



Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

Leveraging User Opinions for Product Catalog Enrichment

Tiago Eugenio de Melo

December, 2018
Manaus, Amazonas



Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

Tiago Eugenio de Melo

Leveraging User Opinions for Product Catalog Enrichment

Tese apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para a obtenção do título de Doutor em Informática.

Advisor: Dr. Altigran Soares da Silva

December, 2018
Manaus, Amazonas

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

M528I Melo, Tiago Eugenio de
Leveraging user opinions for product catalog enrichment / Tiago
Eugenio de Melo. 2018
88 f.: il. color; 31 cm.

Orientador: Altigran Soares da Silva
Tese (Doutorado em Informática) - Universidade Federal do
Amazonas.

1. Sentiment Analysis. 2. Opinion Mining. 3. Product Catalog
Enrichment. 4. Aspect-Base Summarization. 5. Online Reviews. I.
Silva, Altigran Soares da II. Universidade Federal do Amazonas III.
Título



PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA



FOLHA DE APROVAÇÃO

"Leveraging User Opinions for Product Catalog Enrichment"

TIAGO EUGENIO DE MELO

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:


Prof. Altigran Soares da Silva - PRESIDENTE


Prof. Edlener Silva de Moura - MEMBRO INTERNO


Prof. Pável Pereira Calado - MEMBRO EXTERNO


Profa. Viviane Pereira Moreira - MEMBRO EXTERNO


Prof. Luciano de Andrade Barbosa - MEMBRO EXTERNO

Manaus, 03 de Dezembro de 2018

Agradecimentos

Primeiramente, eu gostaria de expressar os meus sinceros agradecimentos ao meu orientador, professor Altigran Soares. O seu apoio contínuo, confiança, orientação, disponibilidade e generosidade me ajudaram a concluir com sucesso o meu doutorado.

Meus agradecimentos também se estendem aos membros da banca examinadora: Edleno de Moura, Luciano Barbosa, Pável Calado e Viviane Moreira. Em especial, gostaria de agradecer ao Pável Calado pela oportunidade de ter realizado doutorado-sanduíche na Universidade de Lisboa. Foi um período de grande aprendizado e as sugestões ao meu trabalho contribuíram para o amadurecimento da minha pesquisa.

Agradeço também aos demais professores e funcionários do IComp/UFAM pelo apoio e por manter esse programa de pós-graduação em nível de excelência.

Meus agradecimentos também à Universidade do Estado do Amazonas (UEA) por proporcionar o suporte financeiro para este trabalho e também aos colegas professores e coordenadores do Núcleo de Computação da Escola Superior de Tecnologia da UEA.

Eu gostaria de agradecer aos meus amigos que colaboraram durante este trabalho: César, Henry, Marcelo e Márcio.

Agradeço imensamente a minha família e, em especial, aos meus pais, Airton e Valdereza, e aos meus irmãos, Diogo e Laila, pelo amor, apoio e torcida em todos os desafios da minha vida.

A minha mulher, Virgínia, gostaria de agradecer pelo amor, atenção e apoio, e também ao meu filho, Dimitri, pela compreensão dos momentos necessários em que estive ausente. Dedico este trabalho a vocês.

Abstract

A large number of people post reviews on the products of all types, which are offered online. In these reviews, people express their opinion regarding these products and their features. Consequently, a large number of opinions are available, which can be a valuable source of knowledge for decision-making for manufacturers as well as customers. From these opinions, manufacturers can obtain immediate feedback to improve the quality of their products and customers are able to obtain assessments from reviews prior to purchasing a product. However, as it is common in many types of social media, the sheer volume of available reviews for each product normally exceeds the human processing capacity and can, thus, become a major barrier to its effective use. The question that now arises is how to structure opinions so that they can be effectively used by customers and manufacturers. Traditional methods of organizing a large number of product reviews aim at creating an opinion summary. However, these methods are inadequate to address customer queries on the most relevant product characteristics. In particular, in current methods, the opinions are arbitrarily clustered by aspect expressions, causing these clusters to not necessarily align with relevant product characteristics. We claim that the most important product characteristics for people are represented by the attributes of the product catalogs and the process of organizing opinions should be guided by the attributes of the product catalogs. Therefore, in this thesis, we formulated and investigated the following problem: *enriching product catalogs with user opinions at the attribute granularity level as a new form of opinion summarization*. Grouping opinions around the attributes of the product catalog also allows the catalog to be enriched with these opinions with the passage of time. To deal with this novel problem, in this thesis, we started by investigating the impacts of attributes of product catalogs on user opinions. In this investigation, we used a large collection of data. The experimental results indicate that user opinions are significantly influenced by product attributes. In addition, we presented a new approach comprising of two phases: opinion extraction and opinion mapping. Based on this approach, we developed two distinct methods. For the first method, named *AspectLink*, an unsupervised strategy has been adopted. For the second method, named *OpinionLink*, a supervised strategy has been adopted. To verify the effectiveness of the methods, an extensive experimental evaluation was conducted which demonstrated the effectiveness of the proposed methods. Furthermore, a bootstrapping strategy was proposed to train the classifiers of *OpinionLink* in order to reduce the dependence on training data. Finally, the supervised method was applied as a full pipeline, and the experimental results demonstrated the feasibility of using this method in real and large-scale applications. To properly evaluate the methods developed in this study, the experimental datasets,

non-existent in the literature, were developed, which are available as another contribution. We also developed a practical application to showcase some proposal ideas in this thesis.

Contents

List of Figures	vi
List of Tables	vii
List of Acronyms	viii
1 Introduction	1
1.1 Problem Statement	3
1.2 Research Questions (RQ)	3
1.3 Our Approach	4
1.4 Contributions	6
1.5 Thesis Organization	6
2 Related Work	8
2.1 Enriching Databases	8
2.2 Opinion Summarization	10
2.3 Product Attributes and User Reviews	12
3 Concepts and Terminology	13
3.1 Product Catalog	13
3.2 Reviews, Sentences and Opinions	14
3.3 Enriching Product Catalogs with Opinions	16
4 <i>AspectLink</i>	18
4.1 Overview	18
4.2 Opinion Extraction	19
4.3 Attribute Descriptors	20
4.4 Matching Aspects and Descriptors	22
4.5 The AspectLink Algorithm	25
4.6 Summary	28
5 <i>OpinionLink</i>	29
5.1 Overview	29
5.2 Opinion Extraction	31
5.2.1 Identifying Direct Opinionated Sentences	31

5.3	Opinion Mapping	34
5.3.1	Opinion-Mapping Algorithm	34
5.3.2	Sentence Core Segments	35
5.4	Bootstrapping Method	36
5.5	Summary	37
6	Experimental Datasets	39
6.1	Overview	39
6.2	BestBuy Collection	40
6.3	Amazon Collection	42
6.4	Summary	45
7	An Analysis on Mentions of Attributes in User Reviews	47
7.1	Background	47
7.2	Research Hypotheses	48
7.3	Results	49
7.3.1	Use of Directed Opinionated Sentences (DOS)	49
7.3.2	Distribution of Sentences among Kinds of Targets	50
7.3.3	Distribution of Aspect Expressions	51
7.3.4	Distribution of Sentences among Product Attributes	51
7.3.5	Diversity of Aspect Expressions over Attributes	52
7.4	Summary	54
8	Experimental Validation	55
8.1	Evaluation Metrics	55
8.2	<i>AspectLink</i> – Experimental Evaluation	55
8.2.1	General Results	56
8.2.2	Common vs. Expanded Descriptors	57
8.2.3	Similarity Functions	58
8.2.4	Estimating parameter Θ_3	58
8.2.5	Discussion	59
8.3	<i>OpinionLink</i> – Experimental Evaluation	60
8.3.1	Identifying Direct Opinionated Sentences	60
8.3.2	Opinion Mapping	62
8.3.3	End-to-End Results	65
8.3.4	<i>OpinionLink</i> in Large Scale	65
8.3.5	Bootstrapping Evaluation	66
8.4	Comparative Analysis	67
8.4.1	Identifying Direct Opinionated Sentences	67
8.4.2	Opinion Mapping	68
8.5	Summary	70
9	The Contender Tool	71
9.1	Motivation	71

9.2	Contender Overview	73
9.3	Demonstration	74
9.4	System Setting	75
9.5	Summary	75
10	Conclusion and Future Work	77
10.1	Future Work	78
	Bibliography	81

List of Figures

1.1	Example of product catalog enrichment with opinions.	4
3.1	Example of product from a typical catalog of the Cell Phone category.	14
3.2	Example of review for product presented in Figure 3.1.	15
3.3	Example of product catalog enriched with opinions presented in Figure 3.2.	17
4.1	<i>AspectLink</i> Overview.	19
6.1	Example of a set of opinions labeled.	42
7.1	Percentage (%) of sentences of each type in user reviews.	50
7.2	Distribution of sentences among targets.	50
7.3	Distribution of the top 100 aspects among the three kinds of targets.	51
7.4	Distribution of sentences among the attributes they refer to. Labels are product attributes.	52
7.5	Distribution of different aspect expressions according the product attributes presented in Table 6.2.	53
8.1	Precision, recall and F_1 results comparing <i>AspectLink</i> to each similarity function applied individually.	59
8.2	Influence of threshold Θ_3 in our method for each category of products.	59
8.3	Confusion Matrices for opinion-mapping task. Each label represents an attribute: Gen (General), Ima (Imaging), Oth (Other), Pro (Processor), Bat (Battery), Dis (Display), Cam (Camera), Pri (Price), Vid (Video), Dim (Dimension), Scr (Screen), and Cov (Coverage Area).	64
8.4	Comparing our proposed methods for the task of identifying DOSs.	68
8.5	Comparing our proposed methods for opinion mapping task. . . .	69
8.6	Comparing our proposed methods for opinion mapping task (without attribute Other).	69
9.1	Compare specs.	72
9.2	Contender overview.	73
9.3	Main modules of Contender.	74
9.4	Different types of charts provided by <i>Contender</i>	75

List of Tables

6.1	Summary of the <i>BestBuy Collection</i>	40
6.2	Set of product attributes for each category in the BestBuy Collection.	41
6.3	Distribution of sentence types by product category in the BestBuy DOS Dataset.	41
6.4	Distribution of opinions among targets.	42
6.5	Summary of the Amazon Review Dataset	43
6.6	Summary of Amazon DOS Dataset and the Amazon-Top100A Dataset.	44
6.7	Summary of datasets created in this thesis.	46
7.1	The 10 most frequent aspect expressions in reviews of each category from the Amazon-Top100A Dataset.	53
8.1	Precision, recall and F_1 for <i>AspectLink</i> and the baseline with and without using stemming.	56
8.2	Results of using common and expanded descriptors in <i>AspectLink</i>	57
8.3	Results of similarity functions applied cumulatively.	58
8.4	Experimental results for the task of identifying DOSs.	61
8.5	Feature ablation study for SVM _{BoW+Feat} classifier.	62
8.6	Results for opinion-mapping task. Subscript _{seg} indicates that classifier uses <i>Sentence Core Segments</i> strategy.	63
8.7	Opinion-mapping results when attribute Other is not considered.	64
8.8	F_1 results of end-to-end evaluation of <i>OpinionLink</i>	65
8.9	Experimental results with data from the Amazon Collection.	66
8.10	Results for the opinion mapping task using the proposed bootstrap- ping strategy.	67

List of Acronyms

NER Named Entity Recognition.....	9
LSTM Long Short-Term Memory.....	9
CRF Conditional Random Fields.....	9
NLP Natural Language Processing.....	9
POS part of speech.....	24
DOS Direct Opinionated Sentence.....	25
BOW Bag-of-Words.....	32
GBT Gradient Boosting Trees.....	34
ME Maximum Entropy.....	34
RF Random Forest.....	34
SVM Support Vector Machine.....	34
CAM cameras.....	40
CEL cell phones.....	40
DVD dvd players.....	40
LAP laptops.....	40
ROT internet routers.....	40

Stimulated by e-commerce websites, hundreds of thousands of people post reviews regarding products of all types, which are offered online. In these reviews, people express their opinion towards these products and their features. Consequently, a large number of opinions are available, which can be a valuable source of knowledge for decision-making for manufacturers as well as customers. From these opinions, manufacturers can obtain immediate feedback to improve the quality of their products, and customers are able to obtain assessments from reviews prior to purchasing a product. However, as it is common in many types of social media, the sheer volume of available reviews for each product normally exceeds the human processing capacity and can, thus, become a major barrier to its effective use (Kwon et al., 2015).

As an example, let us assume that a consumer is interested in the general user opinion related to a particular cell phone’s screen. In most cases, reading all the reviews is impractical. On the other hand, a straightforward query containing the term *screen* is also not effective as people commonly write on different *aspects* of the screen, such as *resolution* or *contrast*, without using the exact word. Moreover, the reviews are typically written by nontechnical users. Consequently, the text is not always correct and frequently contains misspellings and other typing errors. Thus, there have been a number of research initiatives towards organizing the otherwise user-provided reviews in order to facilitate the task of users in examining them.

A traditional method of organizing a large number of product reviews is to create an *opinion summary*. This kind of summary provides a condensed list of product aspects and their corresponding opinions. The most common approach is called *aspect-based opinion summarization* (Hu and Liu, 2004). This approach is commonly performed in three phases (Condori and Pardo, 2017): aspect identification, sentiment prediction, and summary generation. Aspect identification aims at identifying the important topics present in opinions.

Sentiment prediction determines the orientation (polarity) of the opinion on the identified aspects. Finally, summary generation describes what specific information is included in the summary. Recently, a number of techniques have been proposed in the literature to improve the opinion summarization methods (Amplayo and Song, 2017; Rakesh et al., 2018; Zhou et al., 2016).

We claim that current techniques for creating opinion summarizations are inadequate to address customer queries on the most relevant product characteristics. In particular, in current methods, the opinions are arbitrarily clustered by aspects, causing these clusters to not necessarily align with relevant product characteristics. This is illustrated by the aforementioned cell phone screen. For example, there could be several clusters that refer to the screen: a cluster of opinions regarding resolution and color, another one about mixing glossiness with size, and numerous others. Thus, the customer must perform the nontrivial task of identifying which clusters of aspects refer to each product characteristic of interest. Zha et al. (2014) reported that for the *iPhone 3GS*, more than three hundred aspects were identified in the reviews. Summarizing this information can generate hundreds of clusters without identifying what specific aspects refer to the actual cell phone screen. Consequently, following question arises here:

Question 1 (Q1) *How to structure opinions so that they can be effectively used by customers and manufacturers?*

Q1 is the motivating question of this thesis. To address Q1, we start by formulating the first hypothesis in this thesis:

Hypothesis 1 (H1) *The most important product characteristics for people are represented by the attributes of the product catalogs, supplied by the manufacturers, and commonly made available to the customers on e-commerce sites.*

According to Fensel et al. (2001), product catalogs are designed for human readers and their function is to describe products to potential clients, which leads us to the second hypothesis in this thesis:

Hypothesis 2 (H2) *The process of organizing opinions should be guided by the attributes of its product catalogs.*

Grouping opinions around the attributes of the product catalog also allows the catalog to be enriched with these opinions with the passage of time, which makes the user reviews readily available without the requirement for further processing. Another advantage of this approach is that it allows the customer to easily compare people’s opinions on products of the same category since they are all represented by the same attributes. Continuing our example, assuming that *cell phone screen* is one of the characteristics outlined in a product catalog for cell phones, one could easily discover what cell phone has

more positive comments specifically regarding its screen. This comparison of opinions, although extremely useful to buyers, is not possible while using traditional methods of aspect-based opinion summarization, because there is no guarantee of obtaining a single summary of opinions on this specific characteristic (the screen).

1.1 Problem Statement

Motivated by the above mentioned question $\mathcal{Q}1$ and hypotheses $\mathcal{H}1$ and $\mathcal{H}2$, in this thesis, a novel problem formulation for organizing a large number of unstructured user reviews is proposed:

Problem 1 ($\mathcal{P}1$) *Enriching product catalogs with user opinions at the attribute granularity level as a new form of opinion summarization.*

Current opinion-summarization approaches consist of grouping opinions around the aspects, which are not always meaningful. Here, opinions must be aligned to the attributes that are previously defined and are meaningful for customers. Therefore, the ultimate objective of the formulated problem $\mathcal{P}1$ is to enrich the product catalog with opinions extracted from user reviews.

1.2 Research Questions (RQ)

The present thesis aims at contributing to the problem of organizing a large number of product reviews. To achieve this, the following research questions that address problem $\mathcal{P}1$ are formulated:

$\mathcal{R}Q1$ Are there evidences that the most important product characteristics for people are represented by the attributes of the product catalogs?

$\mathcal{R}Q2$ Which approach can be used to address problem $\mathcal{P}1$?

$\mathcal{R}Q2.1$ Which methods are best suited to carry out the proposed approach?

$\mathcal{R}Q2.2$ How to validate the effectiveness of the proposed methods?

The first question provides support to the proposed problem $\mathcal{P}1$. In order to answer $\mathcal{R}Q1$, an empirical study to support hypotheses $\mathcal{H}1$ and $\mathcal{H}2$ on the impacts of attributes of product catalogs on user opinions has been developed. This study used a large collection of data, and the experimental results indicate that user opinions are significantly influenced by product attributes. This study is presented in Chapter 7.

$\mathcal{R}Q2$ is the main question investigated in this thesis, and the two research questions, $\mathcal{R}Q2.1$ and $\mathcal{R}Q2.2$, are the branches of the main question.

1.3 Our Approach

In order to answer $\mathcal{RQ2}$, we are proposing an approach, which comprises of two phases: *opinion extraction* and *opinion mapping*. Both phases are illustrated in Figure 1.1 and are described below.

The first phase consists of identifying opinionated sentences in the reviews of a particular product and extracting the corresponding opinions. In Figure 1.1, eight opinions are extracted from five user reviews. Notice that some sentences are not opinionated and were, therefore, discarded. Furthermore, it is possible to have more than one opinion in a single sentence. The outcome of the first phase is a set of opinions on a product.

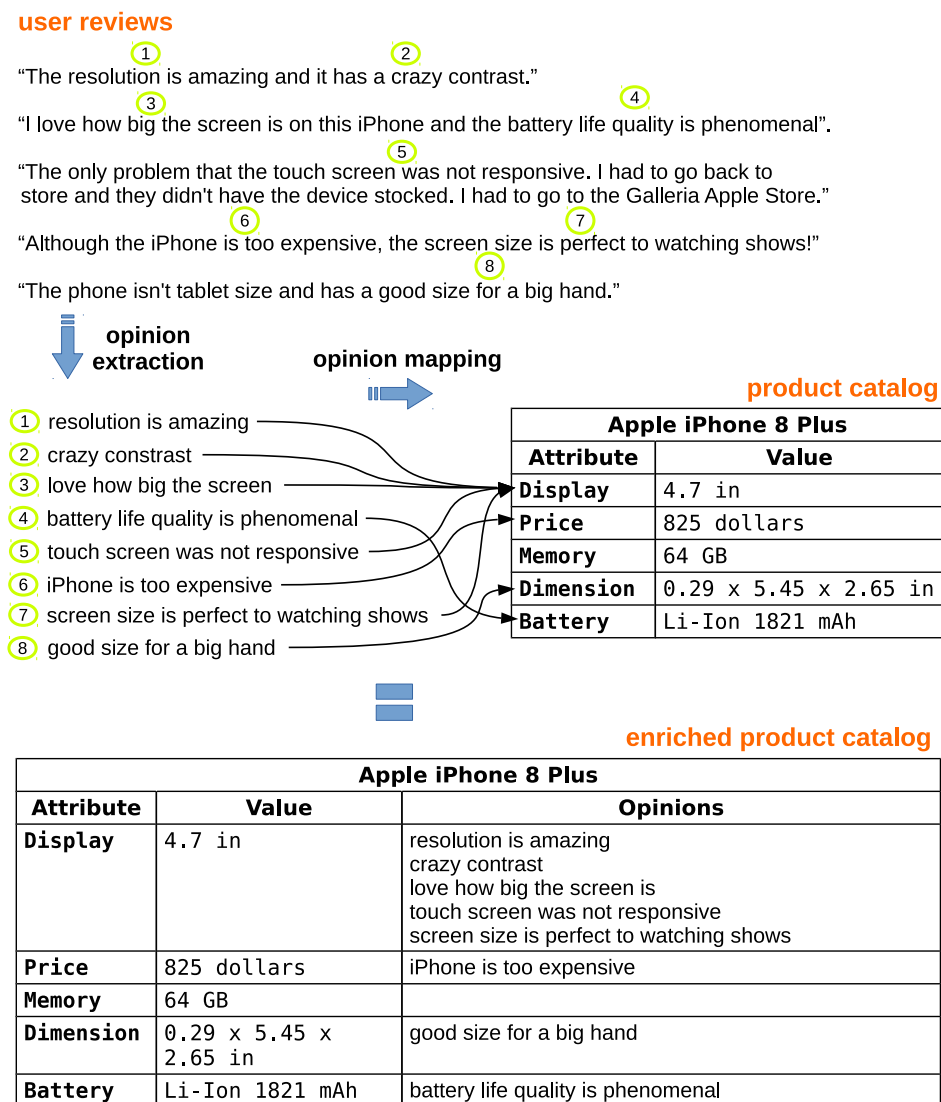


FIGURE 1.1: Example of product catalog enrichment with opinions.

In the second phase, the task is to map the previously extracted opinions

to attributes of the product catalog. The output of the second phase is an *enriched product catalog* where, for each product, each attribute of the catalog is enriched with a set of opinions regarding the attribute. In Figure 1.1, a fragment of a product catalog has been used to illustrate the mapping of the opinions that were previously extracted. The enriched product catalog will have, for each product, the objective values that are normally associated with its attributes, and a new dimension with the *subjective values* represented by the user opinions. For simplicity, in this example, it is assumed that the opinions are represented by sentences. However, in our work the opinions are represented by tuples whose components are, among other things, aspect expressions, and sentiment words. Our representation of opinion was adapted from the classical definition of opinion established by Liu (2015). Concepts and terminology that will be used throughout this thesis are presented in Chapter 3.

In summary, our proposed approach carries out two distinct but related tasks:

Task 1 ($\mathcal{T}1$) *Identifying direct opinionated sentences in the reviews.*

Task 2 ($\mathcal{T}2$) *Mapping opinions to attributes from the product catalog.*

In order to answer $\mathcal{R}Q2.1$ and $\mathcal{R}Q2.2$, in thesis, we have developed and evaluated two distinct methods. For the first method, named *AspectLink*, an unsupervised strategy has been adopted. For the second method, named *OpinionLink*, a supervised strategy has been adopted.

In *AspectLink*, the task $\mathcal{T}1$ is addressed by means of an unsupervised linguistic approach. The task $\mathcal{T}2$ is addressed by means of similarity functions that compare the lexical features of product attributes from the catalog with features from the text of aspect expressions. To verify the effectiveness of the *AspectLink* method ($\mathcal{R}Q2.2$), an extensive experimental evaluation was executed, which allowed the analysis of the impacts of several parameters on the effectiveness of this method. The results obtained in these experiments indicate that *AspectLink* is an effective method to address the problem $\mathcal{P}1$, since it is totally unsupervised.

In *OpinionLink*, for the task $\mathcal{T}1$, a method was devised that uses binary classifiers and a set of statistical features extracted from reviews. Task $\mathcal{T}2$ was modeled as a multi-label classification problem because it was assumed that a single opinion might refer to more than one attribute. To verify the effectiveness of the *OpinionLink* method ($\mathcal{R}Q2.2$), an extensive experimental evaluation was conducted, which demonstrates the effectiveness of the proposed method. Furthermore, a bootstrapping strategy was proposed to train the classifiers in order to reduce the dependence on training data. Moreover, a comparative analysis of the effectiveness of the *AspectLink*, *OpinionLink*, and *OpinionLink* using our proposed bootstrapping strategy was conducted.

Finally, *OpinionLink* was applied as a full pipeline and the experimental results demonstrated the feasibility of using this method in real and large-scale applications.

1.4 Contributions

The main contributions in this thesis are summarized as follows:

1. A novel problem formulation $\mathcal{P}1$ for organizing a large number of product reviews. Chapter 3 presents the concepts and terminologies for this problem.
2. An empirical study on the use of direct and indirect mentions in the user reviews regarding attributes of product catalog ($\mathcal{RQ}1$). Briefly, the results from this study, presented in Chapter 7, indicate that user opinions are guided by the attributes from product catalogs and highlight the influence of attributes of product catalog on the user reviews.
3. An unsupervised method, named *AspectLink*, to address the problem $\mathcal{P}1$ ($\mathcal{RQ}2.1$).
4. A supervised method, named *OpinionLink*, to address the problem $\mathcal{P}1$ ($\mathcal{RQ}2.1$).
5. A comprehensive evaluation which experimentally demonstrates the effectiveness of the *AspectLink* and *OpinionLink* methods and its variations on representative datasets obtained from real e-commerce web sites ($\mathcal{RQ}2.2$).
6. A set of datasets publicly available. To properly evaluate the methods developed in this study, experimental datasets, non-existent in the literature, are needed. This was done using real data collections gathered from on-line sources available on the Web. Chapter 6 describes these datasets and the way they were built.
7. An application named *Contender* for comparing two products at the attribute granularity level based on user opinions. This system was developed to showcase a practical application of some proposal ideas in this thesis and it is available as an Android app. Chapter 9 presents the details on this system.

1.5 Thesis Organization

Chapter 2 presents a review of works in the literature that are related to the current study. Chapter 3 presents the concepts and terminologies that will be used throughout this thesis. In the next two chapters, two methods developed in

this thesis to address the problem ($\mathcal{P}1$) of enriching product catalogs with user opinions are presented. More specifically, Chapter 4 presents an unsupervised method, while Chapter 5 presents a supervised method. Chapter 6 describes the experimental datasets used in the experiments. In Chapter 7 an empirical study is presented on the use of direct and indirect mentions in the user reviews regarding attributes of product catalog. The experiments and results of the proposed methods are presented in Chapter 8. A practical application of the thesis is presented in Chapter 9. Finally, Chapter 10 concludes this thesis and presents future work.

In this chapter, we will review the existing work related to our research. Since the ultimate objective of the formulated problem in this thesis is to enrich a product catalog, which is a representation of databases, with opinions posted on e-commerce websites, we will begin by reviewing research on the problem of enriching databases with information available on the web. We will then discuss several methods and techniques that address the general problem of summarizing opinions regarding a specific target. These methods and techniques are commonly used in sentiment analysis applications to structure opinions in order to enable them to be further processed more easily. Although our research is actually in the realm of opinion mining, we will, for conciseness, limit our review of the related work to the literature on opinion summarization. Finally, we will discuss some related work on the importance of product attributes and user reviews on purchasing decisions. These works are related to our empirical study on the use of direct and indirect mentions in the user reviews to attributes of product catalog.

2.1 Enriching Databases

There is a growing body of literature on the problem of enriching databases with information available from online sources. *InfoGather*, for example, is a system designed for augmenting entities in a database with information gathered from web tables (Yakout et al., 2012). Besides supplying new attributes and their values for existing entities, its ultimate objective is to supply new values for existing attributes. The system first identifies web tables that match a given target table to be augmented. Next, it selects data from web tables that can be used to supply values and attributes to the entities in the target table; it is based on Topic-Sensitive PageRank and an augmentation

framework that aggregates predictions from multiple matched tables. More recently, Zheng et al. (2018) proposed a method called *OpenTag* to supplement product catalogs with missing values for attributes of interest from product descriptions and other related product information, especially with values not previously known. OpenTag builds *Named Entity Recognition* (NER) systems that use bidirectional *Long Short-Term Memory* (LSTM) and *Conditional Random Fields* (CRF).

To enrich a target knowledge base with facts extracted from the web, another related approach was proposed by Dutta et al. (2015). More specifically, this approach uses facts generated by other systems, such as *Nell* (Carlson et al., 2010) and *Reverb* (Fader et al., 2011), to enrich *DBpedia*¹. To accomplish this task, the authors proposed a system that comprises of two phases. Firstly, it applies Markov Clustering to generate groups of relational phrases. Then, the authors proposed an algorithm based on rules and similarity scores that map each group to a DBpedia property.

Whereas the above proposals focus on factual information, the *Surveyor* system (Trummer et al., 2015) mines *dominant opinions* from the web determining whether a subjective property applies to entities of a particular type. The authors assumed that there is a dominant opinion for many entity-property combinations, meaning that a significant number of users agree on whether the property applies to the entity. A subjective property, in this scenario, is an adjective, optionally associated with preceding adverbs. The purpose is to build a knowledge base of subjective properties and entities, given a collection of web documents. The system first applies *Natural Language Processing* (NLP) tools to web documents for identifying mentions of entities in their target knowledge base. It then, applies NLP methods to extract the subjective properties and their sentiment polarities. Finally, the system selects the dominant opinions, based on a probabilistic model, and associates them with the target entity.

Similar to the proposal described above, we also aim at enriching a database with information available from online sources. However, since the database in our target scenario corresponds to a product catalog and the online sources are user reviews, there are several important differences. *InfoGather* functions with structured data, whereas user reviews are not structured. *InfoGather* and *OpenTag* do not address subjective properties (opinions); rather they are concerned only about factual properties. *Surveyor* and Dutta et al. (2015) work at the entity granularity level and cannot process attributes of the product catalog. Therefore, these methods cannot be applied directly to the novel problem formulated in this thesis.

¹<http://dbpedia.org>

2.2 Opinion Summarization

Opinion summarization methods are related to our research in the sense that they are used to organize the large number of user opinions available online. The most common type of opinion summarization technique is *aspect-based opinion summarization* (Kim et al., 2011; Zhou et al., 2016; Condori and Pardo, 2017). This kind of technique generates opinion summaries around a set of aspects — the aspects are extracted from reviews, and the sentiments toward each aspect are identified and summarized. In the simplest case, the summary is a presentation of the positive and negative sentiments toward each aspect (Hu and Liu, 2004; Liu et al., 2005; Li et al., 2010).

Hu and Liu (2004) can be considered the pioneers on aspect-based opinion summarization. Their proposed method has three steps: 1) identification of the aspects of the products in user reviews, 2) identification of the polarity for each sentiment toward each aspect (positive or negative), and 3) summary generation using previous information. Later, Liu et al. (2005) proposed a system called *Opinion Observer*, based on supervised rule mining, to generate language patterns that identify product features. As such, they addressed several linguistic problems that were not well considered by Hu and Liu (2004). Further improvement to the work by Hu and Liu (2004) was proposed by Li et al. (2010), who adopted a machine-learning approach. In their work, the authors proposed using features such as linguistic and syntactic tree structures to train a Conditional Random Fields (CRFs) model for the tasks of extracting and summarizing aspects from reviews.

In a different approach, several authors have used topic models to summarize opinions extracted from online reviews. Recently, Rakesh et al. (2018) proposed an aspect summarization model, called APSUM, which mines fine-grained aspects for user queries by constricting the document and word topic space to create focused topics. Their main goal is to design a model that captures the natural flow of the review writing process. The proposed model outperformed several baselines that are considered state of the art in aspect summarization (Blei et al., 2003; Titov and McDonald, 2008; Jo and Oh, 2011; Wang et al., 2016).

It is important to notice the differences between the above methods and the proposed approach in our research. Firstly, these opinion summarization methods group opinions around automatically discovered aspects, whereas the proposed approach groups opinions around the existing attributes of a product catalog. Consequently, the former methods tend to generate many aspect groups that are generic and difficult to interpret. For example, different groups of aspects generated by these methods can refer to the same characteristic of a target product. Therefore, a new regrouping step continues to be necessary. Even if cohesive aspect groups were obtained, there would remain the task of identifying what features of the products they are related to. The proposed approach addresses this task because we align opinions with the attributes

of a product catalog, which in turn, are the most important features of the products.

Hu et al. (2017) proposed an alternative approach to summarize the essential information from online reviews. The central idea of their method is to identify the top-k most informative sentences and to use these to summarize the reviews. The proposed method starts by collecting reviews from online sites, such as *TripAdvisor*, and performing a cleaning preprocessing over these reviews. Next, metrics, such as *sentence importance* and *sentence similarity*, are calculated relying on features, such as author reliability and review usefulness. The last step is the selection of the top-k most representative sentences, which involves grouping the sentences into k clusters. This is accomplished using the k-medoids algorithm.

Kim and Kang (2018) proposed a method to compare two competing products using opinion summarization. They designed a method which is executed in three steps. Firstly, the method extracts from the reviews of a given product, those words that are representative of this product. These words are called *discriminative attributes*. This is performed for every product. In the second step, the discriminative attribute words are classified using a topic model that identifies similarities between them, forming, as a result, a number of *attribute categories*. The third step is to classify the polarities of the discriminative attributes.

Another line of research to summarize opinions is to group opinions according to a taxonomy (Yu et al., 2011). The method described by Yu et al. (2011) maps opinions to their aspects according to a taxonomy. This taxonomy is not related to a product category; rather it is related to a specific product. It is initially built from information available on the product’s webpage and is then incrementally rebuilt and refined according to the specific aspects identified in a set of reviews on the target product. The method relies on a semantic distance-learning algorithm to group opinions based on their semantic relations, which in turn requires training data.

The approach proposed in our research is related to the method proposed by Yu et al. (2011) in the sense that it uses as input a set of reviews on each specific product, and groups aspect expressions identified in the reviews around features of this product. However, in our approach, the target features are derived from a product catalog — they are fixed and predefined for all products in a given category. Conversely, in the method proposed by Yu et al. (2011), distinct taxonomies can be generated for two products in the same category. In fact, depending on the set of reviews, and even on the order in which reviews are processed, the same product can lead to different taxonomies. For these reasons, although the method proposed by Yu et al. (2011) is effective for the task of building product-oriented aspect taxonomies, it cannot be used for the task of enriching product catalogs.

The method proposed by Carenini et al. (2005) groups aspect expressions into nodes of a taxonomy, where each node represents a feature of products

in some category. This taxonomy is supplied by a user. Except for this, the method is fully unsupervised since it relies on similarity functions to verify whether aspect expression matches features in the taxonomy. In this case, the aspect expression is mapped to the matching feature. Similar to this method, our approach relies on a strategy that groups aspect expressions into attributes. In our case, these attributes are, however, provided by the structure of the product catalog, while in the work of Carenini et al. (2005) the taxonomy must be handcrafted by a user. Our proposed method named *AspectLink* was designed to be unsupervised. For this, we adapted and improved the similarity functions proposed by Carenini et al. (2005). This method is used as the baseline in our experiments with *AspectLink*.

2.3 Product Attributes and User Reviews

In this section, we will describe work that has been done on the importance of product attributes (Maslowska et al., 2017; Lee and Nguyen, 2017; Wang et al., 2018) and user reviews on purchasing decisions (Kostyra et al., 2016; Qazi et al., 2016; Jo and Oh, 2011; Sun et al., 2018; Singh et al., 2017).

There are some studies which investigated the importance of product attributes in e-commerce domain. Lee and Nguyen (2017) investigated the importance of product attributes in purchasing fashion goods and the influence of these attributes on preferences for external versus local fashion brands. Maslowska et al. (2017) studied how review characteristics (i.e., number of reviews), product characteristics (i.e., price) and customer behaviors (i.e., reading reviews) interact with each other to influence purchase decisions. Wang et al. (2018) conducted a study on online reviews to measure how product attributes impact customer satisfaction.

On the other hand, several works have investigated the importance of customer reviews. Kostyra et al. (2016) investigated the impact of online customer reviews on customer’s decisions. Qazi et al. (2016) investigated why some reviews are more helpful as compared to others. Jo and Oh (2011) proposed two distinct models to discovering what aspect expressions are evaluated in user reviews and how sentiments for different aspects are expressed. Sun et al. (2018) proposed a method of using external user generated data to evaluate the relative importance of an entity’s attribute. Singh et al. (2017) developed a method based on machine learning that can predict the helpfulness of the consumer reviews using several textual features such as polarity, subjectivity, entropy, and reading ease.

While prior studies have contributed to understanding the importance of product attributes and user reviews as separate and valuable sources to support purchase making-decisions, to the best of our knowledge, our study presented in this chapter is the first to investigate the relationship between these two kinds of information.

In this chapter, we will present the concepts and the terminology that will be used throughout this thesis. Specifically, in Section 3.1, we will present a formal definition of product catalog. In Section 3.2, we will introduce the concept of review, explain what type of opinionated sentence is used in our work, and present the elements that form an opinion. In Section 3.3, we will explain the representation of the three types of targets to be adopted in the task of mapping opinions to attributes of a product catalog. Moreover, we will present the definition of an enriched product catalog and discuss the differences between a product catalog and its version enriched with opinions.

3.1 Product Catalog

We will consider a catalog as a set of products of the same category (e.g., *cell phones* or *laptop computers*), where each product is represented by its attributes and their corresponding values. More formally, we have defined the concept of product catalog as follows:

Definition 3.1 *A product catalog is a set of products $C = \{p_1, \dots, p_n\}$, where each product is represented as $p = \{\langle A_1, v_1 \rangle, \dots, \langle A_m, v_m \rangle\}$, and each pair $\langle A_i, v_i \rangle$ consists of an attribute name A_i paired with its value v_i for the corresponding product. The value v_i is a set that can be empty or contains one or more elements.*

Figure 3.1 presents an example of a single product within a catalog of the *cell phone* category. In this case, the product is entitled “Apple iPhone 8 Plus”. As can be observed in Figure 3.1, the products in this category are represented using eight attributes, whose names are: **Processor**, **Display**, **Camera**, **Price**, **Memory**, **Dimension**, **Battery**, and **Software**. Notice that in this case, like

other real-life cases, attributes can be divided into sub-attributes. For example, **Display** is further divided into **Size**, **Type**, and **Resolution**. In our work, we will, however, consider only top-level attributes because addressing them is sufficient for our purposes. Thus, the value of each attribute is a set that includes the values of all its sub-attributes. For example, the value of **Display** for this product is given by the set {'4.7 in', 'LED', '750 x 1334 pixels'}.

Apple iPhone 8 Plus		
Attributes	Sub-attributes	Values
Processor		Hexa-core A11 Bionic
Display	Size	4.7 in
	Type	LED
	Resolution	750 x 1334 pixels
Camera	Front	7 megapixels
	Rear	12 megapixels
Price		825 dollars
Memory	RAM	2 GB
	Internal Storage	64 GB
Dimension	Size	0,29 x 5,45 x 2,65 in
	Weight	5.28 ounces
Battery	Type	Non-removable Li-on 1821 mAh
	Talk Time	21 hours
Software		iOS

FIGURE 3.1: Example of product from a typical catalog of the Cell Phone category.

3.2 Reviews, Sentences and Opinions

A *review* is a text posted by a user on an e-commerce website, usually reporting his/her experience regarding a specific product, which we call the *target entity* of the review. Each review is composed of a set of *sentences*. Sentences that express factual information are called *objective* sentences, whereas sentences that express personal feelings or beliefs are called *subjective* or *opinionated* sentences. We are interested in the latter because they represent the user's opinions about a product. A single sentence can have multiple opinions. For example, the sentence "*The screen is bright, but I'm not satisfied with the performance*" has two different opinions; a positive opinion regarding the display and a negative opinion regarding the processor.

An opinionated sentence can be further classified into *comparative* or *direct* sentence. A comparative sentence expresses a relation of similarities or differences between two or more products. The sentence "*the camera of the iPhone is much better than Galaxy*" is an example of a comparative sentence. A

direct opinionated sentence expresses an opinion directly about a characteristic or part of the product, or the product as a whole. The sentence “*The camera of the iPhone is fantastic*” is an example of direct opinion. As our objective, in this thesis, is to enrich each product of the catalog with the opinions of users regarding the specific product, we decided to eliminate comparative sentences. The definition of a direct opinionated sentence (DOS) is more precisely stated as follows:

Definition 3.2 *A direct opinionated sentence (DOS) is a sentence where an opinion is expressed directly about one or more characteristics of a product, or on the product as a whole.*

An *aspect* is any reference made within an opinion to a particular feature of an attribute or the entire product. For example, the attribute **Battery** can have the aspects, *battery life*, *charge time*, and *replaceability*. The same aspect can be expressed using different *aspect expressions*. For example, the sentences “*the battery lasts a long time*”, “*the mobile has a long battery life*”, and “*the battery discharge rate is pretty high even at idle*” contain three different aspect expressions about the same aspect: *lasts*, *battery life*, and *discharge rate*.

An opinion expresses a *sentiment* of the user toward an aspect of a product. A sentiment has a *polarity*, which can be *positive*, *negative*, or *neutral*. The set of words used to express a sentiment are called *sentiment words*. For example, “bright”, “great”, and “fast” indicate positive sentiments, whereas “not satisfied” indicates a negative sentiment.

Opinions are represented by a sextuple $o = \langle a, w, s, st, h, t \rangle$, where a is the aspect of the target entity on which the opinion has been given, w is the sentiment words of the opinion, s is the sentiment polarity of the opinion toward aspect a , st is the sentence from where the opinion was extracted, h is the opinion holder, and t is the opinion posting time. Although adapted to the context of this work, this definition of opinion is derived from that presented in Liu (2015).

In Figure 3.2, we have presented an example of a review written by a user regarding the product in Figure 3.1. Sentiment words are underlined and words comprising aspect expressions are shown in bold face. The symbols near each word are explained in the following section.

*Fiodor on November 1, 2017. I purchased an iPhone 8 64GB from Ting. So far I'm really happy with this **phone**[◇], although it's **expensive**[▲]. **Works**[★] really good and supper **portable**[▲]. Amazing **cell**[◇] and it has fast **charging**[▲]. The **global warranty**[★] is pretty useful for me because I have to travel a lot.*

FIGURE 3.2: Example of review for product presented in Figure 3.1.

3.3 Enriching Product Catalogs with Opinions

In our work, we will address the problem of enriching product catalogs with user opinions extracted from product reviews. Our main goal is to automatically map opinions to specific attributes. However, the reviews frequently also include opinions that do not refer to a specific attribute of a product; rather they refer to the product as a whole. Furthermore, opinions can also target attributes that are not represented in the product catalog. Thus, we will consider these three cases in our work. Our general strategy is to learn, from the text of the sentences, how to group the opinions according to their attributes. Specifically, each opinion is mapped to one or more attributes according to the following three cases:

Case 1: The user posts an opinion referring to one of the attributes of the product catalog. For example, in the sentence *“it has fast charging”*, the user is expressing a positive opinion (“fast”) regarding the battery. Therefore, we should map this opinion to the attribute **Battery** of the product. Examples of opinions that should be associated with one of the attributes present in the original catalog are identified in Figure 3.2 with the symbol ♠. In Figure 3.3, the first eight lines are opinions extracted from the reviews of Figure 3.2 that were mapped to specific attributes of the product.

Case 2: The user posts an opinion about the product as a whole. In the sentence *“So far I’m really happy with this phone”* from Figure 3.2, the user expresses a positive sentiment (“happy”) for the product as a whole, but not for one of its attributes of the product catalog. Therefore, we should map this opinion to the target product. For this, we created a new attribute, called **General**. For convenience, we will consider the value of **General** to be the product title. Examples of expressions that should be associated with the product as a whole are identified in Figure 3.2 with the symbol ◇. This case is illustrated in Figure 3.3, where Line 9 displays the opinions extracted from the reviews of Figure 3.2 that were mapped to the attribute **General**.

Case 3: The user posts an opinion about a specific characteristic of a product that is not explicitly represented as an attribute in the product catalog. Consider the following example: *“The global warranty is pretty useful for me”*. In this sentence, the user posts a positive opinion (“useful”) for a characteristic of the product that is not represented as an attribute in the original product catalog, i.e., its warranty. To address such cases, we created an attribute called **Other**. The value field of **Other** is blank for the product. Examples of aspect expressions that should be associated with **Other** are identified in Figure 3.2

Apple iPhone 8 Plus		
Attributes	Values	Opinions
Processor	{"Hexa-core A11 Bionic"}	-
Display	{"4.7 in"; "LED"; "750 x 1344 pixels"}	-
Camera	{"7 megapixel"; "12 megapixel"}	-
Price	{"825 dollars"}	<"expensive", "-", "negative", "So far ... insanely expensive", "Fiodor", "November 1, 2017">
Memory	{"2 GB"; "64 GB"}	-
Dimension	{"0.29 x 5.45 x 2.65 in"; "5.28 ounces"}	<"portable", "-", "positive", "Works really good ... portable", "Fiodor", "November 1, 2017">
Battery	{"Non-removable Li-on 1821 mAh"; "21 hours"}	<"charging", "fast", "positive", "Amazing cell ... charging", "Fiodor", "November 1, 2017">
Software	{"iOS"}	-
General	{"Apple iPhone 8 Plus"}	<"phone", "happy", "positive", "So far ... insanely expensive", "Fiodor", "November 1, 2017"> <"cell", "amazing", "positive", "Amazing cell ... charging", "Fiodor", "November 1, 2017">
Other	-	<"works", "good", "positive", "Works really good ... portable", "Fiodor", "November 1, 2017"> <"global warranty", "useful", "positive", "The global ... a lot", "Fiodor", "November 1, 2017">

FIGURE 3.3: Example of product catalog enriched with opinions presented in Figure 3.2.

with the symbol ♣. This case is illustrated in Figure 3.3 where Line 10 displays the opinions extracted from the reviews of Figure 3.2 that were mapped to the attribute **Other**.

Considering these cases, given a product catalog C , our ultimate goal is to generate an *enriched catalog* C^+ . More formally, we have defined the concept of an enriched product catalog as follows:

Definition 3.3 *An enriched product catalog is a set of enriched products $C^+ = \{p_1^+, \dots, p_n^+\}$, where each enriched product is represented as $p^+ = \{\langle A_1, v_1, O_1 \rangle, \dots, \langle A_m, v_m, O_m \rangle\} \cup \{\langle \mathbf{General}, v_G, O_G \rangle, \langle \mathbf{Other}, v_O, O_O \rangle\}$, where each A_i and v_i denote the same attribute name and values from the original catalog C (as stated in Definition 3.1), and O_i represents a set of opinions regarding attribute A_i . The triple $\langle \mathbf{General}, v_G, O_G \rangle$ is added to handle opinions on the product as a whole (Case 2), and $\langle \mathbf{Other}, v_O, O_O \rangle$ is added to handle opinions regarding specific characteristics that are not represented in the other attributes (Case 3).*

Notice that there are two main differences between a product p and its enriched version p^+ . The first is that two new attributes have been added: **General**, to represent opinions regarding the product as a whole, and **Other**, to represent opinions related to characteristics not initially represented in p . The second difference is that the enriched product specification considers that each attribute has, in addition to its value v , a set of opinions O regarding each attribute.

In this chapter, we will present our first method for enriching product catalogs with opinions extracted from reviews written by customers. In this method, called *AspectLink*, the products of a given category in a catalog are represented by their attributes, and the opinions are represented by *aspect expressions*, i.e., by the text portion from reviews that defines the aspect (Liu, 2015). Afterwards, the problem of enriching product catalogs is reduced to the task of mapping aspect expressions extracted from user opinions to their corresponding attributes. *AspectLink* addresses this problem by means of similarity functions that compare lexical features of product attributes from the catalog with features from the text of the aspect expressions.

This chapter is organized as follows. Section 4.1 presents an overview of *AspectLink*. Section 4.2 describes the procedure to identify direct opinionated sentences (DOS) from reviews and to extract opinions from these sentences. In Section 4.3, we will describe a proposed representation of product attributes named *attribute descriptors*. In Section 4.4, we will explain the three similarity functions proposed to match aspect expressions and attribute descriptors. Section 4.5 describes the method in details. Finally, in Section 4.6, we will present a summary on this chapter.

4.1 Overview

The main objective of *AspectLink* is to associate opinions with product attributes they refer to. Figure 4.1 presents an overview of *AspectLink*. The method receives as input a product catalog C and a set of reviews R for each product from C . The first step in the method is extracting the opinions from all reviews in R . For this, our method splits each review into sentences and then identifies the direct opinionated sentences (DOSs), as stated in Definition 3.2,

because only this kind of sentence has opinions, in which we are interested. The method then extracts the opinions from the DOSs.

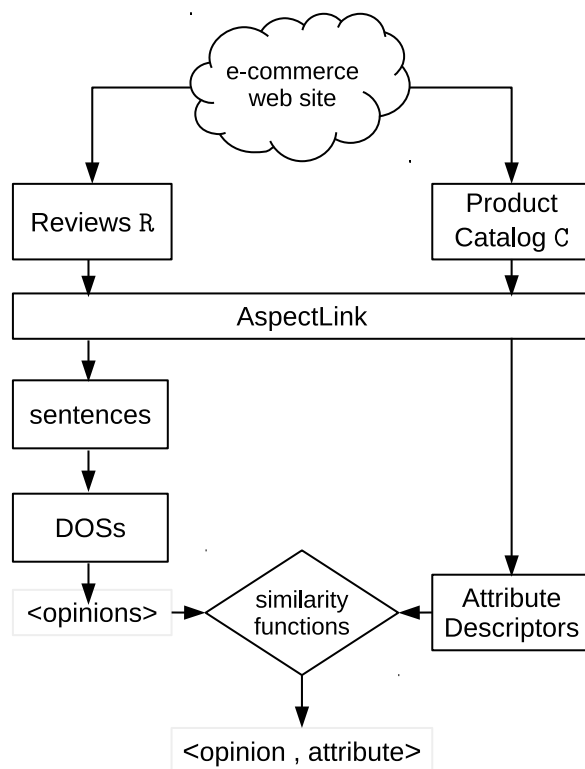


FIGURE 4.1: *AspectLink* Overview.

Next, product attributes are extracted from the product catalog C . For each attribute of each product, a descriptor is built to represent it. Informally, a descriptor is a set of terms that represent an attribute, where this set is composed by the name of the attribute and its values are taken from the product catalog C . The descriptors were adapted for each one of the three cases introduced in Section 3.3.

Finally, the association between an opinion and the product attributes is carried out by matching the aspect expressions identified in opinions with attribute descriptors. *AspectLink* uses similarity functions that compare lexical features of aspect expressions and attribute descriptors to mapping opinion to attribute.

In the remainder of this chapter we will present the details regarding *AspectLink* and its functioning.

4.2 Opinion Extraction

In *AspectLink*, the main goal of the first phase is to extract a set of opinions O_i from a set of reviews R_i for each product p_i in a product catalog C . For this,

our method breaks down each review $r \in R$ in sentences. Next, we implemented the unsupervised method proposed by Qadir (2009), where typed dependency relations, such as open clausal complements or adjectival complements, are used for identifying subjective sentences. Factual sentences are discarded. The opinionated sentences identified by this method can be *comparative* or *direct*. As our goal is to enrich a product catalog with opinions of users regarding the specific product, we must discard comparative sentences. For this, we implemented the unsupervised method proposed by Liu (2010). The output of this method is a set of direct opinionated sentences (DOSs) on the product p_i . Finally, the opinions from the remaining DOS are extracted. In the case of aspect expressions, which in our method guide the mapping of opinions to attributes, we implemented the well-known unsupervised aspect extraction method described in Poria et al. (2014). To keep *AspectLink* unsupervised, we had to use only unsupervised methods in the first phase. Finally, the opinions extracted in the previous step are added to O .

4.3 Attribute Descriptors

As described in the previous chapter (Section 3.3), an opinion can be associated to one or more attributes of the product catalog (Case 1), to the product as a whole (Case 2), or to a specific characteristic of a product that is not explicitly represented as an attribute in the product catalog (Case 3).

To address Case 1, our method tries to match each aspect expression with an attribute, more specifically with a *descriptor* of the attribute. This concept is precisely stated in Definition 4.1.

Definition 4.1 *Let \mathcal{A}_i be an attribute of the products in a product catalog C . We define $\Delta_{p,\mathcal{A}_i} = \{\mathcal{A}_i\} \cup v_i$ as a **descriptor** for \mathcal{A}_i in a product p from C , where, as stated in Definition 3.1, \mathcal{A}_i is a unique name used to refer to the attribute \mathcal{A}_i , and v_i is the set of values of \mathcal{A}_i for the corresponding product p .*

It will be more clear later that the idea of including the attribute N_A and the values of the attribute $V_{p,A}$ together in the descriptor is to allow multiple ways of matching aspect expressions and attributes. For example, the descriptor for the attribute **Software** from product catalog illustrated in Figure 3.1 is formed by the name of attribute (“Software”) and its value (“iOS”). Thus, according to Definition 4.1, the descriptor for the attribute **Software** is {“Software”, “iOS”}.

We notice that in practice, we can apply some common operations for handling sets of words while building descriptors. For example, in our experiments, we considered using a stemming function as alternative for building descriptors, because stemming is widely used in information retrieval systems with the aim of increasing recall (Baeza-Yates et al., 2011).

We will handle Case 2 just like to Case 1. However, in this case, we will use a different kind of descriptor to represent products, and we will try to match each aspect expression with this descriptor. The descriptor concept is precisely stated in Definition 4.2.

Definition 4.2 *Let p be a product in a catalog C . We define $\Delta_p = \{t\}$ as a **descriptor** for p , where, as defined in Section 3.3, t is the title used for this product.*

In this case, the descriptor for the product (cell phone) illustrated in Figure 3.1 is {"Apple iPhone 8 Plus"}.

Finally, our method assumes Case 3 whenever Case 1 and Case 2 do not hold. We stated this kind of mapping in Definition 4.7.

Expanding Attribute Descriptors

To map opinions to attributes, our method relies on matching attributes and descriptors. We use the descriptor $\Delta_{p,A}$ when we want to map an opinion to an attribute A , and we use the descriptor Δ_p when we want to map an opinion to the attribute **General**. Both descriptors consist of words that come from the attributes of the product catalog C or from the title of p .

However, in results obtained from preliminary experiments with *AspectLink*, we noticed that the set of words used to represent the attributes or titles of products in the catalog may be sometimes incomplete. For instance, in the Laptop category, many manufacturers only provide the name of the operating system in the attribute **Software**, while other manufacturers provide a complete list of the softwares that come installed on the laptop, such as applications, anti-virus, browsers etc. In the sentence "*McAfee is always able to protect my personal data*", there is a clear opinion regarding the **Software** attribute, but we need to know that "McAfee" is a kind of software. Therefore, the information that a product p_i has on its values for **Software** can be used for another product p_j .

Another common problem with data in product catalogs is that the manufacturers and stores often represent products that should have the same name with slightly distinct titles. For instance, Apple laptops are presented in many different ways, such as "MacBook", "Mac", "Mac Book", etc. Thus, words appearing in the title of a product p_i from a given category may be useful for describing another product p_j from the same category.

To cope with such problems, we also consider an expanded form of attribute descriptors in our work as defined below.

Definition 4.3 *Let A be an attribute of the products in a product catalog C , and let N_A be the name of the attribute, and let $V(p, A)$ be the set of values of A in product p . We define $\Delta_{*,A} = \{N_A\} \cup V_{p_1,A} \cup \dots \cup V_{p_n,A}$ as an **expanded descriptor** for A , where $\{p_1, \dots, p_n\}$ is the set of all products in C .*

Definition 4.4 Let p be a product in a catalog C , and let t be a unique product title used for product p . We define $\Delta_{p_*} = t_1 \cup \dots \cup t_n$ as an **expanded descriptor** for p , where $\{t_1, \dots, t_n\}$ is the set of titles for all products in C .

We carried out experiments with the two kinds of descriptors and noticed that the use of expanded descriptors led to higher recall values in all categories, with a comparatively small loss in precision. The details about these experiments are presented in Section 8.2.2.

4.4 Matching Aspects and Descriptors

The association between an opinion and the product attributes is done by matching aspect expression identified in opinion with attribute from the product catalog, more specifically with a descriptor of the attribute. A match between an aspect expression and an attribute is defined as follows.

Definition 4.5 Let \mathcal{A}_i be an attribute from an enriched catalog C^+ , which has a descriptor $\Delta_{p, \mathcal{A}_i}$. Let α be an aspect expression from an opinion o . We map o to \mathcal{A}_i , if α matches $\Delta_{p, \mathcal{A}_i}$. We say that α matches $\Delta_{p, \mathcal{A}_i}$, if at least one of its words, say w , matches at least one word, say u from $\Delta_{p, \mathcal{A}_i}$ according to one of the following similarity functions: *str_match*, *syn_score* or *sim_score*.

As an example, the sentence “*the operating system runs really smooth*” has an opinion that would be mapped to the attribute **Software** because the aspect expression “operating system” from this opinion matches at least one word with the descriptor for the attribute **Software**. Recall that the descriptor for the attribute **Software** is {“Software”, “iOS”}.

The three similarity functions referred in Definition 4.5 will be detailed in the next subsection.

A match between an aspect expression and the attribute **General** that represents the product as a whole is precisely stated in Definition 4.6.

Definition 4.6 Consider the attribute **General** from an enriched catalog C^+ , which represents a product from a catalog C , whose descriptor is Δ_p . Let α be an aspect expression from an opinion o . We map o to **General**, if α matches Δ_p . We say that α matches Δ_p , if at least one of its words, say w , matches at least one word, say u from Δ_p according to one of the following similarity functions: *str_match*, *syn_score* or *sim_score*.

For example, the sentence “*iPhone 8+ is awesome*” has an opinion that would be mapped to the attribute **General**, because the aspect expression “iPhone” from this opinion matches at least one word with the descriptor for the attribute **General**. Recall that the descriptor for the attribute **General** illustrated in Figure 3.1 is {“Apple iPhone 8 Plus”}.

Finally, our method assumes Case 3 whenever Case 1 and Case 2 do not hold. We stated this kind of mapping in Definition 4.7.

Definition 4.7 Consider the attribute `Other` from an enriched catalog C^+ , associated with a product p from a catalog C , for representing characteristics of the product that are not represented as an attribute in the product catalog. Let o be an opinion. We map o to `Other`, if o was not mapped to another attribute, according to Definitions 4.5 and 4.6.

For example, the sentence “The global warranty is pretty useful” has an opinion that would be mapped to the attribute `Other`, because the aspect expression “global warranty” from this opinion did not match with the descriptors for the Case 1 and Case 2.

Matching Aspect Expressions

According to Definitions 4.5 and 4.6, three string similarity functions are combined to match aspect expressions with descriptors. In Algorithm 2, a function called *Match* encapsulates the three functions combined. This function is detailed in Algorithm 1.

Algorithm 1: Match Function

```

Input: An aspect expression  $\alpha$ 
Input: A descriptor  $\Delta$ 
1 let  $\Theta_1$ ,  $\Theta_2$  and  $\Theta_3$  be global parameters;
2 foreach word  $w \in \alpha$  do
3   foreach word  $\delta \in \Delta$  do
4      $s_1 \leftarrow \text{max\_str\_match}(w, \delta)$ ;
5     if  $s_1 \geq \Theta_1$  then return TRUE;
6      $s_2 \leftarrow \text{max\_syn\_score}(w, \delta)$ ;
7     if  $s_2 \geq \Theta_2$  then return TRUE;
8      $s_3 \leftarrow \text{max\_sim\_score}(w, \delta)$ ;
9     if  $s_3 \geq \Theta_3$  then return TRUE;
10  end
11 end
12 return FALSE

```

Given an aspect expression α and a descriptor Δ , the algorithm iterates over all words w of α , and for each word δ of Δ , it computes a similarity score value between w and δ , using the three similarity functions. The algorithm terminates and returns TRUE, if one of the pairs w, δ gives a similarity score value higher or equal to some predefined global threshold value for any of the three similarity functions. Otherwise, it terminates and returns FALSE. The

threshold values Θ_1 , Θ_2 and Θ_3 are predefined, and they are global for all calls of the function.

The three similarity functions we use have been adapted from the ones originally proposed in Carenini et al. (2005). In that paper, the authors considered that each product category has a taxonomy that represents the main features of products. They then used these functions to evaluate matching between aspect expressions and terms that identify the product features in the taxonomy. We argue that the product features in the taxonomy play the same semantic role like attributes of a product catalog. However, in our work we generalized the original strategy proposed in Carenini et al. (2005) by matching aspect expressions not only to attribute titles, but also to attribute values of a given product. This is accomplished by the concept of attribute descriptors introduced in Section 4.3. In addition, we checked for matches between aspect expressions and the target product as a whole. In this case, we relied on the concept of product descriptors, also introduced in Section 4.3. We further generalized the original strategy by using the concept of expanded descriptors, which also include information from all products in the catalog. As we will discuss in Chapter 8, both generalized strategies led to improved results compared to the original strategy by Carenini et al. (2005).

In the following paragraphs, we are describing the adapted versions of three similarity functions from Carenini et al. (2005).

Function 1. This function consists of a simple comparison of a word of the aspect expression (w) with a word (δ) of the descriptor, as defined below:

$$\text{max_str_match}(w, \delta) = \begin{cases} 1, & \text{if } w = \delta \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

Function 2. This function employs WordNet and the classification of words into lexical categories or *part of speech* (POS). In WordNet words are grouped into sets of cognitive synonyms called *synsets*. Polysemous words belong to more than one synset. This function verifies whether two words appear in the same WordNet synset, given their POS. If any intersection occurs between the synsets of each word, the function returns 1, otherwise the function returns 0. This function uses another function $\text{syns}(w)$, which returns all synsets to which the word w belongs, considering all senses for w .

$$\text{max_syn_score}(w, \delta) = \begin{cases} 1, & \text{if } \text{syns}(w) \cap \text{syns}(\delta) \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (4.2)$$

Function 3. This function evaluates the degree of similarity between two words using information derived from a semantic network. We implemented the method proposed in Li et al. (2006), which defines the similarity between

two words as a combination of two functions $\ell(\alpha, \delta)$ and $h(\alpha, \delta)$, where ℓ gives the length of the shortest path between two words in WordNet, and h gives the height of the lowest common ancestor of the words in WordNet.

$$\text{max_sim_score}(\alpha, \delta) = \ell(\alpha, \delta) \cdot h(\alpha, \delta) \quad (4.3)$$

Notice that function ℓ alone could be used as a similarity function. However, according to Li et al. (2006), this function may be less accurate when applied to larger and more general semantic nets, such as WordNet. The reason for this is that words at upper layers of hierarchical semantic nets have more general concepts and less semantic similarity between words than words at lower layers. To address this drawback, the authors suggested that the result of ℓ must be adjusted by the function h , which uses hierarchical information. More details about this method can be found in Li et al. (2006).

The function $\text{max_sim_score}(\alpha, \delta)$ returns a normalized value between $[0,1]$, according the suggestion presented in Li et al. (2006).

4.5 The AspectLink Algorithm

Algorithm 2 presents a complete description of *AspectLink*. The algorithm receives as parameters a product catalog C and a set of reviews R for the products from C , and returns an enriched catalog C^+ , where each of its p^+ products are formed by p added with opinions. Our algorithm iterates through the set of products in C (Loop 1–33), and for each product p_i , two sequential phases are performed. In the first phase (Lines 3–10), the algorithm generates, from a set of reviews R_i on p_i , a set of opinions O whose target is p_i . In the second phase (Lines 13–32), the algorithm maps each opinion $o \in O$ to an attribute of the enriched version of p_i , p_i^+ . We will describe the algorithm in detail in the following paragraphs.

In the first phase, our method starts by breaking down each review $r \in R_i$ into sentences. In Line 6, the function $\text{extractSubjSent}()$ is used to extract the subjective sentences from each review r , since, as discussed in Section 3.2, only such kinds of sentences contain opinions. To accomplish this, the function $\text{extractSubjSent}()$ was implemented based on Qadir (2009).

Next, in Line 7, we eliminate comparative sentences through the function $\text{removeCompSent}()$, which accepts the subjective sentences which were extracted in the previous step. This function was implemented based on the method proposed in Liu (2010). This is done because sometimes users compare one product with another product, or one characteristic of one product with another. As our goal is to enrich each product of the catalog with the opinions of users regarding the specific product, we decided to eliminate comparative sentences, even if they are subjective. Therefore, the set of sentences DS will have only *Direct Opinionated Sentence* (DOS). In our experiments, we notice that there are very few sentences of these types of sentences in product reviews.

Algorithm 2: AspectLink

```

Input: A product catalog  $C = \{p_1, \dots, p_n\}$ 
Input: A set of reviews  $R$ 
Output: An enriched product catalog  $C^+ = \{p_1^+, \dots, p_n^+\}$ 
1 foreach product  $p_i \in C$  do
2   let  $p_i = \langle t, \mathcal{A} \rangle$ ;
3   let  $R_i$  be a set of reviews on  $p_i$ ;
4    $O \leftarrow \emptyset$ ;
5   foreach review  $r \in R_i$  do
6      $SS \leftarrow \text{extractSubjSent}(r)$ ;
7      $DS \leftarrow \text{removeCompSent}(SS)$ ;
8      $O_i \leftarrow$  opinions extracted from  $DS$ ;
9      $O \leftarrow O \cup O_i$ ;
10  end
11  let  $p_i^+ = \langle t, \mathcal{A}, \mathcal{S} \rangle$ , where  $\mathcal{S} = \mathcal{A} \cup \{\text{General}, \text{Other}\}$ ;
12  foreach  $S \in \mathcal{S}$  do  $O_{p_i, S} \leftarrow \emptyset$ ;
13  foreach opinion  $o \in O$  do
14    Matched  $\leftarrow$  FALSE;
15     $\alpha_o \leftarrow$  the aspect expression from  $o$ ;
16    foreach Attribute  $A \in \mathcal{A}$  do
17      let  $\Delta_{p_i, A}$  be a descriptor for  $A$ ;
18      if  $\text{Match}(\alpha_o, \Delta_{p_i, A})$  then
19        add  $o$  to  $O_{p_i, A}$  ▷ Case 1
20        Matched  $\leftarrow$  TRUE;
21      end
22    end
23    let  $\Delta_{p_i}$  be a descriptor for product  $p_i$ ;
24    if  $\text{Match}(\alpha_o, \Delta_{p_i})$  then
25      add  $o$  to  $O_{p_i, \text{General}}$  ▷ Case 2
26      Matched  $\leftarrow$  TRUE;
27    end
28    if Not Matched then
29      add  $o$  to  $O_{p_i, \text{Other}}$  ▷ Case 3
30    end
31  end
32   $p_i^+ \leftarrow p_i$  with opinions  $O_{p_i, S}$ ;
33 end
34 return  $C^+$ 

```

This is due to the fact that e-commerce site users focus on writing only about the product of interest, unlike what occurs, for instance, in forums where users usually write comments comparing the products.

In Line 8, the algorithm extracts the opinions from the remaining DOS. Recall from Section 3.2 that there can be more than one opinion per sentence. The extraction of opinions from sentences is carried out using standard methods from the literature. In the case of aspect expressions, which in our method guide the mapping of opinions to attributes, we implemented the well-known unsupervised aspect extraction method described in Poria et al. (2014). Finally, the opinions extracted in the previous step are added to O (Line 9).

In the second phase, opinions $o \in O$ are grouped by our method according to the attributes \mathcal{S} of p_i . For this, each aspect o will be “stored” in a set of opinion $O_{p_i,S}$, where p_i is the product being processed and S is an attribute in \mathcal{S} . In Line 12, we create an empty set of opinions $O_{p,S}$ for each attribute. For each opinion $o \in O$, we have three distinct strategies according to the cases defined in Section 3.3. Our strategy for Case 1 has been implemented in Lines 16 to 22, where we attempted to map each opinion o to some attribute. We used the function $Match()$ to verify whether the aspect expression α_o of opinion o matches the descriptor $\Delta_{p_i,A}$ of attribute A (Line 18). If it matches, the opinion o is added to the set of opinions $O_{p_i,A}$ (Line 19). Notice that the algorithm does not interrupt the loop even if the function $Match()$ returns TRUE. This happens because our method allows the same opinion to be mapped to more than one attribute. Thus, we let the current iteration continue, so that we can try to match the same aspect expression with descriptors of other attributes. We detailed the description of the $Match()$ function in Section 4.4.

Our strategy for Case 2 has been implemented in Lines 23 to 27, where we attempted to map each opinion o to the attribute **General** that represents the product p as a whole. In Line 24, we used the function $Match()$ to verify whether the aspect expression α_o of opinion o matches the descriptor Δ_{p_i} of the product. If it matches, opinion o is added to the set of opinions $O_{p_i,\text{General}}$ (Line 25).

To use these expanded descriptors, the only modifications required in Algorithm 2 are to replace $\Delta_{p,A}$ by $\Delta_{*,A}$ in Line 17 for Case 1, and to replace Δ_p by Δ_{p^*} in Line 23 for Case 2. However, we assume that before running the algorithm, all the expanded descriptors have been generated in a preprocessing step.

The strategy for Case 3 is quite simple. We consider that if there were no match in previous cases, then the opinion o will be added to the set of opinions $O_{p_i,\text{Other}}$ (Line 29).

In Line 32, the algorithm enriches current product p_i with opinions in $O_{p_i,S}$ and set to p^+ . This operation is performed for each p_i from C . Finally, the algorithm produces an enriched catalog C^+ in Line 34.

4.6 Summary

In this chapter, we have presented a method called *AspectLink* based on similarity functions for enriching product catalog with opinions extracted from reviews posted by users on e-commerce websites. An initial version of *AspectLink* was presented in a paper titled “*An Aspect-Driven Method for Enriching Product Catalogs with User Opinions*” (Melo et al., 2018). In this paper, the term *aspect class* was used to represent product attributes. However, we opted for not using that term in this thesis in order to standardize the text and leave it compatible with the terminology employed in our second proposed method.

We have extensively evaluated *AspectLink* by comparing it against a baseline (Carenini et al., 2005), and also analyzed the impacts of several parameters on the effectiveness of our method. The empirical evaluation of *AspectLink* has been reported in Chapter 8.

Although *AspectLink* has already given promising results in the problem of enrich product catalogs with user opinions, we have investigated using machine learning techniques as an alternative solution to this problem. The main motivation for this investigation is exploring the complete sentence of opinions instead of using the aspect expressions. The description of the second method is presented in next chapter.

In this chapter, we will present our second method for enriching product catalogs with opinions extracted from reviews written by the customers. In this method, called *OpinionLink*, the products of a given category in a catalog are represented by their attributes and the opinions are represented by the sentences that contain an opinion. Instead of the similarity functions used in *AspectLink*, *OpinionLink* addresses this problem by means of machine learning techniques.

This chapter is organized as follows. Section 5.1 presents an overview of *OpinionLink*. Section 5.2 describes how *OpinionLink* extracts opinions from opinionated sentences. Section 5.3 describes how our method maps these opinions to attributes of product catalog. Section 5.4 presents a bootstrapping method devised to automatically create training data that can be used for carrying out opinion mapping in a semi-supervised way. Finally, Section 5.5 presents a summary on this chapter.

5.1 Overview

In this section, we are presenting an overview of *OpinionLink*. As described in Algorithm 3, the proposed method is performed in two sequential phases. The main goal in the first phase is to extract a set of opinions O_i from a set of reviews R_i for each product $p_i \in C$. The main goal in the second phase is to map each opinion $o_j \in O_i$ to the attributes in the product catalog to which the opinions refer.

Following the concepts and terminology introduced in Chapter 3, Algorithm 3 receives as parameters a product catalog C and a set of reviews R for the products from C , returning as output an enriched catalog C^+ , where each of its p^+ products is formed by p with added opinions. The algorithm iterates

Algorithm 3: *OpinionLink*

Input: A product catalog $C = \{p_1, \dots, p_n\}$
Input: A set of reviews R
Output: An enriched product catalog $C^+ = \{p_1^+, \dots, p_n^+\}$

```

1 let  $C^+ \leftarrow \emptyset$ ;
2 foreach product  $p_i \in C$  do
3   let  $R_i$  be a set of reviews on  $p_i$ ;
4    $O_i \leftarrow \textit{opinion\_extraction}(R_i)$ ; ▷ first phase
5    $p_i^+ \leftarrow \textit{opinion\_mapping}(p_i, O_i)$ ; ▷ second phase
6    $C^+ \leftarrow C^+ \cup p_i^+$ ;
7 end
8 return  $C^+$ 

```

through the set of products in C (Lines 2 - 7), and for each product p_i , the two sequential phases are performed.

The first phase is encapsulated in function *opinion_extraction* (Line 4). This function receives a set of reviews R_i on product p_i and returns a set of opinions O_i , whose target is p_i . There are two main tasks in *opinion_extraction*. The first task is to identify whether each sentence in a review is a *Direct Opinionated Sentence* (DOS), as established in Definition 3.2, because only this kind of sentence contains a user opinion on the target product. We approached this task as a binary classification problem and investigated four different supervised classifiers for its realization. Further, we proposed a set of features extracted from each sentence to train these classifiers.

Notice that, contrary to what we did in *AspectLink*, in *OpinionLink* we devised a novel strategy specifically for DOS identification. We decided to do so because we wanted to improve the effectiveness of the strategy we have adopted while developing *AspectLink*.

Once the DOSs are identified, the next step is to extract the elements that compose each opinion. For this second task, we adapted and implemented standard methods from the literature. A detailed description of this first phase is presented in Section 5.2.

The second phase is encapsulated in function *opinion_mapping* (Line 5). This function receives a product p_i and a set of opinions O_i , which were extracted in the first phase and returns an enriched version of p , i.e., p_i^+ . The main task in *opinion_mapping* is to map the opinions that were previously obtained to attributes from the product catalog. We approached this task as a multiclass classification problem. We investigated four alternative supervised classifiers to predict one or more classes for each opinion $o \in O_i$, where each class corresponds to an attribute from p_i .

We also introduced a *sentence-segmentation strategy*, based on the idea that rather than the entire phrase, the core of an opinion can best be represented

only by the words occurring between the aspect expression and their sentiment words. We evaluated the effectiveness of this strategy in our experiments. This second phase is detailed in Section 5.3.

Finally, after the two phases are performed, each enriched product p_i^+ is added to an enriched version of C , designated as C^+ (Line 6).

In the following sections, the full process introduced above is detailed.

5.2 Opinion Extraction

In *OpinionLink*, the main goal of the first phase is to extract a set of opinions O_i from a set of reviews R_i for each product p_i in a product catalog C . Recall from Section 3.2 that an opinion is represented by a sextuple whose elements are obtained from a DOS. More formally, we define this task as follows:

Definition 5.1 *Let $R = \{r_1, r_2, \dots, r_m\}$ be a set of reviews on a product p , where each review $r \in R$ contains a set of sentences $ST = \{st_1, st_2, \dots, st_n\}$. The task of an opinionated sentence classifier is to predict whether each sentence st_j is a DOS or not, i.e., to learn a function $\mathcal{F} : st_j \rightarrow \{0, 1\}$ such that*

$$\mathcal{F}(st_j) = \begin{cases} 1 & \text{if } st_j \text{ is DOS} \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

After identifying DOSs, we can obtain the elements that compose each opinion. We assumed that aspects (a), sentiment words (w) and sentiment polarity (s) of opinions can be obtained using well-established existing extraction techniques (see Schouten and Frasincar (2016) for a comprehensive review of these techniques). The opinion holder (h) and posting time (t) can be extracted from reviews using simple parsing. These steps are performed for each sentence of each review until we construct the full set of opinions O_i for each product p_i .

5.2.1 Identifying Direct Opinionated Sentences

The problem of identifying opinionated sentences is frequently called *subjective classification* (Liu, 2012). According to Liu (2015), the majority of the methods for this problem are based on supervised learning (Palshikar et al., 2016; Rajkumar et al., 2014; Chenlo and Losada, 2014). The opinionated sentences identified by these methods can be *comparative* or *direct*. As our goal is to enrich a product catalog with the opinions of users regarding the specific product, we must discard comparative sentences, which mention more than one product. Although there are previous works on specifically identifying comparative sentences (Saritha and Pateriya, 2014; Jindal and Liu, 2006; Park and Blake, 2012), these cannot be used directly in our problem, because there are many sentences that are comparative, yet not subjective.

To address the problem of detecting DOSs, we used supervised learning using a set of features inspired in the works related to subjective classification. More specifically, each sentence in a review is represented by the commonly used *Bag-of-Words* (BOW) features, using TF-IDF values. Further, the set of features described below was added to the sentence representation.

- *Number of Adjectives*: As observed by Palshikar et al. (2016), there is a natural correlation between the presence of adjectives in a sentence and its subjectivity. Our intuition is that a sentence with many adjectives is rarely considered to be factual. In fact, we analyzed the datasets used in our experiments and found out that less than half of the non-direct opinionated sentences have an adjective, whereas nearly 90% of the DOSs have at least one adjective. Therefore, we decided to use the number of adjectives in the sentence as a feature.
- *Number of Words*: Rather than simply reporting a fact, users commonly use more words to describe their opinion in a sentence. Consequently, opinionated sentences tend to have more words than factual sentences. Therefore, as used for other related problems by Palshikar et al. (2016), we decided to use the number of words in the sentence as a feature.
- *Number of Comparative Words*: We considered the number of comparative words used in a sentence as a feature to identify comparative sentences. Recall that, as stated in the Definition 3.2, although comparative sentences are frequently subjective, we decided to discard them because we are interested only in direct sentences.
- *Number of Superlative Words*: We considered the number of superlative words used in a sentence as a feature to identify comparative sentences. As explained above, we decided to discard this kind of sentences.
- *Number of Adverbs*: Since adverbs are used to detect sentiments (Cambria et al., 2013), we decided to use the number of adverbs in the sentence as a feature.
- *Number of Nouns*: Riloff et al. (2003) reported the effectiveness of nouns for detection of subjective sentences. Therefore, we decided to use the number of nouns in a sentence as a feature.
- *Number of Adjectival Modifiers*: An adjective that modifies a noun is called an adjectival modifier (*amod*). For example, the sentence “*This phone has a great display*” contains the adjectival modifier “great”. We considered that an *amod* is an effective indicator of a DOS and we used it as a feature. We analyzed the datasets used in our experiments and determined that approximately 64% of the DOSs have at least one *amod*. Interestingly, none of the sentences that are not DOSs contains an *amod*.

- *Strength of Subjectivity*: The degree of subjectivity of a sentence is a relevant feature to distinguish an opinionated sentence from a factual sentence. Therefore, we decided to compute the subjectivity of a sentence using the method proposed by Smedt and Daelemans (2012). This method leverages WordNet to score subjectivity according to the English adjectives used in the text. It assigns a value between 0.0 and 1.0 to the subjectivity of the sentence. For example, according to this method, the sentence “*I purchased an Iphone 8 64 GB from Ting*” would have a subjectivity score 0.0, whereas the sentence “*Amazing phone and it has fast charging*” would have subjectivity score 0.75.
- *Polarity Score*: The polarity of a sentence is another relevant feature to distinguish an opinionated sentence from a factual sentence. We again used the method proposed by Smedt and Daelemans (2012) to compute the polarity of the sentence as a value between -1.0 and 1.0. For example, according to this method, the sentence “*I purchased an Iphone 8 64 GB from Ting*” would have a polarity score 0.0, whereas the sentence “*Amazing phone and it has fast charging*” would have a polarity score of 0.4.

Algorithm 4 details the *opinion_extraction* function, i.e., Phase 1 of *OpinionLink*.

Algorithm 4: Opinion Extraction (Phase 1)

Input: A set of reviews R_i on product p_i
Output: A set of opinions O_i

```

1 let  $O_i \leftarrow \emptyset$ ;
2 foreach review  $r \in R_i$  do
3   |  $\mathcal{S} \leftarrow \{\textit{sentence } s \mid s \in r \wedge \textit{Opinionated}(s)\}$ ;
4   |  $O_i \leftarrow \bigcup_{s \in \mathcal{S}} \textit{extract\_opinions}(s)$ ;
5 end
6 return  $O_i$ 

```

The algorithm iterates through the set of reviews $r \in R_i$ (Loop 2–5), where each review r is divided into sentences. For each sentence $s \in r$, function *Opinionated*, in Line 3, determines whether s is a DOS. To accomplish this, we adopt the BOW model, using the TF-IDF metric, and a binary classifier. Because of the large number of BOW features, we employed a strategy that selects the top k percentile of features with the highest score, according to a statistical test (ANOVA), to reduce the feature space. Further, we used the nine features described previously in this section to identify DOSs.

In Line 4, the algorithm extracts the opinions from each sentence $s \in \mathcal{S}$ using the function *extract_opinions*. Opinion holder, posting time and sentence having an opinion are extracted using a parser applied to the product’s landing

page. For the other components of opinions (aspect expression, sentiment words, and polarity), we implemented and adapted the unsupervised method proposed by Poria et al. (2014). Finally, the opinions extracted in this step are added to O_i .

5.3 Opinion Mapping

The main goal in the second phase of *OpinionLink* is mapping opinions extracted in the first phase to attributes from the product catalog, as stated in Definition 5.2.

Definition 5.2 Let $O = \{o_1, \dots, o_n\}$ be a set of opinions, A be the set of attributes from a product p , and $A^+ = A \cup \{\text{General}, \text{Other}\}$ be the set of attributes from an enriched version of p , designated as p^+ . The mapping task consists of learning a function $f : O \rightarrow 2^{A^+}$, where each opinion $o_i \in O$ is mapped to one or more attributes in A^+ .

We modeled the mapping task as a multilabel classification problem because we assumed that a single opinion can refer to more than one attribute. We adopted the *binary relevance method* of Zhang and Zhou (2014) for multilabel classification, which transforms a multilabel problem into multiple separate and independent binary problems, one for each label.

More formally, the proposed strategy consists of training a binary classifier $f_{A_j} : o_i \rightarrow \{0, 1\}$ for each attribute $A_j \in A^+$. Opinions that are known to refer to A_j are considered as positive examples to train f_{A_j} and all other opinions are considered as negatives examples. Once the classifiers are trained, function f of Definition 5.2 is specified as:

$$f(o_i) = \{A_j \in A^+ \mid f_{A_j}(o_i) = 1\} \quad (5.2)$$

This means that we map o_i to every attribute A_j where f_{A_j} applied to o_i yields positive.

We conducted a comparative analysis to identify the classifier that best fits the task of mapping opinions to attributes. We considered the following classifiers: *Maximum Entropy* (ME), *Random Forest* (RF), *Support Vector Machine* (SVM), and *Gradient Boosting Trees* (GBT). The discussion about this analysis is reported in Chapter 8.

5.3.1 Opinion-Mapping Algorithm

The opinion mapping process is described in Algorithm 5. The algorithm accepts as input a product $p_i \in C$ and a set of opinions O_i on p_i , and yields as output an enriched product p_i^+ (see Definition 3.3).

In Line 2, the algorithm creates a temporary expanded version of product p_i , designated p' , by adding new attributes $\langle \text{General}, v_G \rangle$ and $\langle \text{Other}, v_O \rangle$.

Algorithm 5: Opinion Mapping (Phase 2)

Input: A product p_i from product catalog C
Input: A set of opinions O_i on p_i
Output: An enriched version of p_i , p_i^+

- 1 **let** $p_i^+ \leftarrow \emptyset$;
- 2 **let** $p' \leftarrow p_i \cup \{\langle \text{General}, v_G \rangle, \langle \text{Other}, v_O \rangle\}$;
- 3 **foreach** $\langle A_j, v_j \rangle \in p'$ **do**
- 4 $O_j \leftarrow \{o \mid o \in O_i \wedge A_j \in f(o)\}$;
- 5 $p_i^+ \leftarrow p_i^+ \cup \{A_j, v_j, O_j\}$;
- 6 **end**
- 7 **return** p_i^+

It then iterates through the set of attributes of p' (Lines 3–6), and in each iteration, it generates a distinct set of opinions O_j , associated with attribute A_j (Line 4). The opinions that compose O_j are those which refer to attribute A_j according to the classifier f . Finally, in Line 5, the triple $\langle A_j, v_j, O_j \rangle$ is added to p_i^+ . Notice that in order to obtain the set of enriched products for each product p_i from the catalog, it is sufficient to execute this algorithm for each product p_i .

5.3.2 Sentence Core Segments

Recall from Section 3.2 that for each opinion o , there is a sentence st , from which this opinion was extracted. Each binary classifier f_{A_j} described above operates on these sentences. More specifically, we represent st as a feature vector, using the BOW model with TF-IDF as the weighting scheme, followed by a feature selection procedure, as discussed in Section 5.2.

However, early experiments have indicated that sentences frequently include numerous irrelevant terms that harm the classification process. Indeed, this same problem occurs in many other text classification problems as described by Khan et al. (2010) and Wang and Chiang (2007). To address this issue, we devised an additional feature selection strategy, based on the observation that the core of an opinionated sentence corresponds only to the segments of the sentence located between the aspect expressions and their sentiment words. This strategy introduces the concept of *Core Segments* of sentences.

Definition 5.3 (Core Segments) *Let $st = \langle w_1 \dots w_n \rangle$ ($n \geq 2$) be a sentence containing an opinion o . A core segment from st is any subsequent $cs = \langle w_b w_{b+1} \dots w_{e-1} w_e \rangle$ ($1 \leq b \leq e \leq n$) of st , where w_b is the leftmost word from either a sentiment or aspect expression from o , and w_e is the rightmost word from either a sentiment or aspect expression from o .*

For example, in the sentence “*The global warranty is pretty useful for me because I have to travel a lot*”, the core segment is “*global warranty is pretty useful*”, where “*global warranty*” is the aspect expression and “*useful*” is the sentiment word. In this case, the core segment has only one third of the words from the sentence, discarding several noisy words.

A single sentence can also contain more than one core segment. For example, the sentence “*So far I’m really happy with this phone, although it’s insanely expensive*” contains two core segments: $cs_1 = \text{“happy with this phone”}$ and $cs_2 = \text{“expensive”}$. In cs_1 , the core segment begins with a sentiment word (“*happy*”) and ends with an aspect expression (“*phone*”). However, cs_2 is comprised solely of an aspect expression. This occurs in the case of aspect expressions with an implicit sentiment, which do not include any explicitly associated sentiment words. Notice that there can be no core segments comprising only of sentiment words because any opinion must have exactly one aspect expression.

The proposed strategy consists of representing sentences only by their core segments. A discussion on the application of this strategy in our experiments is reported in the Chapter 8.

5.4 Bootstrapping Method

As presented in the previous section, *OpinionLink* uses classifiers for mapping opinions extracted in the first phase to attributes from the product catalog. Therefore, the classifiers devised in *OpinionLink* require labeled training data for achieving accurate predictions. In order to ease the labor of manually annotating sentences for generating training data, we proposed an unsupervised bootstrapping strategy to automatically create training data. Our key insight is to use *AspectLink* to create the training data, since it is unsupervised.

Algorithm 6 describes the process of generating a dataset to train the classifiers for the opinion mapping task, relying on the similarity functions described in Section 4.4. The algorithm takes as input a product p_i and a set of opinions O_i on p_i and returns as output a set of pairs $\langle o, A \rangle$, representing matches between opinions (o) and attributes (A).

In Line 1, the algorithm creates a temporary expanded version of product p_i , p' , by adding the new attribute $\langle \text{General}, v_G \rangle$. The attributes from p' will be used to create the attribute descriptors. The algorithm then iterates through the set of opinions $o \in O_i$ (lines 3–16). In each iteration, the algorithm tries to match the opinion o with every attribute A from p' . If there is a match between α_o and $\Delta(A)$, as explained in the Section 4.4, the pair $\langle o, A \rangle$ is added to the training set T (Line 9). Notice that an opinion can be labeled as being associated with more than one attribute.

In Line 13, the algorithm checks whether the opinion o is not associated with any of the attributes from p' . If this is the case, we assume that the

Algorithm 6: Bootstrapping Method

Input: A product p_i from product catalog C
Input: A set of opinions O_i on p_i
Output: A training dataset T

```

1 let  $p' \leftarrow p_i \cup \{\langle \text{General}, v_G \rangle\}$ ;
2 let  $T \leftarrow \emptyset$ ;
3 foreach  $o \in O_i$  do
4    $Matched \leftarrow \text{FALSE}$ ;
5    $\alpha_o \leftarrow$  aspect expression from  $o$ ;
6   foreach Attribute  $A \in p'$  do
7     let  $\Delta(A)$  be a descriptor for  $A \in p'$ ;
8     if  $Match(\alpha_o, \Delta(A))$  then
9        $T \leftarrow T \cup \{o, A\}$ ;
10       $Matched \leftarrow \text{TRUE}$ ;
11    end
12  end
13  if Not Matched then
14     $T \leftarrow T \cup \{o, \text{Other}\}$ ;
15  end
16 end
17 return  $T$ 

```

opinion o refers to some characteristic represented by the attribute **Other** and the pair $\langle o, \text{Other} \rangle$ is added to the training set T (Line 14). Finally, the algorithm outputs the training set T in Line 17. Inevitably, some pairs of the training dataset T will be incorrectly labeled. Since it is not possible to know in advance those pairs, we must use the whole set T to train the classifiers. A discussion on the application of this method is reported in the Chapter 8.

5.5 Summary

In this chapter, we presented a second method called *OpinionLink* for enriching product catalog with opinions extracted from reviews posted by users in e-commerce websites. We presented this method and the results achieved with it in a full paper titled *OpinionLink: Leveraging User Opinions for Product Catalog Enrichment*, which was accepted for publication at the Information Processing & Management Journal (IPM).

In *OpinionLink*, we proposed a novel method for identifying opinionated sentences (DOS), while in *AspectLink* our solution was based on a combination of existing work. Therefore, we consider this to be the first contribution of *OpinionLink*. In addition, *OpinionLink* achieved better results as compared to *AspectLink* in the opinion mapping task. This result was expected because we

used training data in *OpinionLink* and the novel proposed *sentence core strategy*. Therefore, we consider this to be the second contribution of *OpinionLink*.

In order to reduce the labor of manually annotating sentences for generating training data, we proposed a novel bootstrapping method. This method is a viable alternative to reduce the dependence on training data. We will report an empirical evaluation of this method in Chapter 8. This is the third contribution of *OpinionLink*.

The experimental datasets used to validate our proposed methods (*AspectLink* and *OpinionLink*) are reported in the next chapter.

To properly support our claims in this work and to evaluate the methods we developed, we needed experimental datasets that are non-existent so far in the literature. Thus, we had to create them ourselves and we did this using real data collections gathered from on-line sources available on the Web. These datasets will be made publicly available and we regard them as one of the contributions of this work.

This chapter is organized as follows. Section 6.1 presents an overview of the sources of our data used in validating our methods. Next, Section 6.2 and Section 6.3 describe each collection and the datasets we built from them. Section 6.4 presents a summary on this chapter and a summary of our experimental datasets.

6.1 Overview

Our experimental datasets were created from two distinct data collections, the *BestBuy* Collection and the *Amazon* Collection, obtained respectively from BestBuy.com and Amazon.com. As these two stores are very popular among consumers of electronic products, we believe that they are representative to evaluate our work. In particular, data from Amazon.com has been used in many previous studies on opinion mining (e.g., Liu et al. (2017); McAuley and Yang (2016); McAuley et al. (2015b)). Each of these collections is composed of two distinct datasets: a product catalog and a set of reviews. Although essentially similar in terms of structure and contents, we used each collection for different purposes in our work. The datasets from the BestBuy Collection were used for the sake of validating our methods. They are small, which allowed us to manually create reference datasets (golden standard) to be used in measuring the effectiveness of our methods in a very controlled scenario.

The datasets from the Amazon Collection were used to evaluate the feasibility of using our methods in a large scale scenario. In addition, these datasets were used in a study we carried out on the use of direct and indirect mentions in user opinions to attributes of a product catalog (Chapter 7).

6.2 BestBuy Collection

The BestBuy Collection was built using data that we have crawled from the BestBuy Web site¹. To form the product catalog, we crawled a set of products from five different categories along with their attributes and values: *cameras* (CAM), *cell phones* (CEL), *dvd players* (DVD), *laptops* (LAP), and *internet routers* (ROT). These categories are notoriously very popular among consumers of electronic products and they have been explored in many other previous works on opinion mining (Saumya et al., 2018; Kang et al., 2018; Wan and McAuley, 2016).

Regarding attribute values, following the definition in Chapter 4, the value of each attribute may have been built as a set that includes the values of its all sub-attributes, including multivalued attributes.

To form the set of reviews, we randomly selected a subset of the reviews from those available for each of these products on the BestBuy website. Table 6.1 presents the number of products, the number of reviews and the number of sentences extracted from these reviews for each category in the BestBuy Collection.

Category	No. products	No. reviews	No. sentences
CAM	12	291	790
CEL	20	372	1,009
DVD	8	160	388
LAP	20	383	1,135
ROT	10	210	614
Total	70	1,416	3,936

Table 6.1: Summary of the *BestBuy Collection*.

Table 6.2 shows the set of attributes available in the BestBuy website for the products of each of the categories we used.

The product catalog and set of reviews from this collection are named as *BestBuy Product Catalog* and *BestBuy Reviews Dataset* respectively. These datasets were used as inputs to our methods in the experiments we have carried out. Besides these datasets, two reference datasets were built from the BestBuy Collection. We are describing them in the following paragraphs.

¹<https://developer.bestbuy.com>

Category	Product Attributes
CAM	Dimension, Exposure Control, Imaging, Memory, Performance Power, Price, Zoom
CEL	Battery, Camera, Dimension, Display, Memory, Price Processor, Software
DVD	Accessory, Audio, Dimension, Price, Sound, Video
LAP	Battery, Connectivity, Dimension, Graphic, Memory, Price Processor, Screen, Software
ROT	Accessory, Coverage Area, Dimension, Ports, Price Security, Software, Speed

Table 6.2: Set of product attributes for each category in the BestBuy Collection.

BestBuy DOS Dataset

This dataset comprises of all the sentences extracted from the reviews of the BestBuy Reviews Dataset, and each sentence was annotated with the label *DOS* if the sentence is a *direct opinionated sentence*, and *NDOS* otherwise. The annotation process was manual and was based on Definition 3.2, i.e., direct opinionated sentence are subjective sentences that are not comparative.

Table 6.3 shows the distribution of the two types of sentences in this dataset. Notice that types of sentences are quite well-balanced across all product categories.

Category	DOS	NDOS
CAM	339 (56%)	267 (44%)
CEL	449 (44%)	560 (56%)
DVD	195 (52%)	177 (48%)
LAP	488 (47%)	537 (53%)
ROT	283 (46%)	324 (54%)
Total	1,745 (48%)	1,865 (52%)

Table 6.3: Distribution of sentence types by product category in the BestBuy DOS Dataset.

The BestBuy DOS Dataset was used as a reference (golden standard) to evaluate our methods in the task of identifying DOSs. Besides, it was used to train our supervised method for the same task (Section 8.3.1).

BestBuy Opinion Mapping Dataset

This dataset comprises of only DOS sentences from the BestBuy DOS Dataset. Each opinion o in a sentence from this dataset is annotated as follows: (i) if o is an opinion given on an attribute from the product catalog, o is annotated

with the *attribute name*; (ii) if o is an opinion given on the product as a whole, o is annotated with the label *General*; (iii) if o is an opinion given on a product characteristic that is not represented as an attribute in the product catalog, o is annotated with the label *Other*. This annotation is in accordance with our terminology from Chapter 3.

Figure 6.1 illustrates an example of a set of opinions labeled according to our description above. Notice that a DOS may have more than one opinion.

1. This is an ^[General] excellent computer, I'm ^[General] very pleased with this product.
2. This ^[Memory] device has adequate memory and is ^[Dimension] easily portable.
3. ^[Processor] Startup is very fast and the ^[Software] operating system and apps update seamlessly in the background.
4. ^[Price] Retail price is not bad, but if can get on sale for less than \$200 it is a ^[Price] great deal.
5. Despite the ^[Dimension] small size the ^[Other] keyboard is easy and comfortable to use.

FIGURE 6.1: Example of a set of opinions labeled.

Table 6.4 presents a summary of the number of opinions per type of target in the *BestBuy AspectLink Dataset*.

	CAM	CEL	DVD	LAP	ROT	TOTAL
<i>Attribute</i>	140 (34.6%)	266 (42.8%)	75 (27.3%)	329 (48.4%)	114 (31.6%)	810
<i>General</i>	127 (31.4%)	163 (26.3%)	60 (21.8%)	183 (26.9%)	45 (12.4%)	578
<i>Other</i>	138 (34.0%)	192 (30.9%)	140 (50.9%)	168 (24.7%)	203 (56.0%)	841
Total	405	621	275	680	362	1,663

Table 6.4: Distribution of opinions among targets.

The BestBuy Opinion Mapping Dataset was used to evaluate our methods in the task of opinion mapping. Besides, it was also used for training our supervised opinion mapping method (Section 8.3.2).

6.3 Amazon Collection

We built the experimental datasets that comprise of the Amazon Collection, a large collection of about 142 million reviews previously crawled from the Amazon.com website (McAuley et al., 2015a)².

From this collection, we selected all reviews from the same product categories as the BestBuy Collection: CAM, CEL, DVD, LAP, and ROT. These reviews comprise of the dataset called as the *Amazon Review Dataset*. Each

²Available at <http://jmcauley.ucsd.edu/data/amazon>

review in this dataset also identifies the exact product to which it refers. Table 6.5 presents the number reviews and sentences in the Amazon Review Dataset along with the number of products referred in the reviews for each category.

Category	No. products	No. reviews	No. sentences
CAM	8,893	203,836	1,012,077
CEL	7,693	182,491	707,407
DVD	2,503	61,836	243,939
LAP	9,491	115,138	580,955
ROT	1,592	84,059	329,305
Total	30,172	647,360	2,864,683

Table 6.5: Summary of the Amazon Review Dataset

To form the Amazon Product Catalog, we crawled for each product which was the target of at least one review of the Amazon Review Dataset of the corresponding attributes and their values, similar to what was done in the case of the BestBuy Product Catalog.

In summary, the Amazon Product Catalog and the Amazon Reviews Dataset are similar to their counter parts from the BestBuy Collection. These datasets were also used in a study we carried out on the use of direct and indirect mentions in the user opinions to attributes of a product catalog (Chapter 7).

Notice that the volume of data in the Amazon Collection is much larger than the data we crawled to form the BestBuy Collection. Therefore, we could not prepare the reference datasets from the Amazon Collection in fully manual way, as we did in the case of the BestBuy Collection. To remedy this, we built some working datasets with samples for the Amazon Review Dataset. These datasets are described below.

Amazon DOS Dataset

Differently from what we did in the case of the BestBuy Review Dataset, it was impractical to manually select DOSs from the sentences in the Amazon Review Dataset. Thus, to generate the Amazon DOS Dataset, we used the same strategy we applied to select DOSs in *AspectLink*. Specifically, we used function *extractSubSent()* to extract the subjective sentences from each review and then we used the function *removeCompSent()* to eliminate the comparative sentences. These functions were explained in Section 4.5. Out of the 2,864,683 sentences from the Amazon Review Dataset, 1,156,960 sentences were used to form the dataset, called the *Amazon DOS Dataset*.

Amazon-Top100A Dataset

The volume of sentences available in the Amazon DOS Dataset is still too large to allow a more detailed study on the sentences from the Amazon Collection. Thus, we further filtered this dataset as follows.

Firstly, we implemented the aspect extraction method proposed by Poria et al. (2014) and ran it over the Amazon DOS Dataset to extract all aspect expressions from the sentences composing it. Next, we ranked these expressions according to their frequency. To assure that we only use true aspect expressions, we manually inspected the extracted expressions using the ranking order, and removed those that we did not consider as aspect expressions. In the end, only 100 most frequent true aspect expressions were kept for each product category. We named this set of 100 aspect expressions as *Top-100 aspects* in each category.

Following, for each product category, we manually examined each of the Top-100 aspects and annotated each one with the product attributes that are most related to it. We also annotated accordingly cases where the aspect is on the product as a whole (**General**) and when opinion is on a characteristic of the target product that is not represented as attribute from a product catalog (**Other**). These annotations were propagated to each sentence in the Amazon DOS Dataset, wherein at least one of the Top-100 aspects occurs.

Finally, the set of all annotated sentences was used to compose a dataset, called the *Amazon-Top100A Dataset*. This dataset was mainly built to be used in a study we conducted to verify the use of direct and indirect mentions in the user opinions to attributes of product catalog (Chapter 7).

Table 6.6 presents a summary of both the Amazon DOS Dataset and the Amazon-Top100A Dataset. It shows the total number of DOSs (No. DOSs) and the total number of DOSs that include at least one of the Top-100 aspects (No. DOSs *top100*) for each category. As it can be observed, these DOSs represent more than 50% of all DOSs.

Category	No. DOSs	No. DOSs <i>top100</i>
CAM	476,605	249,714
CEL	277,712	138,939
DVD	89,525	48,608
LAP	189,782	126,865
ROT	123,336	73,027
Total	1,156,960	637,153

Table 6.6: Summary of Amazon DOS Dataset and the Amazon-Top100A Dataset.

Amazon-400 Reviews and Amazon Opinion Mapping Dataset

The Amazon-Top100A Dataset described above has all the opinions in it annotated with the attributes to which they refer. Although it is useful for the general characterization of the way customers mention attributes in reviews, it is not suitable for evaluating our opinion mapping methods. This is mainly due to the bias introduced by selecting sentences based on the aspects they contain. Thus, we derived another review dataset from Amazon Review Dataset. This dataset must be small enough to allow a manual annotation effort, but on the other hand, it must be representative to serve as input in our experiments.

This dataset, we called the *Amazon-400 Reviews Dataset*, was created by randomly sampling 400 DOSs from the entire Amazon Review Dataset. This number of sentences is sufficient to allow a confidence level of 95%, within a confidence interval of 5%, in the results of the experiments.

Next, the opinion in each sentence of the Amazon-400 Reviews Dataset was labeled to compose the *Amazon Opinion Mapping Dataset*. The procedure was the same as in the BestBuy Opinion Mapping Dataset. The Amazon-400 Reviews and the Amazon Opinion Mapping datasets were used to evaluate our supervised opinion mapping method. However, in this case we invited two annotators to label the same sentences. The average inter-annotator agreement on classifier prediction annotation was $k = 0.676$ (standard error = 0.0179) according to Cohen's Kappa statistic.

6.4 Summary

In this chapter, we described the datasets we built to perform experimental validations and to support our study on how mentions for product attributes are used in user reviews. We regard these datasets as one of the contributions of our research. Table 6.7 presents a summary of these datasets, where each dataset has specified role and the method which was used for each dataset. All the datasets discussed in this section are publicly available for download at <https://goo.gl/uZQJjb>.

Collection	Dataset	Role	Methods
BestBuy	BestBuy Product Catalog	Input	<i>AspectLink</i> <i>OpinionLink</i> <i>OpinionLink</i> with bootstrapping
	BestBuy Reviews	Input	<i>AspectLink</i> <i>OpinionLink</i> <i>OpinionLink</i> with bootstrapping
	BestBuyDOS	Reference	<i>OpinionLink</i>
	BestBuy Opinion Mapping	Reference	<i>AspectLink</i> <i>OpinionLink</i> <i>OpinionLink</i> with bootstrapping
Amazon	Amazon Product Catalog	Input	<i>OpinionLink</i>
	Amazon Reviews	Analysis	
	Amazon DOS	Analysis	
	Amazon-400 Reviews	Input	<i>OpinionLink</i>
	Amazon-Top100A	Analysis	
	Amazon Opinion Mapping	Reference	<i>OpinionLink</i>

Table 6.7: Summary of datasets created in this thesis.

While making purchasing decisions, customers usually rely on the information from two types of sources: product specifications, provided by the manufacturers, and reviews, posted by the customers. Both kinds of information are often available on e-commerce websites. While researchers have demonstrated the importance of product specifications and reviews as separate and valuable sources to support decision making of a purchase, a mostly uninvestigated issue is: what is the relationship between these two kinds of information?

In this chapter, we will present an empirical study on the use of direct and indirect mentions in user reviews to attributes of a product catalog. This study aims at answering the question $\mathcal{RQ1}$ formulated in Chapter 1: “*Are there evidences that the most important product characteristics for people are represented by the attributes of the product catalogs?*” Briefly, our results indicate that user opinions are indeed guided by the attributes from product catalogs and highlight the influence of attributes of product catalog on the user reviews.

This chapter is organized as follows. Section 7.1 presents a brief background of the importance of product catalogs and user reviews. Section 7.2 describes the hypotheses we have formulated for this study. Then, main results and findings that support our hypotheses formulated on the impacts of attributes of product catalog on user opinions are presented in Section 7.3. Finally, Section 7.4 concludes by discussing our results.

7.1 Background

In typical e-commerce websites, descriptions of products in the catalog usually consist of objective (factual) data provided by the manufacturers informing

customers about product's characteristics, which are represented as a set of previously defined product attributes. For instance, for laptops, the brand, the weight and the processor model are commonly available to help potential customers make their purchase decisions. On the other hand, with the rise of the so-called Web 2.0, there is also a large amount of subjective (opinionative) information available about products and their characteristics. In most cases, this subjective information is provided by the opinions issued by other customers in reviews.

The importance of considering subjective information in addition to objective (factual) information has been verified in many e-commerce related applications (PwC, 2016; Smith and Anderson, 2016). Indeed, considering opinions issued by other people before purchasing a product is a common practice, especially since there are plenty of opinions available on the Web. On the other hand, product attributes also comprise of valuable sources to support decision making of a purchase (Kostyra et al., 2016; Park et al., 2012). According to Park et al. (2012), product attributes on the websites encourage consumer browsing behavior, which can often lead to impulse buying behavior. Therefore, product attributes are a crucial element that influences customer product choice (Kostyra et al., 2016).

Despite the substantive importance of user opinions and product attributes for customer decisions, the relationship between these two kinds of information has been generally overlooked. Motivated by the above observations, in this chapter, we will empirically study the impacts of product attributes on user opinions. Specifically, we will analyze the use of direct and indirect mentions to attributes of product catalogs, defined by the manufactures in product specifications, and written by customers in the reviews. Our ultimate goal is to extend previous research studies that assessed the importance of product attributes and user opinions in a separate way. To this end, we executed an extensive experimental evaluation using a large number of user reviews using datasets from the Amazon Collection described in Chapter 6.3. The results of this study revealed evidences that the most important product characteristics for people are represented by the attributes of the product catalogs.

7.2 Research Hypotheses

In order to answer the question $\mathcal{RQ1}$ proposed in Chapter 1, we will formulate and investigate the following hypotheses in this study:

HS1. E-commerce websites are a valid source of opinions on *target products*.

As explained in Section 3.3, we observed that reviews also include opinions that do not refer to an attribute of product catalog. Opinions may refer to the product as a whole (**General**) or to a characteristic of the target product that

is not represented as an attribute of product catalog (**Other**). In spite of that, we do expect that attributes of a product catalog have a persuasive effect on online user reviews. Hence, we hypothesize:

HS2. Most of the user opinions posted on e-commerce websites is about attributes of product catalogs.

HS3. According to the user opinions, there are certain attributes of product catalogs that are more relevant than the other attributes.

HS4. Customers often make some of them indirect mentions to attributes of product catalogs.

7.3 Results

In this section, we are presenting the main results and findings that support our hypotheses formulated on the impacts of attributes of product catalog on user opinions. In this study, we used the *Amazon-Top100A Dataset* described in Section 6.3.

7.3.1 Use of Directed Opinionated Sentences (DOS)

Our first result is on the use of directed opinionated sentences (DOS) in user reviews. For this, we compared the full set of sentences from the Amazon Collection, i.e., the Amazon Review Dataset, with the set of the sentences considered as DOS, which correspond to the Amazon DOS dataset.

Figure 7.1 presents the percentage of sentences among the three types: *factual*, *comparative*, and *DOS*. We can observe that on the average, 51.31% of the sentences were considered as factual, 40.05% were classified as DOSs, and only 8.64% were comparative. It is noticeable that there are very few comparative sentences in product reviews. This is probably due to the fact that e-commerce site users focus on writing only about the product of interest, unlike what occurs, for example, in forums where users usually write comments comparing the products.

The fact that a large fraction of the sentences is DOS across all categories supports our hypothesis *HS1* that e-commerce websites, such as Amazon.com, are indeed useful as a valuable source of opinions on target products.

It is also interesting to notice that from what was shown in Table 6.6, more than 55% of the DOSs contain at least one of the Top-100 aspects from the Amazon Review Dataset. This finding corroborates our assumption that handling a few top frequent aspect expressions is more valuable than showing every single aspect expression from a potentially huge list.

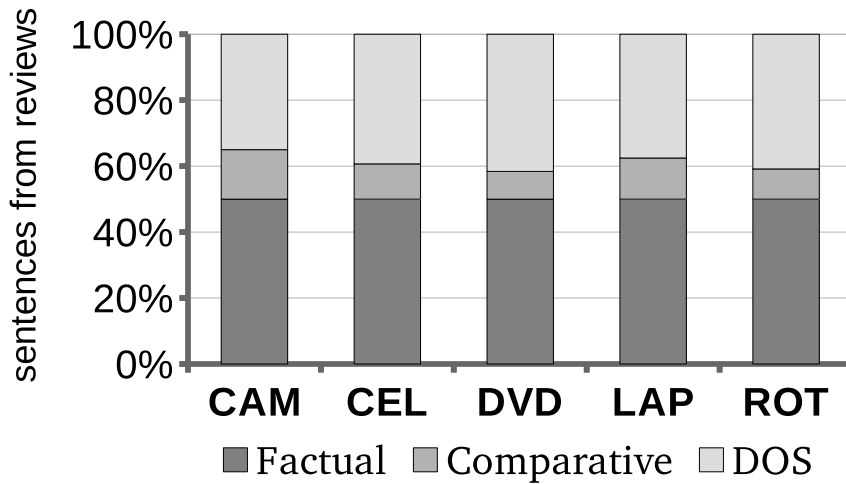


FIGURE 7.1: Percentage (%) of sentences of each type in user reviews.

7.3.2 Distribution of Sentences among Kinds of Targets

Figure 7.2 summarizes the distribution of sentences among the three kinds of targets: **Attributes**, **General**, and **Other** in Amazon-Top100A Dataset. As explained in Section 3.2, a single sentence may contain more than one opinion, and each opinion can refer to a different kind of target. Thus, the sum of percentage of all kinds of targets may be greater than 100%.

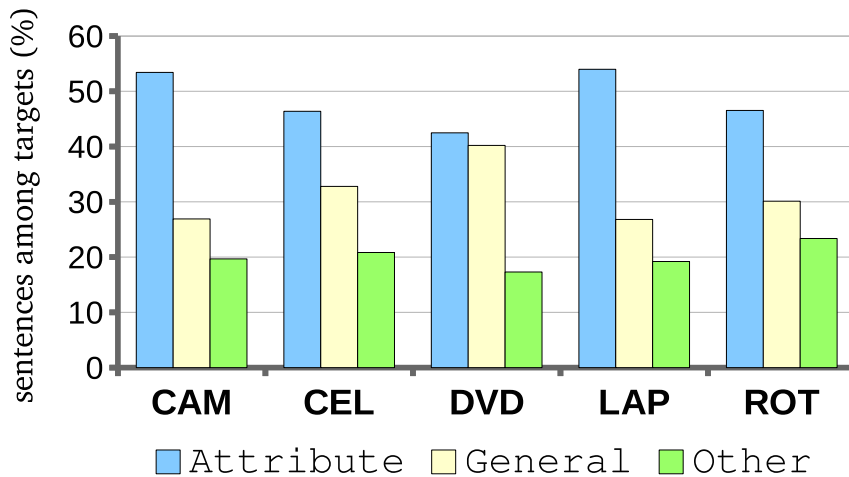


FIGURE 7.2: Distribution of sentences among targets.

Again we observe that most of the DOSs includes aspect expressions that refer to attributes of product catalog, identified as **Attribute**. For example, in CAM and LAP categories they account for more than half of the sentences.

Also, a large share of the sentences contains opinions referring to the target **General**.

7.3.3 Distribution of Aspect Expressions

Figure 7.3 shows the distribution of the Top-100 aspects among the three kinds of targets. Notice that for all categories the fraction of aspect expressions that represent the target **Attribute** is higher than 50%. This supports our hypothesis $\mathcal{HS2}$; most of the user opinions posted in e-commerce websites is about the attributes of product catalog. In addition, it can be observed that a larger share of the aspect expressions refers to the target **Other**. For example, in CAM, CEL and DVD, almost one quarter of the aspect expressions are in opinions which were annotated as the target **Other**.

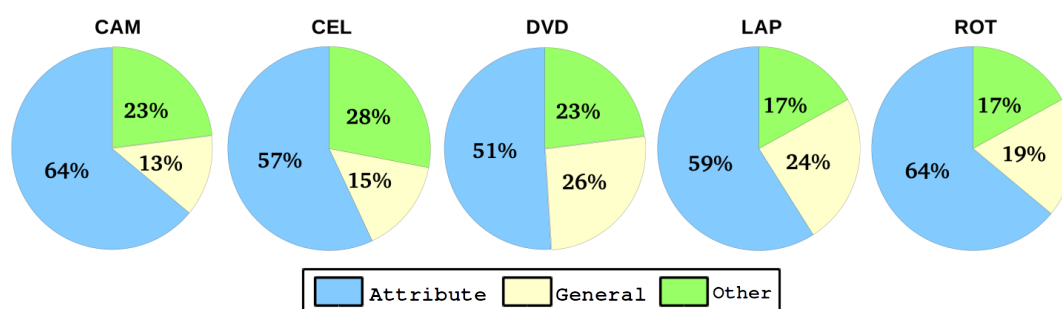


FIGURE 7.3: Distribution of the top 100 aspects among the three kinds of targets.

An intriguing problem we left for future work is to further analyze cases such as these to look for specific latent characteristics that, although not represented by some attribute of product catalog, are of interest for users. For example, “keyboard” is the second most frequent aspect expression in LAP category, but typically, there is no attribute referring to it in the attributes of product catalog.

In sum, Figure 7.3 suggests that users comment more frequently on the specific characteristics of the products than on the product as a whole. This shows the relevance of properly addressing references to attributes in user reviews.

7.3.4 Distribution of Sentences among Product Attributes

Figure 7.4 shows the distribution of sentences among the product attributes for each category. In these graphs, each vertex in the polygon represents a product attribute defined by the manufacturers in product specifications. The graph shows the percentage of sentences that contain an aspect expression

that corresponds to a given attribute. In each graph, the attributes are placed in clockwise order, from the most to the least frequently referred. For example, more than 40% of the sentences that include at least one of the Top-100 aspects in the DVD category refer to the attribute **Accessory**.

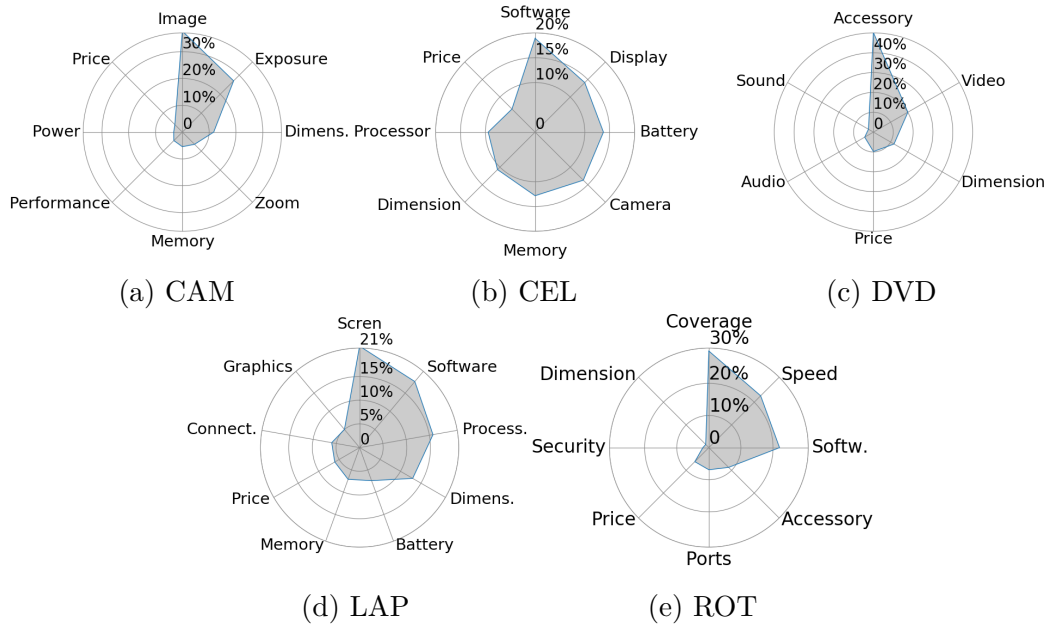


FIGURE 7.4: Distribution of sentences among the attributes they refer to. Labels are product attributes.

There are some attributes that are more frequently referred to in reviews than others from the same category. For example, in the CEL category, users comment twice more on **Battery** than on the **Price** of cell phones. These results support our hypothesis $\mathcal{HS3}$; there are certain attributes that are more relevant than the other. Interestingly, in the five categories in this experiment, the price is not the most commented attribute.

7.3.5 Diversity of Aspect Expressions over Attributes

Figure 7.5 shows the distribution of aspect expressions extracted from user reviews over product attributes in each category. In these graphs, we show the quantity of unique aspect expressions that refer to the same attribute. For example, in the LAP category, we found ten different aspect expressions that refer to the attribute **Software**. Analyzing the sentences, we found that users do indeed employ several different terms such as “apps”, “system”, “vista”, and “program” to refer the attribute **Software** in the LAP category. This experimental evaluation supports our hypothesis $\mathcal{HS4}$; customers often make either direct and indirect mentions to product attributes.

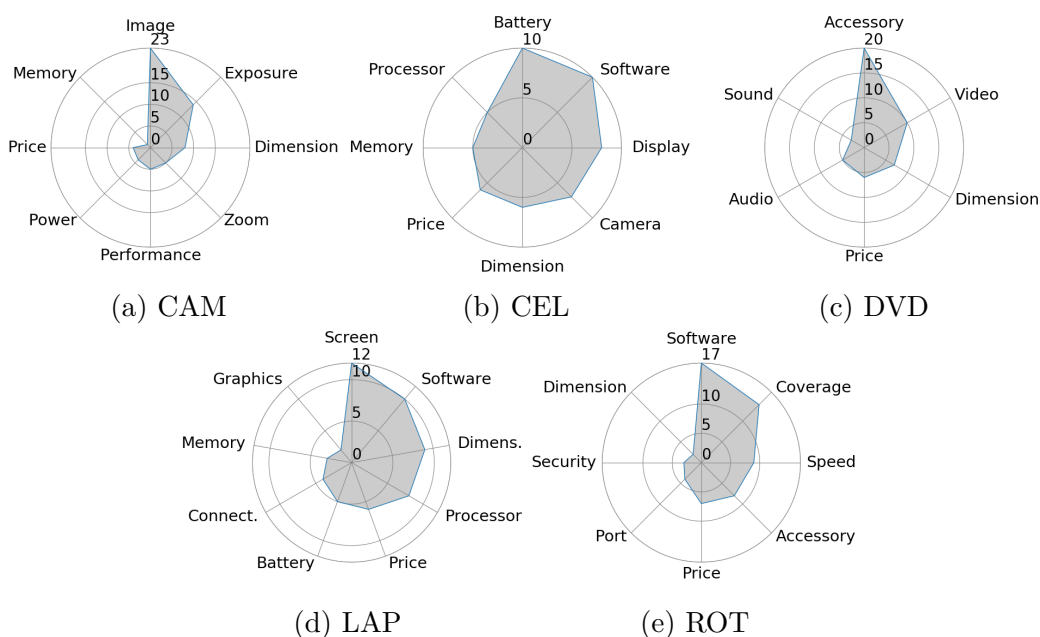


FIGURE 7.5: Distribution of different aspect expressions according the product attributes presented in Table 6.2.

To give an idea of the Top-100 aspects mentioned, Table 7.1 illustrates the ten most frequent aspect expressions extracted in the reviews of each category along with the attribute name, when they refer to **Attribute** target, or the target name (**General** or **Other**). From these results, it is apparent that the ten most frequent aspect expressions extracted are quite representative of each product category and, more importantly, the results show which are the most commented aspects related to attributes. Notice that most of aspect expressions do not match exactly with the name of the product attribute.

CAM	CEL	DVD	LAP	ROT
camera (General)	phone (General)	unit (General)	laptop (General)	router (General)
quality (General)	easy (Other)	dvd (General)	keyboard (Other)	easy (Other)
picture (<i>Imaging</i>)	quality (General)	easy (Other)	computer (General)	instructions (Other)
lens (<i>Exposure Control</i>)	feature (Other)	quality (General)	software (<i>Software</i>)	software (<i>Software</i>)
shots (<i>Exposure Control</i>)	card (<i>Memory</i>)	player (General)	fast (<i>Processor</i>)	unit (General)
features (Other)	size (<i>Dimension</i>)	picture quality (<i>Video</i>)	size (<i>Dimension</i>)	speed (<i>Speed</i>)
size (<i>Dimension</i>)	software (<i>Software</i>)	instructions (Other)	card (<i>Memory</i>)	device (General)
card (<i>Memory</i>)	camera (<i>Camera</i>)	features (Other)	screen (<i>Screen</i>)	internet (<i>Coverage Area</i>)
settings (General)	screen (<i>Display</i>)	picture (<i>Video</i>)	easy (Other)	network (<i>Coverage Area</i>)
pics (<i>Imaging</i>)	keyboard (Other)	product (General)	graphics (<i>Graphics</i>)	settings (Other)

Table 7.1: The 10 most frequent aspect expressions in reviews of each category from the Amazon-Top100A Dataset.

7.4 Summary

The main goal of the study presented in this chapter was to investigate the impacts of product attributes on user opinions. Based on an empirical evaluation carried out over a representative collection of real user reviews, we were able to verify hypotheses we have formulated on this issue. Our results were drawn from a large experimental dataset described in Section 6.3 with more than 1 million of direct opinionated sentences (DOSs) in five product categories.

In our study, we verified that a large fraction of sentences in reviews is composed of direct opinionated sentences, which validates our hypothesis that e-commerce websites are a valuable source of opinions on target products ($\mathcal{HS1}$). We used a large number of sentences and, by means of a well-defined protocol we could verify that the most of user opinions posted in e-commerce websites is on one of the product attributes ($\mathcal{HS2}$). Furthermore, we could verify that there are certain attributes that are more relevant for users than the other attributes ($\mathcal{HS3}$). Finally, we could conclude that customers often make either direct and indirect mentions to product attributes using several distinct expressions ($\mathcal{HS4}$).

This study contributes to understanding the impacts of product attributes on user opinions. In sum, the results of this study indicate that user opinions are indeed guided by the attributes from product catalogs and highlight the influence of attributes in user reviews.

Some limitations are associated with this study, which, however, can provide directions for future research. Firstly, we considered only aspect expressions to represent user opinions. Future research could extend the current study for examining other components of opinions, such as star ratings of reviews, opinion polarity, and opinion posting time. Secondly, our analysis is restricted to products, which in turn, have well established set of attributes provided by manufacturers. However, domains such as restaurants or hotels do not have clear attributes. Therefore, a future research could extend the current study to these domains.

We have reported this study in a short paper submitted to a major international conference, which is currently under revision.

In this chapter, we will present an empirical evaluation of our proposed methods *AspectLink* (Chapter 4) and *OpinionLink* (Chapter 5). This chapter is organized as follows. We will start this chapter by defining the metrics used for evaluating the results of the experiments we carried out (Section 8.1). Then, we will report the results of empirical evaluations of *AspectLink* in Section 8.2 and *OpinionLink* in Section 8.3. Next, Section 8.4 presents a comparative analysis between *AspectLink* and *OpinionLink*. Finally, Section 8.5 presents a summary on this chapter.

8.1 Evaluation Metrics

We used the well-known *precision*, *recall*, and F_1 evaluation metrics (Baeza-Yates et al., 2011). These metrics are calculated as follows. Let A be the set of correct answers, according to a reference set, and let B be the set of answers generated by the method being evaluated. We define precision (P), recall (R) and F_1 as:

$$P = \frac{|A \cap B|}{|B|} \quad R = \frac{|A \cap B|}{|A|} \quad F_1 = \frac{2 \times (P \times R)}{(P + R)}$$

8.2 *AspectLink* – Experimental Evaluation

This experiment consists of first running *AspectLink* using the BestBuy Reviews Dataset and the BestBuy Product Catalog as input. This produces as a final result a mapping of the opinions in each review to one of the attributes of the products in the catalog. Then, we evaluate this result having the BestBuy Opinion Mapping Dataset as a reference (golden standard). The datasets used

in this experiment were described in Section 6.2. We notice that by the time we run experiments with *AspectLink*, we did not have the data corresponding to the CEL category. Thus, we only report results in the remaining four categories. Nevertheless, in Section 8.4, data from all the five categories were used in the comparison we made between *AspectLink* and *OpinionLink*.

We used the method proposed by Carenini et al. (2005) to serve as a baseline for comparison in the opinion mapping task. Recall that this method uses the original versions of the word similarity metrics, which we adapted for *AspectLink*. This method requires the input of a taxonomy of product features for a particular category. The purpose is to map each discovered aspect expression to a node in the taxonomy based on similarity functions.

We implemented this method according to the paper, assuming that the product features in the taxonomy play the semantic role like that of the attributes of a product catalog. As this method works by matching aspect expressions to attribute names, we used the most significant attribute names in the catalog of each product category to ensure a fair comparison. Also, as this method does not generate mappings to the attributes **General** and **Other**, we excluded the mappings for these attributes while evaluating the baseline.

8.2.1 General Results

Table 8.1 compares the results achieved by *AspectLink* and the baseline in the experiment. With both methods, we experiment building descriptors with and without applying stemming functions. These functions were performed through the traditional Porter algorithm (Porter, 1997). We use the symbol ^S to indicate when the method uses stemming.

Method	CAM			DVD			LAP			ROT		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Baseline	0.63	0.17	0.26	0.75	0.46	0.57	0.64	0.34	0.44	0.41	0.33	0.37
<i>AspectLink</i>	0.87	0.82	0.84	0.89	0.82	0.86	0.75	0.57	0.65	0.56	0.49	0.52
Baseline ^S	0.59	0.13	0.21	0.83	0.28	0.41	0.87	0.28	0.43	1.00	0.27	0.42
<i>AspectLink</i> ^S	0.88	0.82	0.84	0.88	0.71	0.78	0.88	0.52	0.65	0.77	0.59	0.67

Table 8.1: Precision, recall and F₁ for *AspectLink* and the baseline with and without using stemming.

Our method achieved higher F₁ values in all categories compared to the baseline. As expected, this is mainly due to the high increase in recall values. On the average, the recall values obtained by our method are almost three times higher than those obtained by the baseline. Interestingly, in the majority of the cases, our precision values are also higher. In a single case, the baseline achieved a higher precision, but with a very poor recall. These results indicate that using attribute values available in the product catalog decisively contributed to the improvement in recall. For instance, the sentence “*The Intel i7 works flawlessly*

with all my application programs including PhotoShop” has an opinion whose aspect expression “Intel i7” refers to the processor of the laptop. Thus, we should map this opinion to the attribute `Processor`. We argue that our method could map it correctly because *AspectLink* uses the brand value of the `Processor` attribute available in the catalog. Using only the attribute names would not yield the correct mapping.

Another issue that we analyzed was the influence of using stemming functions in the method effectiveness. As demonstrated in Table 8.1, in general, using stemming functions helped improving precision and recall for both methods, and we had a slight reduction in precision in just a few cases. Notice that in general, the goal of stemming is to increase recall, and in practice, it may lead to a reduction in precision as a side effect. This undesirable effect was not observed in our experiments, because when we used stemming in similarity functions the method returns a smaller amount of possible matches between an aspect and the descriptors when stemming is not used. For example, the “range” aspect in the ROT category is mapped only to the attribute `Coverage Area` when the method is using stemming, but this same aspect is mapped to `Dimension` and `Coverage Area` attributes when the method is not using stemming. In this example, the precision would be not decreased. In the case of *AspectLink*, in a single case, in the DVD category, the results obtained using stemming functions had a noticeable impact on recall, and as a consequence on F_1 . In this particular case, this is due to the fact that users commonly use acronyms as aspect expressions, and the stemming functions are generally unable to handle acronyms properly.

8.2.2 Common vs. Expanded Descriptors

In Section 4.3, we discussed how product descriptors are built. Two options were considered: common descriptors, which use values of the attributes and the title of the target product only, and *expanded* descriptors, which use values of the attributes and titles of all the products in the catalog that are from the same category as the target product. Table 8.2 compares the results obtained by *AspectLink* using common and expanded descriptors.

Method	CAM			DVD			LAP			ROT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
<i>AspectLink</i> _{common} ^S	0.88	0.82	0.84	0.88	0.71	0.78	0.88	0.52	0.65	0.77	0.59	0.67
<i>AspectLink</i> _{expanded} ^S	0.83	0.85	0.84	0.88	0.72	0.79	0.84	0.63	0.72	0.74	0.62	0.68

Table 8.2: Results of using common and expanded descriptors in *AspectLink*.

Using expanded descriptors led to higher recall values in all categories, with a comparatively small loss in precision. As a consequence, F_1 values with expanded descriptors are higher or equal to those obtained with common

descriptors. This experiment corroborates our motivation for considering expanded descriptors. As discussed in Chapter 4, by using this kind of descriptor, we were able to enrich the representation of attributes or the product as a whole, approximating it from the attribute domain. This explains the increase in recall observed in Table 8.2. From this point on, we will use expanded descriptors in the remaining experiments reported in this section.

8.2.3 Similarity Functions

To better understand the results achieved with *AspectLink*, it is interesting to take a deeper look at each similarity function used in our method. Remember from Chapter 4 that *AspectLink* applies the functions *str_match*, *syn_score* and *sim_score* in sequence. Initially, it uses function *str_match* to map the aspect expressions. Then it uses function *syn_score* to map the aspect expressions, which were not mapped in the previous step. Finally, it uses function *sim_score* to map the aspect expressions that were not mapped by the two previous functions. We ran a specific experiment to verify the cumulative effect of applying the function according to this sequence. The results are shown in Table 8.3. The precision decreases slightly or remains the same, after each function is used, but there is a significant gain in the recall in almost all the categories. This demonstrates that combining various similarity functions has a positive impact on the overall performance.

	CAM			DVD			LAP			ROT		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
<i>str_match</i>	0.86	0.83	0.84	1.00	0.54	0.70	0.92	0.55	0.69	0.88	0.40	0.55
<i>str_match+syn_score</i>	0.86	0.83	0.84	1.00	0.60	0.74	0.88	0.56	0.69	0.88	0.41	0.56
<i>str_match+syn_score+sim_score</i>	0.83	0.85	0.84	0.88	0.72	0.79	0.84	0.63	0.72	0.74	0.62	0.68

Table 8.3: Results of similarity functions applied cumulatively.

In addition to this experiment, we also analyzed the performance of each similarity function individually in comparison to using all of them sequentially as in *AspectLink*. The results are presented in Figure 8.1. Although *str_match* and *syn_score* achieved better values for precision individually, *AspectLink* achieved higher values for recall and F₁ in all the categories. On the average, our gains in F₁ over *syn_match*, *syn_score* and *sim_score* considered alone were about 0.06, 0.42, and 0.17, respectively. This demonstrates that no single similarity function has better results than *AspectLink* and our method is able to perform well in different categories.

8.2.4 Estimating parameter Θ_3

The three similarity functions use global threshold parameters that determine when a given aspect expression α and descriptor Δ match according to the

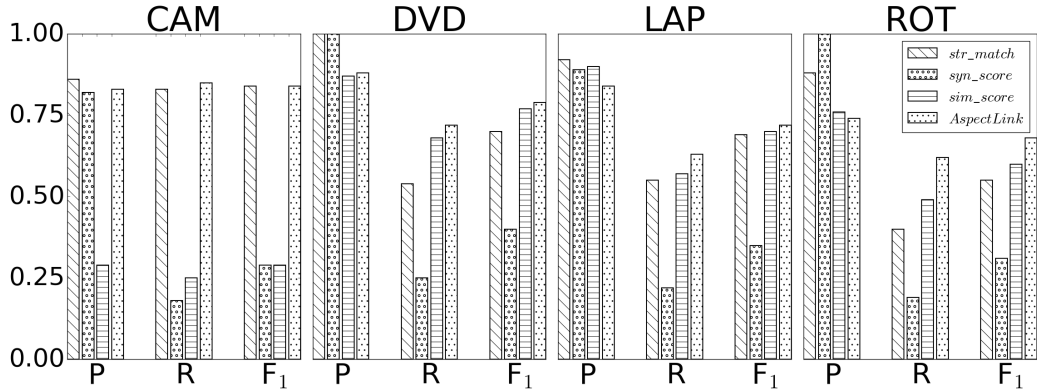


FIGURE 8.1: Precision, recall and F_1 results comparing *AspectLink* to each similarity function applied individually.

function. In the case of functions *str_match* and *syn_score*, their respective threshold values Θ_1 and Θ_2 must be equal to 1, since they only allow exact matches. In the case of function *sim_score*, we must have $0 < \Theta_3 \leq 1$. The experiments described so far all use $\Theta_3 = 0.5$. This value is the same as suggested in Carenini et al. (2005). To corroborate this choice, we performed experiments with different values of Θ_3 . The results are presented in Figure 8.2, where we plot F_1 values obtained while varying Θ_3 from 0.1 to 0.9. As shown, $\Theta_3 = 0.5$ produces the best average result among 4 categories of products.

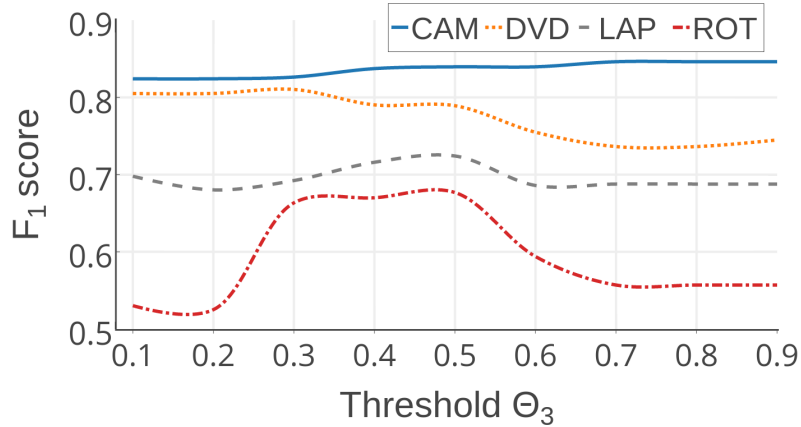


FIGURE 8.2: Influence of threshold Θ_3 in our method for each category of products.

8.2.5 Discussion

The results obtained in this experiment indicate that *AspectLink* is effective for the task of mapping opinions to product attributes based on the aspect expres-

sions that form the opinion. In particular, the results from Tables 8.1 and 8.2, and from Figure 8.2, suggest that the configuration that uses *AspectLink* with stemming functions, expanded descriptors and parameter $\Theta_3 = 0.5$ can be used in practice to perform this task.

8.3 *OpinionLink* – Experimental Evaluation

In the case of *OpinionLink*, we evaluated the two phases described in Section 5.1: opinion extraction (Section 5.2) and opinion mapping (Section 5.3). For the opinion extraction phase, we focused on its core task, which is identifying *direct opinionated sentences* (DOS) (Section 5.2.1). For extracting the other elements that compose each opinion (e.g., aspect, sentiment words, polarity), we relied on well-established methods in the literature (Schouten and Frasincar (2016)). These methods will not be further commented, since they are out of scope of this thesis. For the opinion mapping phase, we focused on the evaluation of the classifiers used to map opinions extracted in the first phase to attributes from the product catalog.

In addition to evaluating these two phases separately, we also evaluated *OpinionLink* as a realistic end-to-end application, where the results generated in the first phase influence the performance of the second phase. Moreover, we reported and discussed the results achieved by *OpinionLink* when applied to a large-scale dataset (Section 8.3.4). Finally, we evaluated our bootstrapping strategy proposed to automatically create training data for the classifiers.

8.3.1 Identifying Direct Opinionated Sentences

To evaluate our proposed method for the task of identifying DOSs, we used as input the BestBuy Review Dataset and evaluated its result using the BestBuy DOS Dataset as a reference. The datasets used in this experiment were described in Section 6.2.

We conducted this experiment using the following classifiers: Maximum Entropy (ME), Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting Trees (GBT). Moreover, we reported three different representations for the sentences used as input to the classifiers, namely:

- *BoW*: Traditional “bag of words” with the TF-IDF weighting scheme. We removed stop-words and applied unigrams plus bigrams.
- *Feat*: Each sentence was represented as a vector with only the nine features proposed in Section 5.2.1.
- *BoW+Feat*: The features used in *Feat* were added to *BoW* as a representation of the sentences.

Table 8.4 displays the classifiers’ performance in terms of precision (P), recall (R), and F_1 measure (F_1). The highest values for each category are marked in bold. In this experiment, each result denotes an average of 10-fold cross-validation.

	CAM			CEL			DVD			LAP			ROT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ME _{BoW}	0.82	0.78	0.80	0.90	0.82	0.86	0.85	0.77	0.81	0.80	0.75	0.77	0.88	0.79	0.83
ME _{Feat}	0.86	0.70	0.77	0.92	0.87	0.89	0.85	0.77	0.81	0.84	0.71	0.77	0.82	0.80	0.81
ME _{BoW+Feat}	0.85	0.86	0.85	0.94	0.88	0.91	0.86	0.86	0.86	0.90	0.79	0.84	0.91	0.82	0.86
RF _{BoW}	0.75	0.86	0.81	0.89	0.81	0.85	0.84	0.76	0.80	0.78	0.71	0.74	0.87	0.76	0.81
RF _{Feat}	0.83	0.76	0.79	0.91	0.82	0.86	0.85	0.81	0.83	0.83	0.76	0.79	0.85	0.79	0.82
RF _{BoW+Feat}	0.82	0.87	0.85	0.93	0.87	0.90	0.85	0.84	0.84	0.86	0.79	0.82	0.89	0.82	0.85
GBT _{BoW}	0.78	0.78	0.78	0.88	0.80	0.84	0.85	0.74	0.79	0.76	0.71	0.73	0.86	0.77	0.81
GBT _{Feat}	0.81	0.79	0.80	0.90	0.84	0.87	0.85	0.83	0.84	0.83	0.76	0.79	0.85	0.82	0.83
GBT _{BoW+Feat}	0.85	0.86	0.86	0.92	0.87	0.89	0.87	0.86	0.86	0.84	0.81	0.83	0.89	0.83	0.86
SVM _{BoW}	0.79	0.83	0.81	0.88	0.84	0.85	0.79	0.83	0.81	0.77	0.76	0.77	0.85	0.80	0.82
SVM _{Feat}	0.81	0.74	0.78	0.86	0.84	0.85	0.84	0.78	0.81	0.84	0.72	0.78	0.83	0.79	0.81
SVM _{BoW+Feat}	0.90	0.83	0.86	0.92	0.88	0.91	0.88	0.84	0.86	0.89	0.81	0.85	0.89	0.83	0.86

Table 8.4: Experimental results for the task of identifying DOSs.

The results in Table 8.4 reveal the following trends. Encoding a sentence as *BoW* or *Feat* representation was not as effective compared to the *BoW+Feat* representation. All the best results, identified in bold in Table 8.4, used this representation. Overall, the best classifier was $SVM_{BoW+Feat}$, which achieved the ten best results from the 15 measures considered. $SVM_{BoW+Feat}$ also obtained results in F_1 score equal to or better than the second best classifier in all product categories. We also note that although the classifiers used in our experiments were considerably different, there is minimal difference between the results obtained while using the *Bow+Feat* representation. For example, the difference between the best ($SVM_{BoW+Feat}$) and worst ($RF_{BoW+Feat}$) classifier was only 0.016 for F_1 when we consider the average of the five product categories. This indicates that choosing the correct set of features for sentence representation is the most important factor in the task of identifying DOSs.

Feature Ablation Study

Given the findings above, to evaluate the importance of the proposed features in the performance of the classifier, we conducted a *feature ablation* study. An ablation study is performed by systematically removing feature sets to identify those with the most influence on the results. For each set of features considered, we retrained and retested the classifier. We focused only on the $SVM_{Bow+Feat}$ because it presented the best performance. Table 8.5 reports the average F_1 score obtained, using 10-fold cross-validation, in all product categories. Each line labeled **All - F** corresponds to the results obtained with all features except feature F . Reduced F_1 values indicate that feature F had a positive contribution to the results. Thus, it should be retained.

We observed that using all the proposed features, together with the BOW representations, led to the best results across all product categories. Interest-

	CAM	CEL	DVD	LAP	ROT
All features	0.86	0.91	0.86	0.85	0.87
All - Adj	0.84	0.89	0.86	0.83	0.86
All - Adv	0.83	0.89	0.85	0.84	0.86
All - Amod	0.81	0.83	0.80	0.75	0.80
All - Comp	0.83	0.89	0.85	0.84	0.86
All - Noun	0.83	0.89	0.86	0.84	0.86
All - Polar	0.83	0.89	0.86	0.85	0.85
All - Subj	0.83	0.89	0.85	0.84	0.86
All - Super	0.84	0.89	0.86	0.84	0.86
All - Word	0.83	0.89	0.86	0.83	0.86

Table 8.5: Feature ablation study for SVM_{BoW+Feat} classifier.

ingly, in some cases, the removal of specific features did not have an influence on the final performance of the classifier. This occurred, for example, with the feature *Adj* in category DVD. However, the same feature *Adj* is the second most important feature in category LAP.

We also observed that the number of adjectival modifiers (*Amod*) was the highest contributor to the performance of the classifier, followed by the number of words (*Word*). The average F_1 of the classifier when using the nine features was 0.868. However, this value decreased to 0.804 when we removed the feature *Amod*. This prompted us to perform SVM_{BoW+Amod} using only this feature with *BoW*, which led to the following F_1 values: 0.82 in CAM, 0.89 in CEL, 0.82 in DVD, 0.81 in LAP, and 0.84 in ROT. This is clearly inferior as compared to the results achieved using all features. We can thus conclude that not a single feature is responsible for the classification performance; rather it is the interaction among all features — each feature can capture aspects of the text that the remaining cannot.

8.3.2 Opinion Mapping

For this experiment, we used as input the BestBuy Product Catalog and the BestBuy DOS Dataset. This means that only directed opinionated sentences were considered in this experiment. We did this in order to isolate this experiment from the results of the DOSs detection method described in previous section. The final result, a mapping of the opinions in each review to one of the attributes of the products in the catalog, was evaluated having the BestBuy Opinion Mapping Dataset as a reference (golden standard). The datasets used in this experiment were described in Section 6.2.

To account for the effects of the training process, the experiment was performed using stratified 10-fold cross-validation to ensure a balance in the proportion of classes within each partition.

General Results

Table 8.6 presents the experimental results of our proposed opinion mapping method. In the first four lines, we have the results using full sentences obtained from the reviews; the last four lines represent the results using the *Sentences Core Segments* strategy, as discussed in Section 5.3. The values in bold indicate the highest value achieved for each metric in each category.

	CAM			CEL			DVD			LAP			ROT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RF	0.77	0.67	0.71	0.83	0.75	0.79	0.73	0.71	0.72	0.78	0.56	0.65	0.78	0.65	0.71
ME	0.67	0.85	0.75	0.80	0.80	0.80	0.73	0.73	0.73	0.72	0.70	0.71	0.81	0.65	0.72
GBT	0.73	0.68	0.70	0.81	0.75	0.78	0.75	0.74	0.75	0.73	0.60	0.65	0.73	0.73	0.73
SVM	0.74	0.77	0.75	0.85	0.76	0.80	0.79	0.73	0.76	0.75	0.63	0.69	0.81	0.70	0.75
RF_{seg}	0.83	0.81	0.82	0.87	0.86	0.86	0.85	0.87	0.86	0.86	0.74	0.79	0.91	0.77	0.84
ME_{seg}	0.84	0.86	0.85	0.88	0.89	0.88	0.87	0.87	0.87	0.87	0.74	0.80	0.90	0.83	0.86
GBT_{seg}	0.82	0.80	0.81	0.87	0.83	0.85	0.89	0.85	0.87	0.84	0.75	0.79	0.86	0.84	0.85
SVM_{seg}	0.82	0.84	0.83	0.89	0.88	0.88	0.88	0.86	0.87	0.84	0.77	0.80	0.89	0.86	0.88

Table 8.6: Results for opinion-mapping task. Subscript _{seg} indicates that classifier uses *Sentence Core Segments* strategy.

As can be observed, the classifiers achieved higher values for precision, recall, and F_1 in all the categories using only Sentence Core Segments. On average, this strategy yielded gains of F_1 of approximately 16.48% for Random Forest, 14.82% for Maximum Entropy, 15.51% for Gradient Boosting Trees, and 13.6% for Support Vector Machines. Moreover, SVM_{seg} and ME_{seg} achieved the same average F_1 score (0.85) across all product categories. However, as observed by Morin and Bengio (2005) and Goodman (2001), a major weakness of Maximum Entropy is the extremely long training time. This leads us to conclude that using Sentence Core Segments and Support Vector Machines is the most appropriate solution for the opinion-mapping task.

Error Analysis

To further understand the results presented in Table 8.6, it is worth examining more closely the cases where the opinions were not correctly mapped. In Figure 8.3, we display the confusion matrices for the errors on the most frequent attributes, i.e., the attributes that are referred to by at least 5% of the opinions in each category. In the matrices, rows represent the actual attributes and columns represent the predicted attributes, i.e., each cell $M_{i,j}$ represents the fraction of opinions from attribute i that were classified as attribute j .

From Figure 8.3, we confirm that the proposed method achieves low error values in the task of mapping opinions to the most frequent attributes, in all product categories. Nevertheless, some errors do occur. An obvious problem is the subjective nature of attempting to classify opinions. For example, the sentence “portable 11” size” from category LAP was annotated as referring to the attribute **Screen**, whereas the proposed method mapped this sentence

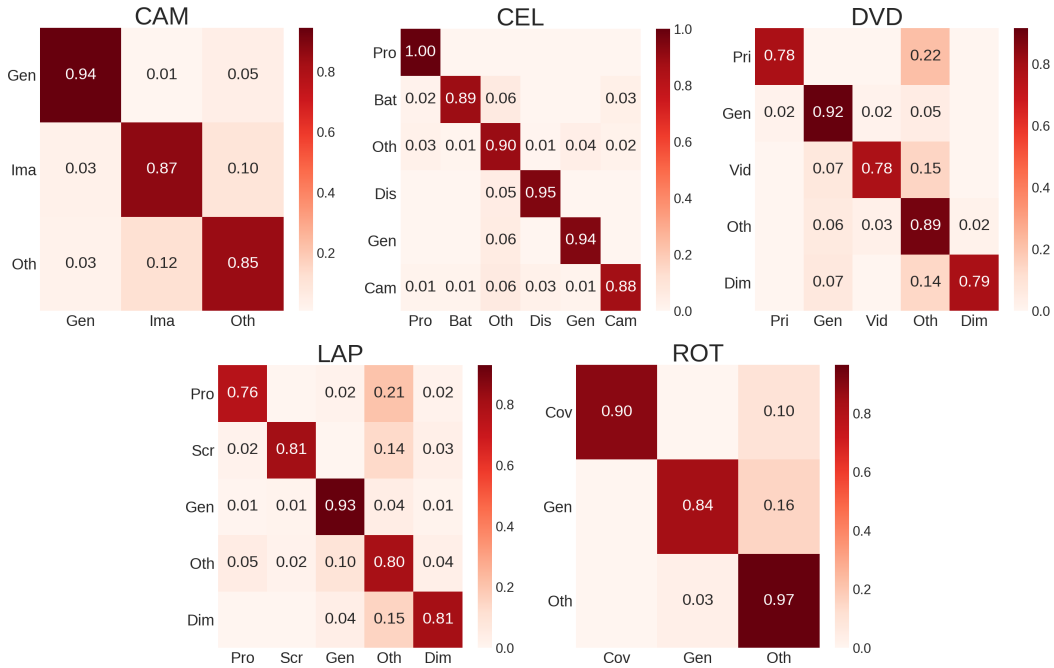


FIGURE 8.3: Confusion Matrices for opinion-mapping task. Each label represents an attribute: Gen (General), Ima (Imaging), Oth (Other), Pro (Processor), Bat (Battery), Dis (Display), Cam (Camera), Pri (Price), Vid (Video), Dim (Dimension), Scr (Screen), and Cov (Coverage Area).

to the attribute `Dimension`. Although we have considered this mapping as incorrect, it would be plausible to accept this opinion as referring to the size (`Dimension`) of the laptop.

Nevertheless, from the confusion matrices, we can clearly observe that the majority of errors were because of the presence of the attribute `Other`. We argue that this is not unexpected because this attribute is the most ambiguous. For example, in the `CAM` category, people frequently comment on the operating manual of the equipment, its accessories, and durability. Although these are considerably different aspects of a product, thus making it more difficult to detect patterns in the data, they must all be assigned to `Other`. To confirm this hypothesis, Table 8.7 presents the results obtained when this attribute is not considered. Entries in boldface indicate the highest value achieved for each metric in each category. We can observe that by removing `Other`, there was indeed an improvement in performance for all product categories.

	CAM			CEL			DVD			LAP			ROT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
\mathbf{RF}_{seg}	0.96	0.84	0.89	0.93	0.85	0.89	0.96	0.79	0.86	0.93	0.78	0.85	0.96	0.72	0.81
\mathbf{ME}_{seg}	0.97	0.87	0.91	0.94	0.87	0.90	0.99	0.80	0.88	0.92	0.80	0.85	0.97	0.77	0.85
\mathbf{GBT}_{seg}	0.94	0.86	0.90	0.92	0.88	0.90	0.96	0.79	0.86	0.90	0.82	0.86	0.93	0.80	0.86
\mathbf{SVM}_{seg}	0.96	0.90	0.93	0.94	0.89	0.91	0.95	0.86	0.90	0.90	0.82	0.86	0.97	0.88	0.92

Table 8.7: Opinion-mapping results when attribute `Other` is not considered.

8.3.3 End-to-End Results

In the previous sections, we verified the effectiveness of the proposed methods to address the tasks of each phase of the *OpinionLink*, when considered in isolation. In this section, we will evaluate the performance of *OpinionLink* as a realistic end-to-end application, where opinion mapping is performed, using as input the sentences identified as DOSs by the proposed Opinion Extraction Algorithm (see Algorithm 4 presented in Section 5.2.1).

In this experiment, opinion extraction was performed using $SVM_{Bow+Feat}$ to identify the DOSs, as described in Section 8.3.1, and opinion mapping was performed using SVM_{seg} , as described in Section 8.3.2. We call this configuration $OpinionLink_{real}$. As a baseline, we used the best configuration from Section 8.3.2, i.e., SVM_{seg} with DOS manually selected. We call this configuration $OpinionLink_{ideal}$.

Table 8.8 presents the performance of $OpinionLink_{ideal}$ in terms of F_1 , compared to $OpinionLink_{real}$. As expected, $OpinionLink_{ideal}$ achieved superior results in all the categories as compared to $OpinionLink_{real}$. The main reason is that the DOSs, which were incorrectly identified as factual in the first phase, were not used for $OpinionLink_{real}$. As a consequence, recall was negatively influenced. Further, sentences that were not DOSs, were rather incorrectly classified as such, had a negative influence on precision. However, the difference in the results of the two methods is small (less than 0.02 on the average), indicating that we can use the proposed method in a real application.

	CAM	CEL	DVD	LAP	ROT
$OpinionLink_{ideal}$	0.83	0.88	0.87	0.80	0.88
$OpinionLink_{real}$	0.82	0.87	0.85	0.79	0.84

Table 8.8: F_1 results of end-to-end evaluation of *OpinionLink*.

8.3.4 *OpinionLink* in Large Scale

In this experiment, we evaluated the feasibility of using *OpinionLink* on a large volume of reviews. For this, we used data from the Amazon Collection, which is considerably larger than BestBuy Collection. Specifically, we used the Amazon Product Catalog and the DOSs from the Amazon-400 Reviews Dataset as input to *OpinionLink*. The Amazon Opinion Mapping Dataset was used as the reference (golden standard) for evaluation. These datasets have been described in Section 6.3.

For the opinion mapping task, we used the SVM_{seg} classifier using sentences core segments strategy, because it presented the best performance previously. We trained this classifier with the Amazon Opinion Mapping Dataset using

10-fold cross-validation. Table 8.9 presents the classifiers' performance in terms of precision (P), recall (R), and F_1 .

	P	R	F₁
CAM	0.90	0.93	0.92
CEL	0.93	0.92	0.92
DVD	0.93	0.96	0.95
LAP	0.92	0.80	0.85
ROT	0.90	0.90	0.90

Table 8.9: Experimental results with data from the Amazon Collection.

As can be observed, the classifiers achieved good results in all product categories. As specified in Section 6.3, that the Amazon-400 Review and the Amazon Opinion Mapping datasets were built to yield experiments with a confidence level of 95% with 5% of confidence interval. Thus, these results indicate that our proposed method can also be effective for a large scale of data.

8.3.5 Bootstrapping Evaluation

In this experiment, we evaluated the performance of our bootstrapping strategy. Again, we provided the BestBuy Reviews Dataset and the BestBuy Product Catalog as input. The result, the mapping of the opinions in each review to one of the attributes of the products in the catalog, was evaluated having the BestBuy Opinion Mapping Dataset as a reference (golden standard).

Like the previous experiments, we adopted the sentence segmentation strategy, defined in Section 5.3.2, for all classifiers. As in Section 8.3.2, we also evaluated the influence of the attribute `Other` on the performance of classifiers. For this, we repeated the experiment using exactly the same settings while discarding the attribute `Other`.

Table 8.10 presents the experimental results in terms of precision, recall, and F_1 , for each product category. The left side shows the results for all attributes and the right side shows the results without attribute `Other`. We decided to report the results of the SVM classifier only, because as in the previous experiments, it presented the best results.

The classifiers achieved an average of 0.71 in terms of F_1 , when using the bootstrapping strategy and considering all attributes. Although these results are lower than those obtained with manual training, they corroborate our claims that this strategy is a viable and practical alternative, since it avoids the costs of manually creating labeled training data, while still achieving high quality results.

Regarding the impacts of attribute `Other`, we can observe that there was a notable improvement in the performance of the classifier when it is not

	P	R	F₁	P	R	F₁
	all attributes			without Other		
CAM	0.70	0.84	0.77	0.90	0.83	0.86
CEL	0.82	0.81	0.82	0.97	0.86	0.91
DVD	0.68	0.72	0.70	0.91	0.65	0.76
LAP	0.68	0.65	0.67	0.84	0.69	0.76
ROT	0.60	0.63	0.61	0.95	0.65	0.78

Table 8.10: Results for the opinion mapping task using the proposed bootstrapping strategy.

present. The classifier achieved an average of 0.81 in terms of F_1 . As before, this improvement is mainly due to the fact that the attribute **Other** has no values in the enriched product catalog. However, this is more noticeable in the bootstrapping strategy, since the similarity functions cannot be used directly for this attribute. Interestingly, the F_1 results achieved in the CEL category with the bootstrapping strategy are the same as those obtained with manual training data.

In sum, the results achieved in this study are quite satisfactory and indicate that it is possible to automatically map opinions to attributes without manual effort by means our bootstrapping strategy.

8.4 Comparative Analysis

In this section, we will present a comparative analysis of the performance of *AspectLink* and *OpinionLink*. More specifically, we will compare the results obtained from these methods in the task of identifying direct opinionated sentences (DOSs) and mapping opinions to attributes of product catalog.

8.4.1 Identifying Direct Opinionated Sentences

In *AspectLink*, the task of identifying DOSs is performed in two steps. Firstly, we identify the subjective sentences and discard the factual sentences. For this, we implemented the method proposed by Qadir (2009). Next, we identify the direct opinionated sentences and discard the comparative sentences. For this, we implemented the method proposed by Liu (2010). Here, we call this configuration as *AspectLink_{DOS}*.

In *OpinionLink*, we devised a new method based on a classifier with a set of features we proposed for addressing the task of identifying DOSs. According our experiments presented in Section 8.3.1, the best classifier was $SVM_{BoW+Feat}$. Here, we call this configuration as *OpinionLink_{DOS}*.

Figure 8.4 presents a comparison of the results obtained with the strategy used in *AspectLink_{DOS}* and *OpinionLink_{DOS}*. This experiment used the Best-

Buy Reviews Dataset as input and the BestBuy DOSs Dataset as a reference (golden standard).

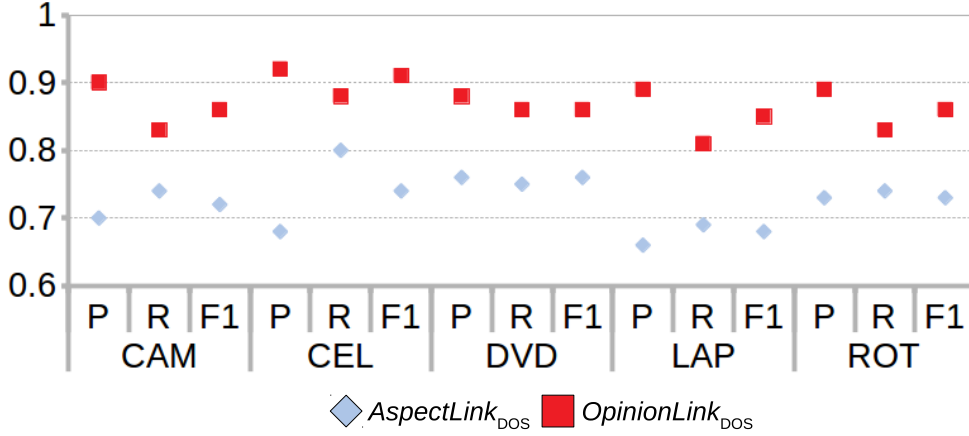


FIGURE 8.4: Comparing our proposed methods for the task of identifying DOSs.

As it can be observed, $OpinionLink_{DOS}$ achieved higher values for precision, recall, and F_1 in all the categories. This classifier yielded gains of F_1 of approximately 19.4% for CAM, 22.9% for CEL, 13.1% for DVD, 25% for LAP, and 17.8% for ROT. This led us to conclude that using $OpinionLink_{DOS}$ is the most appropriate solution for the task of identifying DOSs.

8.4.2 Opinion Mapping

From Section 8.2.2, we observed that using $AspectLink$ with expanded descriptors and stemming functions led to the best results for the opinion mapping task. Here, this configuration is referred simply to as $AspectLink$. On the other hand, we observed, from Section 8.3.2, that SVM_{seg} presented the best results in $OpinionLink$. Here, we named this method as $OpinionLink_{SUP}$, i.e., the fully supervised version of $OpinionLink$. Moreover, we proposed a bootstrapping strategy (Section 8.3.5) applied to SVM_{seg} in order to ease the labor for generating training data. From now, we call this last method as $OpinionLink_{BOOT}$, i.e., the bootstrapping version of $OpinionLink$.

To compare these methods, we used the BestBuy DOS Dataset as input. We had already presented the results of $OpinionLink_{SUP}$ and $OpinionLink_{BOOT}$ using this dataset, as described in Table 8.4 and 8.10, respectively. As $AspectLink_{SUP}$ was previously performed in another dataset, we had to execute this method with the BestBuy DOSs Dataset as input. The BestBuy Opinion Mapping Dataset was used as a reference for the evaluation.

Figure 8.5 presents the results achieved by $AspectLink$, $OpinionLink_{SUP}$, and $OpinionLink_{BOOT}$. As can be observed, $OpinionLink_{SUP}$ achieved higher values for precision, recall, and F_1 in all the categories. On the average, this

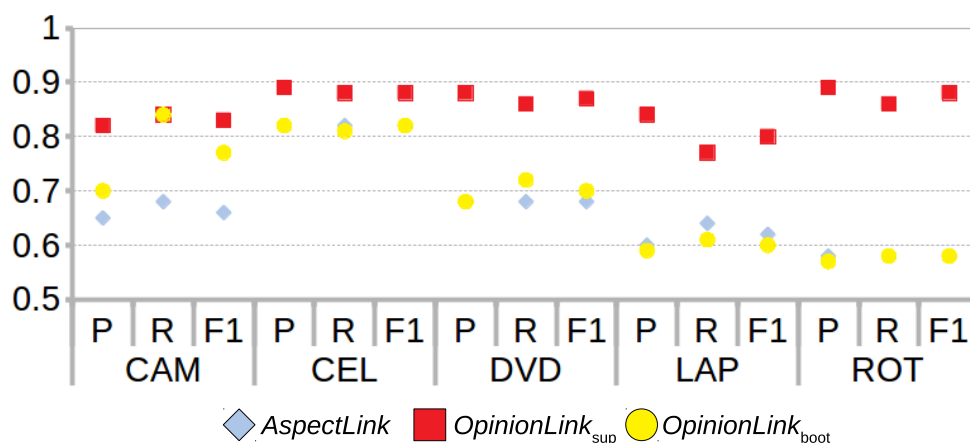


FIGURE 8.5: Comparing our proposed methods for opinion mapping task.

method yielded gains of F_1 of approximately 26,7% for *AspectLink* and 22,7% for *OpinionLink_{BOOT}*. This lead us to conclude that using *OpinionLink_{SUP}* is the most appropriate solution for the opinion mapping task.

As in the previous experiments, we also evaluated the influence of the attribute *Other* on the performance of the methods. For this, we repeated the experiments using exactly the same settings while discarding the attribute *Other*. Figure 8.6 presents the results achieved by *AspectLink*, *OpinionLink_{SUP}*, and *OpinionLink_{BOOT}* when the attribute *Other* is not considered.

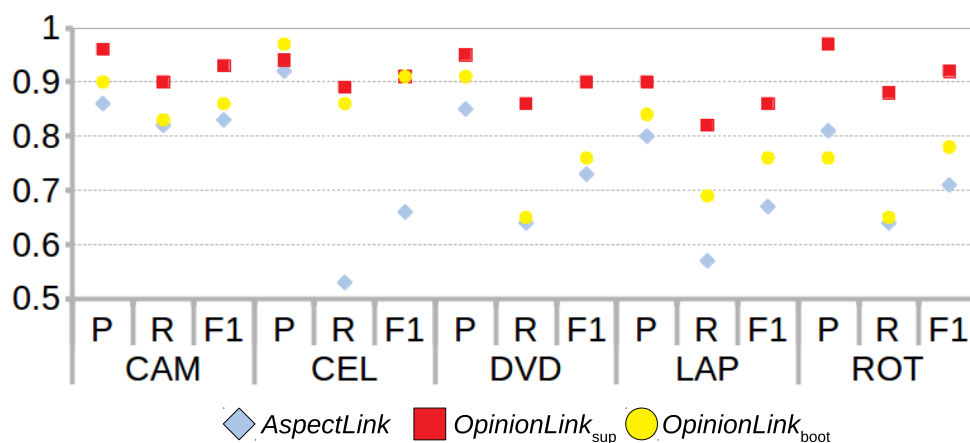


FIGURE 8.6: Comparing our proposed methods for opinion mapping task (without attribute *Other*).

As can be observed, *OpinionLink_{SUP}* achieved higher values for precision, recall, and F_1 in almost all the categories. The only exception occurred for the precision in CEL category, where *OpinionLink_{BOOT}* achieved a higher value than *OpinionLink_{SUP}*. However, in the same category, the *OpinionLink_{SUP}* achieved a higher value than this method using the bootstrapping strategy.

Finally, on the average, *OpinionLink_{SUP}* yielded gains of F_1 of approximately 25,6% for *AspectLink* and 11,1% for *OpinionLink_{BOOT}*. This led us to conclude that using *OpinionLink_{SUP}* is the most appropriate solution for the opinion mapping task when the attribute **Other** is not considered.

8.5 Summary

In this chapter, we presented an empirical evaluation of our proposed methods *AspectLink* and *OpinionLink*. We started the chapter by evaluating *AspectLink*. This method is unsupervised and has shown to be a viable alternative for the problem of enriching product catalog. Then, we reported the evaluation of *OpinionLink*. This method is supervised and presented a superior performance as compared to *AspectLink*. Moreover, we reported an evaluation of our bootstrapping strategy. Finally, we presented a comparative analysis among the main proposed methods for opinion mapping task. In the next chapter, we will present a system to showcase a practical application of some proposal ideas developed in this thesis.


To showcase a practical application of some proposed ideas in this thesis, we developed an Android app for smartphones called Contender. This system is capable of summarizing product opinions aligned to the attributes of these products. Since the opinions are aligned to the same set of attributes, this makes comparing two products at the attribute level granularity based on user opinions possible. To the best of our knowledge, our work is the first that addresses this problem.

This chapter is organized as follows. Section 9.1 presents the motivation for developing the system. Section 9.2 presents an overview of Contender. Section 9.3 demonstrates the main features of the system. Section 9.4 briefly describes the system settings. Finally, Section 9.5 presents a summary on this chapter.

9.1 Motivation

As discussed earlier, while making purchasing decisions, customers usually rely on the information from two types of sources: product specifications provided by the manufacturers, and reviews posted by other customers. Both kinds of information are often available on e-commerce websites. However, in some competitive markets, such as cell phones, many manufacturers make products with very similar characteristics. Figure 9.1 shows a comparison between the product specifications of *Moto Z Play* and *Samsung Galaxy S7 Edge*. We can observe that they have the same screen size, operating system, processor etc. In addition, these products have almost the same price. Therefore, there is almost no difference between them. Especially in such cases, user reviews play an important role in purchase decision-making.

Unfortunately, it is not generally feasible for an ordinary buyer to go



	Samsung Galaxy S7 Edge	Motorola Moto Z Play
Battery	3510 mAh	3600 mAh
Camera	5 (rear) and 16 (front) megapixels	5 (rear) and 12 (front) megapixels
Dimension	6.16 x 3.01 x 0.28 inches	5.94 x 2.86 x 0.30 inches
Display	5.5 inches	5.5 inches
Memory	3 GB	4 GB
Price	\$ 299.00	\$ 289.99
Processor	Qualcomm Snapdragon 625 8953	Qualcomm Snapdragon 820 MSM8996
Software	Android 8.0 Oreo	Android 8.0 Oreo

FIGURE 9.1: Compare specs.

through a large set of reviews to manually compare two similar products. For example, more than 2,000 reviews on *Samsung Galaxy S7 Edge* have been posted at Amazon.com website. A natural approach to handle this problem is to consider the ratings of the products. However, the usefulness of the star-based ratings in the reviews is limited for potential buyers, since a rating represents an average for the product as a whole and can combine both positive and negative evaluations of many single distinct attributes. In addition, this kind of evaluation does not convey any information about why users like a product or which characteristics they like the most. A user looking to buy a cell phone may want to know what user reviews say on battery or screen, not just what is the general rating of the product. Thus, research on how to organize the huge volume of user reviews is a substantial challenge, and it is directly related to the problem addressed in this thesis.

As a concrete example of how the results we obtained in this thesis can be used to address problems such as the one described above, we designed and implemented a system named *Contender*. This system can summarize product reviews aligned to the specification of attributes of cell phones offered in an e-commerce website. The system handles real user reviews from e-commerce websites as a data source. *Contender* extracts the opinions of the reviews and then maps these opinions to the attributes defined for the product specifications. We used our implementation of *AspectLink* to extract and map opinions to attributes.

9.2 Contender Overview

Figure 9.2 presents an overview of Contender. The system has two modules: *pre-processing* (offline), and *user session* (online). In the *pre-processing* module, the product catalog and the opinions on products are extracted from the target websites. Then, the opinions are aligned (mapped) to the product attributes. The *user session* provides an interface so that user can compare two products. We are highlighting and explaining in detail the following major steps: crawling, product catalog extraction, opinion extraction, indexing, searching, and ranking.

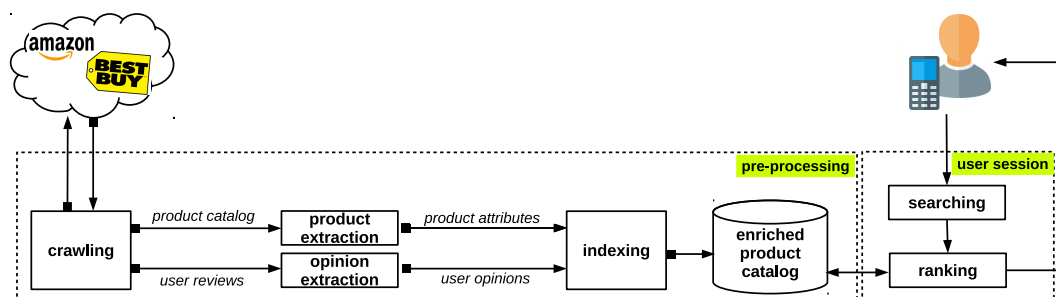


FIGURE 9.2: Contender overview.

Crawling. It is using a crawler to collect web pages of product specifications and user reviews from e-commerce websites such as Amazon.com and BestBuy.com.

Product Catalog Extraction. Extracting product attributes and their corresponding values from web pages collected. Recall from Chapter 3 that we added two new attributes for products: **General** and **Other**, where we considered the value of **General** to be the product title, and the value field of **Other** is blank.

Opinion Extraction. Let $R = \{r_1, r_2, \dots, r_m\}$ be a set of reviews on a product p_i , where each review $r \in R$ contains a set of sentences $ST = \{st_1, st_2, \dots, st_n\}$ and a numeric rating score which takes a value between 1 to 5. Each opinion o extracted from a sentence $st \in ST$ is represented by a triple $\langle p_i, a, rs \rangle$, where p_i is the target product, a is the aspect of the target product on which the opinion has been given, and rs is the numeric rating score of r . In developing of the system, we chose to use a slightly simpler representation of opinion than that presented in Chapter 3. For the aspect identification task, we implemented the unsupervised method proposed by Poria et al. (2014).

Indexing. The system maps each opinion extracted from the reviews in R to specific attributes of the target product. This is the core contribution of our work in Contender and was implemented following the ideas we developed for *AspectLink* described in Chapter 4. Notice that our supervised method, *OpinionLink*, was not complete by the time we started developing Contender.

However, it can be naturally integrated to the system, what we plan to do in the near future.

Searching. Our system allows the user to search for two cell phones to be compared. For this, the user simply types the names of the cell phones. Searches by names are case insensitive. We have also provided an autocomplete search feature.

Ranking. The app presents a ranking of the product attributes. For this, it first aggregates the opinions retrieved in the searching. Then, it calculates an average of the opinion scores for each attribute. We adopted the score normalization similar to that of Amazon and BestBuy, which considers 1-5 stars.

9.3 Demonstration

Figure 9.3 displays screenshots of the main modules of Contender. When the user opens the app, the system shows two search boxes, where the user can type the names of devices. As shown in Figure 9.3 (a), a user inputs the names of two cell phones in the search box: *Motorola Moto X4* and *Samsung Galaxy S7 Edge*. Our system shows the results of the input query in three distinct modules: *Specs*, *Reviews*, and *Charts*.

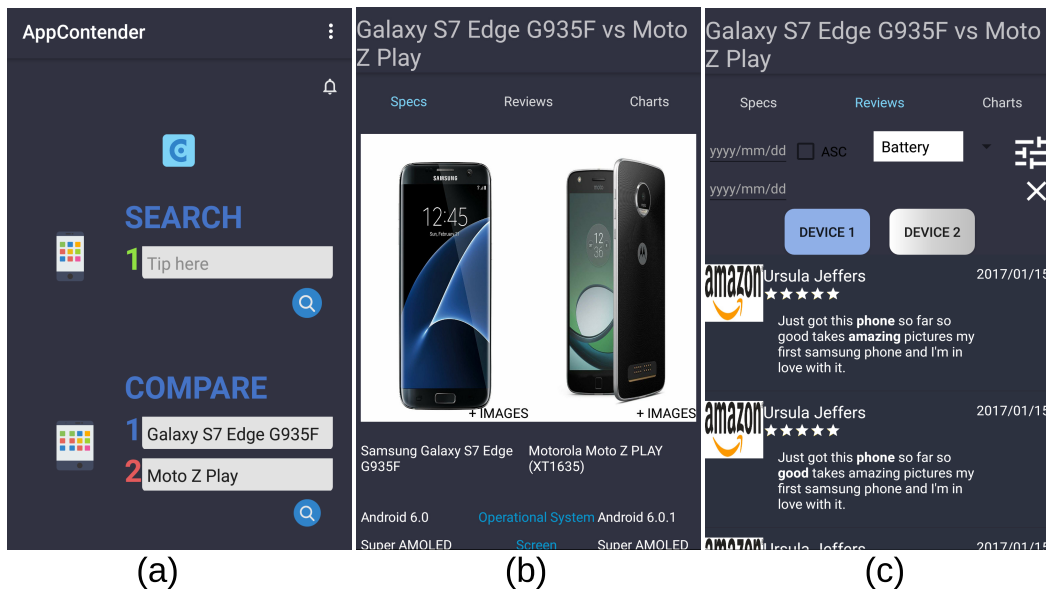


FIGURE 9.3: Main modules of Contender.

Specs displays the specifications of each product provided by the manufacturers and made available by the e-commerce sites (Figure 9.3 (b)). *Reviews* displays user reviews for each product. The reviews can be filtered by the product attributes. For example, if the user selects *Screen*, the system will

only display the reviews on that attribute (Figure 9.3 (c)). For example, if the user selects **Battery**, the system will only show the reviews on that attribute.

Charts displays four different types of charts so that the user can make a comparison between the two products (Figure 9.4): a) bar chart; b) radar chart; c) pie chart; d) rating score. Figure 9.4 (a) displays a bar chart with the quantitative of positive and negative opinions posted monthly in the last 12 months. Figure 9.4 (b) displays a radar chart with the rating score of each attribute for each of the target devices. Figure 9.4 (c) displays a pie chart with the percentage of positive and negative opinions for each of the attributes. Finally, Figure 9.3 (d) displays the attribute scores for each of the products. This is only possible because the system can align opinions to product attributes.

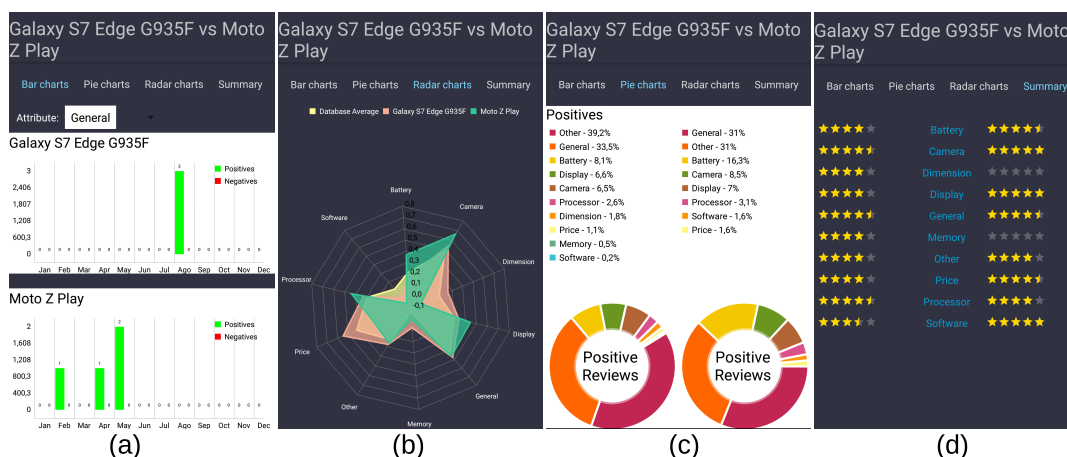


FIGURE 9.4: Different types of charts provided by *Contender*.

9.4 System Setting

The current version uses a review corpus collected from Amazon and BestBuy, with 15,512 products indexed. The number of user reviews currently is over 2 million, but this number is increasing, because our crawler daily collects new user reviews. Contender’s backend is running on a host with the following setting 2.4 GHz Intel Xeon 2 vCPU processor, 2 GB of memory, and 50 GB of SSD.

9.5 Summary

We presented a novel system for comparing two products at the attribute level granularity based on user opinions. To the best of our knowledge, our work is the first that addressed this problem. In future work, we plan to expand the system to other product categories, including, for example, cameras and laptops.

We reported this system in a paper has been accepted as a demonstration paper at the 41st European Conference on Information Retrieval (ECIR 2019).

In the next chapter, we will present our final considerations of our thesis and directions for future research.

In this thesis, we presented contributions to answer the following question ($\mathcal{Q}1$): “*How to structure opinions so that they can be effectively used by customers and manufacturers?*” Firstly, motivated by this question, we proposed a novel problem formulation $\mathcal{P}1$ for organizing a large number of unstructured user opinions: “*Enriching product catalogs with user opinions at the attribute granularity level as a new form of opinion summarization.*”. This problem formulation is our first contribution in this thesis.

To address problem described above, we formulated two research questions. The first question $\mathcal{RQ}1$ is “*Are there evidences that the most important product characteristics for people are represented by the attributes of the product catalogs?*” In order to answer $\mathcal{RQ}1$, we developed an empirical study to analyze the impacts of attributes from product catalogs on user opinions. This study used a large collection of data and the results indicate some conclusions. Firstly, we verified that a large fraction of sentences in reviews is composed of direct opinionated sentences (DOSs), which indicates that e-commerce websites are a valuable source of opinions on target products. Secondly, we used a large number of sentences, which allowed us to verify that the most of user opinions posted in e-commerce web sites are on one of the product attributes. Thirdly, we could verify in our study that there are certain attributes that are more relevant for users than other attributes. Moreover, we could conclude that customers often make either direct and indirect mentions to product attributes using several distinct expressions. Finally, this study contributed to understand the impacts of product attributes on user opinions. In sum, the results of this study allow us to state that user opinions are indeed guided by the attributes from product catalogs. This study is our second contribution in this thesis.

The second research question $\mathcal{RQ}2$ is “*Which approach can be used to address problem $\mathcal{P}1$?*” In order to answer $\mathcal{RQ}2$, we proposed an approach, which comprises of two phases: *opinion extraction* and *opinion mapping*. From

this approach, two sub-questions have emerged.

The first sub-question $\mathcal{RQ2.1}$ is “Which methods are best suited to carry out the proposed approach?” In order to answer $\mathcal{RQ2.1}$, we developed two distinct methods. *AspectLink* is the first method that we have developed. This method, even though being unsupervised, presented a good performance. *OpinionLink* is the second method that we have developed. This method is supervised, and as expected, it presented a better result than *AspectLink*. In addition, we have presented a bootstrapping strategy in order to reduce the dependence on training data. *AspectLink* and *OpinionLink* are two contributions of our thesis.

The second sub-question $\mathcal{RQ2.2}$ is “How to validate the effectiveness of the proposed methods?” In order to answer $\mathcal{RQ2.2}$, we carried out a comprehensive experimental evaluation, which empirically demonstrated the effectiveness of *AspectLink* and *OpinionLink* and its variations on representative datasets obtained from real e-commerce websites. We consider this evaluation as another contribution of this thesis.

By the time we needed to initiate our experiments, no suitable experimental datasets were available in the literature for this purpose. Thus, we had to create them ourselves and we did this using real data collections gathered from on-line sources available on the Web. These datasets will be made publicly available and we regard them as another contribution of this work.

Finally, to showcase a practical application of some of the ideas proposed in this thesis, we developed an Android app for smartphones called Contender. This system is capable of summarizing product opinions aligned to the attributes of these products. Since the opinions are aligned to the same set of attributes, this makes comparing two products at the attribute level granularity based on user opinions possible. To the best of our knowledge, our work is the first that addresses this problem. We regard this application as another contribution of this work.

10.1 Future Work

The results obtained in this thesis open several directions for further research. In this section, we will discuss possible extensions.

- **Diversity of sources.** In Chapter 7, we presented an experimental study for verifying our hypothesis that e-commerce websites are a valuable source of opinions on *target products*. Although we have verified our hypothesis, we plan to experiment our methods on different sources of reviews, such as microblogs and forums, which bring different challenges to be addressed. These sources have also been explored by other researchers in the field of opinion mining (Song et al., 2016; Vieira et al., 2016; Wang et al., 2015; Vieira et al., 2015; Zhao et al., 2014; Biyani et al., 2014).

- **Unlisted attributes.** Another open question that came out from our analysis in Chapter 7 and from the experiments presented in Chapter 8 is the issue of organizing opinions referring to attributes that were not listed in the product catalog. Currently, our methods map such opinions to the attribute `Other`. For example, according our study presented in Chapter 7, *keyboard* is the second most frequent aspect expression in laptop category, but typically, there is no attribute referring to it in product catalog provided by manufacturers. We believe that these opinions could be clustered into meaningful subgroups, and we hope to investigate whether the most relevant subgroups could be transformed into new attributes to be added to the product catalog.
- **New representations for products.** Although our proposed methods, *AspectLink* and *OpinionLink*, have already given promising results in the task of enriching product catalogs with user opinions, we are aware that the so called *product knowledge graphs* may provide a more powerful way of representing products and associated concepts (Liuq et al., 2016). The main motivation for this new representation is the need to enable answering many types of queries about products and related knowledge. The most representative example of this new representation of products is the *Amazon Product Graph* (Dong, 2018). Therefore, a possible future work is to investigate how to enrich product knowledge graphs with subjective information (opinions). For this, it will be necessary to adapt our methods for this new representation of product attributes. We contemplate that the first phase of our methods is independent of the product representation. Thus, this research would focus on the investigation of how to adapt the techniques used in the second phase to associate the opinions with the nodes of the product graph.
- **Product design.** Product designers, while designing new products or newer versions of existing products, have considered user reviews to determine customers' requirements and perceptions pertaining to a given product (Singh and Tucker, 2017). Information retrieved from online data sources, enables effective and efficient product design decisions (Lei and Moon, 2015). On the other hand, the product attributes, commented online, have been proved to be useful during the product development stage (Asur and Huberman, 2010). Since our work is the first, to the best our knowledge, to group user opinions to product attributes, we plan to investigate using the outcome of our methods to generate useful information for product designers.
- **Transfer Learning.** *OpinionLink* is a supervised method. Although we have proposed a bootstrapping strategy to reduce the dependence on training data, we intend to investigate other strategies to alleviate this dependence. More specifically, we plan to extract and transfer knowledge

from some auxiliary data or available in other domains in order to assist the training of *OpinionLink* on the target data. This technique is well-known as *Transfer Learning* (Pan, 2016). For example, we intend to use some datasets that were previously labeled. We could use the BestBuy Opinion Mapping Dataset to train *OpinionLink* and then evaluated it using another dataset as input.

- **Other domains.** Our experimental results in this thesis show that users express a lot of opinions on attributes that already exist in product catalogs. However, in the domains such as hotels and restaurants, there is no representative catalog with attributes that are commented by the users. Therefore, we plan to investigate this issue in a further study.
- **Other Languages** According to Liu (2015), much of the current research on opinion mining has been done in English, and our research is in line with this. However, it would be interesting to investigate the use of our methods in other idioms.
- **Deep Learning** *AspectLink* addresses the task of opinion mapping by means of similarity functions that compare lexical features of product attributes from the catalog with features from the text of aspect expressions. Another possible future line of investigation is using continuous vector representations, also known as *word embeddings* Mikolov et al. (2013), for each word representation.
- **Question Answering (QA).** An emergent problem in *Question Answering* topic is how to respond e-commerce product questions posed by users (Yu et al., 2018). Recently a chatbot for e-commerce sites known as *SuperAgent* has been developed (Cui et al., 2017). This system considers both QA collections and reviews when answering questions. However, it employs separate modules for each of the information sources without mutual coordination. Since many questions posted by users are specifically on product attributes and our approach is able to group opinions around the attributes of products, we plan to investigate how to associate the clustered opinions to user questions.
- **Cold star problem in recommender systems.** A recommender system aims at providing personalized recommendations to users for specific items. A very important issue in this topic is the *cold start problem* (Lika et al., 2014). This problem is related to recommendations for novel users or new items. In case of a new item, the system does not have information about this newly launched product. We believe that our methods can collaborate to address this problem. Since our methods can group opinions around attributes of products and many products have same attributes, we plan to investigate using the clustered opinions to improve current recommendations systems.

Bibliography

- Amplayo, R. K. and Song, M. (2017). An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. *Data & Knowledge Engineering*, 110:54–67.
- Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society.
- Baeza-Yates, R., Ribeiro-Neto, B., et al. (2011). *Modern Information Retrieval*, volume 84. Addison Wesley.
- Biyani, P., Bhatia, S., Caragea, C., and Mitra, P. (2014). Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*, 69:170–178.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.
- Carenini, G., Ng, R. T., and Zwart, E. (2005). Extracting knowledge from evaluative text. In *Proceedings of the 3rd International Conference on Knowledge Capture*, pages 11–18. ACM.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*, volume 5, page 3. Atlanta.
- Chenlo, J. M. and Losada, D. E. (2014). An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences*, 280:275–288.

- Condori, R. E. L. and Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, 78:124–134.
- Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., and Zhou, M. (2017). Supera-gent: a customer service chatbot for e-commerce websites. *Proceedings of ACL 2017, System Demonstrations*, pages 97–102.
- Dong, X. L. (2018). Challenges and innovations in building a product knowledge graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2869–2869.
- Dutta, A., Meilicke, C., and Stuckenschmidt, H. (2015). Enriching structured knowledge with open information. In *Proceedings of the 24th International Conference on World Wide Web*, pages 267–277. International World Wide Web Conferences Steering Committee.
- Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.
- Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., and Flett, A. (2001). Product data integration in b2b e-commerce. *IEEE Intelligent Systems*, 16(4):54–59.
- Goodman, J. (2001). Classes for fast maximum entropy training. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 561–564.
- 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. ACM.
- Hu, Y.-H., Chen, Y.-L., and Chou, H.-L. (2017). Opinion mining from online hotel reviews—a text summarization approach. *Information Processing & Management*, 53(2):436–449.
- Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251. ACM.
- Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

- Kang, M., Ahn, J., and Lee, K. (2018). Opinion mining using ensemble text hidden markov models for text classification. *Expert Systems with Applications*, 94:218–227.
- Khan, A., Baharudin, B., Lee, L. H., and Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20.
- Kim, H. D., Ganesan, K., Sondhi, P., and Zhai, C. (2011). Comprehensive review of opinion summarization. Technical report, University of Illinois at Urbana-Champaign.
- Kim, S. G. and Kang, J. (2018). Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management*, 54(6):938–957.
- Kostyra, D. S., Reiner, J., Natter, M., and Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing*, 33(1):11–26.
- Kwon, B. C., Kim, S.-H., Duket, T., Catalán, A., and Yi, J. S. (2015). Do people really experience information overload while reading online reviews? *International Journal of Human-Computer Interaction*, 31(12):959–973.
- Lee, J. and Nguyen, M. J. (2017). Product attributes and preference for foreign brands among vietnamese consumers. *Journal of Retailing and Consumer Services*, 35:76–83.
- Lei, N. and Moon, S. K. (2015). A decision support system for market-driven product positioning and design. *Decision Support Systems*, 69:82–91.
- Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., and Yu, H. (2010). Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 653–661. Association for Computational Linguistics.
- Li, Y., McLean, D., Bandar, Z. A., Óshea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150.
- Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073.
- Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:627–666.

- Liu, B. (2012). *Sentiment analysis and opinion mining*, volume 5. Morgan & Claypool Publishers.
- Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, pages 342–351. ACM.
- Liu, M., Fang, Y., Choulos, A. G., Park, D. H., and Hu, X. (2017). Product review summarization through question retrieval and diversification. *Information Retrieval Journal*, 20(6):575–605.
- Liuq, L., Duan, H., and et al. (2016). Knowledge graph construction techniques. *Journal of Computer Research and Development*, 53(3):582–600J.
- Maslowska, E., Malthouse, E. C., and Viswanathan, V. (2017). Do customer reviews drive purchase decisions? the moderating roles of review exposure and price. *Decision Support Systems*, 98:1–9.
- McAuley, J., Pandey, R., and Leskovec, J. (2015a). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.
- McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015b). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.
- McAuley, J. and Yang, A. (2016). Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pages 625–635. International World Wide Web Conferences Steering Committee.
- Melo, T., Silva, A., and Moura, E. (2018). An aspect-driven method for enriching product catalogs with user opinions. *Journal of the Brazilian Computer Society*, 24(1):15.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252.

- Palshikar, G. K., Apte, M., Pandita, D., and Singh, V. (2016). Learning to identify subjective sentences. In *13th International Conference on Natural Language Processing*, page 239.
- Pan, W. (2016). A survey of transfer learning for collaborative recommendation with auxiliary data. *Neurocomputing*, 177:447–453.
- Park, D. H. and Blake, C. (2012). Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the workshop on detecting structure in scholarly discourse*, pages 1–9. Association for Computational Linguistics.
- Park, E. J., Kim, E. Y., Funches, V. M., and Foxx, W. (2012). Apparel product attributes, web browsing, and e-impulse buying on shopping websites. *Journal of Business Research*, 65(11):1583–1589.
- Poria, S., Cambria, E., Ku, L.-W., Gui, C., and Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. *SocialNLP 2014*, page 28.
- Porter, M. (1997). An algorithm for suffix stripping. *program*, 14 (3): 130–137, 1980. reprinted in k. sparck jones, and p. willet, readings in information retrieval.
- PwC (2016). They say they want a revolution – total retail 2016.
- Qadir, A. (2009). Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 38–43. Association for Computational Linguistics.
- Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., and Alghazzawi, D. (2016). A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior*, 58:75–81.
- Rajkumar, P., Desai, S., Ganguly, N., and Goyal, P. (2014). A novel two-stage framework for extracting opinionated sentences from news articles. In *Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*, pages 25–33.
- Rakesh, V., Ding, W., Ahuja, A., Rao, N., Sun, Y., and Reddy, C. K. (2018). A sparse topic model for extracting aspect-specific summaries from online reviews. In *Proceedings of the 2018 World Wide Web Conference*, pages 1573–1582. International World Wide Web Conferences Steering Committee.
- Riloff, E., Wiebe, J., and Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference*

- on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.
- Saritha, S. and Pateriya, R. (2014). Methods for identifying comparative sentences. *International Journal of Computer Applications*, 108(19).
- Saumya, S., Singh, J. P., Baabdullah, A. M., Rana, N. P., and Dwivedi, Y. K. (2018). Ranking online consumer reviews. *Electronic Commerce Research and Applications*, 29:78–89.
- Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- Singh, A. and Tucker, C. S. (2017). A machine learning approach to product review disambiguation based on function, form and behavior classification. *Decision Support Systems*, 97:81–91.
- Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. (2017). Predicting the “helpfulness” of online consumer reviews. *Journal of Business Research*, 70:346–355.
- Smedt, T. D. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.
- Smith, A. and Anderson, M. (2016). Online shopping and e-commerce. <http://www.pewinternet.org/2016/12/19/online-shopping-and-e-commerce>. Accessed 08 September 2018.
- Song, K., Chen, L., Gao, W., Feng, S., Wang, D., and Zhang, C. (2016). Persentiment: A personalized sentiment classification system for microblog users. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 255–258. International World Wide Web Conferences Steering Committee.
- Sun, S., Yang, D., Zhang, H., Chen, Y., Wei, C., Meng, X., and Hu, Y. (2018). Important attribute identification in knowledge graph. *arXiv preprint arXiv:1810.05320*.
- Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.
- Trummer, I., Halevy, A., Lee, H., Sarawagi, S., and Gupta, R. (2015). Mining subjective properties on the web. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1745–1760. ACM.

- Vieira, H. S., da Silva, A. S., Calado, P., Cristo, M., and de Moura, E. S. (2016). Towards the effective linking of social media contents to products in e-commerce catalogs. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 1049–1058, New York, NY, USA. ACM.
- Vieira, H. S., da Silva, A. S., Cristo, M., and de Moura, E. S. (2015). A self-training crf method for recognizing product model mentions in web forums. In *European Conference on Information Retrieval*, pages 257–264. Springer.
- Wan, M. and McAuley, J. (2016). Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 489–498. IEEE.
- Wang, J., Yu, C. T., Yu, P. S., Liu, B., and Meng, W. (2015). Diversionary comments under blog posts. *ACM Transactions on the Web (TWEB)*, 9(4):18.
- Wang, S., Chen, Z., Fei, G., Liu, B., and Emery, S. (2016). Targeted topic modeling for focused analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1235–1244. ACM.
- Wang, T.-Y. and Chiang, H.-M. (2007). Fuzzy support vector machine for multi-class text categorization. *Information Processing & Management*, 43(4):914–929.
- Wang, Y., Lu, X., and Tan, Y. (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications*, 29:1–11.
- Yakout, M., Ganjam, K., Chakrabarti, K., and Chaudhuri, S. (2012). Infogather: Entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 97–108.
- Yu, J., Zha, Z.-J., Wang, M., Wang, K., and Chua, T.-S. (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. Association for Computational Linguistics.
- Yu, Q., Lam, W., and Wang, Z. (2018). Responding e-commerce product questions via exploiting qa collections and reviews. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2192–2203.

- Zha, Z.-J., Yu, J., Tang, J., Wang, M., and Chua, T.-S. (2014). Product aspect ranking and its applications. *IEEE transactions on knowledge and data engineering*, 26(5):1211–1224.
- Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.
- Zhao, X. W., Guo, Y., He, Y., Jiang, H., Wu, Y., and Li, X. (2014). We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1935–1944. ACM.
- Zheng, G., Mukherjee, S., Dong, X. L., and Li, F. (2018). Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, pages 1049–1058. ACM.
- Zhou, X., Wan, X., and Xiao, J. (2016). Cminer: opinion extraction and summarization for chinese microblogs. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1650–1663.