Hierarchical clustering of aspects for opinion mining: a *corpus* study

Clusterização hierárquica de aspectos para mineração de opinião: um estudo de córpus

Francielle Alves Vargas
Thiago Alexandre Salgueiro Pardo

Abstract: This paper consists of an empirical study on the problem of clustering and hierarchically organizing opinion aspects in product reviews in order to support aspect-based opinion mining applications. We performed a *corpus* study for characterizing and understanding the involved tasks, looking for linguistic patterns and convergences and divergences across domains. The process has been manually performed and resulted in reference data for future developments and evaluation of automatic methods in the area.

Keywords: Natural Language Processing. Opinion Mining. *Corpus* Linguistics.

Francielle Alves Vargas – Msc in Computer Science and Mathematics Computational (ICMC), University of São Paulo (USP) – francielleavargas@usp.br.
Thiago Alexandre Salgueiro Pardo – Professor and researcher at Institute of Mathematical and Computer Sciences (ICMC) in University of São Paulo (USP), Phd in Computer Science and Mathematics Computational, University of São Paulo (USP) – taspardo@icmc.usp.br.

Resumo: Este artigo consiste em um estudo empírico sobre o problema de agrupamento e organização hierárquica de aspectos a partir de revisões de usuários sobre produtos na *web*, a fim de apoiar aplicações de mineração de opinião baseada em aspectos. Realizamos um estudo de córpus para caracterização e compreensão das tarefas envolvidas, buscando padrões linguísticos, além de convergências e divergências entre os domínios. O processo foi realizado manualmente e resultou em dados de referência para pesquisas futuras e avaliação de métodos automáticos na área.

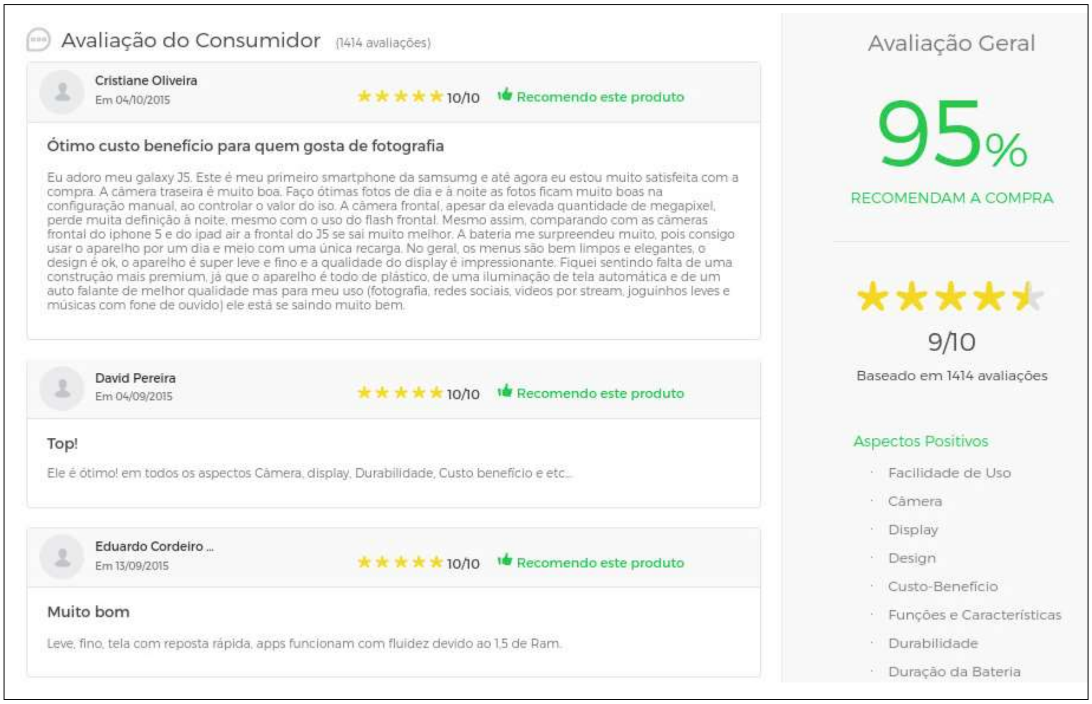
Palavras-chave: Processamento de Linguagem Natural. Mineração de Opinião Baseada em Aspectos. Linguística de *Corpus*.

1 Introduction

The expansion of the social networks and e-commerce services resulted in the growth of on-line reviews in the web. Websites as Amazon and Buscape encourage users to write reviews for products, where users may do objective or subjective descriptions for a product and its aspects or properties. Subjective descriptions are characterized by a personal language, with opinions, sentiments, emotions and judgments. The research area in charge of identifying, extracting and summarizing subjective information in texts is called opinion mining or sentiment analysis (PANG et al., 2002). According to Zhao e Li (2009), this area is different from the traditional text mining area, which is mostly based on objective topics rather than on subjective perceptions. Many searches used reviews of movie, book, and electronic product domains (HU et al., 2004), because these domains have relevance to both companies and consumers. For companies, it is important to evaluate their reputation, acceptability and evaluation of their products. For consumers, summarizing reviews from other users makes it easier to make decisions at the time of purchase. Therefore, providing relevant subjective content in reviews, among these various domains, is an important task in the current context, because it provides a better utilization of those data, for this purpose consumers and for private and governmental organizations.

2 Background

Opinion mining or sentiment analysis is the field of research in intersection between linguistics and computation, responsible for proposing methods of analysis, processing, summarization and classification for large volumes of data, mostly text type, for the extraction of subjective content. According to Liu (2012), there are three main granularity levels of analysis for opinion mining. These are: (i) document level, (ii) sentence level and (iii) aspects level. At the document level, the set of the opinions expressed in the document is accounted. For example, a document composed by subjective content may be classified as positive, negative or neutral, according to the accounting of relevant content that express sentiment. According to Liu (2012), these contents are usually expressed through adjectives. In Figure 1, we present a set of reviews on a smartphone.



Avaliação do Consumidor (1414 avaliações)

Cristiane Oliveira
Em 04/10/2015
★★★★★ 10/10 Recomendo este produto

Ótimo custo benefício para quem gosta de fotografia

Eu adoro meu galaxy J5. Este é meu primeiro smartphone da samsumg e até agora eu estou muito satisfeita com a compra. A câmera traseira é muito boa. Faço ótimas fotos de dia e à noite as fotos ficam muito boas na configuração manual, ao controlar o valor do iso. A câmera frontal, apesar da elevada quantidade de megapixel, perde muita definição à noite, mesmo com o uso do flash frontal. Mesmo assim, comparando com as câmeras frontal do iphone 5 e do ipad air a frontal do J5 se sai muito melhor. A bateria me surpreendeu muito, pois consigo usar o aparelho por um dia e meio com uma única recarga. No geral, os menus são bem limpos e elegantes, o design é ok, o aparelho é super leve e fino e a qualidade do display é impressionante. Fiquei sentindo falta de uma construção mais premium, já que o aparelho é todo de plástico, de uma iluminação de tela automática e de um auto falante de melhor qualidade mas para meu uso (fotografia, redes sociais, vídeos por stream, joguinhos leves e músicas com fone de ouvido) ele está se saindo muito bem.

David Pereira
Em 04/09/2015
★★★★★ 10/10 Recomendo este produto

Top!

Ele é ótimo! em todos os aspectos Câmera, display, Durabilidade, Custo benefício e etc...

Eduardo Cordeiro ...
Em 13/09/2015
★★★★★ 10/10 Recomendo este produto

Muito bom

Leve, fino, tela com reposta rápida, apps funcionam com fluidez devido ao 1,5 de Ram.

Avaliação Geral

95%

RECOMENDAM A COMPRA

9/10

Baseado em 1414 avaliações

Aspectos Positivos

- Facilidade de Uso
- Câmera
- Display
- Design
- Custo-Benefício
- Funções e Características
- Durabilidade
- Duração da Bateria

Figure 1 – Set of reviews on a smartphone and extracted from Buscapé

Note that 1.414 reviews have been issued on the smartphone product. Each of the 1.414 reviews refers to a document for opinion mining systems. Therefore, at document level, a positive, negative or neutral score is issued for each document. In this level of granularity, it is not possible to know precisely what the user liked or did not like. At the sentence level, the objective is to determine the opinion expressed in each sentence of the document. Therefore, a set of documents is segmented into sentences, and then a score is issued for each of the sentences. For example, in a document composed by X sentences, for each of those sentences there will be a positive, negative or neutral score. See again the Figure 1. In the third review, the user issues the following evaluation "Leve, fino, tela com resposta rápida, apps funcionam com fluidez devido ao 15 de Ram". Note that at this level, it is still not possible to know accurately the properties of the product evaluated by the user. To solve this problem, Liu (2012) argues that it is necessary a deeper level of analysis: aspect-based opinion mining. Aspects represent properties or parts of entities that are evaluated by users, in reviews, such as comments on websites and blogs on the web (LIU, 2012). However, the distinction between "attributes" and "aspects" is not clear in the literature. Mostly, they are used with synonyms.

For example, in the review, "A qualidade da imagem da câmera é excelente", the term "qualidade da imagem" is an attribute of the "imagem" aspect. Within this paper scope, we limit our contributions to the aspect level.

For Bhuiyan et al. (2009) opinion mining surveys may be divided into two main tasks: the *sentiment classification* and the aspect-based opinion mining. The sentiment classification consists in to recognize the general sentiment present in a document or a sentence. Typically, this task is simplified, classifying a document or a sentence into 3 classes: positive, negative or neutral (AVANCO; NUNES, 2014). Aspect-based opinion mining is usually focused on the three tasks: (i) aspects identification, (ii) polarity identification, and (iii) summarization (LIU, 2012). In Figure 2, we illustrate these tasks, and then we describe .

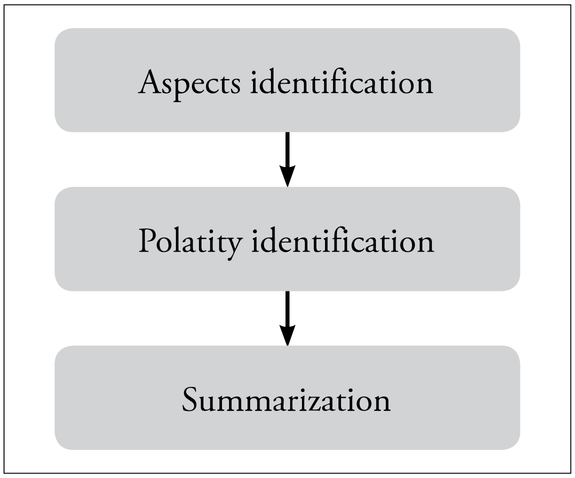


Figure 2 – Aspect-based opinion mining tasks

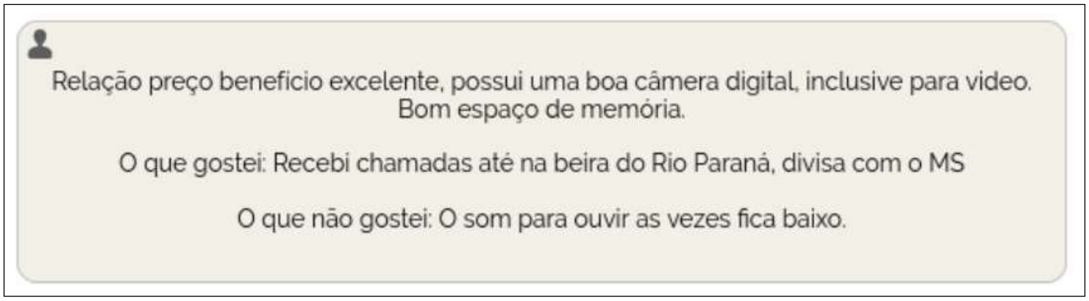
Aspect identification: features evaluated by users are extracted on the target of the opinion. For example, in the review "The Iphone 6 screen is amazing", the aspect evaluated is "screen"'.

Polarity identification: the sentiment associated with the aspects are extracted. For example, in the review "The camera battery is bad", the sentiment expressed about "bateria" aspect is negative, so the polarity of this aspect, considering this review is negative.

Summarization: the most relevant content is displayed through summaries, usually *extractive summaries*, which display the summarized content through the sentences clustering; or the *abstractive summaries*, which not only select the most relevant sentences of the source texts, but analyze the document and automatically generate new sentences. This approach attempts to produce new texts from the original fragments identified as relevant.

In addition to the tasks of aspects identification, polarity identification and summarization, according to Taboada (2016), another task of opinion mining responsible for determining whether a text, or most of them, is subjective or objective. According to the author, textual content may contain objective information (facts, actions) or subjective information (perceptions, opinions, sentiments). Moreover, subjective texts express a positive or negative view and this direction of opinion - whether positive or negative - is also known as semantic orientation.

Aspect-based opinion mining, according to Liu (2012), represents a "delicious challenge". Natural languages are very rich and allow to express subjectivity in different ways. Not every opinion is directly expressed and not every aspect appears in a explicit way in the text. For example, in "The camera is expensive", the evaluated aspect is "price", but it is implicit, not being explicitly said in the sentence and, therefore, must be inferred from the context. Therefore, the aspects may be found explicitly and/or implicitly. Explicit aspects are explicit evaluations of one or more properties of the object / target of opinion. For example, the review showed in Figure 3 (kept in Portuguese, the original language), *"Amazing price-benefit relation, it has good digital camera, inclusive for video. Good memory space. What I liked: I received calls up to the riverside. What I did not liked: the sound sometime is low."*The review aspects are: "*price-benefit*", "camera", "video", "memory space", "sound" and "signal", however, the "signal" aspect is implicit. The users used the expression "r*I received calls up to the riverside*" to evaluate the "signal" aspect of a smartphone.



Relação preço beneficio excelente, possui uma boa câmera digital, inclusive para video. Bom espaço de memória.

O que gostei: Recebi chamadas até na beira do Rio Paraná, divisa com o MS

O que não gostei: O som para ouvir as vezes fica baixo.

Figure 3 – Review about a smartphone

Another challenge consists in identifying different aspects that refers to the same object attribute/property. Users recurrently make reference to services or products attributes/properties using different terms. For example, consumers may use the terms "value", "cost", "price" and "investment" to designate the price of a smartphone, or to employ the terms "screen" "glass" and "display" to qualify a same smartphone property. Furthermore, users may employ implicit aspects cues. For example, the expressions "*I received calls up to the riverside*" and "It working anywhere", all these were employed to evaluate the signal property of a smartphone. Another example is the term "compatibility", which was employed to evaluate the operating system of a smartphone. Concomitant, the terms "program", "system" and "application" were also employed. In addition, there is a significant portion of proper nouns applied to refers to the same property of the object. For example, "edward", "edward cullen", "noelle page", "larry" and "bella" are employed to evaluate the "protagonist"' and the terms "josé saramago" and "thalita rebouças" were employed to evaluate the "author" in the book domain. In the camera domain, the terms "sony", "nikon", "fuji" and "benq" are applied to evaluate the "brand" of a camera. Therefore, in opinion mining systems, the aspects clustering is very important, because this task provides the results the results over the veritable properties evaluated by users.

Explicit and implicit aspects clustering task has great relevance for opinion mining systems. However, this task is not trivial. For example, the speakers of a natural language may employed distinct lexical items to refer to the same object in the world. The words "price", "cost", "price-benefict", "value", "cheap", "expensive", may be employed to refer to the same smartphone propriety, for example. To illustrate how this phenomenon affects reviews, see the diagram shown in Figure 4. In this figure, we presented a portion of the groups of aspects identified in the smartphone domain.

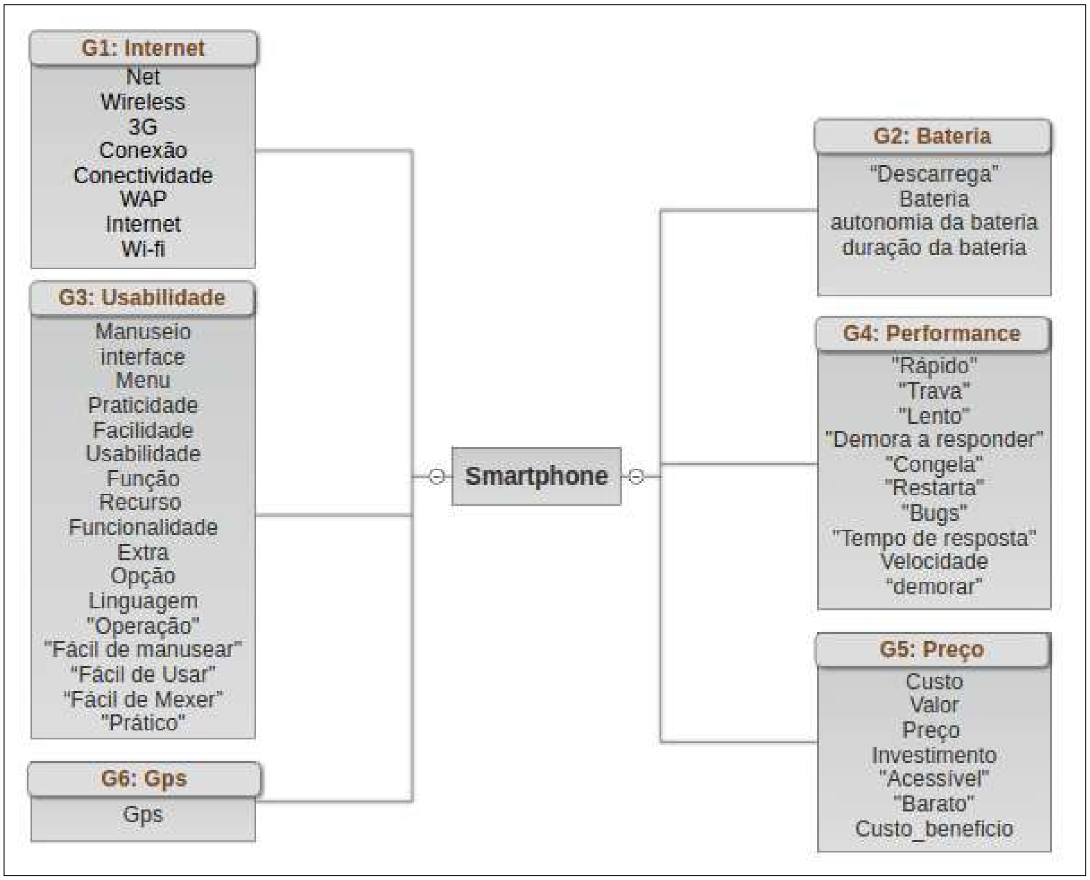


Figure 4 – Some smartphone domain aspects groups

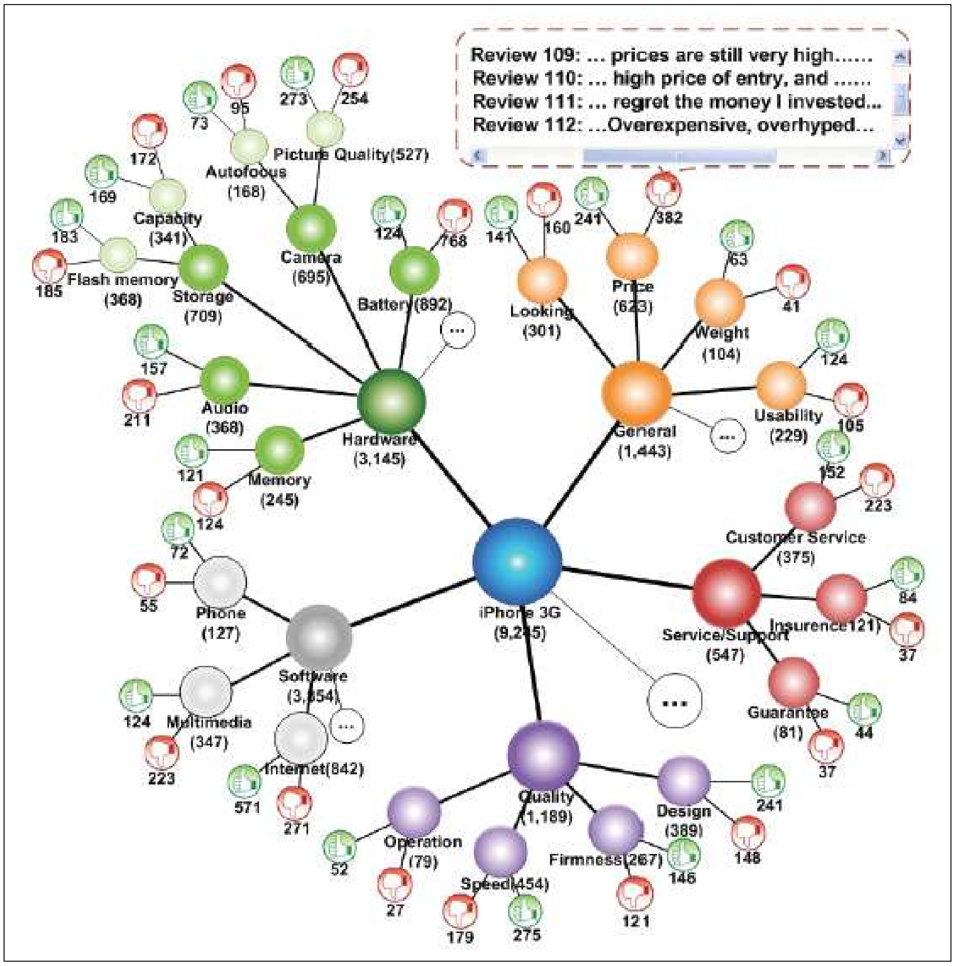
In G1 group, the object property is “internet”. Note that users may use the terms “3g”, “wifi”' and “wireless”, which are internet connection types, to evaluate this property of the device. Note also that these terms, because they represent the domain specificities, are not always found in linguistic-computational resources of the language as wordnets[1]. As well as these terms, other terms may be found, for example, “net”, “internet”, “connection” and “connectivity” to evaluate the same ownership of the mobile device. In G2 group, we find a recurring phenomenon in the *corpus* of user reviews: *aspect attributes*. According to Liu, (2012), aspects have attributes that have aspect properties. For example, the expressions “battery life” is a property of the “battery” aspect. In this case, there is an intrinsic relation between the lexical units. It is also a relation of substring[2]. The G3 group consists of aspects used to evaluate the “usability” property of a smartphone. See how the aspects clustering task is not simple, since many of them are terms that denote vageza (for example, “option”', “function”, “resource”, “extra”'). According to Zipf (1949), most words have multiple definitions, however, more frequent words tend to be more ambiguous. Still on the items of G3 group, we may see expressions indicative of implicit aspects. For example, the terms “It easy use” and “It easy handle”, in addition to the terms “operation” and “practice” are used to designate the aspect of the smartphone. Notice the difficulty for items clustering from distinct nature (verbs, nouns, adjectives) in the same group. In G4 group, users evaluated the performance property of the smartphone. The term “bugs”, derived from foreignism (in portuguese language) and the terms “response time”' and “take time to respond” are indicative of aspects implicit and they are used to evaluate the “performance” propriety. In G5 group, it is interesting to observe 2 behaviors in particular. The first behavior consists of the terms “accessible” and “cheap”, which are terms applied to indicate implicit aspects. See that the terms “accessible” and “cheap” are terms highly ambiguous, and an inference mechanism in the domain is required for correct interpretive correspondence of these items. The second behavior is represented by the term “investimento”. We observe a semantic neologism to which the added value “cost” or “price” is inserted. Finally, the G6 group represents the unit groups in the user review *corpus*. The unit groups represent unique units without semantic correspondence and may be localized in the content plan in reviews. For example, we did not find in the *corpus* another similar aspects with the “gps”' aspect of the smartphone, so this aspect constitutes a unit group. Therefore, the aspects clustering task may be defined by the

[1] Wordnets are large lexical database of a language in which nouns, verbs, adjectives and adverbs, for example, are grouped into synonym synsets, each expressing a distinct concept (MILLER et al., 1990).

[2] Substring is a string that appears within words in the text. For example, the string “ando” is a substring of “walking”.

recognition of correlated aspects semantically, in other words, all of which have interpretive correspondence in a given domain.

Another significant challenge for opinion mining, according to Yu et al. (2011), is that the product reviews are numerous and disorganized. For example, at the *Buscape.com*, the *Smartphone Samsung Galaxy J5 SM-J500M* product has 931 reviews and, for each review, several aspects are evaluated. Thus, consumers will hardly learn all the other consumers' opinions about the product. According to the author, the hierarchical organization of aspects in product reviews would allow a better structuring of this data, so that it becomes intelligible for both machines and humans. A example of hierarchical organization of aspects in product reviews is shown in Figure 5. This work was proposed by Yu et al. (2011). The authors proposed a method based in linguistic and statistical knowledge in order to provide hierarchical organization of aspects from product reviews. There were considered 9.245 reviews on an iPhone 3G. Note that the hierarchical organization of the evaluated aspects about the iPhone 3G product is clear, unambiguous and may be easily understood for other consumers.



Review 109: ... prices are still very high......
Review 110: ... high price of entry, and ......
Review 111: ... regret the money I invested...
Review 112: ...Overexpensive, overhyped...

Figure 5 – Hierarchical organization of aspects (YU et al., 2011)

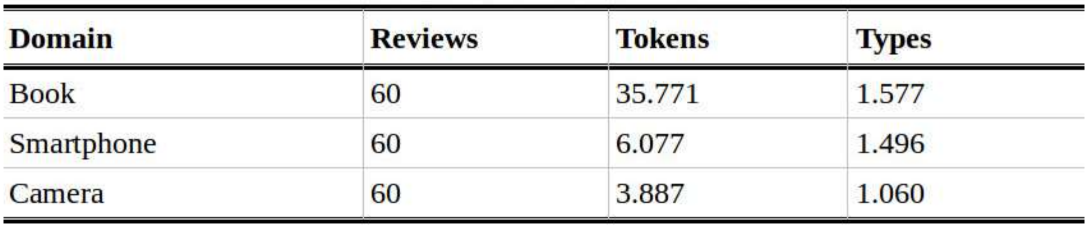
In this scenario, due to all the challenges, we present a *corpus* study of opinion aspects, regarding both their clustering and hierarchical organization. We analyze reviews for books and electronic products, looking for linguistic patterns and convergences and divergences across these domains. We expect that such investigation may help characterizing the involved tasks and provide valuable reference data for future developments and evaluations in the area. The rest of this paper is organized as follows. Section 3 describes the dataset and the analysis method. Section 4 presents the achieved results and learned lessons. Some final remarks are made in Section 5.

3 *Corpus* study

3.1 The dataset

The dataset overview is shown in Table 1. According to Zhao e Li (2009), most of the existing opinion mining initiatives are based on product reviews because reviews usually focus on specific products and contain little irrelevant information. Therefore, we randomly selected 60 smartphone and 60 camera reviews from the Buscape *corpus* (HARTMANN et al., 2007) and 60 book reviews from the ReLi *corpus* (FREITAS et al., 2012). We manually analyzed the data behavior in each domain. The Buscape *corpus* is composed of product reviews in Portuguese language for cameras, notebooks, telephones, TVs etc. In this *corpus*, the reviews are partially structured, with sections for "overall impression", "what I liked" and "what I did not like". For example, see the following camera review: *"Amazing, even today everyone is impressed by its size and beauty beyond perfect pictures that can be taken up 6.3 megapixels! What I liked: slim, practice and light. What I did liked: None!"* . One may note that several aspects were evaluated in this review, but some are not explicit. For example, the terms "beauty", "slim", "practice" and "light" are *clues* that indicate the implicit aspects "design", "size", "usability" and "weight", respectively. The ReLi *corpus* consists of book reviews, that are also in Portuguese language. As an example, one may find the following review: *"Amazing book, very different of what I imagined. Despite being old, it is good reading with the very modern language."*. We chose to select only 60 reviews for each domain, mainly because ours is a study carried out manually, from the qualitative and quantitative approach. Our main hypothesis is that for each domain there are different linguistic behaviors and phenomena.

Table 1



Domain	Reviews	Tokens	Types
Book	60	35.771	1.577
Smartphone	60	6.077	1.496
Camera	60	3.887	1.060

In the book domain, according to the Table 1, there was a significant spike in the number of tokens when compared to smartphone and camera domains. In this domain, we characterize an expressive number of irrelevant content. There were identified 52,01% relevant content and 47,98% irrelevant content. Smartphone and camera domains, the irrelevant content was not statistically significant.

3.2 The analysis method

In this work, the main purpose is to investigate the clustering and hierarchical organization of opinion aspects. Our main motivation with this *corpus* study was to understand the characteristics and challenges in the process of recognizing groups of aspects and the semantic organization of these groups. Our goal is to propose linguistically motivated solutions for opinion mining systems. We have selected 3 distinct domains: smartphone, camera and book in order to understand the convergence and divergence of behavior between domains. The empirical analysis was performed manually and will serve as a reference (human) for the evaluation of the proposed automatic methods, as well as a resource for the future research. In our study, we presented several quantitative and qualitative data on the aspects clustering task, besides some empirical evidence that the linguistic behavior varies between domains and that these variations have strong linkages with the knowledge specificities of a domain and with the profiles of the writer/ user which produces the content. Figure 6 illustrates the clustering process, which was manually performed.

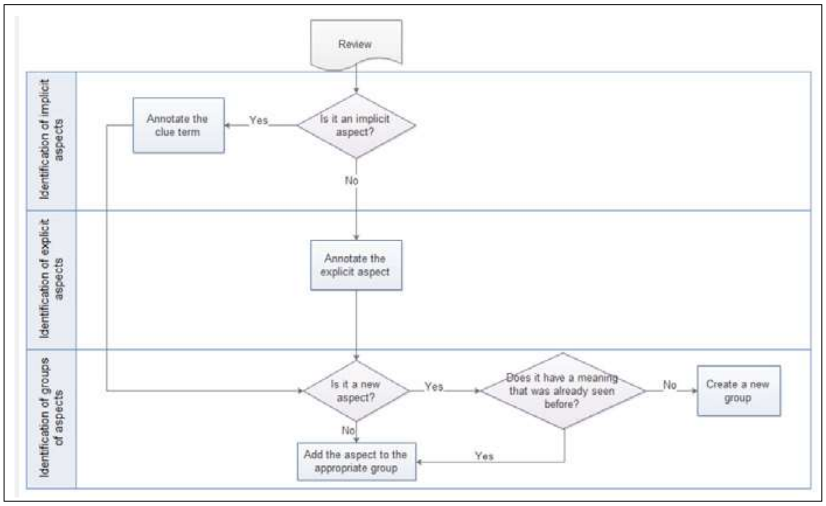

Review
Identification of implicit aspects
Annotate the clue term
Yes
Is it an implicit aspect?
No
Identification of explicit aspects
Annotate the explicit aspect
Identification of groups of aspects
Is it a new aspect?
Yes
Does it have a meaning that was already seen before?
No
Create a new group
No
Add the aspect to the appropriate group
Yes

Figure 6 – Clustering product aspects

For the 180 reviews, a human labeled the implicit and explicit aspects. In the identification of implicit aspects were labeled the clue terms that indicated the aspects. For example, in "*This camera is expensive"*, the evaluated aspect is "price", but it is implicit. The term "expensive" is the clue term. The identification of explicit aspects were directly labeled the aspects. For example, in "*The history of the book is bad"*, "history" is an explicit aspect. In the last stage, the aspects were clustered the that had similar meaning but with different wording, in order to identify groups. For example, the "cost", "value", "price" and "investment" aspects form an unique group.

We also modeled the progression of this process of clustering product aspects, looking for a "learning curve" (shown in Appendix 1). Our objective was to measure the behavior of the aspects clustering task in order to identify the *stabilization point* for identifying new groups of aspects in a domain. The curves are shown in Figures 7, 8 and 9. The ***X*** axis of the learning curves represents the number of reviews analyzed and the ***Y*** axis the number of new groups identified. For example, The ***X*** axis, the number 1, shown in Figure 7, there were recognition 8 groups of aspects, as shown by ***Y*** axis. After the analysis of the first 10 reviews, 33 groups of aspects were recognized, and so on. For smartphone, digital camera and book domains, an average of 40 reviews are required for learning groups of representative aspects of the domain.

Once clustering was ready, the obtained groups were manually organized in hierarchies (one for each domain, shown in Appendix 2). We compared our obtained hierarchies with other available hierarchies in the area. We also identified the groups of aspects with the highest number of evaluations in the smartphone, camera and book domains, looking for a "prototypical groups" (shown in Appendix 3). For example, for the smartphone domain, some groups of prototypes are: "smartphone", "usability", "design", "value", "battery", "brand" etc.

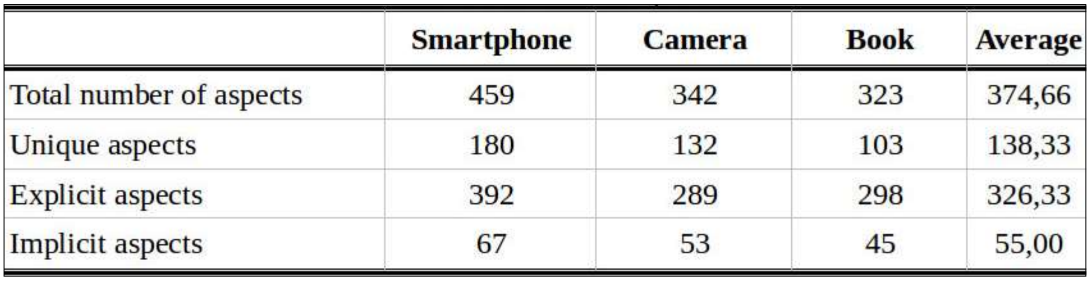
4 Results

As explained before, we manually analyzed the product reviews and could observe some very interesting things. The results demonstrated that product reviews may contain portions of irrelevant information, i.e., information that is not directly related to the opinions about the products. The book domain showed 47.98% of irrelevant content, when users comment about the books but do not express any opinion or sentiment. However, for smartphone and camera domains, there was no significant value of irrelevant content.

We could notice that the user profile influences the review informational status[3]. We observed that the smartphone and camera domains present more aspects and groups of aspects than the book domain, as shown in Table 2. Note that the total aspects number in each domain and the average number of aspects in reviews seems to us to be relevant empirical evidence for the relationship between user profile and informational status. Smartphones and cameras are popular technological products and their aspects are more easily identified by non-expert users. Therefore, "expert users" have a greater level of information, in other words, these users have more knowledge about the domain, which allows them to evaluate a larger number of aspects of the evaluated entity. In the book domain, the users often are "just readers" and non-expert users in literature or literary critic. Therefore, they usually do not care about the book technical aspects (such as "size" or "paper type"). These users have been able to evaluate a limited number of product aspects, generally prototypical aspects of the books. It is also interesting that the vocabulary in book reviews are not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level. More "adult" books have more sophisticated reviews, with better language, while the reviews of "teenager" books are more often marked by the orality and informal language. These results demonstrate how complex the tasks of opinion mining are, especially aspect-based opinion mining. Opinion mining systems that does not consider the linguistic behavior and/or domain specificities as a processing criterion, for example, incurs the risk of classifying aspects that were not evaluated by the user, so they will return a result that is not in accordance with the reality presented in the review. In addition, we noticed in reviews content that had opinion/sentiment explicitly, they was accompanied mainly by psychological verbs, as occurs, for example, in "I found the history a little stopped", "I loved the book" and "Although I did not like the book", without necessarily having adjectives.

[3] According to Koch (2009), the informativeness of a text is associated to its ability to present new and unexpected information.

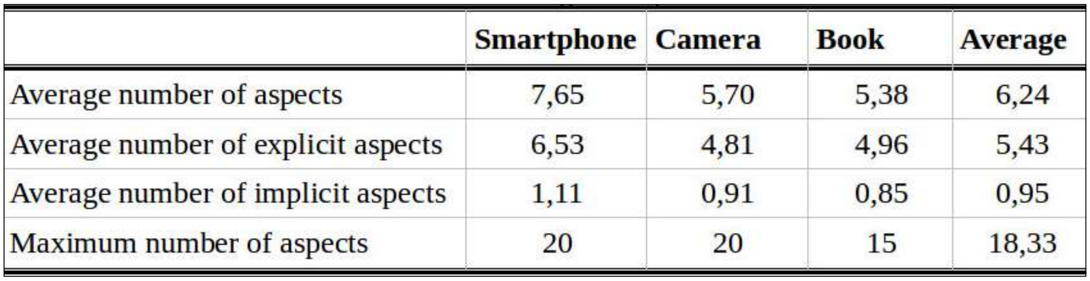
Table 2



	Smartphone	Camera	Book	Average
Total number of aspects	459	342	323	374,66
Unique aspects	180	132	103	138,33
Explicit aspects	392	289	298	326,33
Implicit aspects	67	53	45	55,00

Overall, 87.08% of the aspects are explicit and 12.91% are implicit in the domains. Furthermore, a product review is composed of, on average, 6 aspects, and it may have at least 1 implicit aspect (see Table 3). We also identified product reviews with the maximum of 20 aspects and the maximum of 5 implicit aspects.

Table 3

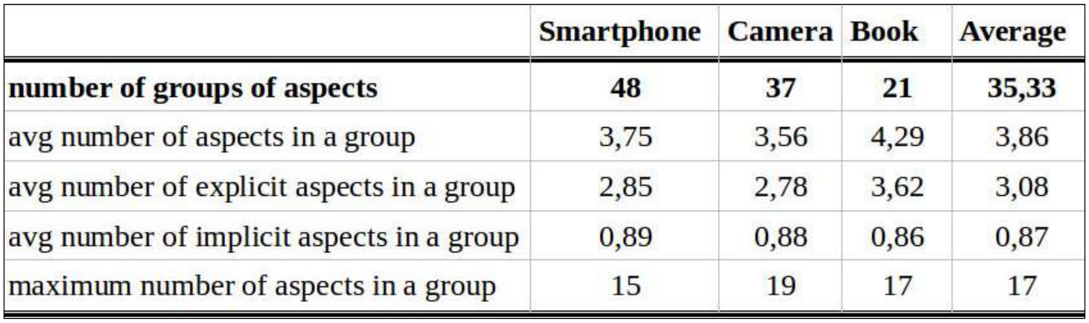


	Smartphone	Camera	Book	Average
Average number of aspects	7,65	5,70	5,38	6,24
Average number of explicit aspects	6,53	4,81	4,96	5,43
Average number of implicit aspects	1,11	0,91	0,85	0,95
Maximum number of aspects	20	20	15	18,33

We also perform a mapping of the grammatical classes of the terms indicative of implicit aspects. We divide the indicative terms of aspects into 2 classes, *nominal* and *verbal,* in order to measure the proportion of each one of these classes in the analyzed domains. In the "nominal" class , we framed "non-verbal" lexical items, in other words, it is nouns, adjectives, adverbs etc. In the verbal class, verbal lexical items were framed, in other words, it is verbs. In the smartphone domain, 73,68% are nominal implicit aspects and 26,31% are verbal implicit aspects. In the camera domain, 69,56% are nominal implicit aspects and 30,43% are verbal implicit aspects. Lastly, in the book domain, 50% are nominal implicit aspects and 50% are verbal implicit aspects.

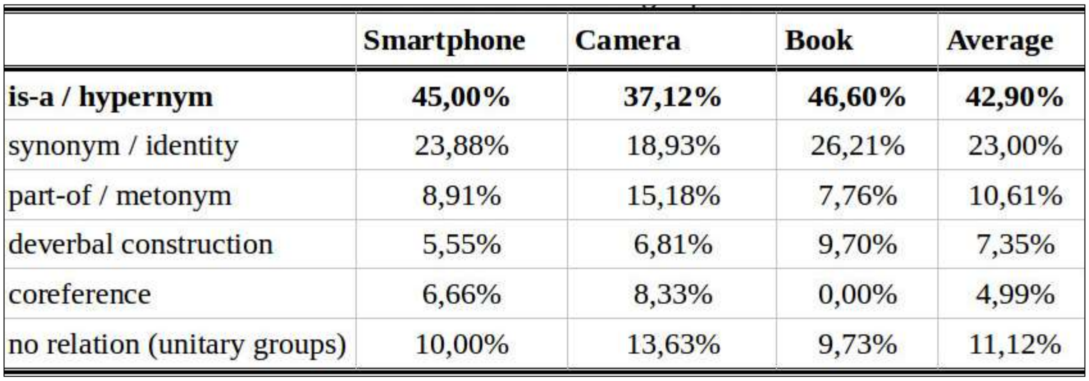
Regarding the clustering step, we identified, on average, 3,08 explicit aspects and 0,77 implicit aspects per group. Some groups presented the maximum of 19 aspects, as shown in Table 4. In these groups (those that are not unitary, i.e., that contain more than one aspect), the predominant relation between 2 aspects is of the *is-a* / hypernym (or hyponym, dependending of the direction of the relation) type (e.g., between the aspects "equipment" and "product"), followed by synonym ("price" and "cost") or identity (when there is a single aspect without a direct corresponding synonym in the group), part-of / metonym (or holonym) (“key” and “keyboard”), deverbal construction (“reflect” and “reflection”) and coreference (“manufacturer” and “brand”). The remaining cases are formed by unitary groups, with only one aspect (without relations, therefore). Table 5 shows the distribution of these relations.

Table 4



	Smartphone	Camera	Book	Average
number of groups of aspects	**48**	**37**	**21**	**35,33**
avg number of aspects in a group	3,75	3,56	4,29	3,86
avg number of explicit aspects in a group	2,85	2,78	3,62	3,08
avg number of implicit aspects in a group	0,89	0,88	0,86	0,87
maximum number of aspects in a group	15	19	17	17

Table 5



	Smartphone	Camera	Book	Average
is-a / hypernym	**45,00%**	**37,12%**	**46,60%**	**42,90%**
synonym / identity	23,88%	18,93%	26,21%	23,00%
part-of / metonym	8,91%	15,18%	7,76%	10,61%
deverbal construction	5,55%	6,81%	9,70%	7,35%
coreference	6,66%	8,33%	0,00%	4,99%
no relation (unitary groups)	10,00%	13,63%	9,73%	11,12%

We found several challenges in the analysis: *(i) the inherent ambiguity of the natural languages*, occurring, for example, for the terms “function”, “resource” and “application”, that are used to refer to the same smartphone application; *(ii) the specificities of the domain*, as each domain requires specific background knowledge; *(iii) the implicit aspects,* as the implicit aspect identification task is not always intuitive; *(iv) the aspects outside the domain,* as the terms “delivery”, “technical assistance” and “SAC”, which, although have been evaluated, are not directly related to the products. Our study also showed that it is necessary the analysis of 40 reviews, on average, to learn/identify most of the relevant aspects in a given domain. The “learning curves” (shown in Appendix 1), represent the learning behavior of groups of aspects for the analyzed domains, that is the amount of new groups of aspects learned at each review. We also hierarchically organized the identified groups of aspects (see in Appendix 2) and compared our hierarchies with the hierarchies proposed by Condori (2015), Acir et al. (2006) and Goulart and Montardo (2007). In the hierarchies in the literature, the relations of the type *is-a* are more often used. However, we observed that reviews are predominantly composed by *part-of* relations. Furthermore, the hierarchies in the literature do not represent all the domain specificities.

5 Final remarks

As shown above, clustering product aspects and building their hierarchical organizations are not simple tasks. There are several challenges to overcome. The results demonstrated that product reviews may contain a significant portion of irrelevant content and that informational status may be influenced by the user profile. The vocabulary in book reviews is not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level, which results in varied writing behavior. In addition, it was found that, for a good domain coverage, at least 40 reviews are required, on average. We also observed that, on average, some domains may have more identifiable aspects. The aspect groups and the hierarchies will be made available for research purposes. We expect that automatic methods for opinion mining may be trained and/or evaluated over such datasets.

Acknowledgments

The authors are grateful to FAPESP and CAPES for supporting this work.

References

ACIR, S.; ZHANG, D.; SIMOFF, S.; DEBENHAM, J. Recommender system based on consumer product reviews. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, 2006. *Proceedings*... Washington: [s/n], 2003, p. 719-723.

AVANCO, L.; NUNES, G. M. V. Lexicon-based Sentiment Analysis for reviews of products in Brazilian Portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS. *Proceedings*... São Carlos: [s/n], 2014, p. 277-281.

BHUIYANET, T.; XU, Y.; JOSANG, A. state-of-the-art review on Opinion Mining from online customers feedback. In: ASIA-PACIFIC COMPLEX SYSTEMS CONFERENCE, 9. *Proceedings*... Tokyo: [s/n], 2009, p. 385-390.

CONDORI, R. E. L. *Sumarização automática de opiniões baseada em aspectos*. Dissertação (mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015.

FREITAS, C.; MOTTA, E.; MILIDIU, R.; CESAR, J. Vampiro que brilha... rá! desafios na anotação de opinião em um córpus de resenhas de livros. In: ENCONTRO DE LINGUÍSTICA DE *CORPUS*, 11. *Anais*... São Carlos: [s/n], 2012.

GOULART, R. R. V.; MONTARDO, S. P. Os mecanismos de busca e suas implicações em Comunicação e Marketing. In: CONGRESSO NACIONAL DE HISTÓRIA DA MÍDIA, 5. *Anais*... São Paulo: [s/n], 2007, p. 478-514.

HARTMANN, N.; AVANÇO, L.; BALAGE, P.; DURAN, M.; NUNES, M. D. G. V.; PARDO, T.; ALUÍSIO, S. A large *corpus* of product reviews in Portuguese: tackling out-of vocabulary words. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9. *Proceedings*... Reykjavik: [s/n], 2007, p. 3865-3871.

KOCH, I. G. V. *Introdução a Linguística Textual*. 2. ed. São Paulo: Martins Fontes, 2009.

LIU, B. *Sentiment Analysis and Opinion Mining*. 1. ed. San Rafael: Morgan & Claypool Publishers, 2012.

MILLER, G. A., BECKWITH, R., FELBAUM, C., GROSS, D. and MILLER, K. WordNet: An on-line lexical database. *International Journal of Lexicography*, v. 3, p. 235-244, 1990.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment classification using machine learning techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. *Proceedings*... Vol. 10. Stroudsburg: [s/n], 2002, p. 79-86.

TABOADA, M. Sentiment Analysis: an overview from Linguistics. *Annual Review of Linguistics*, v. 2, p. 325-347, 2016.

YU, J.; ZHA, Z.; MENG, W.; WANG, K.; CHUA, T. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. *Proceedings*... Stroudsburg: [s/n], 2011, p. 140-150.

ZHAO, L.; LI, C. Ontology based Opinion Mining for movie reviews. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE SCIENCE, ENGINEERING AND MANAGEMENT, 3. *Proceedings*... Berlin: Springer-Verlag, 2009, p. 204-214,

ZIPF, G. K. *Human behavior and the principle of least effort*. 1. ed. Cambridge: Addison-Wesley Press, 1949.

Appendix 1

We present below the learning curves for the identification of groups of aspects. As an illustration of how to interpret these graphics, in Figure 2, one may see that, after have analyzed 2 smartphone reviews, we could identify 10 groups of aspects; after 60 reviews, we end up with 48 groups.

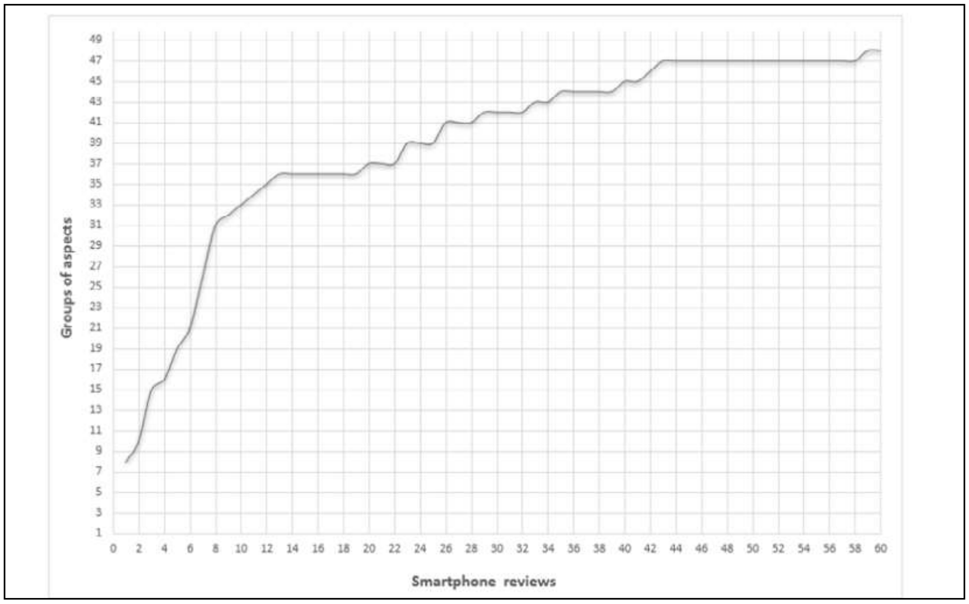


Figure 7 – Learning curve for the smartphone domain

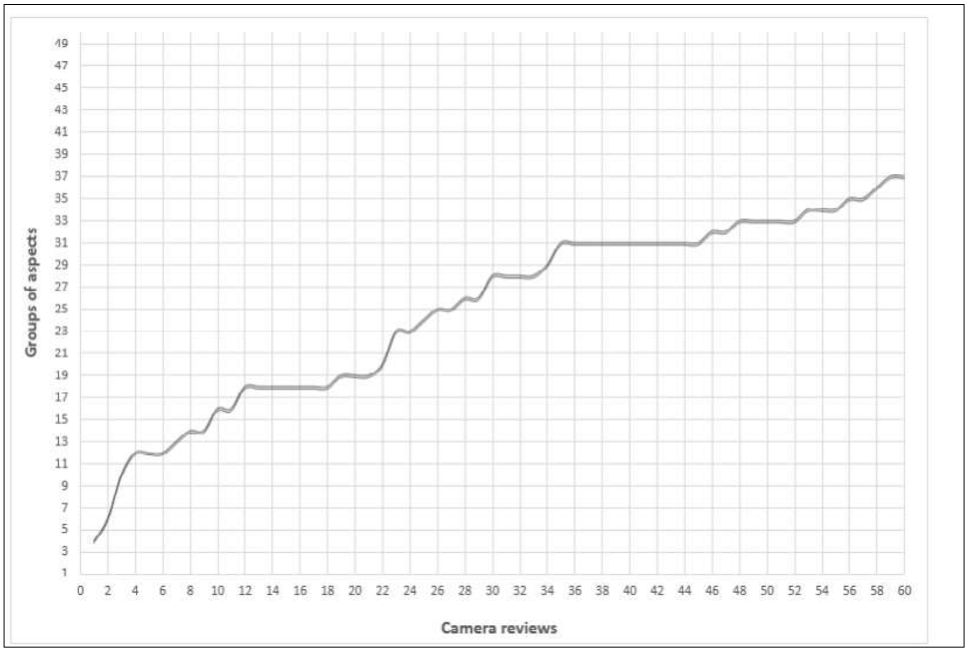


Figure 8 – Learning curve for the camera domain

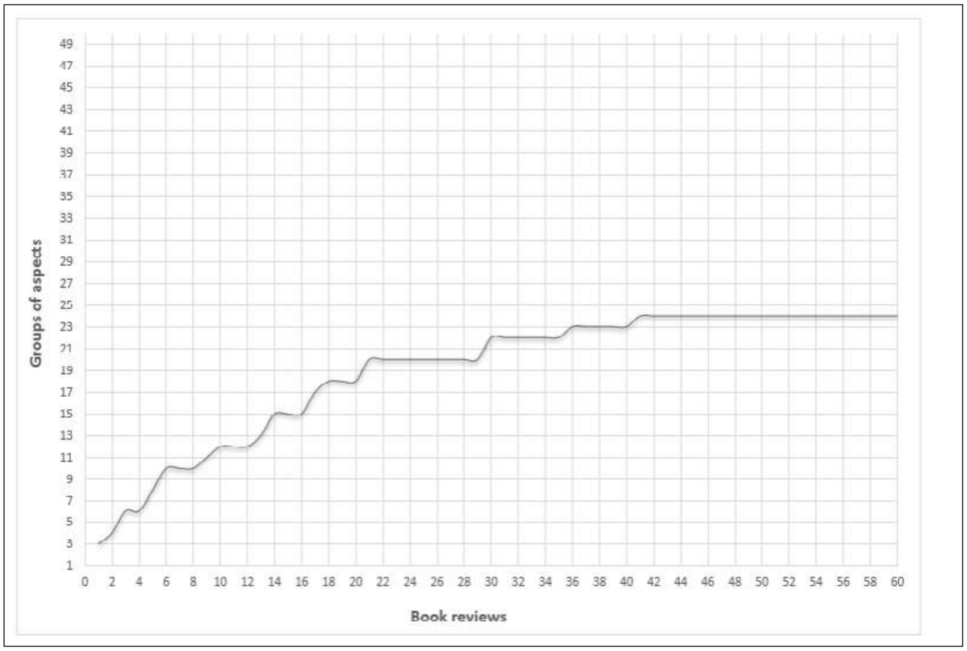


Figure 9 – Learning curve for the book domain

Appendix 2

We present below the hierarchies obtained for the smartphone, camera and book domains, where each circle represents a group of aspects. For each group, we show only the most representative word. We show them in Portuguese because the *corpus* is in this language.

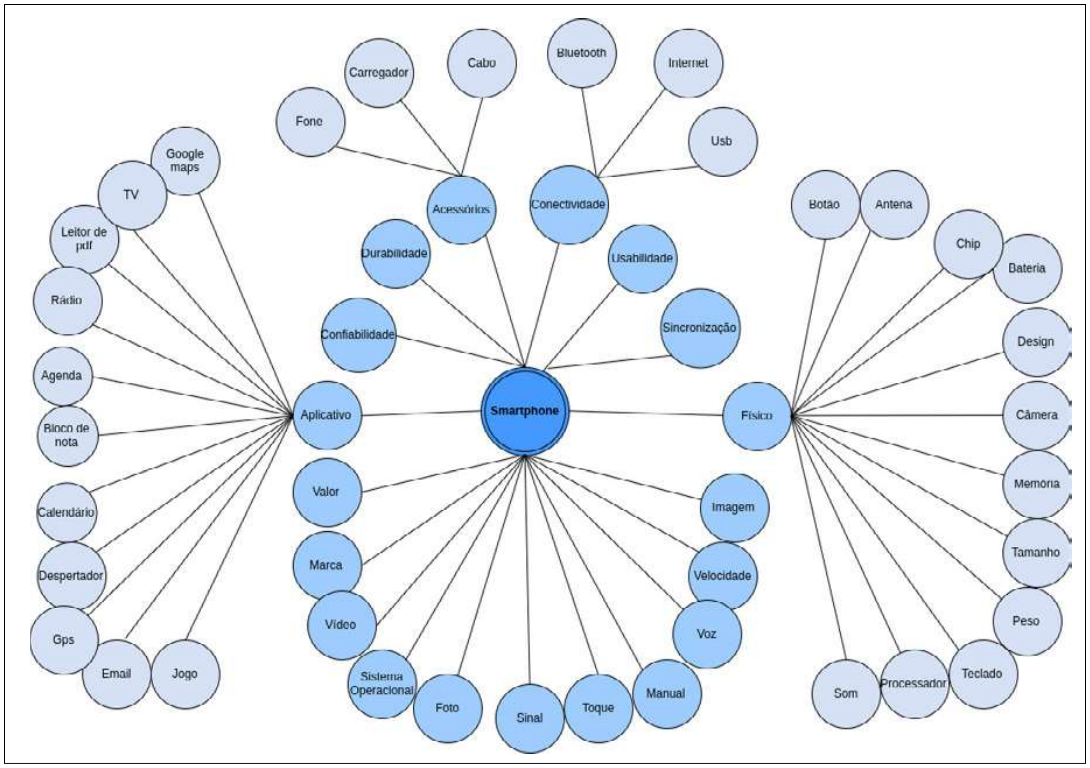


Figure 10 – Hierarchy for the smartphone domain

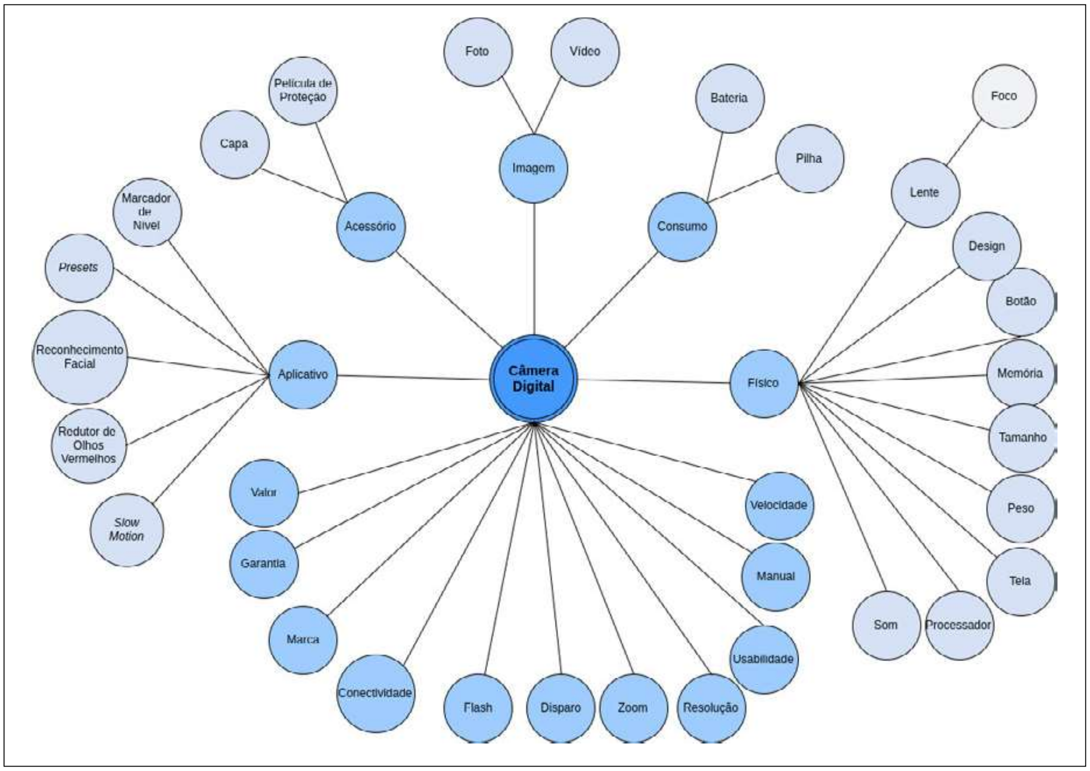


Figure 11 – Hierarchy for the camera domain

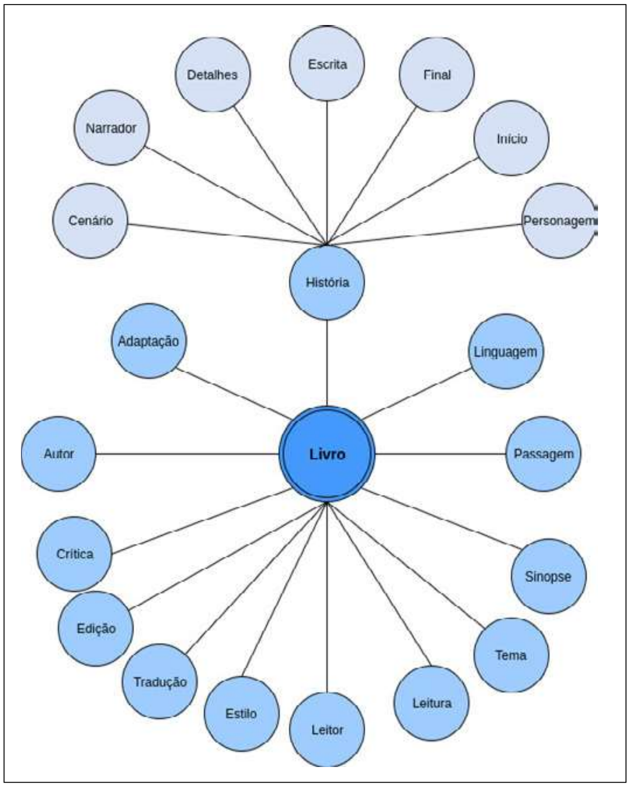


Figure 12 – Hierarchy for the book domain

Appendix 3

We present below the prototypical groups on the smartphone, camera and book domains. We show them in Portuguese because the *corpus* is in this language. Items marked with black color represent prototypical groups.

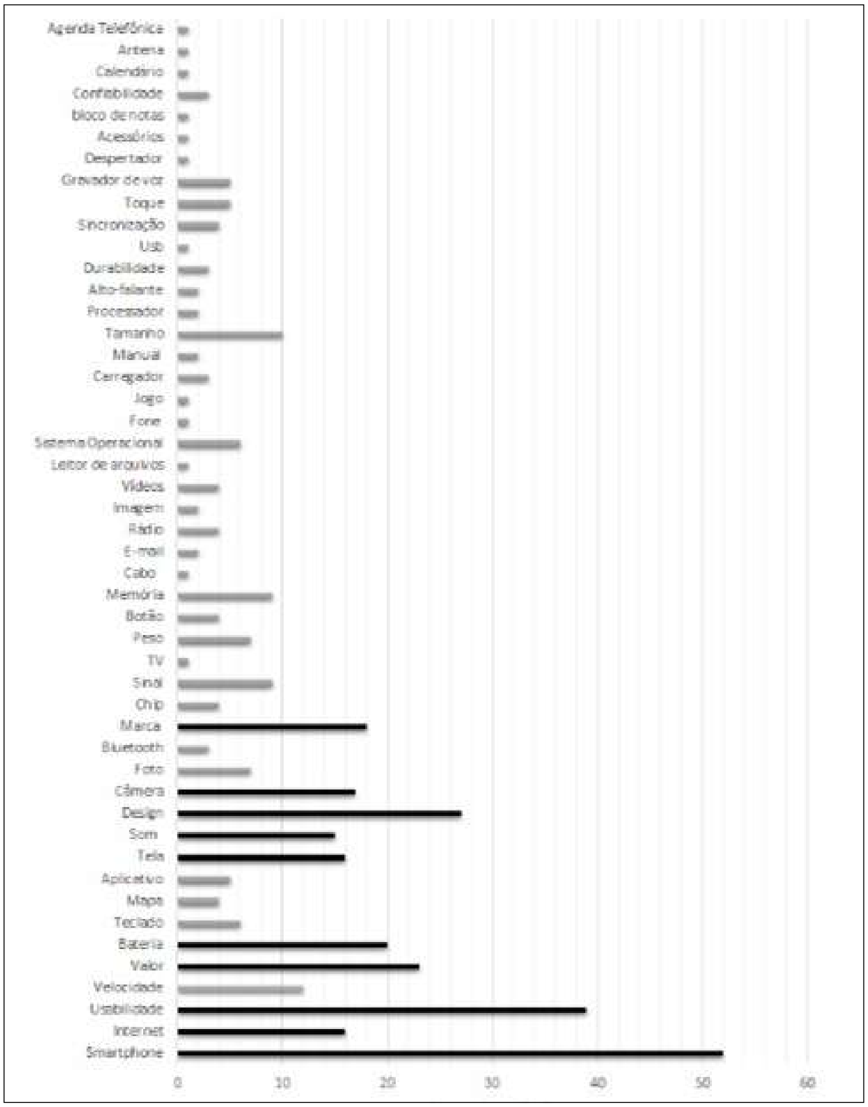


Figure 13 – Prototypical groups for the smartphone domain

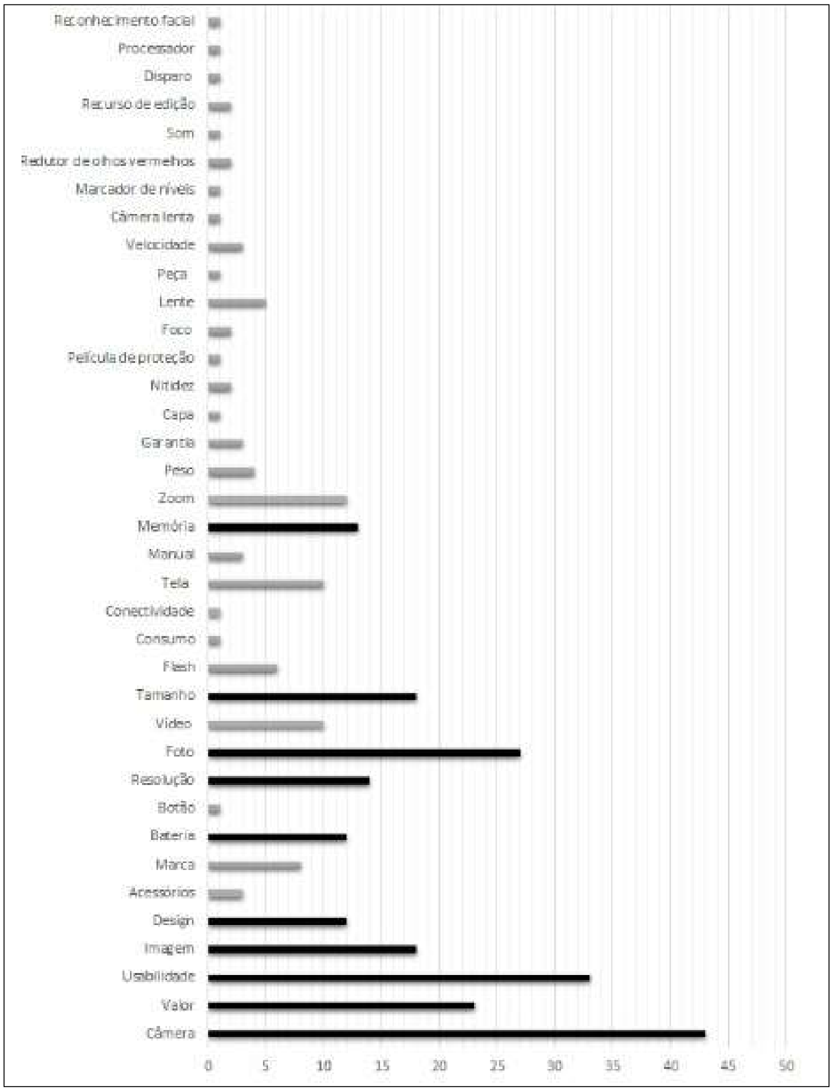


Figure 14 – Prototypical groups for the camera domain

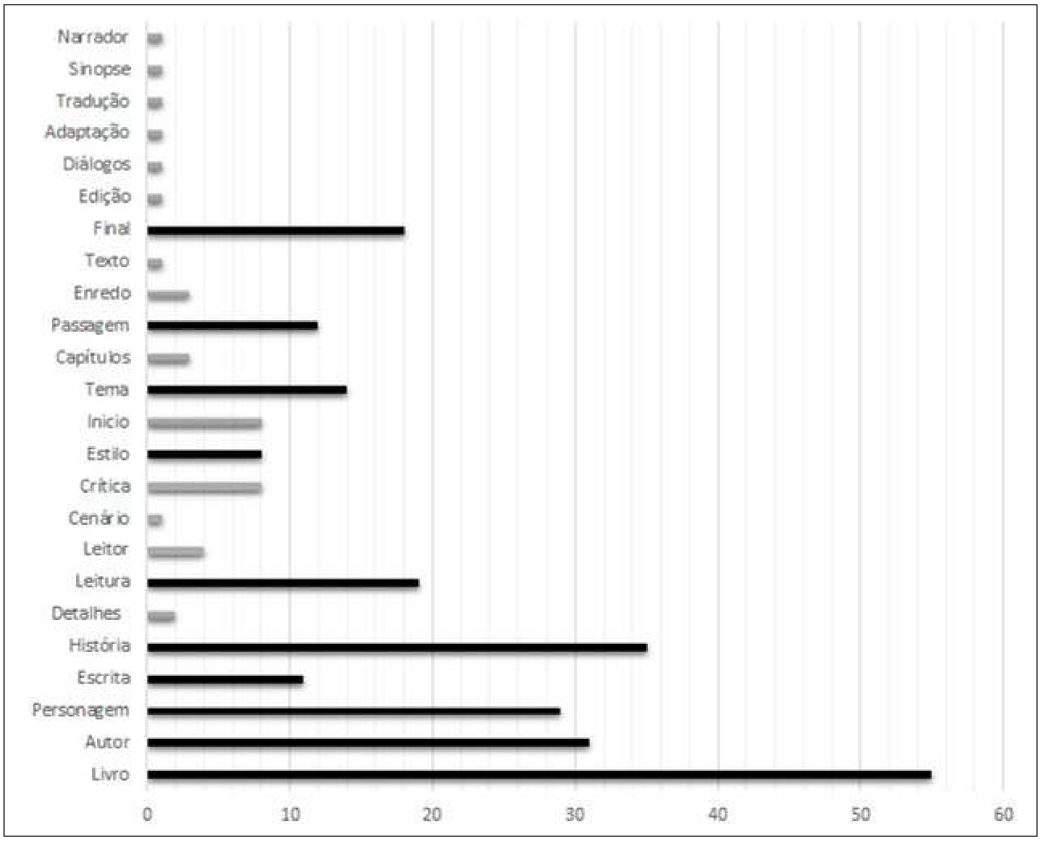


Figure 15 – Prototypical groups for the book domain