desse tipo de texto produzido por usuários.

5.1.3. Especificidades de 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 se trata 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 também é usada pelos usuários de forma intercambiável com o termo "megapixels". Os usuários de 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 "quadriband". Essa propriedade diz respeito ao tipo de sinal de comunicação do aparelho, ou seja, o usuário está avaliando o aspecto "sinal".

5.1.4. Aspectos implícitos

Também tivemos o objetivo de mensurar a ocorrência de aspectos implícitos nos domínios de smartphone, câmera digital e livro. Um panorama dos aspectos implícitos identificados nos três domínios encontra-se na Tabela 5. Um dado interessante é a proximidade de comportamento entre os domínios de smartphone e câmera digital e o distanciamento destes dois domínios em relação ao domínio de livros. A quantidade de aspectos implícitos varia bastante. Nos domínios de câmera digital e smartphone, houve um salto no número de aspectos e grupos de aspectos identificados em relação ao domínio de livro (ver Tabela 2).

Tabela 5. Panorama sobre a classificação de aspectos implícitos

Aspectos por revisão	Smartphone	Câmera	Livro
Número total de aspectos implícitos	67	53	25
Número máximo de aspectos implícitos por revisão	4	5	3
Número máximo de aspectos implícitos por grupo	8	10	9
Número médio de aspectos implícitos por revisão	1,11	0,88	0,41
Número médio de aspectos implícitos por grupo	0,89	0,78	0,66

5.1.5. Aspectos fora do domínio

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

5.1.6. Relações entre aspectos

Os grupos obtidos com a tarefa de identificação de grupos de aspectos encontram-se nos Apêndices 1, 2 e 3. As relações entre aspectos identificadas nos três domínios foram de *hiperonímia/hiponímia, meronímia/holonímia, sinonímia (incluindo identidade, ou seja, as mesmas palavras ocorrem mais de uma vez - mas aparecem uma vez só no grupo), construção deverbal e correferência*. É interessante observar que, para o domínio de livro, houve 46,6% de relações de hiperonímia/hiponímia entre os aspectos nos grupos identificados; para o domínio de smartphone, obtivemos 45%; e, no domínio de câmera digital, também obtivemos um número considerável de relações de hiperonímia/hiponímia, sendo de 37,12%. Esse resultado é interessante, pois encoraja a utilização de léxicos de hiperônimos e hipônimos (como wordnets) para a tarefa de agrupamento de aspectos. O panorama geral, contendo o percentual de relações entre aspectos nos grupos identificados para os três domínios, é exibido na Tabela 6.

Tabela 6. Panorama sobre as relações entre aspectos dos grupos identificados

Relação	Smartphone	Câmera	Livro
Hiperonímia/hiponímia	45,00%	37,12%	46,60%
Sinonímia	23,90%	28,90%	18,30%
Meronímia/holonímia	19,50%	26,00%	6,10%
Construção deverbal	5,55%	6,81%	9,70%
Correferência	6,66%	8,33%	0,00%

Um exemplo de relação de hiperínima/hiponímia ocorre entre os aspectos "aparelho" e "produto"; para sinonímia, pode-se citar "preço" e "custo"; para metonímia/holonímia, "tecla" e "teclado"; para construção deverbal, "refletir" (que é um *termo pista* de um aspecto implícito) e "reflexão"; para correferência, "fabricante" e "marca".

5.1.7. Grupos de aspectos mais avaliados

Observamos também que, para cada domínio, há grupos de aspectos que são mais avaliados pelos usuários. Esses aspectos formam, normalmente, os *grupos de aspectos prototípicos* do domínio. Por exemplo, para o domínio de smartphone, os grupos de aspectos mais avaliados foram: "smartphone", "usabilidade", "design", "valor", "bateria", "marca", "câmera", "som", "tela" e "internet'. No domínio de câmera, os grupos mais avaliados foram: "câmera", "usabilidade", "foto", "valor", "tamanho", "imagem", "resolução", "bateria", "design" e "memória". Por fim, no domínio de livro, os grupos de aspectos mais avaliados foram: "livro", "história","autor", "personagem", "leitura", "escrita", "estilo", "passagem", "tema" e "final". Os gráficos contendo o número de avaliações para os grupos de aspectos nos três domínios analisados encontram-se no Apêndices 4. Um exemplo para o domínio de livro é exibido na Figura 14, em que as barras na cor preta indicam os aspectos mais avaliados.

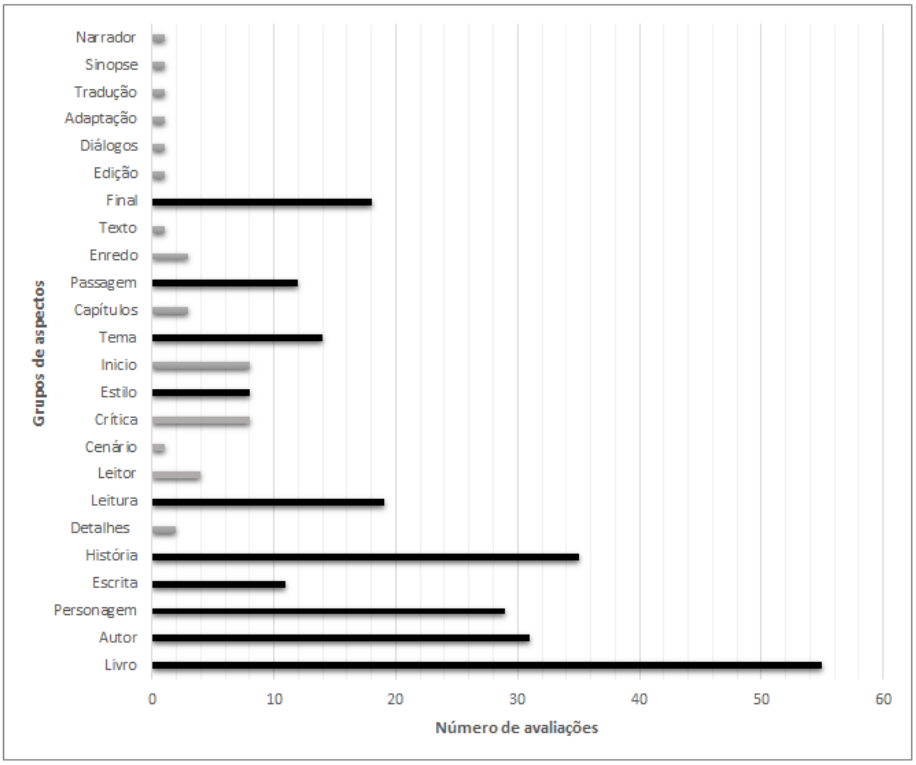

Figura 14. Número de avaliações para os grupos de aspectos do domínio de livro

5.1.8. Curvas de aprendizagem

Com o objetivo de observar o comportamento do agrupamento de aspectos de opinião e identificar o ponto de estabilização para identificação de novos grupos, descrevemos o que chamamos de *curvas de aprendizagem*, resultantes do processo de identificação de novos grupos de aspectos para os dominios de smartphone, câmera digital e livro, exibidas nas Figuras 15, 16 e 17, respectivamente. O eixo X das curvas de aprendizagem representa a quantidade de revisões avaliadas e o eixo Y a quantidade de grupos de as-

pectos identificados. Por exemplo, após análise da revisão número 1 da Figura 15 (no eixo *X*), houve a identificação de 8 grupos de aspectos (como mostra o eixo *Y*). Após a análise das dez primeiras revisões, houve a identificação de 33 grupos de aspectos, e assim sucessivamente.

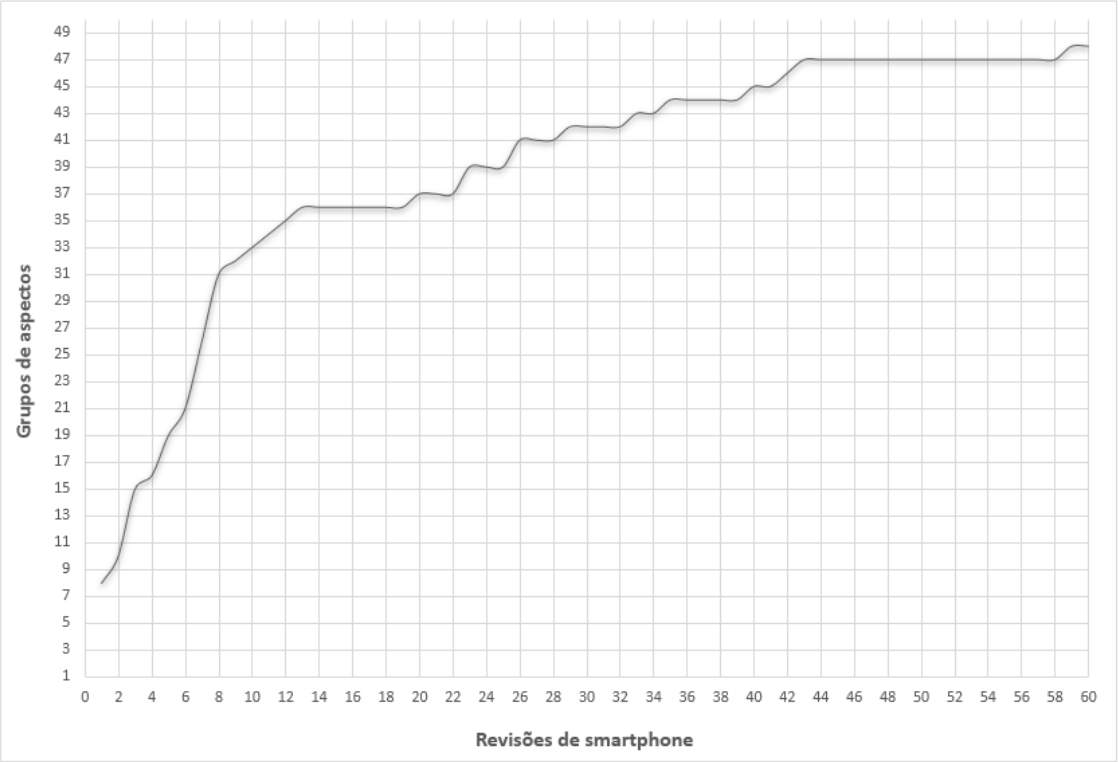

Figura 15. Curva de aprendizagem do domínio de smartphone

Chamamos de *ponto de estabilização* o ponto no qual a quantidade de revisões analisadas gerou a quantidade satisfatória de grupos de aspectos para se ter uma boa cobertura sêmantica de um domínio. Para os domínios analisados, houve convergência do ponto de estabilização médio para identificação de novos grupos. Nos três domínios, o ponto convergiu entre 37-43 revisões. O ponto médio de estabilização para o domínio de smartphone foi de 43 revisões; para o domínio de câmera digital, o ponto de estabilização foi de 35 revisões; e, para o domínio de livro, 41 revisões. Nesses pontos, percebe-se que se tem um platô na curva, não havendo identificação de novos grupos na região.

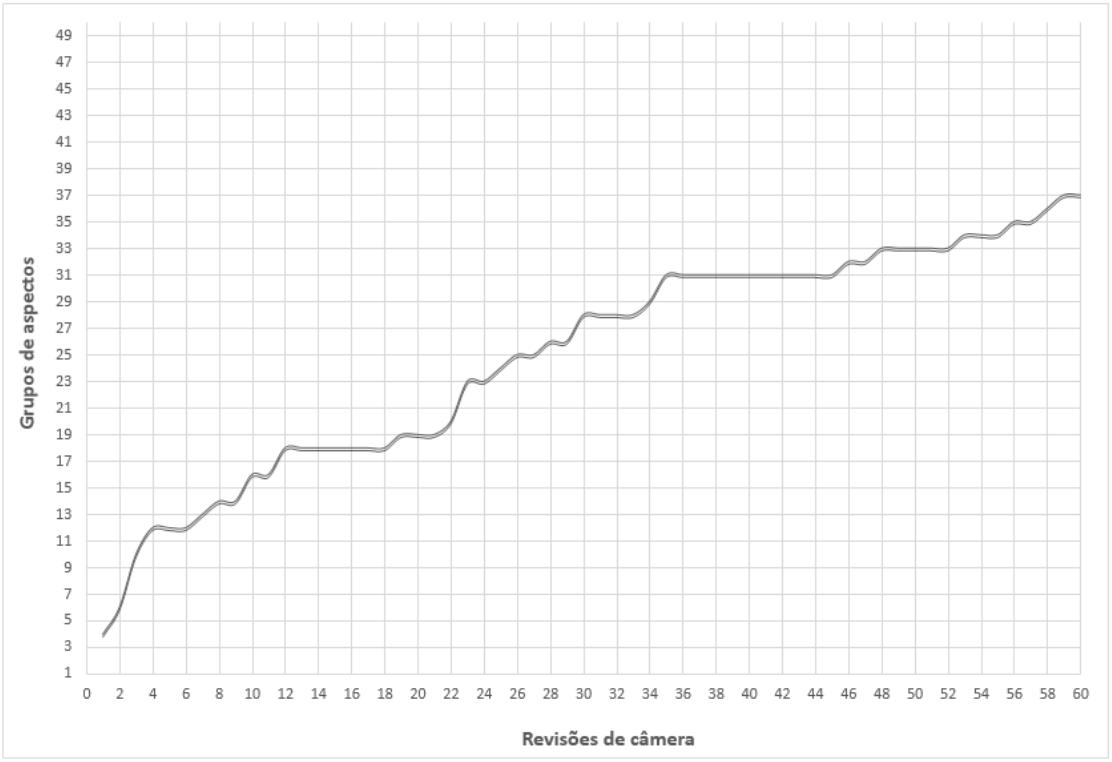

Figura 16. Curva de aprendizagem do domínio de câmera digital

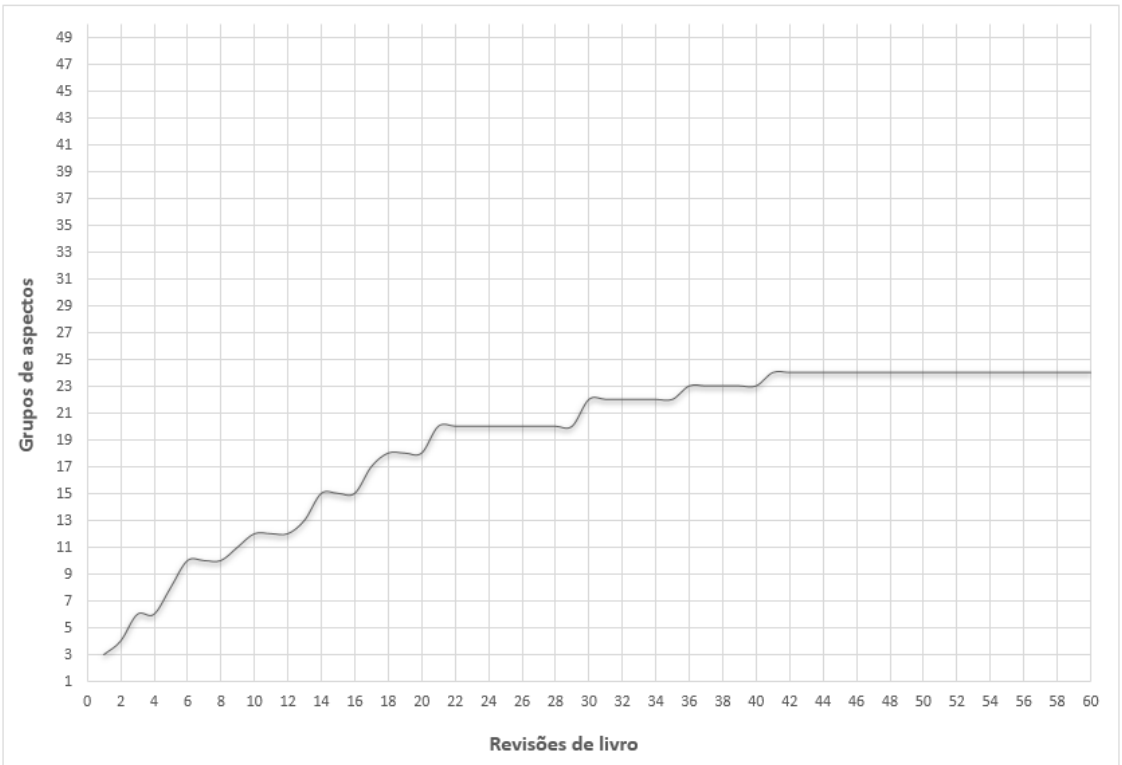

Figura 17. Curva de aprendizagem do domínio de livro

5.1.9. Compartilhamento de características linguísticas

Os itens dos grupos de aspectos parecem compartilhar características linguísticas. Por exemplo, há grupos de aspectos que possuem itens que compartilham características morfológicas. Alguns deles têm em comum a mesma raiz, por exemplo, {função, funções, funcionalidade}, {livro, livrinho}, {romance, romancezinho}, {foto, fotografia}, {leve, leveza, levinho}, {prático, praticidade}, {fácil, facilidade} e {manuseio, manusear}. Os grupos de aspectos também parecem compartilhar características morfossintáticas: cerca de 95% dos aspectos fazem parte dos grupo dos substantivos. Os 5% restantes são referências a aspectos implícitos. Grande parte dos aspectos são núcleos de sintagma nominial. Características semânticas também são compartilhadas entre os grupos. Por exemplo, os aspectos de cada grupo fazem parte do mesmo campo semântico. Segundo [Lyons 1977][6], os campos semânticos são estruturas subjacentes de representação ideológica pelas palavras. Por exemplo, o grupo de aspecto "valor" é composto pelos aspectos "preço", "custo", "investimento" e "custo-benefício". Todos esses aspectos compartilham traços semânticos relacionados e que compõem uma mesma ideologia.

5.1.10. Compartilhamento de características extralinguísticas

De acordo com [Zhao and Li 2009], revisões de usuários sobre produtos contêm menor grau de conteúdo irrelevante, no entanto, é necessário especializar esse conceito. Constatamos que, para o domínio de livro, existe uma parcela significativa de conteúdo irrelevante, cerca de 48% (ver seção 5.1.1). Além disso, outras variáveis influenciam no status informacional de uma revisão de usuário e, portanto, no grau de conteúdo relevante e irrelevante de um domínio. De acordo com [Weinreich et al. 1968] e [Labov 1994], vários fatores regulam a escolha entre um ou outro fenômeno linguístico em um domínio, chamados por ele de "condicionadores". É o controle rigoroso desses fatores que permite ao linguista sugerir em que tipo de ambiente, tanto linguístico quanto extralinguístico, um fenômeno linguístico tem maior probabilidade de ser escolhido em detrimento de outro. De acordo com [Labov 1994] e [Weinreich et al. 1968], os condicionadores são divididos em dois grandes grupos, em função de serem mais ligados a aspectos internos ao sistema linguístico ou externos a ele. No primeiro caso, são chamados de condicionadores linguísticos. Alguns exemplos são a ordem dos constituintes, a categoria das palavras ou construções envolvidas, aspectos semânticos, etc. Os condicionadores externos são chamados de condicionadores extralinguísticos ou sociais. Entre eles, os mais comuns são o sexo/gênero, o grau de escolaridade, a faixa etária do informante e o nível sócio-econômico.

Durante o processo de análise das revisões dos domínios analisados, observamos comportamentos linguísticos distintos entre os domínios de smartphone e câmera digital em relação ao domínio de livro. Para o domínio de smartphone e câmera, as revisões apresentaram certo "padão de linguagem", diferentemente do domínio de livro. No domínio de livro, constatamos que os usuários emissores das revisões não compartilhavam as mesmas características de perfil, diferentemente dos usuários dos domínios de smartphone e câmera digital. Por exemplo, revisões do livro *Crepúsculo* possuíam vocabulário e com-

[6] Trier (1931) apud Lyons (1977, p. 253)

portamento diferentes das revisões emitidas para o livro *1984*. A maioria dos usuários consumidores da câmera digital, por exemplo, possuem proximidade de faixa etária, escolaridade e padrão sócio-econômico e cultural. Todas essa variáveis implicam diretamente nos fenômentos linguísticos presentes nas revisões emitidas por esses usuários e, necessariamente, influenciam o grau de complexidade de análise e processamento, além da escolha dos metódos a serem aplicados. Por exemplo, usuários com alta escolaridade e alto perfil sócio-econômico tendem a produzir conteúdo com maior adequação à variante padrão da língua. Contudo, fenômenos como ironia e sarcasmo também são comuns em conteúdos gerados por esse grupo de usuários. Em contrapartida, usuários com perfil de baixa escolaridade tendem a produzir conteúdo com menor adequação à variante padrão da língua. No domínio de livro, selecionamos vários livros, de vários autores, e, portanto, de gêneros variados e para públicos com características distintas. Por exemplo, observamos que, em revisões do livro *Fala sério, Amiga!*, cujo público é majoritariamente formado por adolescentes de sexo feminino e classe média, havia marcas expressivas de oralidade e itens lexicais de senso comum, além de diminutivos e vagueza. Tais características aumentaram a complexidade da análise e interpretação humana. Um exemplo de revisão do livro *Fala sério, amiga!* é mostrado na Figura 5. Nas revisões do livro *1984*, o tema é político e voltado para um público adulto e de alta escolaridade. As revisões eram emitidas de forma clara e com maior adequação à variante padrão da língua. Além disso, o uso de itens lexicais de alta cultura eram recorrentes, além da clareza e organização de ideias. Portanto, este último não apresentou tantas dificuldades de análise e interpretação humana.

Esta análise demonstrou o quão importante é a análise de fatores intrínsecos e extrínsecos às línguas natuais, principalmente tratando-se do domínio de opinião, cujos textos possuem marcas expressivas de subjetividade.

5.2. Organização hierárquica dos grupos de aspectos

Os grupos identificados foram organizados de forma hierárquica, resultando em três hierarquias, cada uma delas para um dos domínios analisados. As hierarquias obtidas, todas árboreas, são exibidas nas Figuras 20, 21 e 22.

O *nó* raiz das três hierarquias representa o domínio analisado, ou seja, smartphone, câmera digital e livro. Também tomamos algumas decisões pontuais de organização para gerar hierarquias mais claras. Por exemplo, organizamos os grupos de aspectos que se referiam às características físicas dos domíninos no *nó* filho denominado *físico*; para o domínio de smartphone, organizamos todos os grupos de aspectos que remetiam a conectividade em um *nó* chamado "conexão". É interessante notar que tais decisões acabam por ser *ad hoc*, dando margem para a criaçao de hierarquias diferentes. Esse tipo de questão é recorrente quando se lida com a criação de estruturas taxonômicas e ontológicas.

Em alguns casos, identificar a posição de alguns grupos de aspectos na hierarquia era um desafio, por, muitas vezes, haver ambiguidade ou múltiplas possibilidades. Por exemplo, o grupo de aspectos de "nitidez", do domínio de câmera, pode tanto ser um *nó* filho do *nó* raiz câmera em uma relação *parte-todo*, como pode ser um *nó* filho de "imagem". O mesmo fenômeno ocorreu com vários outros grupos. Por exemplo, o *nó* "resolução" pode ser relacionado com o *nó* raiz câmera, ou com o *nó* "imagem". Para o domínio de livro, os *nós* "estilo" e "escrita" podem ter relação *parte-todo* tanto com o *nó*

raiz livro quanto com “história”, ou até mesmo com “autor”. No domínio de smartphone, os *nós* “rádio”, “e-mail” e “jogos”, por exemplo, poderiam ter relação de *parte-todo* com o *nó* “aplicativos” ou com o *nó* raiz smartphone.

Fizemos também a comparação entre as hierarquias geradas manualmente pelo nosso processo e outras três ontologias: (i) a ontologia do domínio de smartphone disponível no repositório OntoLP [7]; (ii) a ontologia proposta por [Aciar et al. 2006] de revisões de usários do domínio de câmera digital; e (iii) a ontologia parcial para o domínio de livro proposta por [Condori 2014]. As ontologias usadas para comparação são ilustradas nas Figuras 10, 18 e 19.

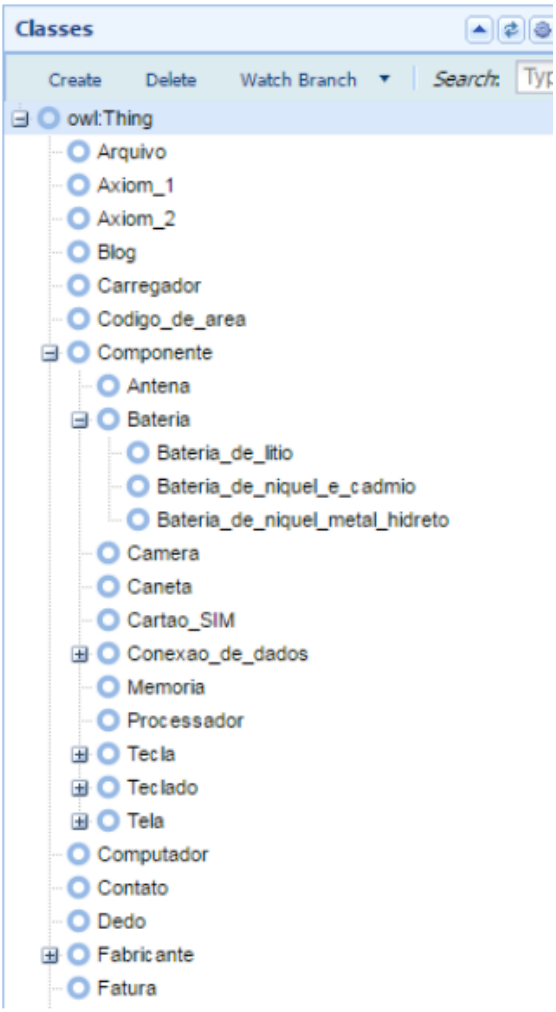

Figura 18. Organização hierárquia do domínio de smartphone disponível no Onto-LP

A Figura 18 ilustra a hierarquia para o domínio de smartphone, proposta por [Goulart and Montardo 2007] e disponível no repostório OntoLP. Essa hierarquia,

[7]Este repositório se propõe a divulgar ontologias disponíveis na língua portuguesa (incluindo desde bases terminológicas e vocabulários controlados até ontologias mais complexas representadas em OWL-DL), bem como ferramentas e recursos relacionados à pesquisa na área. O respositório pode ser acessado em http://ontolp.inf.pucrs.br/

diferentemente da hieraquia que propomos neste trabalho, é constituída apenas de relações de hiperonímia/hiponímia, caracterizada pelo tipo *é-um*. Por exemplo, os *nós* "bateria-de-litio", "bateria-de-niquel-e-cadmio" e "bateria-de-niquel-metal-hidreto", possuem relação de *é-um* com o *nó* "bateria". Esse tipo de hierarquia é interessante para vários domínios. Entretanto, para o domínio de opinião, não traduz com especifícidade o modelo discursivo predominante em revisões de usuários. Usuários, ao emitirem uma revisão, podem decompor a entidade ou alvo da opinião em características ou aspectos e avaliarem cada um dos aspectos de forma conjunta ou separadamente. As revisões do domínio de smartphone analisadas neste trabalho tiveram aspectos avaliados pelos usuários, em sua maioria, com relação de *parte-todo* com a entidade ou alvo da opinião, como pode ser visto na Figura 20.

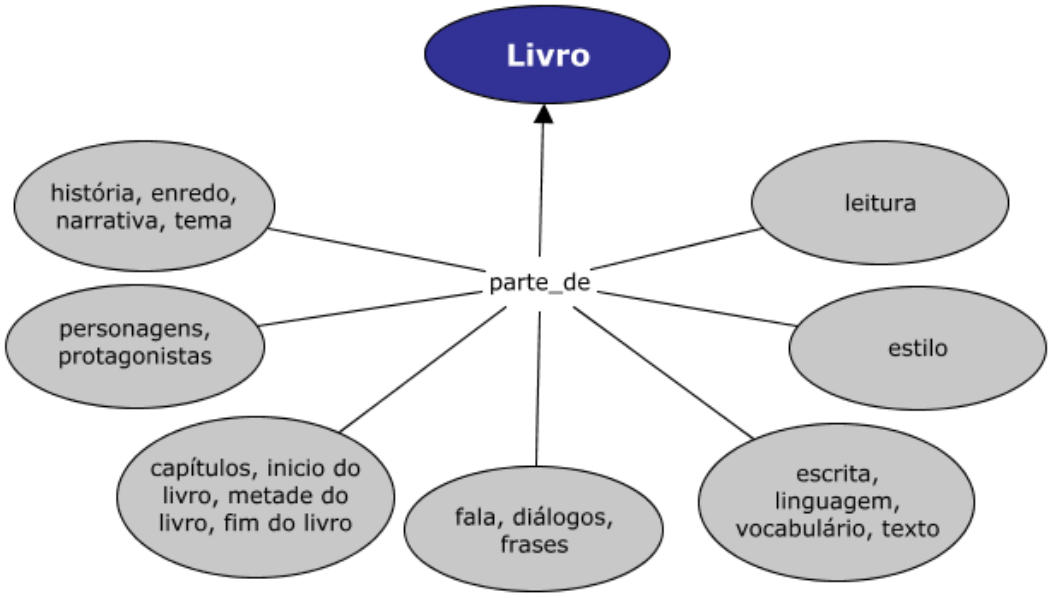

Figura 19. Organização hierárquica do domínio de livro proposta por [Condori 2014]

Retomenos a Figura 10, que ilustra a hierarquia proposta por [Aciar et al. 2006] para o domínio de câmera digital. Neste trabalho, os autores propõem a organização hierarquica de aspectos, a partir de revisões de usuários, para melhorar a acurácia de sistemas de recomendação. Nesta hierarquia, diferentemente da hierarquia obtida em nossa proposta para o domínio de câmera, os autores apresentam apenas os *aspectos físicos* de câmera em consonância com a *qualidade da imagem*. Observamos que, em revisões de usuário do domínio de câmera digital, são avaliados muitos outros aspectos de câmera e não apenas os aspectos físicos, como é proposto na hierarquia dos autores. Em nossa hierarquia, obtivemos 40 grupos de aspectos, sendo que a hierarquia dos autores possui apenas 12. As relações entre aspectos, assim como na nossa hierarquia, também são do tipo *parte-todo*, o que implica que, para esse domínio, as relações predominantes são de meronímia/holonímia.

Também fizemos a comparação com a hierarquia proposta no trabalho de [Condori 2014]. Neste trabalho, a hierarquia é proposta para a tarefa de sumarização de

opiniões. A hierarquia do autor é concisa e objetiva, entretanto, não parece representar as especificídade de conhecimento do domínio. Por exemplo, no trabalo de [Condori 2014], os aspectos "história", "enredo", "narrativa" e "tema" foram agrupados em um mesmo *nó*. Contudo, essas características podem ser avaliadas separadamente pelo usuário, por se tratar, cada uma delas, de uma especificídade da entidade. Por exemplo, "tema", que também pode ser "assunto" ou "essência" do livro, são aspectos que poderiam ser agrupados, por todos eles tratarem apenas de uma especificidade da entidade e possuírem o mesmo contexto de uso. Na hierarquia proposta neste trabalho, para o domínio de livro, organizamos hierarquicamente todos os grupos de aspectos identificados no domínio, como pode ser visto na Figura 22. Nossa hierarquia para o domínio de livro, assim como a hierarquia proposta por [Condori 2014], apresentou relações apenas do tipo *parte-todo*.

Um fato interessante sobre a hierarquia do domínio de livro obtida nesta proposta, em relação aos outros domínios analisados neste trabalho, é a ocorrência apenas de relações do tipo *parte-todo*, enquanto que, para os outros domínios, nós também tivemos relações do tipo *é-um*, mesmo que em número não muito expressivo.

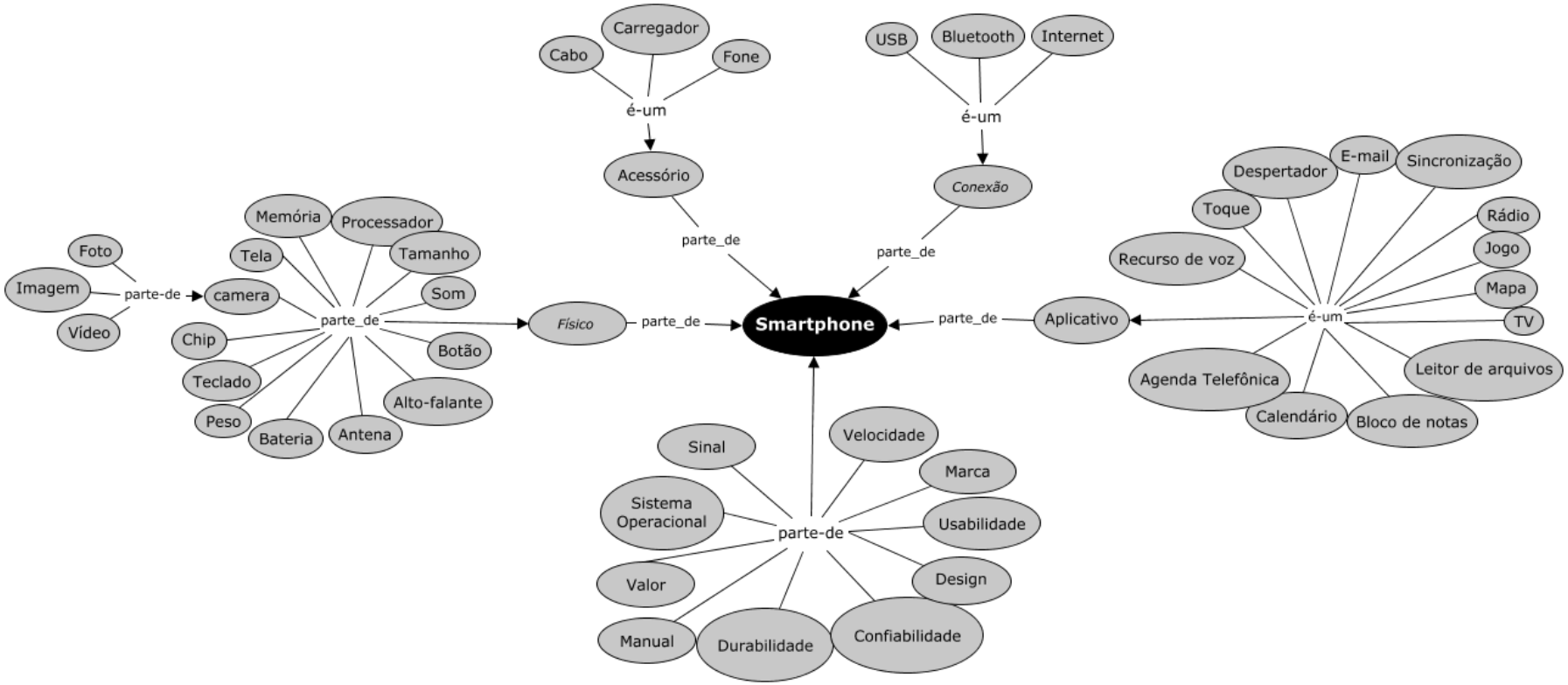

Figura 20. Hierarquia obtida para o domínio de smartphone

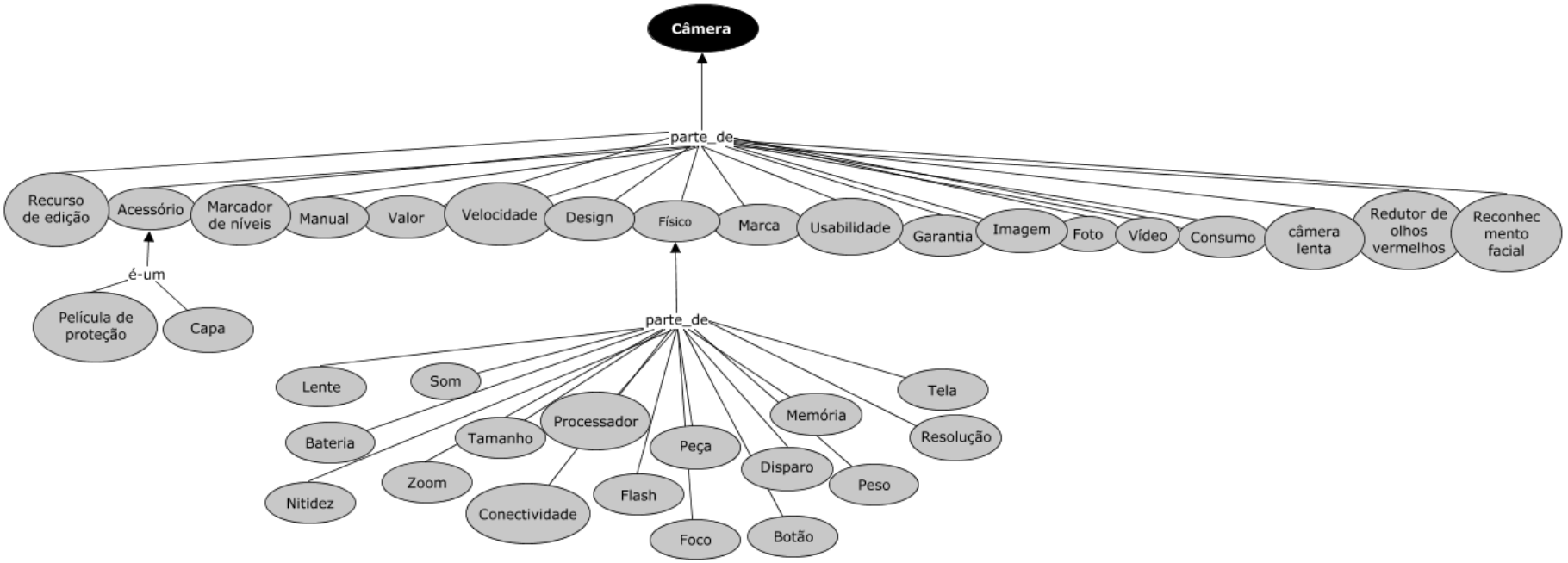

Figura 21. Hierarquia obtida para o domínio de câmera digital

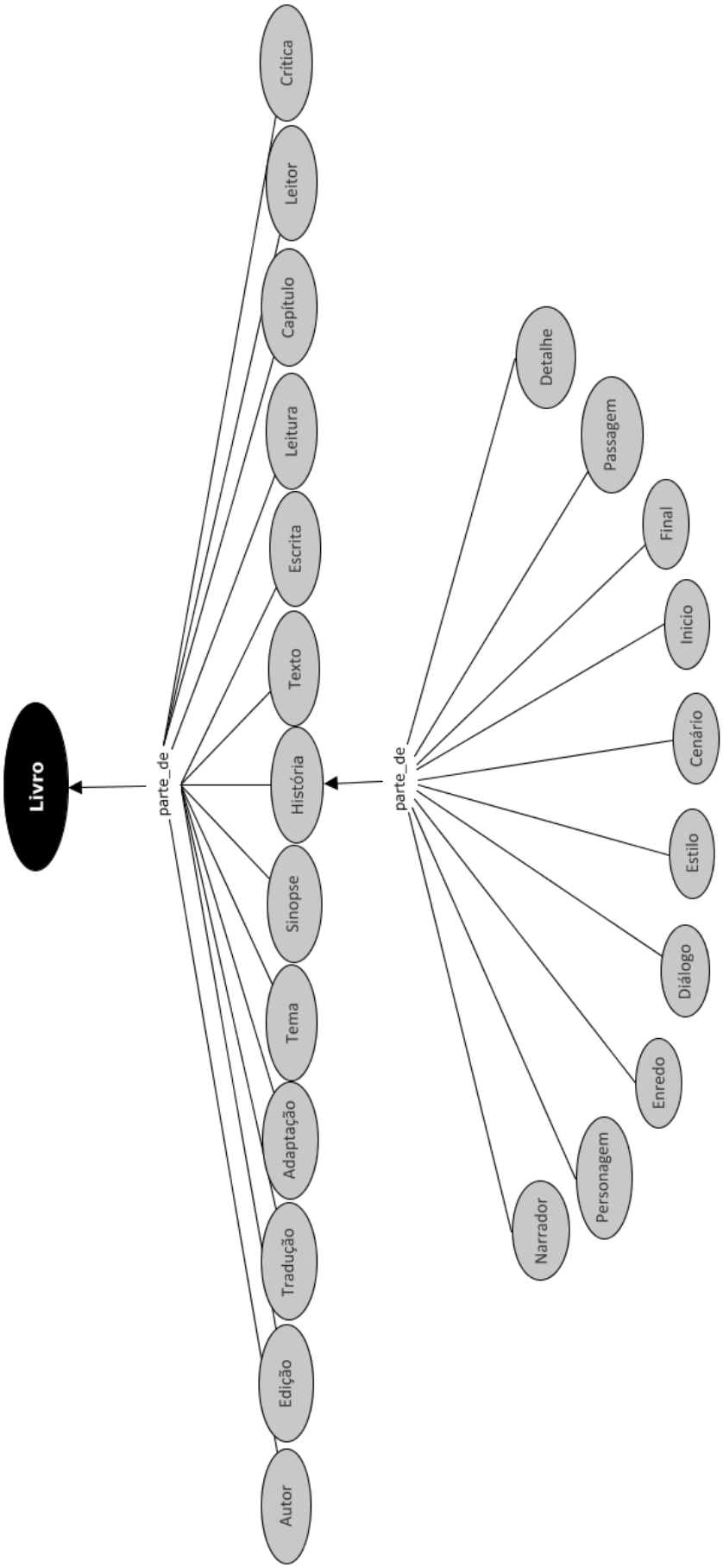

Figura 22. Hierarquia obtida para o domínio de livro

6. Considerações finais

Neste trabalho, apresentamos um estudo téorico e empírico para a tarefa da mineração de opinião baseada em aspectos. Apresentamos um estudo empírico e teórico sobre os processos de agrupamento de aspectos e a organização hierárquica dos grupos identificados. Analisamos três domínios diferentes de revisões de usuários e constatamos que é possível encontrar tanto conteúdo descritivo-objetivo quanto conteúdo descritivo-subjetivo, e que o grau desses conteúdos tem fortes ligações com o domínio. Segundo [Zhao and Li 2009], revisões de usuários sobre um produto específico contém pouca informação irrelevante. No entanto, constatamos que revisões de usuários sobre livros podem conter uma parcela significativa de conteúdo irrelevante e que o status de informatividade de uma revisão sofre influência de fatores linguísticos e extralinguísticos. Observamos que, para esse domínio, diferentemente dos outros domínios, o perfil dos usuários é distinto. Portanto, diferente do que se acreditava em [Zhao and Li 2009], a parcela de conteúdo relevante e irrelevante de um domínio pode sofrer influência de múltiplas variáveis, entre elas variáveis linguísticas e extralinguísticas. Além disso, constatamos que, para a cobertura de um domínio, é necesária, em média, a leitura de quarenta revisões. Entretanto, isso pode variar de acordo com o perfil do usuário. Observamos que, quanto mais conhecimento o usuário emissor da revisão possuir sobre a entidade/alvo avaliado, maior a probabilidade deste usuário avaliar um número mais expressivo de grupos de aspectos. Portanto, identificar o perfil desses usuários é interessante, pois permite classificar revisões "pontenciais" para identificação de um número maior de grupo de aspectos. Observamos também certo "padrão" de vocabulário e comportamentos linguísticos para cada domínio e que esse padrão sofre grande influência de variáveis extralinguísticas.

Os dados produzidos nesta investigação devem ser disponibilizados publicamente. Espera-se que eles possam servir de dados de referência para desenvolvimento e avaliação de métodos computacionais de base para a mineração de opinião.

Agradecimentos

À FAPESP e à CAPES pelo apoio a este trabalho.

Referências

Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2006). Recommender system based on consumer product reviews. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 719–723, Washington, USA.

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

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

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

Biemann, C. (2005). Ontology learning from text: A survey of methods. *LDV Forum*, 20(2):75–93.

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

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

Condori, R. E. L. (2014). *Sumarização automática de opiniões baseada em aspectos*. Dissertação de Mestrado em Ciência da Computação e Matemática Computacional, Universidade de São Paulo, São Carlos, Brazil.

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

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

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

Goulart, R. R. V. and Montardo, S. P. (2007). Os mecanismos de busca e suas implicações em comunicação e marketing. In *Anais do V Congresso Nacional de História da Mídia*, pages 478–514, São Paulo, Brazil.

Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M. D. G. V., Pardo, T., and Aluísio, S. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3865–3871, Reykjavik, Iceland.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, USA.

Koch, I. G. V. (2009). *Introdução à Linguística Textual*. Martins Fontes, 2nd edition.

Labov, W. (1994). *Principles of linguistic change: Internal Factors*, volume 1. Oxford, Cambridge: Blackwell.

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

Lyons, J. (1977). *Semantics*. Cambridge University Press, 1st edition.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

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

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education, 2nd edition.

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

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347.

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

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

Weinreich, U., Labov, W., and Herzog, M. I. (1968). *Empirical foundations for a theory of language change*. University of Texas Press.

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

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

A. Apêndice 1

Nesta seção, apresentamos os grupos de aspectos identificados para o domínio de smartphone.

Na Tabela 7, os elementos das colunas se relacionam com o elemento da coluna "etiqueta" pela relação indicada. Por exemplo, os elementos sob a relação de "sinonímia" encontram-se em relação de sinonímia ou identidade com a etiqueta. No caso da coluna de "hiperonímia", os elementos estão tanto em relação de hiperonímia quanto hiponímia (que são os dois sentidos da relação). O mesmo se aplica para as demais colunas. Os símbolos das chaves são usados para agrupar aspectos iguais, mas que apresentam alguma variação de escrita (nem sempre adequada, segundo a variante padrão da língua). Essa mesma forma de interpretação se aplica às tabelas dos apêndices seguintes.

Tabela 7: Grupos de aspectos para o domínio de smartphone

grupo	**etiqueta**	**sinonímia**	**é-um**	**parte-todo**	**deverbal**	**correferência**
G1	*smartphone*	smartphone	aparelho, produto, celular, telefone	-	-	”Porcaria”
G2	*conexão de internet*	internet, net, conexão, conectividade	wifi, wi-fi, wi fi, wireless, 3G, WAP	-	-	-
G3	*usabilidade*	usabilidade, praticidade, ”prático”, facilidade, funcionalidade	interface, menu, função, recurso, extra, opção, manuseio, ”linguagens”, ”operação”,	-	“fácil de Usar”, “fácil de Mexer”, ”fácil de manusear”	-
G4	*velocidade*	velocidade	”rápido”, ”lento”	-	”congela”, ”restarta”, ”trava”	”bugs”, ”tempo de resposta”, ”demora a responder”,
G5	*valor*	custo, preço, valor, investimento	{Custo-benefício, custo-beneficio}, preço-benefício, ”acessível”, ”barato”	-	-	-

G6	*bateria*	bateria	-	autonomia da Bateria; duração da Bateria	”Descarrega”	-
G7	*teclado*	teclado	-	tecla, tecla de atalho	-	-
G8	*mapa*	-	google Maps; gps	-	-	-
G9	*aplicativo*	aplicativo	-	-	-	
G10	*tela*	tela, visor, display	touch, touch screen, touchscreen	tamanho do visor		vidro, ”sensibilidade”;
G11	*som*	áudio, sonorização; som	música; mp3; mp3 player	qualidade do áudio; qualidade do som; qualidade sonora, volume do áudio	-	volume
G12	*toque*	toque	hit polifônico, toque polifônico	-	-	-
G13	*alto-falante*	alto-falante	-	-	-	
G14	*design*	{designer, designe, desing, design}, estética, modelo	estilo, beleza, ”arrojado”, elegância, ”chique”, ”atual”, “bonito”, ”moderno”, ”robusto”, ”lindo”	-	-	-
G15	*câmera*	câmera	câmera digital; câmera imbutida, filmadora, ”módulo de filmar”	foco da câmera; resolução da câmera; flash da câmera; luz do flash, zoom da câmera;	-	-

G16	*foto*	fotografia, foto	foto panoramica	qualidade da foto	-	,
G17	*bluetooth*	Bluetooth	-	-	-	
G18	*marca*	-	empresa; nokia; motorola; LG; sony; sony ericson; siemens, marca	-	-	fabricante; fábrica
G19	*chip*	sim	dual chip	-	-	-
G20	*sinal*	sinal	quadriband, ligação	-	-	recepção; ”recebi chamadas até na beira do rio são francisco”; “funciona em qualquer lugar”
G21	*tv*	tv	-	-	-	-
G22	*peso*	pesado, ”versátil”	”leve”, ”leveza”, ”levinho”	-	-	-
G23	*botão*	-	Botões Liga/Desliga, Botões de toque	-	-	-
G24	*memória*	memória	sd de memória; memória interna; cartão de memória; expansão de memória; cartão de expansão; espaço de memória	-	-	-

G25	*cabo*	-	Cabo para TV; {Cabo de dados, Cabo-de-dados}	-	-	-
G26	*e-mail*	e-mail	-	-	-	-
G27	*rádio*	rádio	-	-	-	-
G28	imagem	imagem	-	definição de imagem	-	-
G29	vídeo	vídeo	-	-	-	-
G30	*leitor de arquivo*	-	Leitor de PDF	-	-	-
G31	*sistema operacional*	sistema operacional, programa operacional, software	”Sistema Symbiam/Java”	-	-	”falta de compatibilidade”
G32	*sincronização*	-	sincronização com o pc, sincronização de dados	-	-	”acesso aos dados”.
G33	*fone*	fone	fone de ouvido	-	-	-
G34	*jogo*	jogo	-	-	-	-
G35	*carregador*	carregador	carregador de carro	-	-	
G36	*manual*	manual	-	-	-	-
G37	*tamanho*	tamanho	”pequeno”, ”compacto”, ”volumoso”	-	-	-
G38	*processador*	processador	-	-	-	-
G39	*durabilidade*	durabilidade	-	-	-	-
G40	*usb*	usb	-	-	-	-
G41	*recurso de voz*	-	Gravador de Voz; {Viva Voz; Viva-Voz}	-	-	-
G42	*despertador*	despertador	-	-	-	-
G43	*acessório*	acessório	-	-	-	-
G44	*bloco de notas*	bloco de notas	-	-	-	-

G45	*confiabilidade*	confiabilidade	-	-	-	“Confiança na marca”
G46	*calendário*	calendário	-	-	-	-
G47	*antena*	antena	-	-	-	-
G48	*agenda telefônica*	agenda telefônica	-	-	-	-

B. Apêndice 2

Nesta seção, apresentamos os grupos de aspectos identificados para o domínio de câmera.

Tabela 8: Grupos de aspectos para o domínio de câmera

grupo	**etiqueta**	**sinonímia**	**é-um**	**parte-todo**	**deverbal**	**correferência**
G1	*câmera*	câmera, camerazinha	produto, máquina, camera semiprofissional, câmera digital, máquina digital	qualidade da câmera	-	
G2	*valor*	custo, preço, valor, investimento	”barato”, custo-benefício, custo/benefício	-	-	-
G3	*imagem*	imagem	”modo noite”	qualidade da imagem; cores da imagem; filtros de imagem; modo de imagem	-	-

G4	*usabilidade*	praticidade, ”prática”, funcionalidade, usabilidade, facilidade	menu, função, recurso, opções, manuseio, ”intuitiva”, ”auto explicativa”	-	”facilidade de uso”; ”fácil de utilizar”; ”fácil de manusear”; ”fácil de operar”,”fácil de usar”; ”fácil uso”; ”facilidade de mexer”	-
G5	*design*	design, aparência	-	acabamento, beleza, ”material”	-	”bonita”, ”linda”, ”elegante”
G6	*acessório*	acessório	-	-	-	-
G7	*marca*	marca	sony, fujy, empresa	-	-	-
G8	*bateria*	bateria	bateria reserva, pilha	-	-	-
G9	*botão*	botão	-	-	-	-
G10	*resolução*	resolução	megapixel	-	-	
G11	*foto*	foto, fotografia	foto dentro d’água, foto noturna, nigth shot	resolução da foto ; qualidade de fotos; opção de foto; cor da foto; navegação na foto	-	-
G12	*vídeo*	vídeo	filme, filmagem; gravação	resolução do vídeo; qualidade de vídeo; qualidade do filme; gravação de vídeo	”filma em HD”	

G13	*tamanho*	tamanho, volume	”medidas”, ”pequeno”, ”compacto”, ”grande”, ”fino”	-	-	
G14	*flash*	flash	-	-	-	”se não tiver luz boa”
G15	*consumo*	-	consumo de energia	-	-	-
G16	*conectividade*	conectividade	-	-	-	-
G17	*tela*	tela, visor, display	visor Ocular, touchscreen	Tamanho da tela	-	-
G18	*manual*	manual	manual em português, manual de instruções	-	-	
G19	*memória*	memória, cartão, sd card	memória interna; cartão de memória; cartão sd	capacidade de memória	-	armazenamento
G20	*zoom*	zoom	ultrazoom	-	-	-
G21	*peso*	peso	”leve”, ”leveza”, ”versatil”	-	-	-
G22	*garantia*	garantia	-	-	-	-
G23	*capa*	capinha	-	-	-	-
G24	*nitidez*	nitidez	-	-	-	-
G25	*película de proteção*	película de proteção	-	-	-	-
G26	*foco*	foco	-	-	-	-
G27	*lente*	lente, objectiva	lente auxiliar	-	-	-
G28	*peça*	peça	-	-	-	-
G29	*velocidade*	velocidade	-	-	”demora para responder”	”rápida”
G30	*câmera lenta*	slow motion		-	-	-
G31	*marcador de níveis*	-	Mostrador de níveis de memória e bateria	-	-	-

G32	*redutor de olhos vermelhos*	redutor de olhos vermelhos		-	-	-
G33	*som*	som	-	-	-	-
G34	*recurso de edição*	recurso de edição, presets	-	-	-	-
G35	*disparo*	disparo	-	-	-	-
G36	*processador*	processador	-	-	-	-
G37	*reconhecimento facial*	reconhecimento facial		-	-	-

C. Apêndice 3

Nesta seção, apresentamos os grupos de aspectos identificados para o domínio de livro.

Tabela 9: Grupos de aspectos para o domínio de livro

grupo	**etiqueta**	**sinonímia**	**é-um**	**parte-todo**	**deverbal**	**correferência**
G1	*livro*	livro, obra, livrinho	”Bestseller”	-	-	
G2	*autor*	escritor, autor	Jorge Amado; Stephenie Meyer; Thalita Rebouças; Tatá; Sidney Sheldon; Saramago; Sidney; Orwell; George Orwell; Jorge Amado; José Saramago; Sheldon	-		

G3	*personagem*	personagem	protagonista, herói; Isabella Swanchega; Bella; Isabella Swan; Malu; Pedro Bala; Alice; Noelle; ; Holden; Edward Cullen; Edward; Larry, ”garota”, ”mocinha”	-	-	
G4	*escrita*	escrita	linguagem	técnicas de escrita; Estilo de Escrita	”escrever”, ”escreve de forma envolvente”	-
G5	*história*	história, estória	narrativa, crônica, suspense, ficção, romance, aventura, literatura, romancezinho	tipo de história	-	-
G6	*detalhe*	detalhe	-	riqueza de detalhes	-	-
G7	*leitura*	leitura	-	-	-	-
G8	*leitor*	leitor	-	-	-	-
G9	*cenário*	cenário	-	-	-	-
G10	*crítica*	crítica, pensamento, reflexão	crítica social	-	”refletir”	-
G11	*estilo*	estilo	-	-	-	-
G12	*início*	início, começo	-	-	”começa”	-

G13	*tema*	tema, mensagem, assunto, questão, ”essência”, ”ponto”	”caricatura cega do amor ou ódio”, ”sociedade do big brother”, ”mentes sombrias”, ”sobre”	-	”remetendo”, ”explora a”, ”conta coisas ...”, ”é mostrado”, ”Fala de”	-
G14	*capítulo*	capítulo	-	-	-	-
G15	*passagem*	passagem, página	acontecimento, caso, situação	-	-	
G16	*texto*	texto	-	linguagem, frase, palavra, expressão	-	
G17	*enredo*	enredo, trama	-	-	-	-
G18	*adaptação*	adaptação	-	-	-	-
G19	*tradução*	tradução, idioma	-	-	-	-
G20	*final*	final, fim, desfecho	-	-	”termina”	-
G21	*edição*	edição	-	-	-	-
G22	*diálogo*	diálogo	-	-	-	-
G23	*sinopse*	sinopse	-	-	-	-
G24	*narrador*	narrador	-	-	-	-

D. Apêndice 4

Nesta seção, exibem-se os gráficos contendo os grupos de aspectos e o número de avaliações relacionadas a cada grupo. Os grupos de aspectos mais avaliados são os grupos que chamamos de "prototípicos" do domínio. A Figura 23 representa os grupos mais avaliados para o domínio de smartphone; a Figura 24 para o domínio de câmera digital; e a Figura 25 para o domíno de livro. Novamente, as barras em preto são utilizadas para destacar esses grupos.

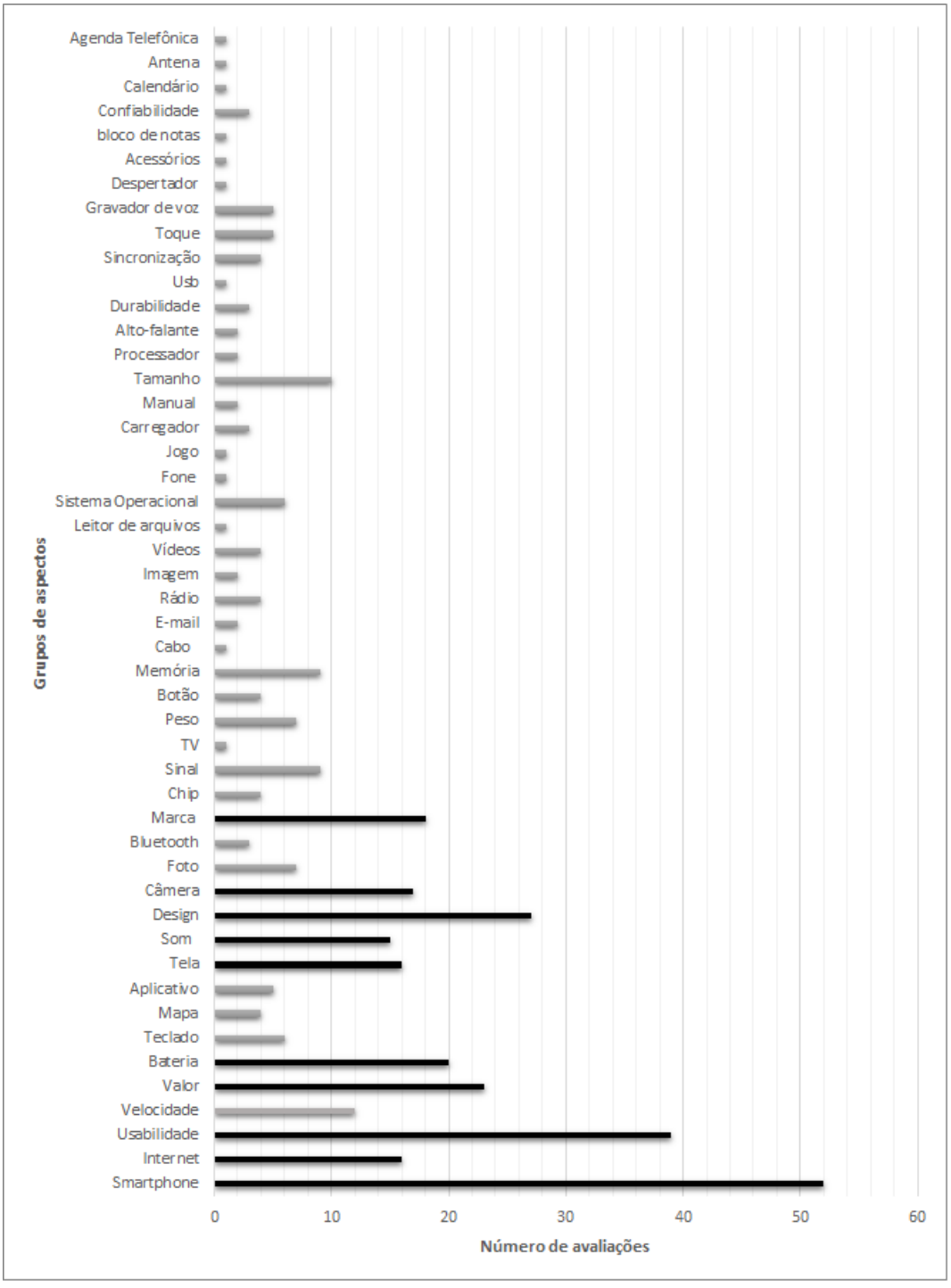

Figura 23. Número de avaliações para os grupos de aspectos do domínio de smartphone.

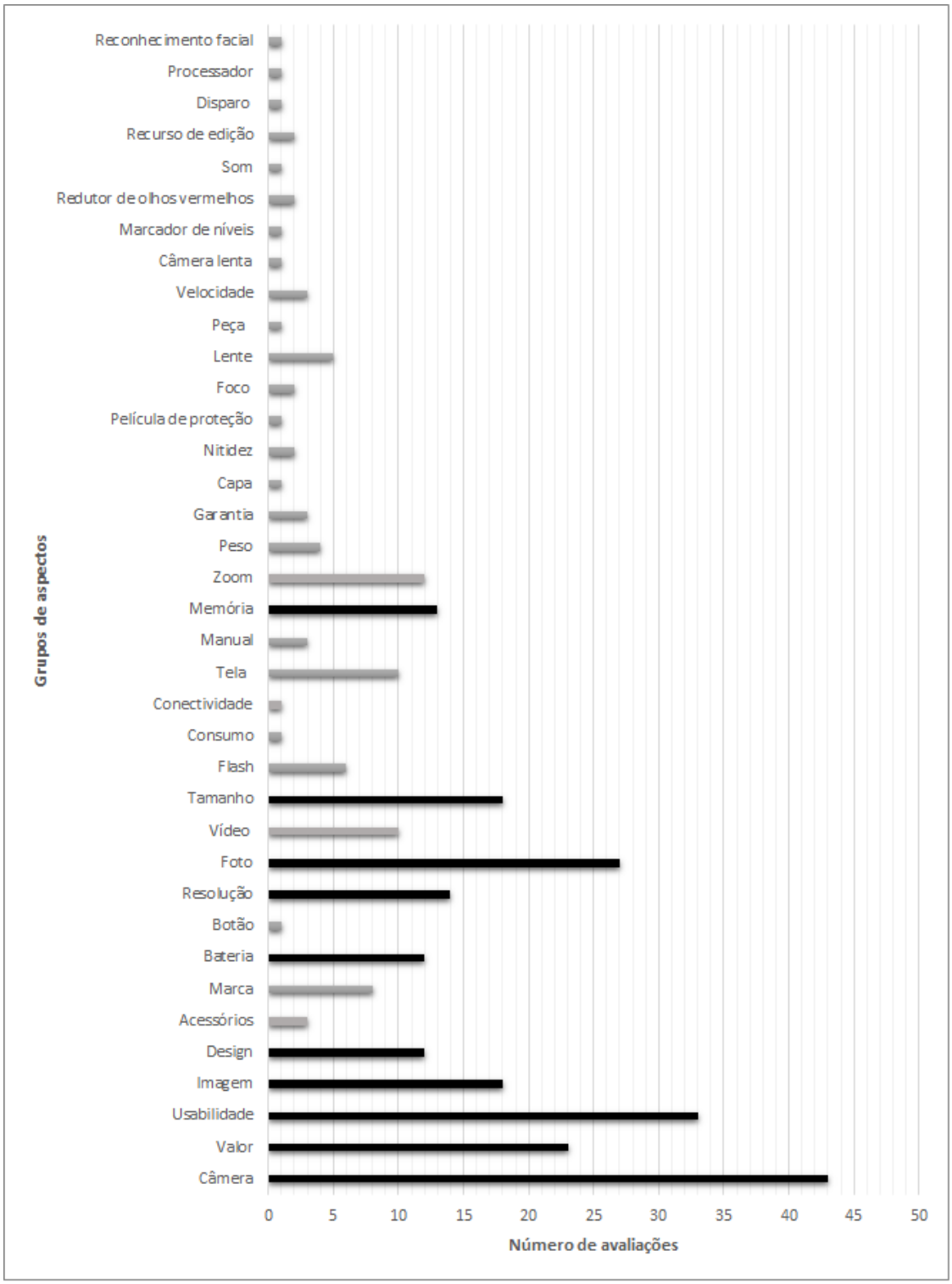

Figura 24. Número de avaliações para os grupos de aspectos do domínio de câmera.

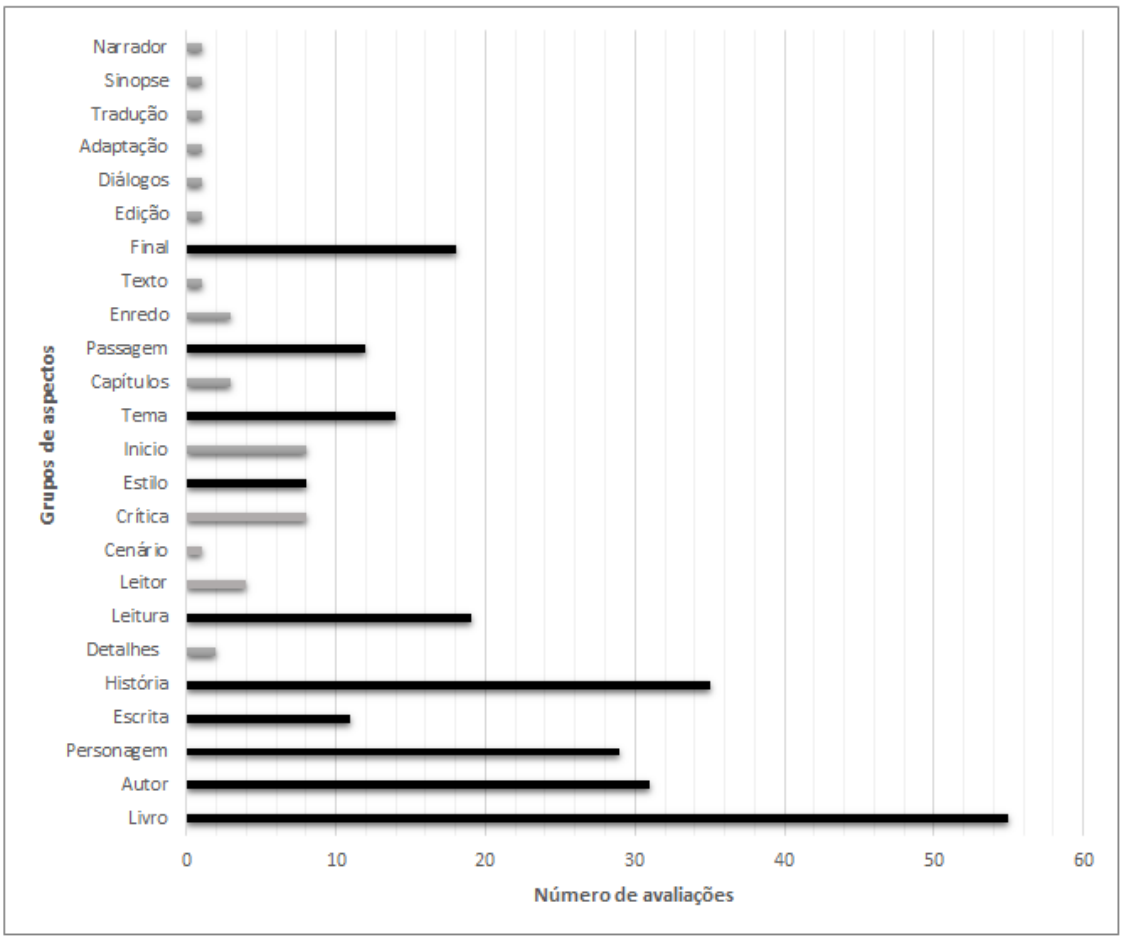

Figura 25. Número de avaliações para os grupos de aspectos do domínio de livro.