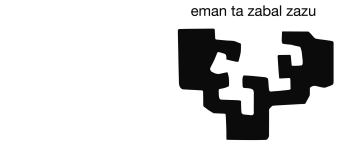


Multilingual Opinion Mining

Aitor García Pablos

German Rigau Claramunt

Director

Montserrat Cuadros Oller

Co-Director

PhD Thesis

Donostia, June, 2017

This work has been funded by Vicomtech-IK4. This work started under and was partially supported by European Commission project OpeNER (Open polarity enhanced Named Entity Recognition) (ICT-296451).

Acknowledgements

To be honest, I do not know where to start or what to say. This has been a three-year process. A personal voyage. A little odyssey. I can hardly believe that I am finally reaching the shore, just like Ulysses arrived at Ithaca. Somehow, I have also managed to leave my own Scylla and Charybdis behind. But none of this would have happened without an immense amount of help. Words will hardly suffice to honour that debt, but let us try.

First things first, I would like to express my gratitude to Dr German Rigau and Dr Montse Cuadros for their support and guidance. They have had an almost infinite patience helping me in this process. I want to thank the wise and useful corrections, ideas and advice from German. I admire his ability to summarise in a few sentences what I have written in a full page, without loss of information. I also want to make a special mention to Montse and her ability to cheer people and make them feel better, even in difficult moments. If you know her, you know what I am speaking about.

German and Montse have been the people more directly involved in taking me to the end of this PhD process. But there has been much more people involved in making this possible in one way or another.

I would not have started this process without the support of Vicomtech-IK4. Vicomtech as an organisation has supported me in many ways, but I want to put the focus on the great people that work there, from researchers to administrative staff. I am glad I was part of that team of people during these years.

However, there are some people that deserve a special mention. Big thanks to Seán Gaines, my first "boss" when I started working at Vicomtech. I still remember his words, just two or three weeks after I started working there,

asking me *"if I would like to continue my formation"*. I think that it was, in some way, the inception of the path that reaches here.

Other big thanks to Mariate Linaza, my boss until very recently. She has been always helping and encouraging me to continue. I could not have accomplished this without her support. Thank you Mariate!

Of course, I would also like to thank Arantza del Pozo, my current boss, for kindly taking me in her department. I extend this gratitude to my department colleagues, both from my former *Smart Environment and Energy* department and from my current *Human Speech and Language Technologies* department. Thank you, guys.

In addition, I would like to thank professor Piek Vossen and the people of the *Computational Lexicology & Terminology Lab* (CLTL) at VUA, for being so kind to me during my three months stay with them, back in 2014. That happened about three years ago, but I still have very good memories of Amsterdam. And I still use the mug they gave me for coffee!

And talking about coffee, I would like to mention Izaro and the other mates from the university for those morning coffee-breaks during the last months. I really needed them.

Also, I want to thank Esther for offering help to revise the English-wise correctness of this document. I rejected her offer because I know how busy she always is, but I appreciate the intention.

My thanks to Dr Oier López de Lacalle for his initial advice and help with topic models and LDA related questions.

And of course, I thank my family and friends for being there and for bearing with me when I was not feeling confident about what I was doing, or when I was tired, busy or stressed (i.e. the 99% of the time...) ;)

I prefer to stop here to avoid the risk of forgetting some name in the last moment if I enter into too much detail. I really appreciate all the help, attention and kind words received. You know it :)

Thank you very much!

Muchas gracias!

Moltes gràcies!

Eskerrik asko!

Contents

List of Figures

List of Tables

INTRODUCTION

CHAPTER 1

Introduction

This chapter serves as the introduction to the content and purpose of this thesis. The chapter is structured as follows. Section 1.1 describes the research framework and the main concepts related to the field of sentiment analysis and opinion mining. Next, section 1.2 introduces the main goals of this research. Section 1.3 enumerates the main contributions resulting from the work carried out in this thesis. And finally, section 1.4 presents how the rest of this document is organised.

1.1 Research framework

Natural Language Processing (NLP) is a field of computer science which aims at automatically processing interactions expressed in a natural language. A natural language is any (human) language that has evolved and continues evolving naturally, as opposite to computer languages based on a strict set of unambiguous rules and grammars. Natural languages are harder to process because they use relaxed grammar rules, are based on continuously evolving vocabulary and present a many ambiguous phenomena and requires commonsense knowledge.

Natural languages can take different forms, like speech or writing. With regard to written language, the vast amount of digital content generated

from the popularisation of the Web has boosted the interest in automatically processing texts. There are a lot of opportunities and challenges coming from turning unstructured texts into structured and valuable information.

One of such areas of interest in text processing is sentiment analysis, also known as opinion mining (Pang and Lee, 2008; Liu, 2012). Every day tonnes of opinions are submitted to online websites, available for everyone. Specialised websites offer a channel for customers to share their experiences and thoughts about products and services. Social networks point new trends every few minutes and quickly react to any event that happens every day. Companies and governments have a big interest in monitoring those sources, and the regular users are interested in the opinion of their peers before making a purchase.

The vast amount of content calls for automatic tools to provide support for gathering, filtering, classifying and aggregating this information. That is where sentiment analysis and opinion mining tools come into action.

There are different ways to categorise or stratify sentiment analysis approaches. One way is focusing on their granularity. Document-based sentiment analysis treat a full document as the basic sentiment polarity bearing unit. This assumes that a global sentiment can be derived from the document. However, in the reality sentiment is usually associated to smaller pieces of textual units (Zhang and Liu, 2014). Sentence-based sentiment analysis increases the granularity to sentences, so each sentence is assigned with an individual sentiment polarity value. Despite this granularity is closer to reality, it is still losing valuable information for those sentences that mention two or more elements with opposite polarities. For instance, in a sentence like *"awesome food and drinks, but the waiters were too slow"*, the overall sentence sentiment is neither positive or negative. A more fine-grained analysis is required to completely understand the sentiment expressed in the text.

Aspect-based sentiment analysis is closer to the ideal case in which the scope of a sentiment value is an individual aspect of the evaluated entity. Such an aspect is one of the domain-dependent features that help to describe the referred entity. With this level of granularity, the resulting information is more valuable since it provides insight of which elements/aspects a causing satisfaction and which ones are causing dissatisfaction. In the previous example, *"awesome food and drinks, but the waiters were too slow"*, an aspect-based sentiment analysis would reveal that the customer is satisfied with the *food* and the *drinks*, and dissatisfied with the *waiters*. The most recent senti-

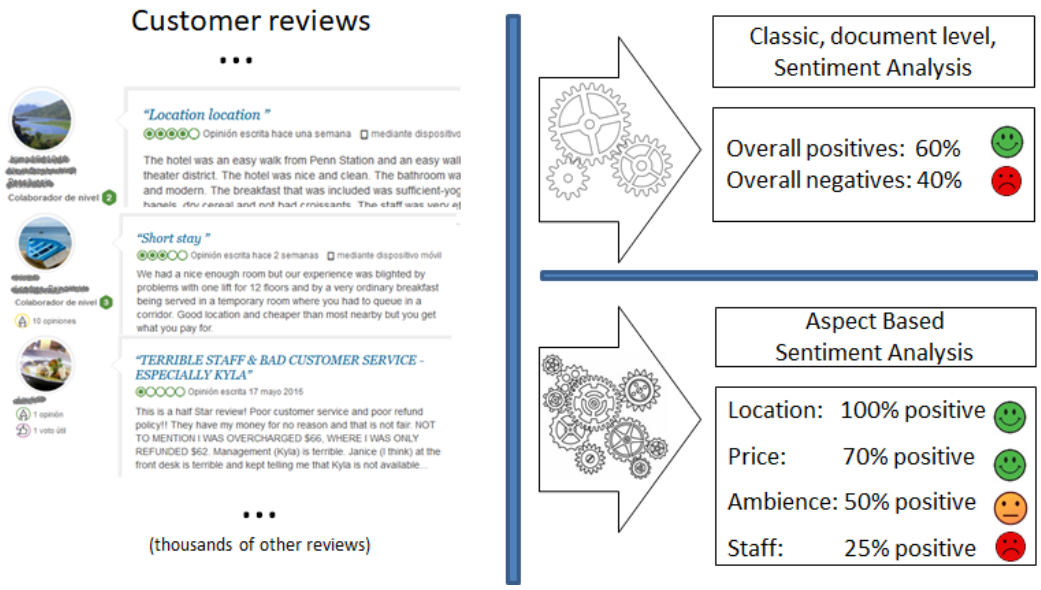


Figure 1.1: Classic document-level sentiment analysis vs. Aspect Based Sentiment Analysis (ABSA).

ment analysis are aspect-based, and in fact, this research sub-area receives its own name, Aspect Based Sentiment Analysis (ABSA). Figure 1.1 shows an example of ABSA, and why it is more informative than a classical, document level, sentiment analysis.

Some sentiment analysis approaches are based on some computation over the count of words with a certain polarity. In order to determine the polarity of each word, a so-called sentiment lexicon is created. A sentiment lexicon contains a sentiment polarity value for each word of interest (positive, negative, etc.). Sentiment lexicons can be manually created, or they can be (semi)automatically bootstrapped from pre-existing resources. The main drawback of methods that involve manual intervention is that they can be hard to maintain or to adapt to different languages or domains.

Other approaches for sentiment analysis make use of supervised machine learning approaches. Supervised machine learning methods train a statistical model that *learns* to assign a sentiment polarity value to a given word, sentence, etc. The main drawback of such methods is that they heavily rely on labelled training data that is not always easy or even feasible to obtain for the target domain, language or particular task (Mohammad, 2016).

Unsupervised or weakly supervised approaches try to alleviate the need of labelled data and other lexical resources. Such resources usually rely on big corpora to derive lexical associations and recurrent patterns. This way unsupervised or weakly supervised methods are able of bootstrapping prominent domain vocabulary and calculate associations and/or similarity among terms.

The interest in weakly-supervised methods is motivated because the need of manually labelled data or other language or domain dependent resources hinders the portability of sentiment analysis systems to other languages or domains. Systems that obtain very good performance for a particular language (typically English) and domain may not be directly applied to another language or domain.

1.2 Main goals

As introduced in the last part of the previous section, there is a motivated interest on weakly-supervised methods. This thesis focus on exploring weakly-supervised methods to perform sentiment analysis using approaches that require almost no language dependent resources or linguistic tools. Such methods could be used to process content for different languages and domains without relying on the availability of certain resources.

This does not mean at all that supervised methods or manually labelled training datasets are not useful tools. Currently, supervised methods based on deep-learning are beating the state-of-the-art results for many of the sentiment analysis related tasks. But they usually need a large amount of labelled data for training, which limits their application to those domains/tasks for which there is not such data available. Resources like manually labelled training data are often expensive to obtain.

This is the reason why we focus on certain tasks related to sentiment analysis using only weakly supervised methods. Our aim is to perform Aspect Based Sentiment Analysis (ABSA) using weakly supervised methods. We want to reduce the requirements of linguistic tools and resources to a minimum, and assess which performance we get. In the end, the overall objective is to obtain a weakly-supervised ABSA system that can be used for different languages and domains with a minimal adaptation effort.

In summary, the research work described in this thesis pursues these main goals:

1. Review of the state of the art related to sentiment analysis, including the motivation and relevance of the automatic detection of opinions in texts and relevant approaches published during the last years.

2. Explore methods to generate sentiment analysis related resources, like lists of domain aspect terms and opinion words and the calculation of a sentiment polarity value for opinion words, requiring the less possible amount of supervision, tools and resources.

3. Build a system capable of performing Aspect Based Sentiment Analysis using the less possible amount of supervision and resources. The objective is to obtain a system that estimates the domain aspects and sentiment polarities of customer reviews for any language and domain. The resulting system is aimed at working for many languages and domains requiring almost no adaptation effort.

1.3 Main contributions

Sentiment analysis, and more precisely Aspect Based Sentiment Analysis (ABSA), tries to detect, classify and sort subjective pieces of text by domain aspect and sentiment polarity.

Domain aspects are relevant coarse-grained categories for a particular domain. Domain aspects are explicitly referred in a text by words that usually receive the name of *aspect terms*. Typical examples of aspect terms in restaurant customer reviews domain are the words *waiter*, *waitress*, *waitstaff* that refer to the coarser domain aspect *service*.

During the last fifteen years, thousands of works have been published about sentiment analysis (Mäntylä et al., 2016). However, it is frequent that in order to use a particular sentiment analysis algorithm or method, it is necessary to have language or domain dependent tools and resources available. The research community has devoted most of its attention to process texts in English, and many of the published approaches cannot be easily ported to other languages.

This is particularly relevant for supervised systems that make use of manually labelled data to train a classification algorithm. Such manually labelled datasets tend to be of many thousand sentences, and they only serve for the language and domain they were generated for. This kind of algorithms usually offer a good performance, but in order to train an analogous algorithm for a different language or domain, or even for a different set of targets (e.g. if the categories to be detected change over time), a new training set has to be manually labelled. This is a time consuming and expensive task, that often becomes impractical or infeasible (Mohammad, 2016).

For this reason, we focus our attention on exploring weakly-supervised approaches that require almost no language dependent resources and that work directly on unlabelled data. This objective is hard, and the results can hardly beat the performance that can be obtained by a fully supervised system trained on a large training dataset. But the advantage is that the proposed approaches can be applied to different languages and domains with almost no adaptation effort.

Our contributions in this thesis are the following:

- A review of the state of the art related to sentiment analysis and Aspect Based Sentiment Analysis. Sentiment Analysis is such a wide research area that it is hardly possible to cover it from all the perspectives. We provide a review that covers the motivations and relevance of automatic Sentiment Analysis for the society in the digital era, as well as the most relevant approaches and methods.

- A description of a double-propagation based approach (Qiu et al., 2011) to bootstrap aspect terms and opinion words for a particular domain, using only two seed words, a few extraction rules based on syntactic dependencies and a graph-based algorithm. We show results of the described approach for several application domains (restaurant reviews, laptop reviews, hotel reviews). The proposed system has been evaluated with the participation in the SemEval 2014 task 4, consisting on detecting aspect term occurrences in sentences from restaurants and laptops reviews.

- A description of an approach to calculate a sentiment polarity value for words of a particular domain, using only two seed words and continuous word embeddings (Mikolov et al., 2013c; Pennington et al., 2014) word

similarity, effectively building a sentiment lexicon. We show examples of resulting sentiment polarity values for several different languages (English, Spanish) and domains (restaurants, laptops, hotels). We perform the evaluation of the proposed approach comparing the resulting polarity values against other well-known resources and methods for sentiment lexicon creation.

- A description of an approach to separate opinion words from the rest of non-opinion words in order to clean the resulting sentiment lexicon, just by using a single extra seed word, keeping the overall method almost unsupervised. We perform the evaluation of the opinion word separation using lists of known opinion words as ground truth, and compare different method variants (using Brown clustering (Brown et al., 1992) vs. using Word2Vec based word clusters).

- A combination of the previously devised approaches into a more complex system based on topic modelling. The resulting system is an extension of the well-known Latent Dirichlet Allocation model (LDA) (Blei et al., 2003), which includes extra latent variables to model the aspect term and opinion word separation and the sentiment polarity value. Again, the only required resources are a few seed words (one per desired domain aspect and polarity). The system classifies each sentence into one of the pre-defined domain aspects together with its polarity, performing aspect-based sentiment analysis without the need of manually labelled data for training. We evaluate the proposed weakly-supervised ABSA system against other LDA-based approaches. In addition, we evaluate the domain aspect and sentiment polarity classification performance in a multilingual setting, using the SemEval 2016 task 5[1] restaurant reviews datasets in four languages (English, Spanish, French, Dutch).

- A publicly available source code containing the proposed ABSA system implementation[2].

[1] `http://alt.qcri.org/semeval2016/task5/`
[2] `https://bitbucket.org/aitor-garcia-p/w2vlda-last`

1.4 Organization of the document

This thesis presents the research we have carried out in weakly supervised sentiment analysis, in an incremental way. We start building upon the ideas from other research works, adding some extra elements, and implementing and evaluating them. From the first step consisting of bootstrapping domain aspect terms and opinion words, to a combination of methods into a weakly-supervised topic modelling approach for performing aspect-based sentiment analysis. The remaining of this document is organised as follows:

- Chapter 2: State of the art

  This chapter presents a review of the state of the art on sentiment analysis and opinion mining. Sentiment analysis is a very wide research field that has received a lot of attention during the last two decades. We introduce what sentiment analysis is and why it is so important. After setting the context we briefly describe some of the most relevant tasks related to sentiment analysis and some of its associated research works. Then, since topic models and continuous word embeddings are closely related to the work carried out in this thesis, we describe these techniques in more detail including references to how they have been applied to sentiment analysis.

- Chapter 3: A framework for weakly supervised opinion mining

  This chapter defines the research framework on this thesis by introducing some of the concepts that are later used in the other chapters. This chapter also describes the structure and objective of each individual chapter, and how they are related with the final objective of obtaining a weakly-supervised Aspect-Based Sentiment Analysis system that can be easily adapted to work on corpora of different languages and domains.

- Chapter 4: Domain aspect-terms and opinion-words extraction

  This chapter describes the approach followed in this thesis to bootstrap a list of domain aspect terms and a list of opinion words from

a domain corpus and some seed words. The method is an extension of the double-propagation approach proposed by Qiu et al. (2011), including new elements and processes. It is based on a set of rules that are recurrently applied over a domain corpus to expand the initial seed words. We extend this method by building a graph during the expansion process and performing a graph-based algorithm to obtain a score for each word. We evaluate the results of the method in the SemEval 2014 ABSA dataset. In spite of using only few seed words and an unlabelled text corpus, this approach still requires supervision in the form of syntax-based rules. Since the need of a syntactic parser of a reasonable performance is not a minor requirement, in the next chapters we relax these requirements even more.

- Chapter 5: Unsupervised domain sentiment lexicon generation

  This chapter explores the use of continuous word embeddings to obtain a sentiment polarity value for the words of a domain. The proposed approach only needs a single positive word and a single negative word, plus a representative unlabelled domain corpus. In particular, we rely on the well-known Word2Vec algorithm (Mikolov et al., 2013a) but we also try other word embedding calculation approaches like GloVe (Pennington et al., 2014). The result is a domain dependent sentiment lexicon. We evaluate and compare the proposed method against other sentiment lexicon generation approaches. In addition, we propose a simple approach to separate opinion words from the rest of the words, obtaining a cleaner sentiment lexicon, just by adding an extra seed word. The described approaches are further reused and integrated into an aspect-based sentiment analysis (ABSA) system in the following chapter.

- Chapter 6: Weakly unsupervised ABSA

  This chapter integrates the ideas and approaches depicted in the previous chapter into a topic modelling system. This topic modelling system is an extended Latent Dirichlet Allocation model (LDA), which includes additional variables to model the aspect-term / opinion-word separation, and the polarity calculation, together with the domain aspects modelling. The system performs aspect-based sentiment analysis

requiring just one seed word per desired domain aspect. The only requirement to apply the system on a corpus from another language or domain is to adapt these few seed words, resulting ABSA system easy to adapt to different languages and domains. We evaluate the performance of the resulting system for several languages and domains using the SemEval 2016 task 5 datasets.

- Chapter 7: Conclusions and further work

  Finally, this chapter summarises the main conclusions obtained after this research and outlines some ideas for future work.

CHAPTER 2

State of the art

This chapter presents a review of the state of the art related to sentiment analysis and opinion mining, which is the main research area of this thesis. It is structured as follows. Section 2.1 and 2.2 provide a brief definition and motivation for the sentiment analysis task, including the description of some application domains in which sentiment analysis is relevant. Section 2.3 lists some recently funded research projects about sentiment analysis, while section 2.4 describes some international sentiment analysis competitions. Section 2.5 provides a general review of the sentiment analysis literature. Section 2.6 enters a bit more in detail and introduces continuous word embeddings and how they are used for natural language processing and sentiment analysis. Finally, section 2.7 is focused on topic modelling, in particular, on Latent Dirichlet Allocation (LDA) extensions and variations as unsupervised approaches for sentiment analysis.

2.1 Sentiment Analysis and Opinion Mining: a definition

Sentiment analysis, or opinion mining, is a sub-area of Natural Language Processing research field that focuses on detecting, classifying and quantifying affective states and subjective information (Balahur, 2011). Both terms,

sentiment analysis and opinion mining, are used interchangeably in the literature to refer to the same concept (Pang and Lee, 2008).

An opinion can be defined as a subjective statement, view, attitude, emotion, or appraisal about an entity or an aspect of an entity (Hu and Liu, 2004). Sentiment analysis aims at detecting, classifying and measuring the *sentiment* present in opinions expressed in human language. This sentiment expresses the inclinations, satisfaction and/or dissatisfaction of somebody towards something else. For example:

I would definitely recommend this place, it is amazing! → Satisfaction

What a scam! I will never buy in this store again. → Dissatisfaction

A more fine-grained scope for sentiment analysis from a psychological point of view, is the emotion analysis, that tries to go beyond a *positive/negative* paradigm to detect universal emotions, like *anger*, *joy*, *fear*, *surprise*, etc. Automatically detecting these emotions and affective states is interesting for many different purposes like customer satisfaction measure at call centres or to measure the effect of TV commercials (Kanjo et al., 2015). In many situations, like in text analysis, detecting and classifying fine-grained emotions is very hard, and the task is often simplified to measure a degree of general *positiveness* or *negativeness*. However there exist several works about capturing emotions from text (Calvo and D'Mello, 2010) and resources that map words to emotional dimensions (Strapparava and Mihalcea, 2008; Valitutti et al., 2004).

Broadly speaking, sentiments can be expressed using any of the human communication capabilities, such as body language, speech or written text. In this thesis, we deal with the sentiment expressed in digital texts. More precisely, we will refer to on-line texts written by users or customers giving their opinion about certain entities, like a purchased product or service.

The sentiment expressed in this context is commonly treated as a single dimension ranging from very positive (indicating happiness or satisfaction) to very negative (indicating anger, sadness or dissatisfaction). Opinions expressed towards a particular entity may contain a mix of positive and negative sentiments referring to different aspects of the evaluated entity, like in the following example:

*I like the place because of the **music** and the affordable **prices**. However the **staff** is slow and unfriendly and the **food** could be better.*

{music, prices} → positive

{service, food} → negative

Detecting the sentiment in a text can be broken down into several subtasks focused on detecting relevant information, like *who* is giving *which* opinion, about *what*, and *when*. A more formal definition can be found at (Liu, 2012), which defines an opinion as a tuple:

$$(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$$

where:

- e_j is a target entity being evaluated (e.g. a restaurant)
- a_{jk} is an aspect/feature of the entity e_j (e.g. the service)
- so_{ijkl} is the sentiment value from the opinion holder h_i on the aspect a_{jk} of the entity e_j at time t_l (e.g. happy/unhappy)
- h_i is the opinion holder (i.e the person emitting an opinion) (e.g. the author of the text, or a third person)
- t_l is the time when the opinion was expressed (e.g. last week, last month, etc.)

For example:

I bought the phone XYZ two weeks ago and my wife said that she does not like the new design.

This example would result, ideally, in a tuple like the following:

$$(\textit{phone XYZ}, \textit{design}, \textit{negative}, \textit{author's wife}, \textit{two weeks ago})$$

In the example, the entity would be *the phone XYZ*. The evaluated aspect is the design. The sentiment could be interpreted as negative (dissatisfaction with the evaluated aspect). However, this is the truly subjective part of the task and can be interpreted from many points of view (Maks and Vossen, 2013). In this case, the opinion holder is the *wife* of the author. Detecting the holder of the opinion is important in order to segment the users for an accurate market targeting. In this case, the needs of the author of the reviews (i.e. presumably a man) may differ from the needs of his wife with regard to the design of the product. This information could be useful to determine the root cause of the dissatisfaction. However, it is very difficult to automatically perform such a fine-grained analysis in many real cases. The time is *two weeks ago*, back from the authoring time of the review. The time is important in order to analyse trends and the evolution of the satisfaction over time.

The general objective of a sentiment analysis tool or algorithm is to fill the elements of this tuple (or a subset of them) for every opinion contained in a piece of subjective text, with the most fine-grained and accurate information possible. However, for practical reasons, most of the published sentiment analysis approaches are focused on the first three elements, entity, aspect and sentiment value, at least for on-line product reviews. The other two elements, opinion holder and time, can be usually obtained from meta-data rather than from text analysis (e.g. the username and publication date of an on-line comment).

2.2 Motivation for sentiment analysis

Sentiment analysis has always been an important element in the society (Pang and Lee, 2008). Companies want to know the market and the feelings of their customers, customers want to know the first-hand opinion about products or services before purchasing, and governments need to measure the satisfaction level of the citizens. However, compiling such information from a representative amount of people was mostly a manual process. With the advent of the digital era, the situation has changed dramatically (Mäntylä et al., 2016).

The explosion of the Web 2.0 platforms (forums, blogs, customer review portals) and social networks (Twitter, Facebook, Instagram) have caused a

constant flow of on-line user opinions. These opinions come and go in digital format, and can be easily reached, gathered and analysed.

This is a great opportunity but also a challenge. Companies need to react faster than ever to increase the satisfaction of their clients and to early detect problems and dissatisfactions. Customers have also the chance of reading recommendations and experiences from other users before making a purchase decision. Everything is subject to opinion in this new digital world: products, services, personal or corporative reputation, tourist destinations, etc.

The amount of on-line opinions generated every day makes it unfeasible to read and digest all of them without the help of automated systems. At this point is where sentiment analysis comes in, providing support to analyse, classify, group and sort this big amount of digital content, making it more manageable (Mata et al., 2012). Some remarkable application domains for sentiment analysis are:

- **Products and services:** all sort of products and services are evaluated on-line every day by experts and customers. To name a few: books, films, music, digital cameras, laptops, smartphones, phone companies, cars, fashion, etc. Each of these products families have their own particular aspect/features than can be evaluated and discussed, like image quality for cameras or performance for laptops (Pang and Lee, 2008; Cambria et al., 2013; Liu, 2012).

- **Hospitalities, restaurants, tourism destinations:** tourism has become a key industry for many countries and regions, and it is directly influenced by opinions that can be found on-line. Specialised customer review websites like TripAdvisor[1], Trivago[2] or Yelp[3], play a critical role at the moment of making a hotel or restaurant reservation. Again, hotels and restaurants are domains with their own elements, like cleanliness, service or price (Lak and Turetken, 2014; Jabreel et al., 2017; Höpken et al., 2017; Neidhardt et al., 2017).

- **eGovernance:** citizens react to their governors' actions and policies, and these reactions are represented almost in real time in on-line communication channels. Twitter and similar social networks have become

[1] `https://www.tripadvisor.es/`
[2] `https://www.trivago.es`
[3] `https://www.yelp.es`

critical tools to measure the level of satisfaction of a society, and the need of tools to analyse and interpret the content flowing at social media is increasing every day (van Son et al., 2014; Ceron and Negri, 2015; Wang et al., 2016; Inyang et al., 2017).

These are only a few examples of domains for which automatic analysis of on-line opinions is important. Other application domains that are gaining more and more relevance are cyberbullying prevention, stock markets prediction and medicine. As digital devices pervade the society the volume of content that needs to be analysed will only grow in the coming years. Automatic analysis tools will need to handle different types of content, about many different topics, and in several different languages (Balahur and Perea-Ortega, 2015).

2.3 Sentiment analysis related projects

Sentiment analysis is a research area of interest for the society. For this reason, during the last years, several projects have received funding to carry out research and development closely related with sentiment analysis, applied to different domains. The following list cites some projects related to sentiment analysis funded by the European Commission:

- **OpeNER**[4] (Open Polarity Enhanced Named Entity Recognition) was an European Commission 7th Framework Programme project spanning from 2012 to 2014. Its goal was easing the reuse of existing language resources and data sets to provide a set of base NLP technologies to the community. OpeNER provided Named Entity Recognition tools and Sentiment Analysis tools for Spanish, English, French, German, Dutch and Italian. The modular architecture based on KAF format (Bosma et al., 2009) allows easy reuse of existing modules and extension. In addition, a big dataset of social media opinions about hospitalities was gathered, resulting in an interesting resource called Tour-pedia[5] (Cresci et al., 2014) that contains thousands of customer reviews in several languages with rating and geo-positioning, as shown in figure 2.1. We did participate in the OpeNER project (García-Pablos et al., 2015a, 2013).

[4] `http://www.opener-project.eu/`
[5] `http://tour-pedia.org/about/`

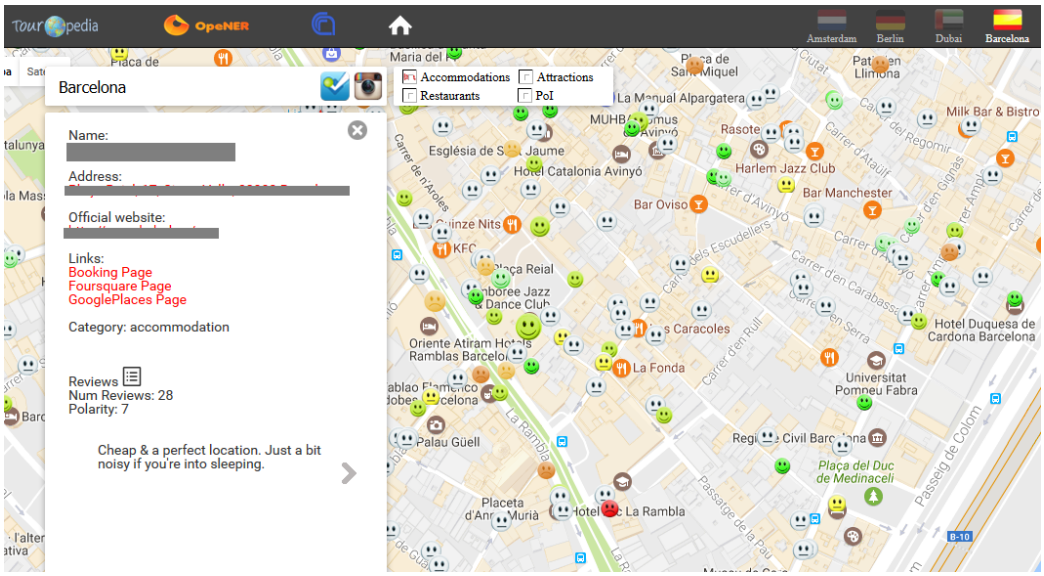


Figure 2.1: Tourpedia GUI screenshot (Cresci et al., 2014).

- **EuroSentiment**[6] (Language Resource Pool for Sentiment Analysis in European Languages) was a European Commission 7th Framework Programme project spanning from 2012 to 2014. EuroSentiment aimed at developing a large shared data pool for language resources meant to be used by sentiment analysis systems, in order to bundle together scattered resources. The project specified a schema for sentiment analysis and normalised the metrics used for sentiment strength. The project covered 6 languages: English, Catalan, German, Italian, Portuguese and Spanish.

- **MULTISENSOR**[7] (Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation) was a European Commission 7th Framework Programme (FP7) project spanning from 2013 to 2016. The project was aimed at mining heterogeneous data from TV, radio, mass media websites and social media and apply multidimensional content integration, including sentiment and context analysis of content and social interactions.

[6] http://eurosentiment.eu/
[7] https://www.multisensorproject.eu/

- **SSIX**[8] (Social Sentiment analysis financial IndeXes) is an European Commission Horizon2020 project spanning from 2015 to 2018. SSIX aims at providing European SMEs with a collection of easy to interpret tools to analyse and understand social media users attitudes for any given subject. These sentiment characteristics can be exploited to help SMEs to operate more efficiently resulting in increased revenues.

In addition, with the increase of on-line radicalisation and religious extremism sentiment analysis has become an element of interest for cybersecurity and terrorism prevention related projects. For instance, an estimation from Autumn 2014 pointed that between 40k and 90k Twitter accounts were supporting terrorist groups (Ghajar-Khosravi et al., 2016) and between January and June 2015, about 25k Twitter accounts were reported for supporting terrorist groups (Ferrara et al., 2016). Sentiment analysis, in combination with other approaches, can help detecting cases like these in a more systematic way. Due to that, the application of sentiment analysis to this context is an element of active research (Agarwal and Sureka, 2015; Bouchard et al., 2014; Choi et al., 2014; Zubiaga et al., 2015; Scanlon and Gerber, 2014; Munezero et al., 2014).

2.4 International shared tasks and competitions

This section briefly describes some relevant international competitions related to sentiment analysis. In those competitions and shared tasks research teams from all around the world compete to solve sentiment analysis related problems using real datasets. Some of these competitions are sponsored and involve actual money prizes, revealing the relevance of sentiment analysis both for the industry and the academia.

2.4.1 SemEval ABSA shared task

Sentiment Analysis has motivated a number of competitions and shared tasks. One of the most directly related to sentiment analysis is SemEval.

[8] `https://ssix-project.eu/`

The SemEval (Semantic Evaluation) competition holds every year one or more shared tasks related to sentiment analysis. During the SemEval editions in 2014[9], 2015[10] and 2016[11] there has been a task devoted to Aspect Based Sentiment Analysis (ABSA) (Pontiki et al., 2014, 2015, 2016). The participation, as in other SemEval tasks, is open to any group or individual that wants to participate with an ABSA system. The participants are given a manually labelled training dataset to develop and train their systems. At the evaluation time, the organisers provide a test set without the gold annotation, and the participants submit the resulting annotations after applying their methods to the test set.

The objectives of the ABSA analysis have subtle differences from one SemEval edition to the next, but in general, they consist of identifying explicit aspect terms, classifying sentences into coarse-grained domain aspects/categories and detecting sentiment polarity. A more detailed explanation of each subtask is the following:

- **Aspect term extraction:** given a set of sentences with pre-identified entities (e.g., restaurants), this subtask is about identifying the aspect terms present in the sentence and return a list containing all the distinct aspect terms.

  For example, *"I liked the service and the staff, but not the food"*, or *"The food was nothing much, but I loved the staff"*.

  Multi-word aspect terms (e.g., *"hard disk"*) should be treated as single terms (e.g., in *"The hard disk is very noisy"* the only aspect term is *"hard disk"*).

- **Aspect category detection:** given a predefined set of domain aspect categories (e.g., price, food), this subtask is about identifying the aspect categories discussed in a given sentence.

  For example, given the set of aspect categories {food, service, price, ambience, anecdotes/miscellaneous}:

  "The restaurant was too expensive" → {price}
  "The restaurant was expensive, but the menu was great" → price, food

[9] http://alt.qcri.org/semeval2014/task4/
[10] http://alt.qcri.org/semeval2015/task12/
[11] http://alt.qcri.org/semeval2016/task5/

Domain	Train	Test	Total
Restaurants	3041	800	3841
Laptops	3045	800	3845

Table 2.1: Sizes of datasets provided in SemEval 2014 (number of sentences).

- **Polarity classification:** for a given set of aspects or aspect terms within a sentence, the objective of this subtask was to determine whether the polarity of each aspect term was positive, negative, neutral or conflict (i.e., both positive and negative).

  For example:

  "I loved their fajitas" → {fajitas: positive}
  "I hated their fajitas, but their salads were great" → {fajitas: negative, salads: positive}
  "The fajitas are their first plate" → {fajitas: neutral}
  "The fajitas were great to taste, but not to see" → {fajitas: conflict}

2.4.1.1 SemEval ABSA datasets

SemEval ABSA competitions provide the datasets as XML files. Each sentence contains the original texts, and in their corresponding fields or attributes appears the information about the labelled aspects or polarities. Figure 2.2 shows an example of a fragment of one of these XML files containing annotations for restaurants reviews.

SemEval 2014 and SemEval 2015 ABSA task editions provided datasets of restaurant and laptop reviews, only for English. SemEval 2015 included a small hotel reviews dataset only for testing domain adaptation. Tables 2.1 and 2.2 show the size of these datasets measured in sentences. SemEval 2016 provided multilingual datasets, in particular, restaurant customer reviews in English, Spanish, French, Dutch, Russian and Turkish and other datasets scattered among several languages and domains.

Table 2.3 shows the sentence distribution among the different languages and domains. All datasets and more detailed information can be found in their corresponding SemEval task web pages.

```
<sentence id="1700205:1">
    <text>From the spectacular caviar to the hospitable
    waitstaff, I felt like royalty and enjoyed every second of
    it.</text>
    <Opinions>
        <Opinion target="caviar" category="FOOD#QUALITY"
        polarity="positive" from="21" to="27"/>
        <Opinion target="waitstaff" category="SERVICE#GENERAL"
        polarity="positive" from="46" to="55"/>
    </Opinions>
</sentence>
<sentence id="1700205:2">
    <text>Considering we were the last patrons there and it was
    after the closing time, the waitstaff did not rush us at
    all and made us feel comfortable and relaxed.</text>
    <Opinions>
        <Opinion target="waitstaff" category="SERVICE#GENERAL"
        polarity="positive" from="82" to="91"/>
    </Opinions>
</sentence>
<sentence id="1700205:3">
    <text>I highly recommend Caviar Russe to anyone who wants
    delicious top grade caviar and fantastic service.</text>
    <Opinions>
        <Opinion target="caviar" category="FOOD#QUALITY"
        polarity="positive" from="72" to="78"/>
        <Opinion target="service" category="SERVICE#GENERAL"
        polarity="positive" from="93" to="100"/>
    </Opinions>
</sentence>
```

Figure 2.2: Fragment of a SemEval annotated dataset.

2.4.2 WASSA 2017 shared tasks

The 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2017) includes a shared task related to emotion detection (Mohammad and Bravo-Marquez, 2017). The objective of the shared task is described as: *"Given a tweet and an emotion X, determine the intensity or degree of emotion X felt by the speaker, a real-valued score between 0 and 1."*. Several datasets and baselines are provided[12]. Provided training and test datasets are annotated using four different emotions: joy, sadness, fear, and anger. Another shared task about emotion linking and classification is also planned for the workshop.

[12] `http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html`

Domain	Train	Test	Total
Restaurants	1315	685	2000
Laptops	1739	761	2500
Hotels	-	266	266

Table 2.2: Sizes of datasets provided in SemEval 2015 (number of sentences).

Language	Domain	Train	Test	Total
English	Restaurants	2000	676	2676
Spanish	Restaurants	2070	881	2951
French	Restaurants	1733	696	2429
Dutch	Restaurants	1711	575	2286
Russian	Restaurants	3490	1209	4699
Turkish	Restaurants	1104	144	1248
English	Laptops	2500	808	3308
Arabic	Hotels	4802	1227	6029
Chinese	Phones	6330	3191	9521
Chinese	Cameras	5784	2256	8040
Dutch	Phones	1389	308	1697
French	Museums	-	686	686
Turkish	Telecom	3000	310	3310

Table 2.3: Sizes of datasets provided in SemEval 2016 (number of sentences).

2.4.3 Other competitions

Kaggle[13] is a platform for predictive modelling and analytics competitions. It was founded in 2010, and hosts competitions to build models for predictive analysis in a wide range of tasks. Kaggle is a crowd-sourcing platform in which organisations publish datasets and competitions, and teams from all around the world compete to obtain the reward. There are different types of rewards, like money prizes or a job position, which motivates the participation of a lot of people trying to solve each problem. Kaggle has also held competitions about sentiment analysis. For instance, a competition about

[13] `https://www.kaggle.com/`

predicting the user ratings of movie reviews using a dataset from the website Rotten Tomatoes took place in 2015. Another competition was about predicting sentiment in US Airline tweets. The competitors were requested to predict the sentiment polarity value (positive, negative, neutral) and classify the reasons for negatives (e.g. a delay or rude service).

Another competition worth mentioning is the Yelp dataset challenge[14]. In this case, the centric resource is the so-called Yelp dataset, which contains several millions of customer reviews about places to eat (primarily, but not only, restaurants) in English. The competition has had several rounds and phases and has awarded more than 50k dollars in money prizes. The competition is not only about sentiment analysis. There are more elements that have to be modelled and predicted like: seasonal and cultural trends, social graphs, events that affect businesses, location mining, etc.

2.5 Overview of sentiment analysis approaches

The last two decades have been very prolific to what comes for sentiment analysis systems and methods. Sentiment analysis techniques can be arranged into different categories depending on the subtasks they solve or the type of algorithm they use. There are numerous surveys that cover a wide range of works about sentiment analysis from different perspectives (Pang and Lee, 2008; Moghaddam and Ester, 2013; Liu, 2012; Zhang and Liu, 2014; Ravi and Ravi, 2015; Balahur and Jacquet, 2015; Schouten and Frasincar, 2016; Rana and Cheah, 2016). In this section we briefly point to some relevant sentiment analysis systems, to outline the diverse approaches that have been attempted for aspect detection, sentiment classification, and for other challenges that are important for an accurate sentiment analysis, like dealing with negation or interpreting sarcasm and irony.

2.5.1 Aspect detection

Sentiment analysis can be done at different granularity levels. Sentiment can be classified at document level, sentence level or aspect level (Liu, 2012; Moghaddam and Ester, 2013).

[14] `https://www.yelp.com/dataset_challenge`

Document-level sentiment analysis provides a single sentiment value for an entire piece of text, such as a full customer review, a news article or a complete blog entry. In this case, it is assumed that the whole document is expressing a single sentiment value about for only one entity or topic. Most of the time this is not the case. A piece of text may exhibit different sentiments about several entities or aspects of an entity, so document level approaches miss a lot of potentially useful information.

Sentence level sentiment analysis makes a similar assumption but the analysis unit are sentences instead of entire pieces of text. This increases the granularity and the information that can be obtained, but it is still assuming that each sentence contains only a single sentiment value about a single entity or feature. However, it is often the case that within the same sentence different elements are compared or several features are evaluated at the same time.

Aspect-based sentiment analysis aims at obtaining a sentiment value for every detected opinion target. Opinion targets are the features that are being evaluated, like the *size* of an element when it is described as being *small* or *huge*. This includes the additional step of detecting those potential opinion targets. Each application domain has its own collections of opinion targets that refer to rateable elements. Opinion targets are commonly grouped into more coarse-grained categories, called *aspects*.

Aspect-based sentiment analysis (ABSA) is more informative and useful because of its finer granularity, and it has been the main trend in sentiment analysis during the last years (Zhang and Liu, 2014). Aspect detection approaches can be grouped according to the main set of techniques they use to perform the task.

2.5.1.1 Frequency based approaches

Frequency based approaches assume that within a corpus of a certain domain, some domain-related words are significantly more frequent than the rest of the vocabulary. These words, usually nouns or noun phrases, have a high probability of being aspects. Despite its simplicity, counting the frequency of occurrence of some words has proven to be a valid heuristic to identify some aspect terms.

An obvious drawback of these approaches is that not all frequent nouns or noun phrases are aspect terms, decreasing the precision of the results. On

the other hand, some very specific aspects may not be mentioned frequent enough to be captured by frequency-based methods.

Hu and Liu (2004) propose a frequency-based system to mine aspect terms. This system only considers single or composed nouns as potential candidates to be aspect terms. After obtaining a list of aspect term candidates based on frequency, the authors use hand-crafted rules to prune the lists and reduce the noise produced by false positives. Long et al. (2010) improved this approach by incorporating some grammatical dependencies to find infrequent aspects terms.

Hai et al. (2011) introduce association rules to mine aspects terms. These rules restrict the aspect terms to appear in association with sentiment words. In order to alleviate the problem of frequent non-aspect nouns polluting the result Li et al. (2009) compare the frequency of vocabulary words with a baseline corpus of 100 million English words. Comparing the out-of-domain frequencies provides useful information for discarding some of the incorrect results based on how are they distributed across different domains.

2.5.1.2 Syntax based approaches

An improvement over just paying attention to word frequencies is to include syntax-based rules and methods to help to detect aspect terms from domain texts.

The most common way is to focus on simple syntactical relations like an adjectival modifier, as in *wonderful service* or *horrible design*. Since the frequency is no longer the only heuristic to find aspect terms, less frequent aspect terms have more chances of being extracted. However, syntax based approaches have their own shortcomings.

On the one hand, syntax based approaches rely on the performance of available syntax analysis tools for the target language. This is especially relevant for languages other than English, but also for informal texts which tend to contain grammatical and orthographical errors that reduce syntax analysis tools accuracy, especially when such tools were trained on formal texts.

On the other hand, if the syntactic relations used for extraction are too restrictive, it may result in a low recall problem. To alleviate this low recall some generalisation techniques have been applied. Zhao et al. (2010b) do

not apply syntax trees directly to extract new aspect terms. Instead, syntax trees are split into several substructures and a similarity measure is calculated based on these substructures to decide when to extract a term or not.

Qiu et al. (2011) combine aspect term extraction with opinion words extraction in a so-called double-propagation approach. A set of syntax-based rules is applied to a domain dataset to extract target words (aspect terms or opinion words depending on the triggered rule) starting from some seeds. Each extracted word is added to the set of words that can trigger a rule. The method runs over the dataset until no more words can be extracted.

2.5.1.3 Supervised machine learning approaches

Supervised machine learning algorithms use training data to fit a model capable of classifying unseen data into some of the learned categories. Hence, a key point of the performance of this kind of methods is the quality of the available training data. If there is enough representative training data of the target domain, then supervised machine learning approaches obtain a good performance, usually outperforming unsupervised approaches. Of course, such training data must be obtained from somewhere. Most of the times it requires a manual labelling task carried out by a domain expert. This is often a tedious, time-consuming and expensive task, and difficult to reuse for other languages or domains.

Another important factor is how these training instances are represented, i.e. the feature engineering that is used to represent each training example as the input for the machine learning algorithm. Common features are based on word and n-gram frequencies, bag-of-words, Part-of-Speech, and other salient features present in the texts.

Jakob and Gurevych (2010) use Conditional Random Fields (CRF) (Lafferty et al., 2001). CRFs are a common approach in natural language processing for sequence labelling. In sequence labelling, the context of a word and the labels assigned to previous words are used to determine the probability of each possible label for the current word. Jakob and Gurevych (2010) use, among other features, the current word, its Part-of-Speech, the presence of a direct dependency relation and the proximity to a sentiment expression to describe each word in the sequence.

Agerri and Rigau (2016) use a perceptron based classifier, but they enrich the number of features using different types of unsupervised word clusters,

n-grams, the presence or absence of a word in domain based gazetteers, etc. Despite the system is primarily aimed at Named Entity Recognition, the authors also use it for aspect term extraction obtaining the best results at SemEval 2015 shared task (Pontiki et al., 2015). Their participation in that shared task is described by San Vicente et al. (2015).

More recent approaches use neural networks and deep learning (Qian et al., 2017; Socher et al., 2013; Shin et al., 2016). Apart from the algorithms, one of the main difference is in how the data is encoded. Words are transformed into the so-called continuous word embedding spaces, which are high dimensional vector spaces capable of encoding some linguistic and semantic regularities. These vectors serve as the input to complex architectures of neural networks.

2.5.1.4 Unsupervised machine learning approaches

As opposed to supervised machine learning approaches, unsupervised machine learning approaches do not require labelled data for training. They only require labelled data to evaluate the performance of the resulting model. However such methods need a large amount of unlabelled data to calculate meaningful statistics necessary to build a robust model. Fortunately, unlabelled data is much easier and cheaper to obtain because it is continuously generated by users.

Etzioni et al. (2004) present the KnowItAll system to extract Web-based and domain-independent information. Starting from a set of relations of interest, KnowItAll induces relation-specific extraction patterns to find aspect term candidates. Popescu and Etzioni (2007) build upon KnowItAll to present OPINE, another Web-based information extraction system. To score the candidates they use metrics like Pointwise Mutual Information (PMI) combined with a Naive Bayes classifier.

Quan and Ren (2014) propose another unsupervised method to extract product aspects. The domain aspects are extracted by measuring the similarity distance of some calculated domain vectors. These domain vectors are derived from the association values calculated for each feature in the domain corpus. They introduce a new similarity measure named PMI-TFIDF to evaluate the association of candidate aspects and domain entities.

Many other unsupervised approaches rely on Latent Dirichlet Allocation (LDA) based extensions. LDA (Blei et al., 2003) is a generative probabilistic

model, where documents from a text dataset are modelled as a finite mixture over an underlying set of topics. Since the original LDA is designed to work on the document level, the aforementioned extensions implement changes to model topics at sentence and/or aspect level (Titov and McDonald, 2008). A more detailed description of LDA-based approaches for aspect and sentiment analysis is carried out at section 2.7.

2.5.2 Sentiment classification

The other part of aspect-based sentiment analysis, once the target aspect has been determined, is the sentiment classification itself. This step determines if the mentioned target is being evaluated positively or negatively. A practical approach to measuring the sentiment polarity is to use a dictionary to obtain the sentiment polarity value of the words contained in a piece of text. Obviously, a naive dictionary based approach has its limitations, like word ambiguity, context dependent sentiment, complex expressions, negations, sarcasm, etc. Sentiment polarity can be also determined using supervised machine learning based methods. And also several unsupervised methods have been proposed to mitigate the fact that supervised methods need training data that may not be available for given language and domain.

2.5.2.1 Dictionary based approaches

Dictionary-based approaches use a dictionary to map words to a sentiment polarity value. Such dictionaries are called sentiment lexicons. A sentiment lexicon can be used to find word occurrences in texts, counting and weighting the total number of positive and negative words. Another use for sentiment lexicons is to serve as input features for more complex methods, like supervised machine learning algorithms.

A sentiment lexicon can be general, meaning that it contains the most probable sentiment polarity for each word in a regular use of the language. There are several general polarity lexicons available, like the old General Inquirer (Stone et al., 1968) or the one from (Hu and Liu, 2004) commonly named "Bing Liu's sentiment lexicon". Most of the attention and resources are generally focused on English, but there are equivalent approaches to obtain sentiment polarity lexicons for other languages, like Spanish. Molina-González et al. (2013) presented eSOL, an equivalent to Liu's polarity lexicon

for Spanish. The authors translate the terms, and perform several word expansion in Spanish datasets, using some word pruning rules to clean and fine-tune the resulting polarity lexicon. Similar efforts to improve domain-adapted sentiment lexicons for Spanish can be found at (Cruz et al., 2014) or (Jiménez-Zafra et al., 2016a).

There are several methods to generate sentiment lexicons from lexical and semantic databases like WordNet (Fellbaum, 1998). Examples of methods that use WordNet to calculate a sentiment polarity value for words are SentiWordNet (Baccianella et al., 2010), SentiWords (Guerini et al., 2013) or QWordNet-PPV (San Vicente et al., 2014).

The main drawback of general sentiment lexicons is that the polarity of some words may depend on each particular application domain. Another obvious problem is that it is likely that general lexicons will miss specific words that are only of common use for a certain domain (e.g. technical words, jargon, slang).

2.5.2.2 Supervised machine learning approaches

Dictionary based methods rely mainly on the information present in the dictionary itself, or in the sources used to generate the dictionary (e.g. a semantic or lexical database). Supervised machine learning approaches use labelled data to learn sentiment classification models. Sentiment lexicons are often used as features for these classifiers.

Blair-Goldensohn et al. (2008) present a system that detects, summarises and aggregates the sentiment of reviews for services like a restaurant or a hotel. The system uses maximum entropy classifiers trained over the user-generated ratings for the reviews.

Yu et al. (2011) introduce a system to automatically identify opinions from on-line consumer reviews. Their system first extracts and ranks domain aspects, and then they use extracted aspects as features for a supervised sentiment classifier trained on customer reviews.

Martínez-Cámara et al. (2015) combine different resources from training, in this case, a sentiment classification method for Spanish. They show that combining information coming from different sources lead to an improvement of the sentiment classification accuracy.

Furthermore, during the last few years, the deep-learning trend has also reached to sentiment analysis. Neural Networks and Deep-learning based approaches are beating state-of-the-art supervised approaches, especially when there is enough training data available.

Socher et al. (2013) explore recurrent neural networks for sentiment analysis proposing the Recursive Neural Tensor Network (RNTN) to learn to classify sentiment and its scope within sentences. They use a manually labelled dataset, Sentiment TreeBank, which mixes syntactic dependencies with the sentiment of each node of the syntactic tree. Since obtaining this kind of manually labelled data is difficult, they use Amazon Mechanical Turk to outsource the annotation efforts, obtaining more than 200k phrases for training. RNTN recursively traverses the syntactic tree learning the expected sentiment value for each node according to the words and their syntactic relations. This syntactic level granularity allows the model to eventually learn the implicit scope of negations and other modifiers.

Tai et al. (2015) propose an approach using Long Short-term Memory neural networks (Hochreiter and Schmidhuber, 1997) (LSTMs). LSTMs are a special type of RNNs, which neurones (individual units that form the network) are more complex and include a few additional logical gates to operate in each step. LSTMs are designed to better preserve the information when processing long sequences, avoiding the gradient vanishing problems that affect to other simpler RNNs (Hochreiter, 1998). The approach proposed by Tai et al. (2015) combines LSTMs in a tree structure.

Qian et al. (2017) present another method based on LSTMs for sentiment classification. In this case, the model works at sentence-level, and try to incorporate information from sentiment lexicons, negations or intensifiers.

Kim (2014) uses Convolutional Neural Networks (CNN), a special neural network architecture inherited from image processing, for sentence-level sentiment classification. Each sentence is modelled as a stack of word vectors coming from pre-computed word embeddings. To handle sentences of variable length a special *padding* token is used. Several variations of this architecture are proposed. CNN-rand uses randomly initialized word vectors (i.e. no pre-computed word embeddings). CNN-static freezes the word embeddings, so they are not fine tuned during the training of the CNN. CNN-non-static allows word embeddings to change during the CNN training. And finally, CNN-multichannel uses two sets of word embeddings, one static and the other non-static.

Shin et al. (2016) propose an extension that integrates sentiment lexicon information in the CNN training. They look at several sentiment lexicons and normalise the polarity score for each word in a range [-1,1]. Then they concatenate each polarity value into a vector, and this vector is combined with the regular word embedding for each word of a sentence. Using this combination as input they are injecting sentiment information.

2.5.2.3 Unsupervised machine learning approaches

There are methods to bootstrap the polarity of the words directly from a corpus and a few seed words. These methods expand the initial polarity of the provided seeds using different techniques, from statistical counts to simple language rules and heuristics or graph-based algorithms. This kind of methods is unsupervised in the sense that they do not rely on a set of manually labelled examples to train a statistical model. Instead, the results are obtained exploiting linguistic cues and regularities.

Hatzivassiloglou and McKeown (1997) bootstrap positive and negative adjectives from a large corpus using conjunctions. Starting from some positive and negative seed adjectives, an iterative process extracts other adjectives based on some simple heuristics. If an adjective is related to a known positive adjective by a copulative conjunction (e.g. the staff was attentive *and* polite), then that adjective polarity is assumed as positive (same case for the negatives). Moreover, if an adjective is related to another adjective with a known sentiment polarity by the word *but* the polarity is reversed (e.g. the staff was polite *but* clumsy).

Turney (2002) calculates the semantic orientation (SO) of all the words in a corpus using two words with a significant polarity value (excellent and poor) and the hits count on the now deceased web search engine Altavista. Measuring the Pointwise Mutual Information (PMI) of each word and the chosen polarity words and continuous value is obtained for each word in the vocabulary.

Qiu et al. (2011) apply a set of heuristics similar to the ones from (Hatzivassiloglou and McKeown, 1997) to assign a polarity to each word while bootstrapping words from a corpus using syntactic dependency based extraction rules. A very similar technique was also used in (Brody and Elhadad, 2010).

In (Kiritchenko et al., 2014) the authors use a similar technique but the co-occurrence counts are based on domain datasets instead of web search hits

count. They also filter out some low-frequency terms from the vocabulary to reduce the noise caused by those elements.

More recently, Hamilton et al. (2016) use continuous word embeddings to induce the sentiment polarity value of words. They build a graph with weighted edges based on word-vector-based similarity and perform random walks to expand the polarity of some initial seeds with known sentiment polarity. The resulting word scores in the graph are used to rank the words polarity.

Other unsupervised or weakly supervised approaches rely on Latent Dirichlet Allocation (LDA). A more detailed list of such systems can be found at section 2.7.

2.5.3 Other sentiment analysis related sub-problems

Apart from the main tasks of aspect and sentiment classification, sentiment analysis must deal with some other nuances and details for a correct and complete understanding of the sentiment expressed in a text. Most sentiment analysis approaches ignore these details or try to solve them with very basic heuristics. In this section, we briefly outline and describe some of them and mention some research works related to each of them.

2.5.3.1 Sentiment negation and augmentation

Within a sentence, there may appear words that affect the sentiment polarity of other nearby words. One of the names used to refer to such words is *contextual valence shifters*. A theoretical discussion about contextual valence shifters can be found at (Polanyi and Zaenen, 2006). There are different types of valence shifters according to how they affect the sentiment polarity value.

Negation shifters revert the polarity. Some examples of negation shifters for English are *no*, *not*, *nothing* or *neither*. When a negation modifies a positive opinion expression it becomes negative and vice-versa.

Example:

This is a good place→ *positive*
This is **not** *a good place*→ *negative*

Intensifiers are other type of valence shifters. Some examples of intensifiers shifters for English are *very*, *really*, *completely* or *incredibly*. Instead of reverting the polarity, they change its intensity.

Example:

This is a good place→ positive
This is a **very** *good place→ very positive*

From these two types of valence shifters, negations are the most important, because they completely change the meaning of an opinion.

Example:

The food is good and the service is attentive. Good choice! → Satisfied customer
The food is ***not*** *good and the service is* ***not*** *attentive.* ***Not*** *a good choice! → Unsatisfied customer*

The above examples just differ in few words, however, their meaning with regard to customer satisfaction are opposite. Ignoring negations may lead to a total misunderstanding of the sentiment contained in a piece of text.

Moilanen and Pulman (2007) encoded some of the ideas theoretically described by (Polanyi and Zaenen, 2006) in explicit rules. The result is a pipeline, which includes a Part-of-Speech tagger and a syntactic parser, in which the rules are applied systematically to deal with negation.

A very important issue with the negation is to determine its scope, i.e. to which parts of the discourse does affect the occurrence of a negation expression. Most sentiment analysis approaches deal with negation and their scope using a very basic heuristic, consisting of a simple word distance. According to this distance-based heuristic, any word within a certain context window is affected by the negation reversing its polarity. Despite being widely used due to their simplicity, such fixed hand-crafted rules are not as effective as other statistical learning counterparts, as described in (Kiritchenko and Mohammad, 2016). An example of a system that uses machine-learning based information to determine the negation scope is (Socher et al., 2013). The proposed approach combines the syntactic-dependency tree of each analysed sentence with a set of rules to decide the polarity.

Jiménez-Zafra et al. (2016b) study negation in the context of Spanish texts (Jiménez-Zafra et al., 2015). For other sentiment shifters, like intensifiers, the scope can be delimited using similar techniques.

In addition, other indicators present in texts can be used to modify the sentiment polarity value or intensity, like the punctuation(e.g. *'!!!!'*, *'!?!?!'*), repetition of letters within the same word (e.g. *"yessss"*, *"woooow"*), capitalization (e.g. "this is a BAD choice") or the presence of *emoticons*. These kind of cues are specially relevant for the analysis of texts in the context of social media and micro-blogging platforms (Go et al., 2009; Kouloumpis et al., 2011; Sarlan et al., 2014).

2.5.3.2 Multiword terms

Each application domain has its own vocabulary, the set of words to refer to the different aspects and features. In order to accurately manage the terminology of a domain, a system has to deal with multiword terms and expressions (Sag et al., 2002).

Multiwords can be of different types, from composed names to expressions that involve more than one word. Examples of multiwords are: *air conditioning*, *happy hour* or *touch pad*.

Multiwords are important, especially for systems based on bag-of-words, because some multiword expressions can be misleading if their individual components are processed separately. For example, the expression *happy hour* is a concept that does not necessarily bear any sentiment, as in the sentence *"The happy hour starts at ten o'clock."*. But since it contains the word *happy*, a word bearing a positive sentiment, a bag-of-words based system may process the sentiment of the sentence incorrectly.

Capturing multiword expressions is also important for domains with a lot of specialised terminology, like electronic devices or computer: *touch screen*, *touch pad*, *hard disk drive*, *graphics card*, *sound card*, etc.

Common ways to deal with multiword expressions are based on dictionaries and, depending on the language, simple heuristic rules (Anastasiou, 2010). A gazetteer of multiword expressions can be created manually or by looking for composed terms in encyclopaedic sources (e.g. Wikipedia), lexical databases, ontologies or thesaurus (e.g. WordNet, ConceptNet), etc. This

approach provides high precision because all the included expression are implicitly verified. But this comes at the cost of a lower recall since new expressions, slang, jargon, and other language phenomena may not be present in the more formal sources (Sag et al., 2002).

An alternative to discover multiword expressions directly from a domain corpus is using word co-occurrence based statistics. A general heuristic states that if the probability of n words appearing in sequence in a corpus is significantly higher than the probability of finding the words independently, then such sequence is probably a valid multiword expression. There are different measures, like Pointwise-Mutual-Information (PMI), or log-likelihood ratio (LLR) (Dunning, 1993). Brown clustering can be also used to capture composed expressions (Brown et al., 1992).

Finlayson and Kulkarni (2011) introduce jMWE, a Java library for detecting multi-word expressions in a text that bundles several algorithms for this task, evaluated in the context of word sense disambiguation.

2.5.3.3 Comparative sentences

Comparative sentences are those which present two or more entities or aspects of an entity, making a comparison between them and stating a preference of one over the other.

For example:

The camera X has a better image quality than Y

In the above example, X and Y are two digital cameras, and their image quality is being compared.

Jindal and Liu (2006) study the problem of detecting comparative sentences in customer reviews. They first categorise sentences into several types and then present a method that combines pattern detection with supervised learning to classify comparative sentences.

Ganapathibhotla and Liu (2008) propose a method to mine opinions from comparative sentences. They enumerate a set of comparative sentence types and rules to infer the sentiment from them.

2.5.3.4 Conditional sentences

Conditional sentences state something that is not immediately true but conditioned to something else. For example:

If the laptop X has the best performance I will buy it.

In the above example the excerpt *the laptop X has the best performance* can be interpreted as positive. However, in the context of the conditional both the interpretation and the sentiment of the sentence change.

Narayanan et al. (2009) present a study and an approach to deal with conditional sentences. First, they introduce the different types of conditional sentences and propose some rules and heuristics to process them. The proposed approach is mostly based on the detection of certain connectives and conditional particles, and Part-of-Speech based patterns.

2.5.3.5 Ironic and sarcastic sentences

Finally, another nuance to be taken into account when processing subjective texts in sentiment analysis is the irony and sarcasm. Sarcasm and irony are forms of communication in which the intended message is opposite to the one distilled from a literal interpretation of the message itself. It is a common phenomenon present in sentiment analysis, especially when people make jokes or try to express their anger or dissatisfaction in a more subtle way.

For example:

Their new phone is only 800$? Give me a dozen!

The best part is the cover (about a book)

I felt so happy when I saw the end credits! (about a film)

It has all the features I always wanted. Too bad none of them work!

In the above example, a naive interpretation would lead to the belief that mentioned products have positive aspects that satisfy the customer. However, it is obvious that the intention is precisely the opposite, to satirise about them. Such interpretation is really challenging for a machine because it is a task in which the amount of data to train models on does not seem to help computers to perform better (Wallace and Kertz, 2014). A correct interpretation of a sarcastic sentence requires a deep world knowledge, and even humans have difficulties identifying sarcasm.

Tsur et al. (2010) present a semi-supervised approach to identify sarcastic sentences in product reviews. Their approach has two stages. the first is a semi-supervised pattern acquisition. The second stage performs the sarcasm classification.

Davidov et al. (2010) propose a semi-supervised sarcasm identification on two different datasets: one composed of tweets from Twitter and another composed by customer reviews from Amazon. They create a gold standard using Amazon Mechanical Turk as the tool to label the datasets. The proposed approach consists of a semi-supervised pattern extraction, pattern selection and pattern matching.

Reyes et al. (2013) describe a set of textual features for recognising the irony in short texts like tweets from Twitter. They identify a set of discriminative features to automatically differentiate an ironic text from a non-ironic one.

A recent approach described at (Ghosh and Veale, 2016) uses *deep learning* to tackle the problem. In particular, the proposed method combines convolutional and sequential neural networks to train a supervised model for sarcasm detection. The system outperforms the previous state-of-the-art methods for sarcasm detection.

2.6 Continuous word embeddings

Historically, one of the most common ways to represent words in Natural Language Processing algorithms was treating them as symbols. In this context *word* means any individual token resulting from the application of a particular text pre-processing and segmentation algorithm like, for example, a simple white-space tokenisation. Each word/token is treated as an atomic symbol, a literal, represented by an index over a vocabulary, with no additional meaning, context or associated information (Manning et al., 1999).

Continuous word embeddings are related to the area of distributional semantics, in which words are described with regard to their co-occurrence with other words. Hence they are not so different in essence to well-known methods like Latent Semantic Indexing (Deerwester et al., 1990). However neural word embeddings achieve far better results in most Natural Language Processing tasks (Baroni et al., 2014).

Continuous word embeddings map words from a vocabulary to dense vectors in a continuous multidimensional space. More formally, a continuous word embedding is any function that maps words from a fixed vocabulary to some vector space, $W : words \rightarrow \mathbb{R}^n$. The obtained word vectors are capable of encoding interesting word information and properties. The nature and robustness of the properties encoded depend on the particular algorithm used to obtain the word embedding.

Word embeddings can be computed directly from a big corpus of unstructured data, like Wikipedia articles or news text. Some methods include extra resources and information (e.g. manually labelled data) to specialise the resulting embeddings for a certain task.

In addition, word embeddings serve as a good representation for the artificial neural networks and deep learning approaches applied to Natural Language Processing, including sentiment analysis. Neural networks handle better information-rich dense vectors of continuous values rather than sparse one-hot encodings.

This section describes some of the most popular continuous word embedding general approaches, as well as some methods to obtain more specific word embeddings.

2.6.1 General purpose word embeddings

Word2Vec is probably the best-known method to obtain continuous word embeddings from a unlabelled text. It was published by Mikolov et al. (2013a) and Google holds the patent for the algorithm. However, the original code, written in C programming language is open source and free to use. Multiple implementations exist for a wide variety of toolkits and programming languages.

Word2Vec introduces two different models to compute dense vectors for each word according to its context words. There are two variants, Continuous Bag of Words (CBOW) and Skip-grams.

CBOW and Skip-grams are opposed models in the sense that CBOW builds its word embedding model trying to predict the current word from its context words. On the contrary, Skip-grams builds the model of a word trying to predict its context words. Figure 2.3, borrowed from Mikolov et al. (2013a), shows both variants. Word vectors obtained using Word2Vec has

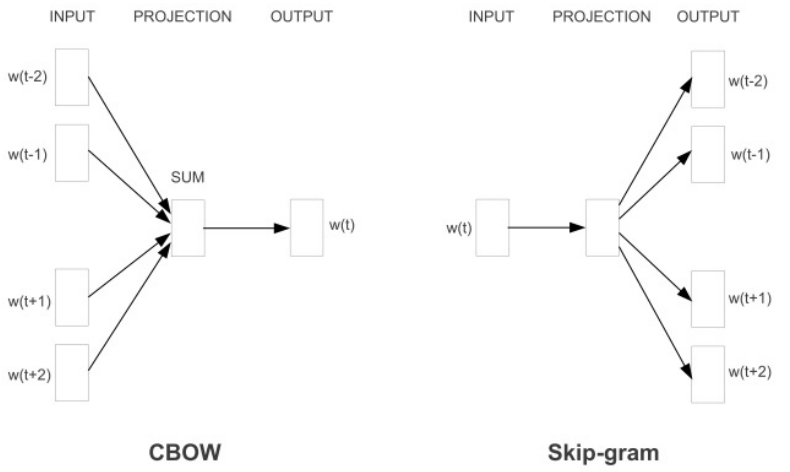


Figure 2.3: Word2Vec variants: Continuous Bag of Words (CBOW) and Skip-grams. Image borrowed from Mikolov et al. (2013a).

shown interesting properties and linguistic regularities (Mikolov et al., 2013c) in tasks like word similarity or word analogies.

Furthermore, GloVe (Pennington et al., 2014) is another popular implementation of an algorithm to compute general purpose continuous word embeddings. The source code is also open source. Its name stands for Global Vectors because GloVe combines both global context (capturing global document statistics from the corpus) and local context using context windows around the target word like Word2Vec does. This combination provides additional information to the model, resulting in vectors that outperform other representation when evaluated in several Natural Language Processing tasks.

There are further attempts of improving the information carried by word embeddings. Instead of using only a plain text corpus, word embedding generation is combined with knowledge databases and monolingual or multilingual dictionaries.

Goikoetxea et al. (2015) explore the combination of neural network based language models with random walks over the structure of knowledge bases like WordNet (Fellbaum, 1998). The resulting word representations achieve better results evaluated on word relatedness and word similarity tasks.

Chen et al. (2015) also make use of WordNet to improve words representation. First, they learn word sense embeddings from WordNet glosses using

convolutional neural networks and then they use the learned sense embeddings to generate distributed representations of word senses.

Ammar et al. (2016) introduce a method that uses dictionaries and monolingual data to obtain multilingual word embeddings without the need of parallel data. Multilingual word embeddings are a valuable resource for machine translation and other cross-lingual NLP tasks.

2.6.2 Word embeddings for sentiment analysis

Word embeddings are also used in the context of sentiment analysis. Word embeddings can be trained exploiting sentiment labelled data, adapting the function to predict the sentiment label so the learned model captures sentiment-related regularities. In addition, the correlation between a specific set of words can be used to fine-tune word embeddings for specific tasks through transformation functions that project word embeddings into a different vector space.

Tang et al. (2014a) present a method to learn word embeddings in the context of sentiment classification for Twitter. They call it Sentiment Specific Word Embedding (SSWE). The key difference is that instead of computing the word embedding from unlabelled data, they incorporate supervision in the form of sentiment polarity coming from the sentiment label assigned to tweets. In order to obtain a large dataset of tweets with their corresponding sentiment label, they carry out a distant supervision labelling. For that purpose, they leverage positive and negative emoticons, happy or unhappy smiles like *:)* or *:(* , as sentiment indicators to automatically label tweet polarity. The resulting embeddings outperform other general embeddings like Word2Vec when used for sentiment classification.

Cardoso and Roy (2016) present a method to obtain a list of words with their sentiment combining continuous word embeddings and a supervised training of a Multilayer Perceptron (MLP) algorithm using sentiment annotated words. The words are first mapped into a continuous vector space, and the MLP is trained to predict the sentiment polarity of the words.

Rothe et al. (2016) introduce DENSIFIER, a transformation method to map regular embeddings obtained with another method to a new ultradense subspace (of smaller dimensionality). DENSIFIER learns an orthogonal transformation that takes a word embedding of a certain size with the

encoded information scattered in some or all of the vector elements, and obtains a small vector representation, with the information relative to a certain task concentrated in those few dimensions. In the case of sentiment polarity, a word embedding of N positions in which the polarity information is not explicitly represented by any of the dimensions, it can be transformed into a one-dimension vector containing the polarity information in the value of its only dimension. In order to learn this transformation, DENSIFIER needs two lists of words of opposite polarity. The transformation parameters are trained to maximise the distance among words from opposite lists and to minimise the distance among words from the same list. They use several existing sentiment lexicons for the task.

Hamilton et al. (2016) follow a similar idea of using a set of seed words to tune the word embeddings to create domain-specific word embeddings but using a different approach. They combine word embeddings with a label propagation method constructing a graph and performing random walks. The graph edges are created between words that are close from a semantic perspective (measured by cosine similarity of the involved word vectors in the embedding space). The edges are also weighted according to this semantic distance notion. Using a set of polarity seeds to label the initial words in the graph, a set of random walks propagate each corresponding polarity among the connected words.

Xiong (2016) proposes increasing the granularity of the sentiment values when calculating sentiment-aware word embeddings. Instead of assuming that every word within a tweet shares the same sentiment polarity value than the tweet itself, they add an additional word-level polarity value. They include it in the supervised calculation of the embeddings, using a sentiment lexicon to set the polarity labels for individual words during the supervised training.

2.7 Topic modelling

Topic modelling is the name that receives the set of statistical techniques focused on discovering the so-called *topics* in a collection of documents. Topic modelling is a useful tool for text mining that helps to find hidden semantic structures in documents of a corpus.

These modelled topics can be described as a cluster of words that recurrently co-occur together. This frequent co-occurrence usually happens because those words belong to the same topic from a semantic point of view.

Topic modelling approaches build these topics from the documents of a corpus, and at the same time estimate the topic composition of each document. This helps to sort and to aggregate documents according to the information provided by the discovered topics.

Probabilistic topic models are useful tools for the unsupervised analysis of text, providing both a predictive model for new unseen text and a latent topic representation of the modelled corpus.

2.7.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most popular topic models, being also one of the simplest. A lot of different variations and extensions have been coined during the last fifteen years (Reed, 2012).

LDA provides a generative model that describes how the documents in a dataset were generated. Of course, this generative hypothesis is just a simplification of the reality, but this simplification helps to understand the semantic structure underlying the modelled documents.

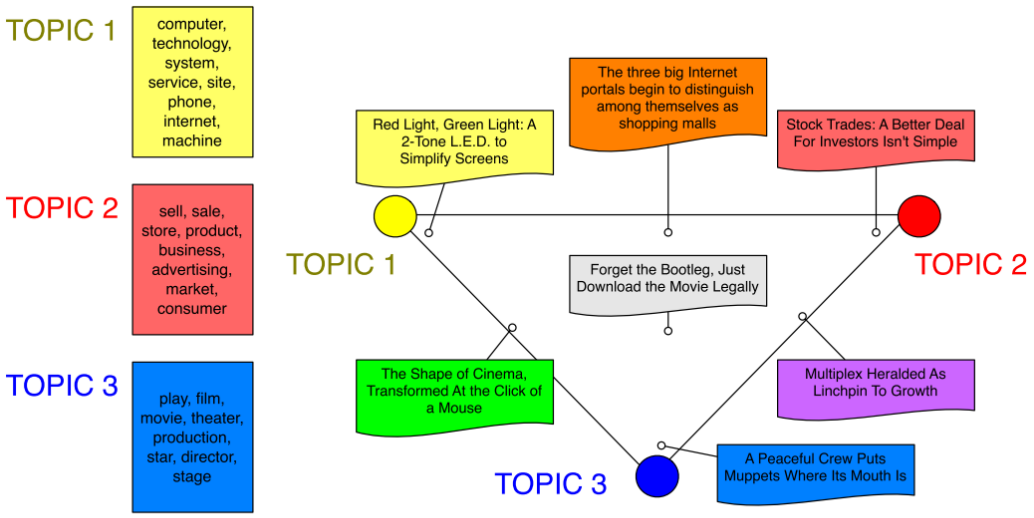
In summary, LDA offers a statistically plausible explanation of the words observed in the documents of a dataset.

Figure 2.4 shows a visual example of modelled topics and documents. The topic inspection provides information about the semantic themes treated in the documents of the dataset, providing a valuable insight into large collections of documents. Documents themselves can be aggregated according to their resulting topic distributions.

2.7.1.1 Generative model

Let D be a collection of documents. A document in this context will a collection of words from a fixed vocabulary. Let V be that vocabulary.

Let K be the number of latent topics that hypothetically are represented in the documents contained in D. K is a number that may vary according to the necessities, chosen by hand or following different heuristics to find



(a) Topics (b) Document Assignments to Topics

Figure 2.4: Visual example of topics (word distributions) and documents (topic distributions). Image borrowed from (Chang et al., 2009).

an optimal number for the each modelled dataset. Each topic $t \in T$ is a multinomial probability distribution over the vocabulary V.

The generative hypothesis modelled by LDA can be formalised as follows:

For each topic $t \in \{1..T\}$:
    sample $\phi_t \sim \text{Dirichlet}(\beta)$
For each document $d \in \{d_1..d_M\}$:
    sample $\theta_d \sim \text{Dirichlet}(\alpha)$
    For each word $w \in \{w_{d,1}..w_{d,N}\}$:
        draw $z_{d,n} \sim \text{Multinomial}(\theta_d)$
        sample $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$

Where ϕ_t is the multinomial word distribution corresponding to topic t, and θ_d is the multinomial topic distribution for document d. Both distributions are sampled from Dirichlet distributions. The Dirichlet distribution is a family of continuous multivariate probability distributions parameterised

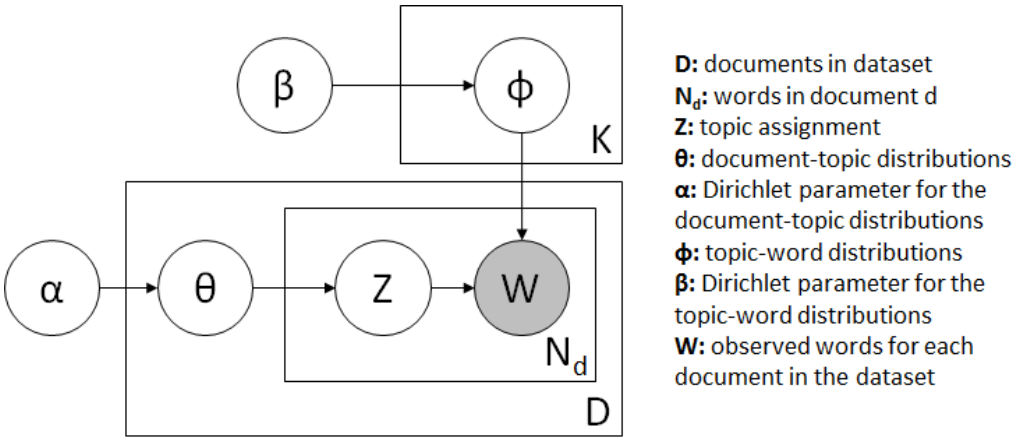


Figure 2.5: LDA model represented in plate notation.

by a vector of positive reals. It is a multivariate generalisation of the beta distribution.

In the LDA, α and β are the vectors of positive reals that control the shape of each of the Dirichlet distributions involved in the process. In LDA all the values of those vectors are equal, so initially, all the topics and word distributions are equally probable. Because of that in the literature about LDA, α and β are treated as if they were scalar values.

Figure 2.5 shows the dependencies among the involved variables in plate notation. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. White circles represent latent variables while grey circles (in this case only the words) are the observed variables.

2.7.1.2 Model inference

Latent variables try to explain the observations. The value of these variables has to be inferred using some statistical inference algorithm. The core inferential problem that LDA is about determining the posterior distribution of the latent variables given the observed documents:

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{2.1}$$

This distribution is intractable because the denominator $p(w|\alpha, \beta)$ cannot be computed. In order to workaround this problem there are some approximate inference techniques available that can be applied, for example, variational inference (Blei et al., 2003) or Gibbs Sampling (Griffiths and Steyvers, 2004).

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) based algorithm. MCMC algorithms generate samples from the posterior distribution, constructing a Markov-chain that converges in the target posterior distribution. This means that after a certain number of iterations sampling from the distribution, it should converge to be close to the posterior distribution of interest. Gibbs sampling needs a reasonable amount of iterations to achieve convergence, which means that it may require a decent amount of computation depending on the number of variables and complexity of the model. But since Gibbs sampling is easier to implement and it is used in many topic modelling approaches to estimate the posterior distribution of the model variables. Variational inference requires less computation time and resources but it is more difficult to implement.

2.7.2 Extensions to LDA

LDA has the advantage of being a very flexible and extensible model. Adding latent variables to the model, changing the generative hypothesis or adjusting the model hyper-priors to inject information coming from different sources are some of the ways of adapting LDA to new tasks.

There are many of such adaptations in the literature. Some of them aim at modelling more specific facets of documents or they aim at obtaining more coherent and cohesive topics. In order to achieve such objective, some information is injected in the process, for example altering the probability of some words appearing together in the same topic. This is done by manually defining sets of words, or using some existing resource as the source of information.

Moreover, other proposed models are focused on sentiment analysis, adding the notion of *sentiment* to the topic modelling process. We describe several of the most relevant approaches.

2.7.2.1 Improving topic semantic coherence

LDA is a straightforward and powerful approach to model and to discover hidden topic structures in textual data. From a statistical point of view topics are distributions of words over a vocabulary, but from a human perspective topics represent a semantic theme of some sort. Vanilla LDA often generates topics that, despite being statistically coherent, present semantic inconsistencies or are difficult to interpret by a human. Different approaches and extensions aim at improving this fact.

Mcauliffe and Blei (2008) introduced supervised latent Dirichlet allocation (sLDA). The idea behind this is that unsupervised topics models like the original LDA do not capture certain *responses*. By *response* the authors refer to the particular information structure to be modelled by the algorithm, for example, the sentiment structure versus the genre structure for movie reviews. In sLDA the desired response variable (e.g. movie rating) is associated to each document. Then documents and response variables are jointly modelled, so the resulting topics will be capable of predicting the modelled response variable. However, this supervised variant requires those response variables to be obtained from somewhere, for example from manual labelling of the documents.

Andrzejewski et al. (2009) present Dirichlet Forest prior for Latent Dirichlet Allocation (DF-LDA). The aim of this method is to improve the semantic topic coherence, preventing some words co-occurring in the same topic and encouraging other words to belong to the same topics. In order to achieve it, they encode two word-sets, *must-link* and *cannot-link*, for words that should share the same topic and words that should not share the same topic. The information contained in these word sets are encoded using Dirichlet tree distributions (Dennis III, 1991) as priors that replace the original Dirichlet priors used in the basic LDA. By changing the words inside *must-link* and *cannot-link* sets the user can inject domain knowledge to generate topics less prone to contain semantically inconsistent words.

In order to obtain more fine-grained topics, Kim et al. (2013) propose a hierarchical topic model (HASM) to discover hierarchical relations in topics. HASM model is a tree that recursively models the topic distributions. It uses the so-called *Chinese Restaurant Process* (CRP) (Griffiths and Tenenbaum, 2004) to build the hierarchical structure and to compute the model parameters. They apply to a dataset of digital devices to capture coarse-grained

topics like *performance* which are further decomposed in the hierarchy as topics containing words about *cpu* or *graphics*.

More recently continuous word embeddings (see section 2.6) are being combined with topic modelling as an additional source of information during the stochastic sampling process. Word embeddings computed on unstructured text have demonstrated to be powerful tools to capture semantic regularities.

Das et al. (2015) propose Gaussian Latent Dirichlet Allocation (GLDA) a topic model extended to exploit these semantic regularities coming from word embeddings. GLDA replaces the basic LDA categorical topics distributions with multivariate Gaussian distributions on the word embedding space. The idea behind this is to encourage the model to cluster semantically related words incorporating the notion of semantic relation coming from the word embeddings. The authors use English Wikipedia as the general corpus to compute word embeddings and evaluate the approach on news domain.

Nguyen et al. (2015) follow a similar idea of improving topic modelling via information from word vector representations to the process. They present the Latent Feature LDA (LF-LDA) which is an extension of the original LDA (Blei et al., 2003) and Latent Feature Dirichlet Multinomial Mixture (LF-DMM) which is an extension of a Dirichlet Multinomial Mixture (DMM) model (Nigam et al., 2000). They evaluate the improvement of the base LDA measuring the topic coherence and the normalised pointwise mutual information (NPMI) (Lau et al., 2014).

Under a similar intuition, Moody (2016) presents *lda2vec*. In this case, the proposed model learns the word embeddings jointly with the latent document mixtures of topics, producing unsupervised and interpretable document representations.

In spite of generating more easily interpretable topics, most of the topic modelling approaches generate anonymous topics, that are represented as a list of words. It is usually an end-user task to interpret each topic assigning a meaningful name to each of them if necessary. Bhatia et al. (2016) propose a method to help in this task. They propose an automatic labelling of topics using neural embeddings for documents and Wikipedia document titles as label candidates. First topic label candidates are generated based on English Wikipedia then a topic label ranking is learnt using a combination of several

algorithms. They evaluate the method on a custom dataset of documents with manually assigned labels.

2.7.2.2 Topic models for sentiment analysis

Topic models have been used in the context of sentiment analysis. A topic can be interpreted in different ways depending on what is the topic modelling algorithm objective. One of such interpretations is taking topics as words distributions across *sentiment polarities*. Some common strategies are used to achieve this objective: the addition of extra latent variables and dependencies among them to model other facets of the documents (e.g. the sentiment) and the injection some apriori information in the inference process. Sentiment modelling is usually combined with the improved topic modelling techniques to better capture the domain aspects that are being evaluated.

Brody and Elhadad (2010) present an early use topic modelling in the context of sentiment analysis, LocalLDA (LocLDA). LocLDA is based on the standard implementation of LDA, treating each individual sentence as a document in the topic modelling process. This is a typical approach when modelling customer reviews for sentiment analysis, because each review may refer to several domain aspects (i.e. topics in the topic modelling context). In order to increase the granularity of the topic modelling for the customer reviews has to be split, and sentence boundaries are a reasonable boundary. In the original LDA, the number of topics to model is a parameter that must be set by the user. LocLDA uses a cluster validation method to check a different number of topics and choose the most consistent according to a clustering consistency function (Niu et al., 2007). Once the optimal number of topics is modelled, they are manually interpreted and mapped to a domain aspect. LocLDA does not model sentiment directly, it is done using a non-LDA bootstrapping approach.

Zhao et al. (2010a) focus their attention on the aspect terms and opinion word separation during the topic modelling process, introducing MaxEnt-LDA (ME-LDA). The authors argue that topics are distributions of words, mixing all kind of words and sometimes obscuring the meaningful domain aspects. In sentiment analysis there are specific roles for some words, like aspect terms (i.e. words that explicitly refer to a certain domain aspect) and opinion words (i.e. word that bear a sentiment polarity value). Separating

these two word-types could be advantageous to ease the topic content interpretation or to improve the sentiment classification. ME-LDA extends the base LDA model with additional latent variables to distinguish three kind of words: *aspect terms*, *opinion words* and *other*. For that purpose, they train a Maximum Entropy classifier on a small dataset manually labelled with the type of each word. They use Part-of-Speech tags as features for the classifier. Once the MaxEnt classifier is trained it is incorporated into the topic modelling process. The result of the process is several word distributions per topic, one for each word type.

Mukherjee and Liu (2012) further develop the idea of separating aspect terms and opinion words within the topic modelling process. They propose a MaxEnt Seeded Aspect Sentiment model (ME-SAS). This topic model elaborates over the DF-LDA (Andrzejewski et al., 2009) and ME-LDA from (Zhao et al., 2010a). On the one hand, ME-SAS include seeds sets that group some aspect-related words for the domain of interest. These seed sets are created by a domain expert with the words that should go together in the same topic (e.g. *bed*, *pillow* and *linens* in a hotel reviews domain), similar to the *must-link* set from DF-LDA. Besides, they reuse the MaxEnt classifier from ME-LDA to separate aspect terms from opinion words, but instead of manually labelling a dataset to train the classifier they use a sentiment lexicon from (Hu and Liu, 2004) to automatically generate training data.

Lin et al. (2011) extend LDA to model sentiment. They propose Joint Sentiment-Topic model (JST) and its equivalent Reverse-JST. In the JST there are extra latent variables to model the sentiment. Sentiment labels are associated with documents, topics depend on sentiment labels, and finally, words are associated with both sentiment labels and topics. The Reverse-JST makes an alternative assumption, where the sentiment labels are dependent on topics. The sentiment polarity estimation during the topic modelling process is biased using MPQA lexicon (Deng and Wiebe, 2015) to modify the priors for some words of the vocabulary. The topic modelling process results in a set of topics with separated distributions of positive and negative words.

Jo and Oh (2011) follow a similar idea introducing the Aspect and Sentiment Unification Model (ASUM). ASUM aims at discovering pairs of aspects and sentiments. It also uses a set of predefined polarity words to bias the modelling of the sentiment. In particular, they use PARADIGM words from (Turney and Littman, 2003), and an extended version, PARADIGM+, containing additional affective words. Again, the outcome consists of several

topics containing two word-distributions, one for positive words and other for negative words.

Kim et al. (2013) follow a similar idea in their HASM. At the same time that HASM infers a hierarchical structure for the topics, it models the polarity of the documents relying on the same PARADIGM+ words used for ASUM. They perform additional pre-preprocessing steps like splitting sentences when they contain conjunctions like *but*, *however* or *yet*.

Alam et al. (2016) propose an extension over previous systems to relate the aspects and sentiment to windows of words instead of full sentences. They call it Joint Multi-grain Topic Sentiment (JMTS). In order to achieve this change of boundary, JMTS introduces an additional sentiment layer in the model. To capture the sentiment information in the model JMTS uses asymmetric priors too.

2.8 Conclusions

In summary, sentiment analysis and opinion mining has attracted a lot of attention from the computational linguistics community and the industry. This attention is well motivated by the growing need of analysing and exploiting the vast amount of content generated every day that involves opinions.

Current literature covers wide range of techniques focused on different facets and challenges of the sentiment analysis. There are a lot of open challenges in the way towards a sentiment analysis system that achieves human-like performance. Besides, much of the effort has been historically focused on English content and some of the existing methods or techniques are not easily portable to other languages or domains.

An ideal sentiment analysis system should deal with all the nuances of text analysis, detect and aggregate opinions by theme or product, deal with negations and other modifiers, understand the irony and sarcasm, etc. But also, such an ideal system should be multilingual, multidomain, unsupervised or weakly-supervised, and combine whatever other characteristics that would make it usable in the wider possible set of situations. These last considerations are part of the main research goals of this thesis.

MULTILINGUAL OPINION MINING

CHAPTER 3

A framework for weakly supervised opinion mining

This chapter presents a summary of the research and development carried out in this thesis and explains some concepts and definitions used in the following chapters. The overall objectives of this thesis are to explore methods to generate resources and tools useful in the context of sentiment analysis (i.e. customer reviews evaluating products or services) and to obtain a system capable of performing Aspect Based Sentiment Analysis (ABSA) in a way easy to port to other languages and domains. In order to achieve this language and domain portability, the explored approaches use the minimum possible supervision or language dependent tools and resources. The content of the thesis is arranged in incremental steps towards an almost unsupervised ABSA system, following the chronological order in which each of the presented parts was explored and developed.

3.1 Domain aspects, aspect-terms and opinion-words

In this thesis, we deal mainly with customer reviews. Customer reviews are short and subjective pieces of text, focused on reviewing and evaluating a

Example of domain aspects for different domains				
Restaurants	**Hotels**	**Films**	**Digital Cameras**	**Laptops**
food	rooms	plot	image quality	performance
service	staff	acting	size	screen
ambience	location	special effects	battery	battery
price	restaurant	music	weight	design

Table 3.1: Examples of typical domains aspects for several different domains.

particular entity, for example, a point of interest, a product or a service. Customer reviews contain opinions about different aspects of the evaluated entity. The analysis of an opinion involves the detection and classification of several elements.

First, each customer review and its associated opinions pertain to a particular *domain*, which influences the meaning and intended interpretation of the words it contains. In this context, by *domain* we refer to the overall set of items being reviewed in a subjective piece of text (i.e. in a customer review), for example, hotels, films, computer, digital cameras or smartphones. Each domain differs from the rest in which features are subject to evaluation by customers, and in the vocabulary used to express satisfaction or dissatisfaction about them. These features are called *domain aspects* (also referred as domain topics or domain categories). Table 3.1 shows some examples of typical aspects reviewed for several different domains.

Domain aspects are referred in a text using a certain set of words called aspect terms (also known as opinion targets).

Aspect terms are words, or groups of words, that explicitly refer to a domain aspect of the product or service under evaluation. For example, in a corpus of restaurant reviews we can find aspect terms like *waitress*, *decoration* or *cheese burger*. Each of those words refers to a coarser domain aspect: service, ambience, food. Aspect terms are domain dependent. Each domain has its own set of domain aspects and aspect terms (e.g. restaurant reviews speaking about food types vs. electronic devices reviews evaluating the battery life). Table 3.2 shows an example of some domain aspects and aspect terms.

Moreover, opinion words are terms or expressions that bear the sentiment towards the evaluated element. Opinion words imply some degree of

Restaurant reviews domain					
Service		**Food**		**Ambiance**	
ATs	**OWs**	**ATs**	**OWs**	**ATs**	**OWs**
waitress	attentive	cheese burger	tasty	decoration	cool
waiter	rude	soup	yummy	atmosphere	nice
owner	fast	rice salad	tasteless	music	dark
staff	polite	chicken	awful	light	dated

Table 3.2: Examples of aspect terms (ATs) and opinion words (OWs) for several domain aspects related to customer reviews about restaurants.

sentiment polarity: positive, negative or neutral.

In customer reviews, aspect terms usually appear in combination with opinion words. Aspect Based Sentiment Analysis (ABSA) systems try to identify the sentiment, denoted by opinion words, towards each individual domain aspect, denoted by aspect terms. For example:

The waiter was rude. $\rightarrow$ {service : negative}

The burgers are amazing. $\rightarrow$ {food : positive}

The sentiment implied by an opinion word is also domain dependent. It may vary across domains, or even across aspects of the same domain. For example, the opinion word *big* can bear a positive sentiment in the context of a restaurant review, as in *"they serve really big burgers"*, but when used in the context of a hotel review, as in *"there was a big noise during the night"*, the implied sentiment becomes negative. That is one of the reasons because the domain adaptation is so important to accurately understand the sentiment present on a piece of text.

3.2 Objectives of this thesis

Taking into account the definitions provided in the previous section, the objectives of this thesis can be enumerated as follows:

- Review of the state of the art related to sentiment analysis, including its relevance in the nowadays digital society, and relevant challenges and approaches related to sentiment analysis. This objective has been undertaken in chapter 2.
- Explore methods to generate sentiment analysis related resources, like lists of domain aspect terms and opinion words and the calculation of a sentiment polarity value for opinion words, requiring the less possible amount of supervision and resources. These objectives are part of chapters 4 and 5.
- Build a system capable of performing Aspect Based Sentiment Analysis using the less possible amount of supervision, language tools and resources. The objective is to obtain a system that estimates the domain aspects and sentiment polarities of customer reviews. The resulting system, described at chapter 6, is aimed at working for many languages and domains without requiring a big adaptation effort.

Figure 3.1 shows a diagram illustrating the methods explored and developed in each chapter and how they are related together in a weakly supervised Aspect Based Sentiment Analysis framework. The relation is also chronological with regard to the order in which each method was developed.

The process has involved several steps. The first step, described in chapter 4 aims at bootstrapping domain aspect terms and opinion words using a graph based algorithm. The resulting system only needs few initial seed words to work but requires syntactic dependencies as part of the process, hindering its application for languages for which there is no reliable dependency-parser available. The next step, described at chapter 5 is focused on calculating a polarity value using only two seed words, easy to adapt to any language or domain. Furthermore, the chapter explores an approach to separate opinion words from aspect terms just by adding an extra seed word. Finally, chapter 6 describes how some of the methods explored in the previous chapter are

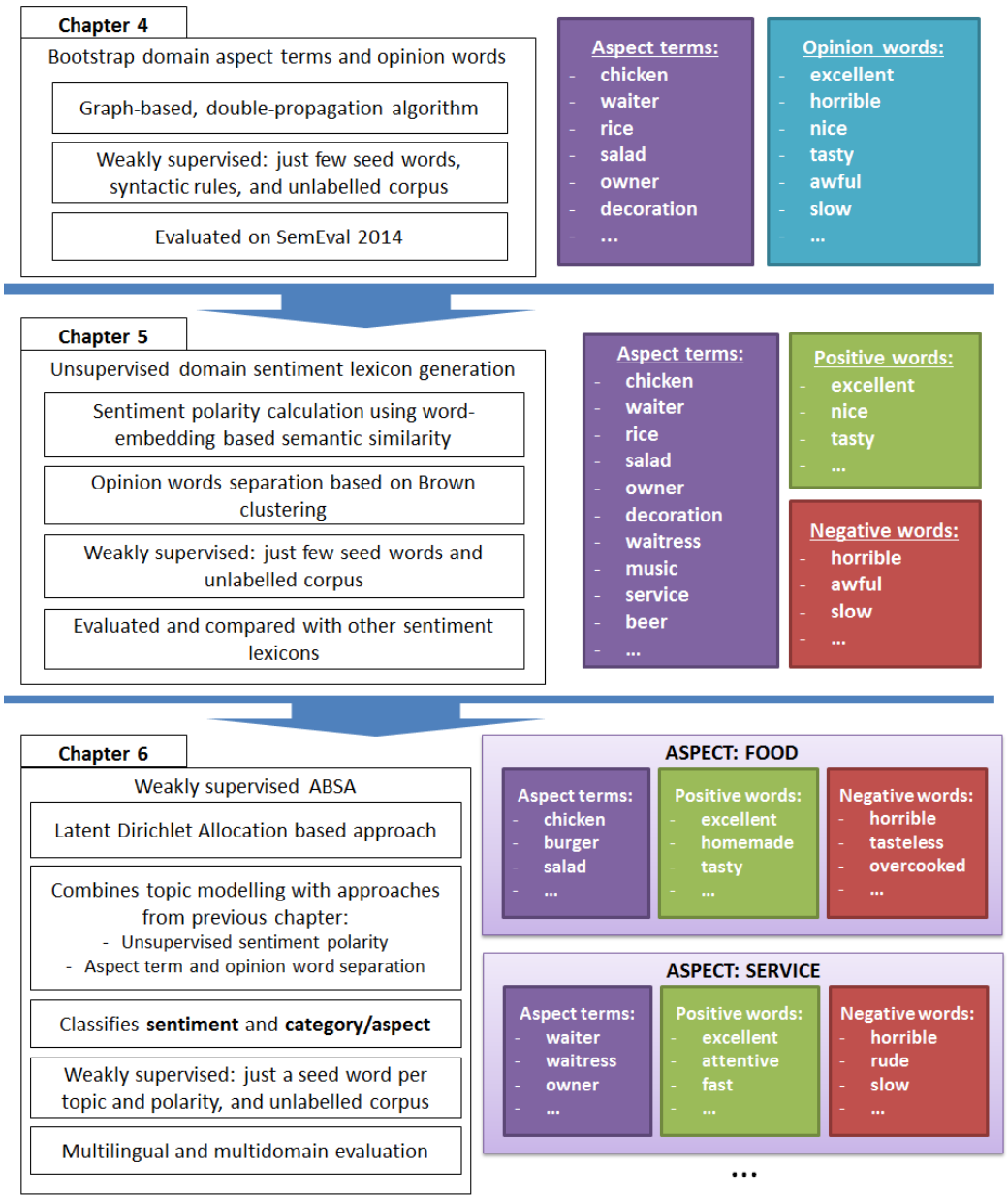


Figure 3.1: Chapters content and relation to sentiment analysis tasks.

combined within a topic modelling approach. The resulting system can model a text corpus separating words by aspect and polarity, and performing Aspect Based Sentiment Analysis (ABSA) with the minimum possible supervision.

3.3 Bootstrapping aspect-terms and opinion-words

We start by exploring an approach to bootstrap domain terminology. In particular, we are interested in obtaining a set of potential aspect terms and opinion words for a particular domain, bootstrapping them from an unlabelled domain corpus and few seed words. This approach is based on the work described at (Qiu et al., 2011).

In this approach a small set of seed words, for example, a known domain aspect term and a known opinion word, provide the starting point to apply a set of word extraction rules over a domain corpus. These extraction rules are based on the syntactic roles and relations of the words in each sentence. The rules propagate the initial set of seed words to obtain more candidates. Each time a word is extracted it is added to the set of seed words. The process iterates over the corpus, using the extracted words to further apply the extraction rules in a recurrent manner. Since some rules extract opinion words from a seed aspect term and vice versa, the approach receives the name of double-propagation extraction.

Instead of simply gathering the bootstrapped words as in the original double-propagation approach, in this proposed extension a graph is built during the double-propagation process using the extracted word as nodes and the extraction rules as edges. Each edge is weighted by the number of times a rule was applied to two words in the corpus, obtaining a weighted graph composed by the domain words that are potential candidates of being aspect terms or opinion words.

Once the graph is built, a graph-based algorithm is used to score each node in the graph. We use the well-known PageRank, but other random walk algorithms could be applied. This process assigns a relevance score to each node. This score is then used to rank the extracted terms. Terms are sorted by score, being the ones with higher score those with higher confidence. The resulting lists are cropped at some threshold, to keep only the term with

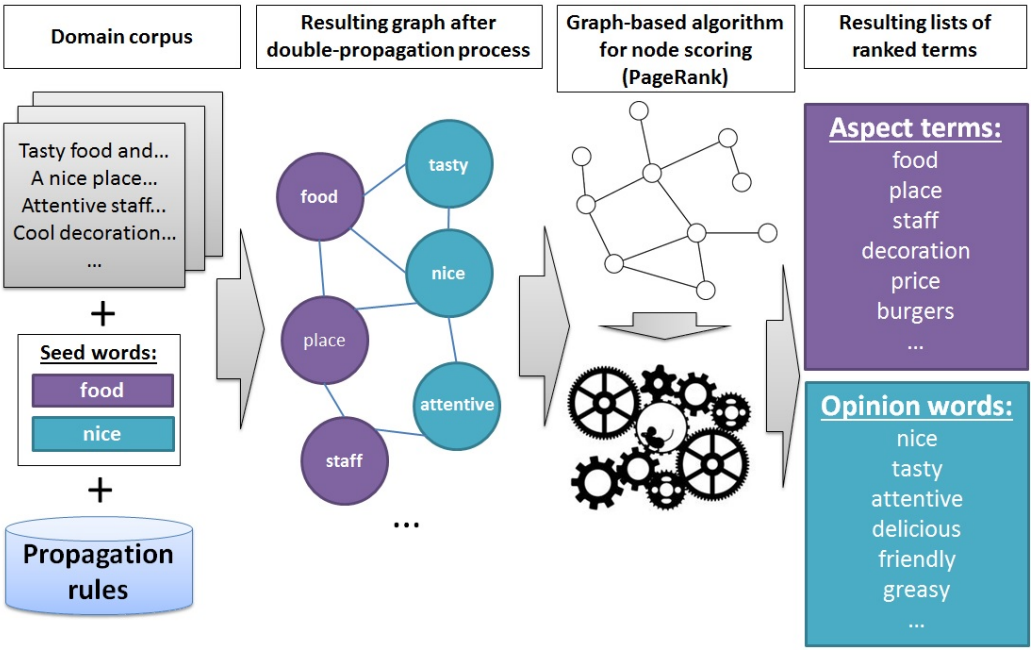


Figure 3.2: A domain corpus modelled as a graph to obtain ranked lists of aspect terms and opinion words.

top confidence. This method has been evaluated for two different domain, customer restaurant reviews and customer laptop reviews, by participating in the SemEval 2014 task 4 shared task.

Figure 3.2 shows a conceptual representation of the described approach, in which a domain corpus, in combination with seed words and propagation rules is turned into a graph relating words. This resulting graph is further processed to score each node, obtaining ranked lists of aspect terms and opinion words. The approach is described in more detail in chapter 4.

3.4 Bootstrapping polarity lexicons

A further step beyond bootstrapping aspect terms and opinion words for a given domain is to assign a sentiment polarity value to the words.

The sentiment polarity of a word indicates if that word expresses a positive or negative (or neutral) feeling towards the element that is being de-

Categorical sentiment values	Continuous sentiment values
excellent: very positive	excellent: +1.0
wonderful: very positive	wonderful: +0.8
good: positive	good: +0.4
bad: negative	bad: -0.5
awful: very negative	awful: -0.7
horrible: very negative	horrible: -1.0

Table 3.3: Example of categorical and continuous sentiment polarity values.

scribed or evaluated. The sentiment polarity can be encoded as a categorical value, for example *very positive*, *positive*, *neutral*, *negative* and *very negative*. It can be also encoded as a continuous value, for example, a numeric value ranging from -1.0 to +1.0, with values below zero indicating the degree of negativeness and values above zero indicating the degree of positiveness. Table 3.3 shows some examples of words with categorical and continuous sentiment polarity values.

A collection of words mapped to their corresponding sentiment polarity value is called a sentiment lexicon. Sentiment lexicons can be crafted manually by a human from scratch, which is time-consuming and hard to adapt to different domains and languages. Besides, sentiment lexicons can be bootstrapped from existing lexical or semantic resources, or directly from a corpus using different techniques.

We explore and experiment with the use of continuous word embeddings to obtain a sentiment polarity value for the words of a domain.

Continuous word embeddings are vector representations of words over a vocabulary. They are related to the field of distributional semantics. In summary, continuous word embeddings refer to any function that maps words from a fixed vocabulary to some vector space, $W : words \rightarrow \mathbb{R}^n$.

In the last few years, many word embedding calculation methods have appeared in the literature, being the Word2Vec (Mikolov et al., 2013a) one of the most popular. Depending on the particular algorithm used to calculate these vectors representations the resulting vectors show different properties.

Figure 3.3 shows an outline of the process. The proposed approach only requires two seed words, a very positive word and a very negative word, and uses word embeddings to calculate a sentiment polarity value for each word

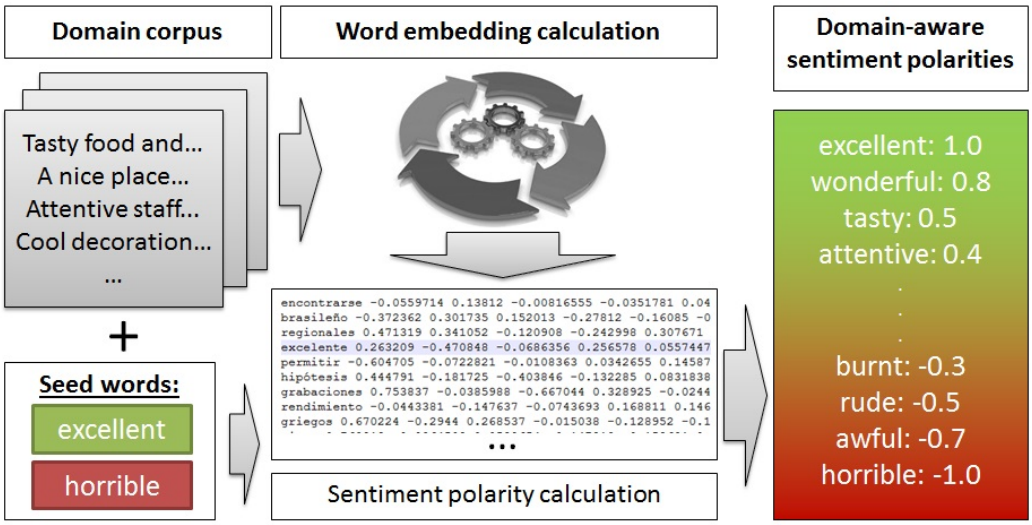


Figure 3.3: A method to calculate sentiment polarity values for domain words using word embeddings.

in the vocabulary. We experiment with different word embedding algorithms. We also compare this approach using the word embeddings calculated on a big general domain corpus (like a text dump from Wikipedia) vs. using word embeddings calculated on a smaller corpus of the target domain. We compare the performance to a variety of other well-known sentiment polarity lexicons.

3.4.1 An approach for unsupervised aspect term and opinion word separation

In chapter 5 we explore an algorithm to assign a sentiment polarity value to the words in the vocabulary. But not all words in the vocabulary bear a sentiment polarity. Aspect terms like *waiter*, *waitress* or *screen* do not carry sentiment information on their own. The sentiment polarity is expressed by the opinion words that accompany them, like *polite*, *attentive* or *awesome*.

Hence, in this chapter, we also explore a way of obtaining separated lists of aspect terms and opinion words without requiring additional resources or labelled data. Just by using an extra seed word (a representative aspect term of the target domain), we develop a method to bootstrap separated lists of aspect terms and opinion words. This, combined with the sentiment polarity

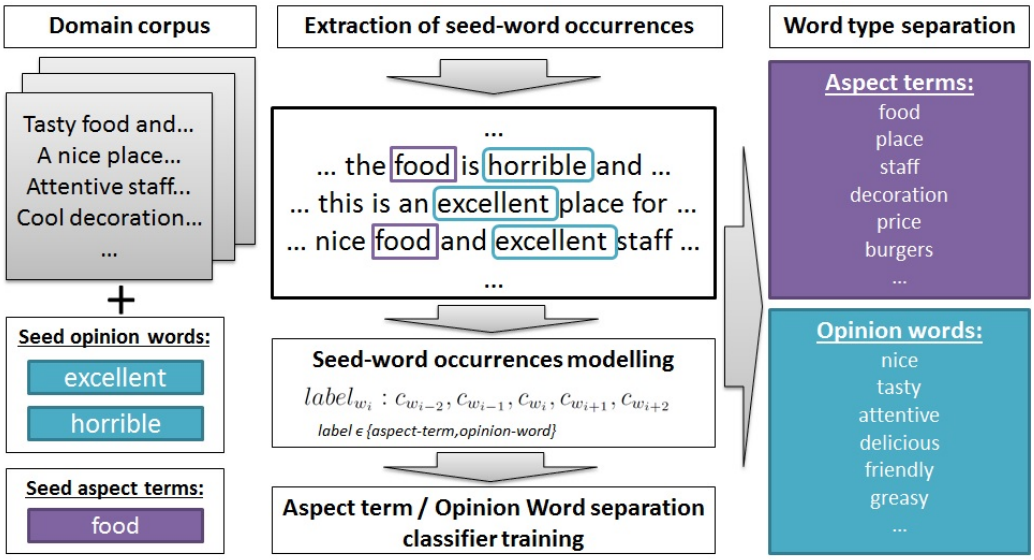


Figure 3.4: Separation of opinion words from aspect terms using only few seed words.

calculation provides a tool to bootstrap a sentiment lexicon from a corpus of any target domain and language.

The process is outlined at figure 3.4. First, Brown clusters (Brown et al., 1992) are computed for the domain corpus. Brown clustering is an unsupervised hierarchical clustering algorithm. Next, an opinion word seed set and aspect term seed set are needed. The same two seed words used for the polarity calculation can be reused as opinion word seeds. An additional seed word, a representative aspect term for the target domain, is used as aspect term seed. Then, the occurrences of the seed words in the domain corpus are extracted. Each seed word occurrence is transformed into a training instance labelled according to the set of seed words it belongs to (i.e. aspect term or opinion word). Each training instance is represented by the context words in a window around the word occurrence. Then the words in context are replaced by their associated Brown cluster. These training instances are them fed to a supervised classifier to learn a model to estimate when a word is likely and aspect term or an opinion word given its context.

Every word occurrence in the domain corpus is classified using this classification model, which outputs a probability of being from one of the two

classes. The intuition behind this idea is that word occurrences that are opinions words will tend to obtain a high probability of being classified as opinion words and vice versa. Words that are neither aspect terms or opinion words will fall somewhere in the middle. The aggregated probabilities for all the occurrences of the same word (i.e. every occurrence of the word *waiter*, every occurrence of the word *wonderful*, etc.) are averaged. The resulting values are used to rank the words as aspect terms or opinion words, pushing unwanted words to the lower part of the rank.

3.5 Weakly supervised ABSA

After exploring methods to bootstrap domain aspect terms and opinion words from a target domain corpus, and to assign a sentiment polarity value to words, we move to the next step. We combine the mentioned methods in a topic modelling approach. The objective of the proposed system is to classify each sentence of a customer review corpus into a set of predefined domain aspects, and, at the same time separate the aspect terms from the opinion words, and finally estimate the sentiment of the sentence.

The process is based on an extended Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model. LDA is a generative model that models documents as multinomial distributions of so-called topics. Topics are multinomial distributions of words over a fixed vocabulary. Topics can be interpreted as the categories from which each document is built-up, and they can be used for several kinds of tasks, like dimensionality reduction or unsupervised clustering. LDA variations can be used for topic and sentiment classification as in this case.

The proposed system includes additional hidden variables to model not only the topic distribution of documents but also the aspect-term, opinion-word and polarity distributions of words among topics. The hyper-parameters that govern how the words are distributed are asymmetric, biased according to word embedding similarities of the corpus and seed words. The separation of aspect-terms and opinion-words is embedded in the topic modelling process and done on a per-topic basis, governed by a bootstrapped classifier using the method described in chapter 5.

The only supervision required by the algorithm to work is the set of desired domain aspects. These domain aspects are chosen apriori according

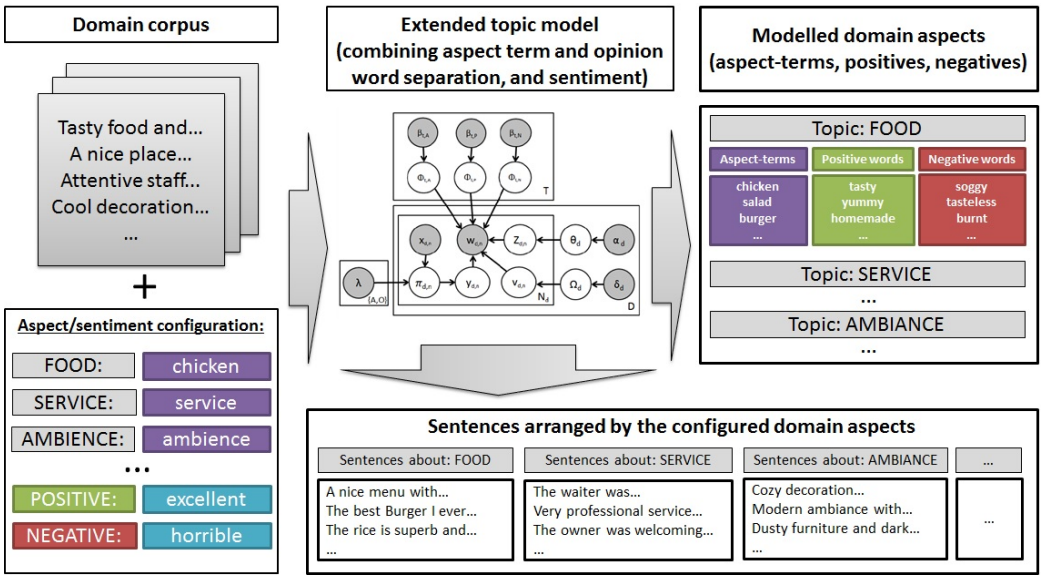


Figure 3.5: Extended topic modelling approach for almost unsupervised aspect based sentiment analysis

to the target domain and defined by one or more seed words. With this simple configuration, the topic modelling process is guided to fit a corpus into the set of defined domain aspects. Everything else is bootstrapped, calculated and modelled from an unlabelled domain corpus. The method can be applied to other domains or languages just by adapting the domain aspect definitions (i.e. the seed words for each domain aspect and polarity).

Figure 3.5 illustrates part of the process and the structure of the outcome. Chapter 6 describes the system in detail together with the experimental results. The experiments include results for several languages (English, Spanish, French, Dutch) and domains (restaurant reviews, hotel reviews, electronic devices reviews).

3.6 Conclusions

This chapter has defined some of the concepts that will be used along the rest of the chapters of this thesis, like *domain aspects*, *aspect-terms* or *opinion-words*. In addition, this chapter has described the structure and objective

of each individual chapter. The next chapters explore methods that generate different resources and outcomes related to sentiment analysis, like bootstrapping domain aspect-terms and opinion-words, calculating domain-aware sentiment polarity values or classifying reviews into a predefined set of domain aspects.

The key idea, and a self-imposed restriction, is to avoid relying on language or domain-based tools or resources. In particular, the explored methods avoid using manually labelled data for training. The aim is to assess the feasibility of performing sentiment analysis for different languages and/or domains without the need for specific adaptations for each language or domain. After exploring approaches to generate domain sentiment lexicons with almost no supervision, we propose a combination of almost unsupervised methods into a system capable of doing Aspect Based Sentiment Analysis. The resulting system can be used to model corpora of different languages and domains with a minimal adaptation effort.

CHAPTER 4

Domain aspect-terms and opinion-words extraction

In this chapter we describe a double-propagation method to bootstrap lists of candidate aspect terms and opinion words from unlabelled text corpora, starting from few seed words and expanding them iteratively following certain expansion rules. The resulting lists of terms are ranked by confidence using a graph based algorithm. The chapter is structured as follows. Section 4.1 introduces the work and the objective of the chapter. Section 4.2 describes the proposed method. Section 4.3 shows some experimental results for several domains. Finally, section 4.4 presents some concluding remarks.

4.1 Introduction

Opinion mining systems can be roughly classified into two types, supervised, and unsupervised or semi-supervised since some level of supervision is almost always required to guide or initialize most of the existing systems. Supervised systems require training data, which usually includes manually annotated data, in order to train a model that can *learn* how to label new unseen data. Supervised systems perform quite well, but they are usually hard to port to different domains or languages due to the cost of obtaining such manually annotated data.

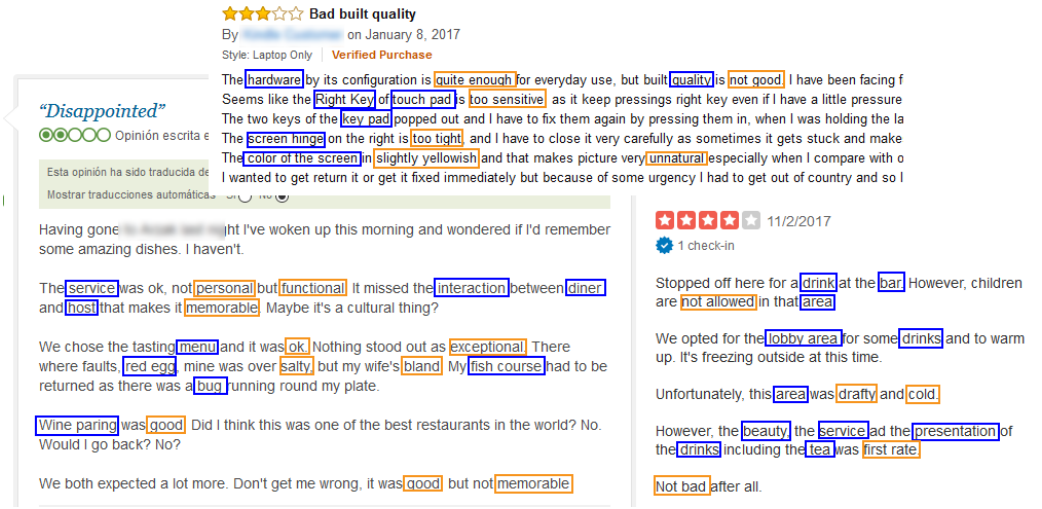


Figure 4.1: Example of on-line customer reviews from several sources and domains. Some aspect terms (blue) and opinion words (orange) have been manually highlighted for illustrative purposes.

Instead of directly relying on manually labelled data which is hard and expensive to obtain, we can try to leverage the unstructured/semi-structured content that is constantly generated over the Internet. Companies like TripAdvisor, Yelp or Amazon run websites that produce many customer-generated texts containing opinions. Figure 4.1 shows some example reviews from real websites. This kind of content is a valuable source of data for both customers and companies. Even if the data is not labelled with the information we may require, it can be used to discover interesting patterns and apply different algorithms to uncover valuable information.

A first thing that can be inferred from a corpus of unlabelled customer reviews is a list of aspect terms and opinion words used in the reviews of products or services. Obtaining these lists of terms is not enough to predict the attitude (i.e. sentiment) of the customer, but it is a first step towards gaining a better insight of what is being said by customers and may provide useful information for further processing.

In this chapter we introduce a double-propagation method to bootstrap lists of candidate aspect terms and opinion words from unlabelled text corpora, starting from few seed words and expanding them iteratively following certain expansion rules. The base double-propagation method is based on

(Qiu et al., 2011). But instead of extracting a plain list with all the bootstrapped term and pruning them with manually designed heuristics, we build a graph structure and use a graph algorithm to rank the resulting lists of terms.

We describe the SemEval 2014 task 4 about Aspect Based Sentiment Analysis (ABSA), in particular, the subtask that deals with aspect terms detection. We show results of applying the proposed approach on the SemEval 2014 datasets, composed by restaurant and laptop customer reviews and the result of our participation in the competition for the aspect terms detection subtask. In this competition, our approach was the only unsupervised system that did not make use of the provided labelled data for training (Pavlopoulos, 2014).

4.2 Bootstrapping aspect terms and opinion words

Our aim is to build a system that is capable of generating a list of potential aspect terms and opinion words for a new domain without any kind of adaptation or tuning. Such a list can be a useful resource to exploit in a more complex system aiming to perform Aspect Based Sentiment Analysis.

Aspect terms, also known as *opinion targets* in the literature, generally refer to parts of features of a given entity. For example, *wine list* and *menu* could be aspect terms in a text reviewing a restaurant, and *hard disk* and *battery life* could be aspect terms in a laptop review. Opinion words are those words that carry some sentiment related information, like *wonderful* or *attentive*. Obviously, each domain has its own set of aspect terms and opinion words, referring to different aspects, parts and features of the entities described in that domain. The only requirement to generate the list of aspect terms and opinion words for a new domain is a, preferably large, set of unlabelled documents or review describing entities of the domain and a few seed words. Our method combines some techniques already described in the literature with some modifications and additions.

4.2.1 Double propagation

We have adapted and extended the double-propagation technique described in (Qiu et al., 2009) and (Qiu et al., 2011). This method consists of using an initial seed list of aspect terms and opinion words and propagate them through a dataset using a set of propagation rules. The goal is to expand both the aspect term and opinion word sets.

Qiu et al. (2009) define opinion words as words that convey some positive or negative sentiment polarities. They only use nouns as aspect terms, and only adjectives can be opinion words. This is an important restriction that limits the recall of the process, but the double-propagation process is intended to extract only explicit aspects (i.e. aspects that are explicitly mentioned in the text, and not aspects implicitly derived from the context). The detection of implicit aspects (e.g. *"The phone fits in the pocket"* referring to the size) requires a different set of techniques and approaches that are described in alternative works in the literature (Fei et al., 2012; Hai et al., 2012).

During the propagation process, a set of propagation rules are applied to discover new terms (aspect terms or opinion words), and the initial aspect term and opinion word sets are expanded with each new discovery. The newly discovered words are also used to trigger the propagation rules, so in each loop of the process, additional words can be discovered. The process ends when no more words can be extracted. Because aspect terms are employed to discover new opinion words, and opinion words are employed to discover new aspect terms, the method receives the name of double-propagation.

The propagation is guided by some propagation rules. When the conditions of a rule are matched, the target word (aspect term or opinion word) is added to its corresponding set.

In the original double-propagation, all the bootstrapped terms are extracted, which leads to a large number of noisy terms (i.e. incorrect or undesired terms extracted by the rules). The authors try to alleviate this noise using certain heuristics and manually defined rules to prune the results. Instead of that, the variation proposed in this chapter uses the double-propagation process to build a graph structure. In this graph, words are modelled as the nodes. Each pair of nodes is connected by an edge if they are related to each other by a propagation rule. Once the complete graph is built, a graph algorithm is run to score each node. In particular, we use the well-known

Rule	Observations	Constraints	Action
R11	O → amod→W	W is a noun	W→T
R12	O→dobj→W1 ←subj←W2	W2 is a noun	W2→T
R21	T ← amod ← W	W is an adjective	W→O
R22	T → subj → W1 ← dobj ← W2	W2 is an adjective	W2→ O
R31	T → conj → W	W is a noun	W → T
R32	T → subj → has *gets* dobj ← W	W is a noun	W → T
R41	O → conj → W	W is an adjective	W→ O
R42	O → Dep1 → W1 ← Dep2 ← W2	Dep1==Dep2, W2 is an adjective	W2→ O

Table 4.1: Propagation rules applied during the double-propagation process.

PageRank (Page et al., 1999) to obtain a score for each node. Then a ranked list of terms is built sorting the words by their node score.

4.2.2 Propagation rules

The propagation rules are based on dependency relations and some part-of-speech restrictions. We have mainly followed the same rules detailed in (Qiu et al., 2011) with some minor modifications. The exact propagation rules used in this approach can be observed in the Table 4.1.

Some rules extract new aspect terms, and others extract new opinion words. In Table 4.1, T means *aspect term* (i.e. a word already in the aspect terms set) and O means *opinion word* (i.e. a word already in the opinion words set). W means *any word*. The dependency types used are *amod*, *dobj*, *subj* and *conj*, which stand for *adjectival modifier*, *direct object*, *subject* and *conjunction* respectively. Additional restrictions on the Part-Of-Speech (POS) of the words present in the rule, it is shown in the third column of the table. The last column indicates to which set (aspect terms or opinion words) the new word is added.

To obtain the dependency trees and word lemmas and POS tags, we use the Stanford NLP tools[1]. Our initial seed words are just *good* and *bad*, which are added to the initial opinion words set. The initial aspect terms set starts empty. This way the initial sets are not domain dependent, and we expect that, if the propagation rules are good enough, the propagation should obtain the same results after some iterations.

[1] `http://nlp.stanford.edu/software/lex-parser.shtml`

Each sentence in the dataset is analysed to obtain its dependency tree. Then the rules are checked. If a word and its dependency-related words trigger the rule, and the conditions hold, then the word indicated by the rule is added to the corresponding set (aspect terms or opinion words, depending on the rule). The process continues sentence by sentence adding words to both sets. When the process finishes processing sentences, if new words have been added to any of the two sets, the process starts again from the first sentence with the enriched sets. The process continues iterating over the dataset using the newly extracted words to trigger the propagation rules to extract additional words. The process ends when no additional words have been extracted during a full dataset pass.

4.2.3 Ranking the aspect terms

Although the double-propagation process populates both sets of domain aspect terms and domain opinion words, we focus our attention in the aspect terms set. Depending on the size and content of the employed dataset, the number of potential aspect terms will be quite large. In our case, the process generates many thousands of different potential aspect terms. Much of them are incorrect, or very unusual aspect terms (e.g. in the restaurant domain, a cooking recipe written in another language, a typo, etc.). Thus, the aspect terms need to be ranked, trying to keep the most important aspects on top, and pushing the less important ones to the long tail.

In order to rank the obtained aspect terms, we have modelled the double-propagation process as a graph population process. Each new aspect term or opinion word discovered by applying a propagation rule is added as a vertex to the graph. The rule used to extract the new word is added as an edge to the graph, connecting the originating word and the discovered word.

Figure 4.2 presents as an example a small part of the graph obtained by the double-propagation process. Each vertex representing a word maintains the count of how many times that word has appeared in the dataset, and also if it is an aspect term or an opinion word. A word is identified by its lemma and its POS tag. Every edge in the graph also maintains a count of how many times the same rule has been used to connect a pair of words. At the end of the double-propagation process, the generated graph contains some useful information: the frequency of appearance of each word in the dataset, the frequency of each propagation rule, the number of different words related

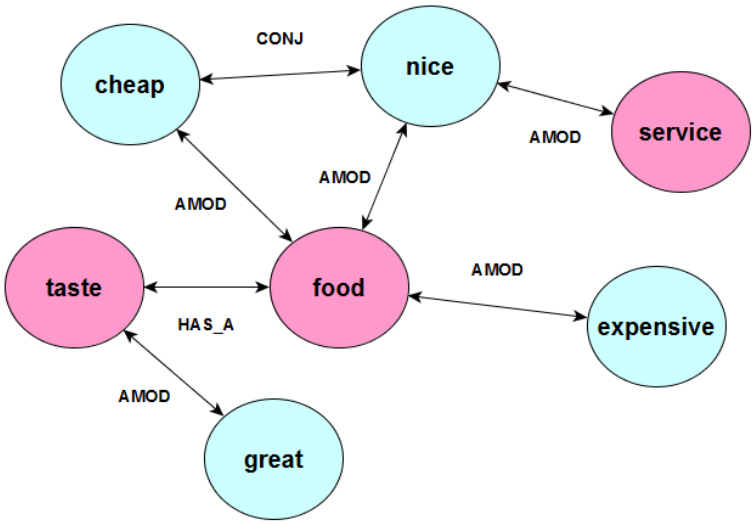


Figure 4.2: Example of a graph fragment constructed with the bootstrapped words and relations. Dark nodes are domain aspect terms and the light nodes are opinion words.

to a given word, etc. We have applied the well-known PageRank algorithm on the graph to score the vertices. To calculate the PageRank scores we have used the JUNG framework[2], a set of Java libraries to work with graphs. The value of the alpha parameter that represents the probability of a random jump to any node of the graph has been left at 0.15 (in the literature it is recommended an alpha value between 0.1 and 0.2).

The graph is treated as an undirected graph because the propagation rules represented by the graph edges can be interpreted in both directions (e.g. A modifies to B, or B is modified by A). The aspect terms are then ordered using their associated score, being the most relevant aspect term the one with the highest score.

4.2.4 Filtering undesired words

The double-propagation method always introduces many undesired words. Some of these undesired words appear very frequently and are combined

[2] `http://jung.sourceforge.net`

with a large number of words. So, they tend to also appear in high positions in the ranking.

Many of these words are easy to identify, and they are not likely to be useful aspect terms in any domain. Examples of these words are: *nothing*, *everything*, *thing*, *anyone*, *someone*, *somebody*, etc. They are extracted during the double-propagation process because they appear in common expressions like *It was a good thing*, *It is nothing special*, *I like everything*. The process also extracts other words, like *year*, *month*, *night*, and other time expressions. Also, some common words, like *boy*, *girl*, *husband* or *wife*.

The main reason behind these cases is that the input texts are customers reviews, and it is quite common to find anecdotes and personal comments like *"I saw a nice girl in the bar"*. It would be interesting to find an automatic method to safely remove all these words, valid for many domains. A TF-IDF weighting of the words could be a possible preprocessing to identify noisy content. In this case, we have chosen to add them to a customizable stop word list. The final list contains about one hundred words that are not likely to be aspect terms in any domain, like the ones mentioned above. The list has been crafted observing the most common unwanted words after running the system, and using intuition and common sense. Our purpose is not to tune a stop word list to work better any specific evaluation domain, so the same stopword list has been used during the evaluation for both restaurants and laptops domains.

4.2.5 Dealing with multiword terms

Many aspect terms are not just a single word, but compounds and multiword terms. For some domains, this is more critical than for others. For example, for laptop reviews domain the top ranked aspect term is *battery life* (see table 4.3). In particular, laptop reviews are part of an especially challenging domain due to the extensive use of technical vocabulary that usually combines several words (e.g. *hard disk drive*, *Intel i7 processor*, etc.). In order to improve the precision and the recall of the generated set of aspect terms, multiword aspect terms must be detected and included in the resulting sets. Next, we describe different approaches. They are rather simple approaches that try to increase the aspect term recall without adding too many incorrect terms.

4.2.5.1 Multiword terms from WordNet

One of the approaches included in the system exploits WordNet[3] (Fellbaum, 1998), and some simple rules. Each time a word is going to be processed during the double-propagation algorithm, the combination of the current word plus the next word is checked. If some conditions are satisfied then we treat both words as a single multiword term. The conditions are the following:

- If word n and word n+1 are nouns, and the combination is an entry in WordNet (or in Wikipedia, see below). E.g.: *battery life*
- If word n is an adjective and word n+1 is a noun, and the combination is an entry in WordNet. E.g.: *hot dog*, *happy hour*
- If word n is an adjective, word n+1 is a noun, and word n is a relational adjective in WordNet (lexical file 01). E.g.: *Thai food*, *Italian food*

4.2.5.2 Multiword terms from Wikipedia

In order to improve the coverage of the WordNet approach, we also check if a combination of two consecutive nouns appears as a Wikipedia article title. Wikipedia articles refer to real word concepts and entities, so if a combination of words is a title of a Wikipedia article it is very likely that this word combination is also meaningful for the domain under analysis (e.g. *DVD player*, *USB port*, *goat cheese*, *pepperoni pizza*). However, since Wikipedia contains many entries that are titles of films, books, songs, etc., that would lead to the inclusion of erroneous multiword expressions, for example *good time*. For this reason, we limit the lookup in Wikipedia titles just to a combination of nouns, avoiding combinations of adjective + noun. This gives a good balance between extended coverage and inclusion of incorrect aspect terms.

4.2.5.3 Multiword terms from simple patterns

In this case, we have limited the length of the multiword terms to just bigrams. But in some cases it is interesting to have word combinations of a bigger size. For that purpose, we have included some configurable patterns

[3] `http://wordnet.princeton.edu/`

to treat longer chains of words as a single aspect term. The patterns are very simple, being expressed with a simple syntax like *A of N*. It means that a known aspect term (represented by the uppercased A) followed by the word *of*, followed by a noun (represented by the uppercased N) must be processed as a single aspect term. Similar patterns would be *N of A*, *A with N*, *N with A*, etc. These patterns are useful to extract expressions like *chicken with onion*, or *glass of wine*. The complete list of used patterns is the following (note that the order matters, the longest go first to avoid partial matches):

- "N with N and A",
- "N with A and N",
- "A with N and N",
- "A with N",
- "N with A",
- "N of A and A",
- "N of A and N",
- "N of N and A",
- "N of A",
- "A and A",
- "A A A",
- "N A A",
- "A A"

4.3 Experiments and results

To evaluate the quality of the resulting aspect term lists, we have used our method to annotate the SemEval 2014 datasets of task 4, *Aspect Based Sentiment Analysis* which provides two datasets, one containing customer reviews about restaurants and another containing customer reviews about laptops. The datasets are composed of individual sentences. Each sentence contains annotated data about the aspect terms present in that sentence. The aspect terms are the span of characters inside the sentence that holds the mention to the aspect.

4.3.1 Used data

Using a web-scraping program we have extracted about 7,000 English reviews from a restaurant review website[4], and a similar amount of English reviews from a laptop review website[5]. We have not performed any kind of sampling or preprocessing on the extracted data, it has been extracted "as-is" from the list of entities (restaurants and laptops) available in the respective websites at the time of the scraping. The extracted reviews have been split into sentences using Stanford NLP tools and stored into an XML file. A subset of 25,000 sentences has been used to acquire the aspect term lists, combined with the already mentioned 3,000 sentences of the SemEval 2014 task 4 datasets. We have also used several thousand sentences from customer reviews about hotels, scrapped from on-line websites. SemEval 2014 does not provide any dataset about hotel reviews for evaluation, but we have run the double-propagation process on our hotel reviews as an additional example of an application domain.

4.3.2 Examining the outcome for several domains

After running the proposed approach on datasets for several domains we examine the top ranked terms. The bootstrapped words provide the first insight into a domain-specific customer review corpus.

Table 4.2 shows some of the top aspect terms and opinion words for customer reviews about restaurants. Aspect terms that appear in the table are the aspect term graph nodes (i.e. bootstrapped words) that have the higher score after the graph ranking. Analogously, opinion words are the highest ranked opinion word graph nodes. A quick look over these top ranked words provides some immediate insight. The most remarkable aspect terms and opinion words refer to domain aspects like *food* (e.g. *food*, *cooking*, *sushi*, *delicious*), *service* (e.g. *service*, *staff*, *quick*, *slow*) and *ambiance* (e.g. *atmosphere*, *people*, *fun*).

Table 4.3 shows equivalent lists of words, aspect terms and opinion words, but for customer reviews about laptops. In this case there are composed terms like *battery life* and *hard drive*. This kind of terms are pretty common for

[4]Restaurant reviews of different cities from `http://www.citysearch.com`

[5]Laptop reviews from `http://www.toshibadirect.com`

Restaurants	
Aspect terms	**Opinion words**
food	good
service	delicious
price	fresh
place	great
atmosphere	quick
staff	excellent
experience	awesome
cooking	greasy
sushi	decent
people	friendly
fun	terrible
restaurant	wonderful
deal	divine
sandwich	authentic
attitude	slow

Table 4.2: Top 15 ranked aspect terms and opinion words for a customer reviews corpus about restaurants

Laptops	
Aspect terms	**Opinion words**
battery life	bright
keyboard	superb
screen	subtle
feature	nice
price	simplistic
machine	clean
toshiba laptop	faulty
windows	inconsistent
performance	powerful
use	unpowered
battery	high
program	awful
speaker	large
key	fine
hard drive	smallish

Table 4.3: Top 15 ranked aspect terms and opinion words for a customer reviews corpus about laptops

Hotels	
Aspect terms	**Opinion words**
room	helpful
bed	clean
service	friendly
hotel	nice
bathroom	excellent
area	comfortable
location	available
staff	old
microwave	spacious
place	courteous
pet	expensive
close	small
tv	gorgeous
view	dirty
facility	beautiful

Table 4.4: Top 15 ranked aspect terms and opinion words for a customer reviews corpus about hotels

this domain (e.g. *graphics card*, *RAM memory*, etc.). In this domain, laptop reviews, the most remarkable aspect are related to *battery* (i.e. *battery life*, *battery*, *unpowered*), *hardware* (e.g. *performance*, *hard drive*, *machine*, *powerful*) and *multimedia devices* (e.g. *keyboard*, *screen*, *speaker*, *large*, *bright*).

Finally, table 4.4 shows aspect term and opinion word lists computed on customer reviews about hotels. For this domain, the most salient aspects are related to the *rooms* (e.g. *room*, *bed*, *bathroom*), *service* (e.g. *service*, *staff*, *friendly*, *helpful*) and *location* (e.g. *area*, *location*, *place*, *view gorgeous*).

Not surprisingly, the provided insight reveals that the vocabulary used to refer to the most salient domain aspects presents remarkable differences among different application domains. This is one of the reasons why sentiment analysis systems are domain dependent, especially when it comes to supervised systems that are trained on labelled data for a particular domain.

```
<sentence id="270">
    <text>From the incredible food, to the warm atmosphere, to the friendly
    service, this downtown neighborhood spot doesn't miss a beat.</text>
    <aspectTerms>
        <aspectTerm term="food" polarity="positive" from="20" to="24"/>
        <aspectTerm term="atmosphere" polarity="positive" from="38" to="48"/>
        <aspectTerm term="service" polarity="positive" from="66" to="73"/>
    </aspectTerms>
    <aspectCategories>
        <aspectCategory category="food" polarity="positive"/>
        <aspectCategory category="service" polarity="positive"/>
        <aspectCategory category="ambience" polarity="positive"/>
    </aspectCategories>
</sentence>
```

Figure 4.3: Example of SemEval 2014 Task 4 dataset sentence.

4.3.3 SemEval 2014 Task 4 evaluation framework

SemEval 2014 task 4[6] *Aspect Based Sentiment Analysis* (Pontiki et al., 2014) provides two training datasets, one of restaurant reviews and other of laptop reviews. The restaurant review dataset consists of over 3,000 English sentences from restaurant reviews borrowed from Ganu et al. (2009). The laptop review dataset consists of over 3,000 English sentences extracted from customer reviews. Figure 4.3 shows an example of one of the manually annotated sentences contained in the SemEval 2014 restaurants dataset.

The task is divided into four different subtasks. Subtask 1 is *aspect term extraction*: given a set of sentences referring to pre-identified entities (i.e. restaurants or laptops), return the list of distinct aspect terms present in the sentence. An aspect term names a particular aspect of the target entity (e.g. *menu* or *wine* for restaurants, *hard disk* or *battery life* for laptops). Subtask 2 focuses on detecting the polarity of a given set of aspect terms in a sentence. The polarity in this task can be one of the following: *positive*, *negative*, *neutral* or *conflict*. The objective of subtask 3 is to classify the identified aspect terms into a predefined set of categories (i.e. domain aspects). In this SemEval task the predefined set of categories for restaurants are: *food*, *service*, *price*, *ambiance* and *anecdotes/miscellaneous*. Categories are not labelled for the laptop reviews. Subtask 4 is analogous to the subtask 2, but in this case, the polarity has to be determined for the aspect categories.

We focus on subtask 1, i.e. *aspect term extraction*. Our aim is to use the

[6] `http://alt.qcri.org/semeval2014/task4/`

SemEval Restaurants	Precision	Recall	F-score
SemEval Baseline	0.539	0.514	0.526
Our system (S)	**0.576**	0.649	**0.610**
Our system (W)	0.555	0.661	0.603
Our system (W+S)	0.551	**0.662**	0.601
SemEval-Best	*0.853*	*0.827*	*0.840*

Table 4.5: Result comparison for SemEval restaurant review dataset. Both SemEval-Best and SemEval Baseline are supervised machine learning approaches trained on the provided training data.

SemEval Laptops	Precision	Recall	F-score
SemEval Baseline	**0.401**	0.381	0.391
Our system (S)	0.309	0.475	0.374
Our system (W)	0.327	0.508	**0.398**
Our system (W+S)	0.307	**0.533**	0.389
SemEval-Best	*0.847*	*0.665*	*0.745*

Table 4.6: Result comparison for SemEval laptop review dataset. Both SemEval-Best and SemEval Baseline are supervised machine learning approaches trained on the provided training data.

bootstrapped lists of aspect terms for each domain to label the SemEval 2014 provided test sets, participating in the shared task.

The SemEval task provides an evaluation script which evaluates standard precision, recall and F-score measures. Both datasets (restaurants and laptops) contain 3,000 sentences each. The restaurant dataset contains 3,693 labelled gold aspect term spans (1,212 different aspect terms), and the laptop dataset contains 2,358 labelled gold aspect term spans (955 different aspect terms). We use these gold aspect terms to evaluate the experiments.

The experiments consists of running the proposed approach to generate aspect term lists (for restaurants and laptops) and using these lists to annotate the sentences. The generated aspect term lists have been limited to the top 550 items, because according to some empirical experiments this number provided the best trade-off between precision and recall. In this particular experiment, we have observed that using longer lists increases the recall, but decreases the precision due to the inclusion of more incorrect aspects terms. The annotation process is a simple lemma matching between the words in

the dataset and the words in our generated lists.

We compare the results against the SemEval baseline provided by the Semeval organisers. This baseline splits the dataset into *train* and *test* subsets and uses all the labelled aspect terms in the *train* subset to build a dictionary of aspect terms. Then it simply uses that dictionary to label the test subset for evaluation. In that sense it performs the same type of labelling than our approach, but its word list comes from manually obtained annotations in the training set, while ours is bootstrapped without the need of a manually labelled training set.

We also show the result of the best system submitted to SemEval (SemEval-Best in the table) for each domain. However, the results are not comparable since our approach is unsupervised and it is not intended for sentence labelling. Our approach was the only SemEval 2014 task 4 participant not using the annotated data from the training set.

Tables 4.5 and 4.6 show the performance of our system with respect to the baselines in both datasets. "Our system (S)" stands for our system only using the SemEval provided data (as it is unsupervised it learns from the available texts for the task). (W) refers to the results when using our own dataset scraped from the Web. Finally (W+S) refers to the results using both SemEval and our Web dataset. In the restaurant dataset, our system outperforms the baseline and it obtains quite similar results on the laptop dataset. Interestingly, the results are quite similar even if the learning datasets are very different in size. Probably this is because it only leverages more documents if they include new words that can be bootstrapped. If the overall distribution of words and relations does not change, the resulting aspect term list would be ranked very similarly.

Apart from the non-recognized aspect terms (i.e. not present in the generated list) another important source of errors is the multiword aspect term detection. In the SemEval training dataset, about the 25% of the gold aspect terms are multiword terms. In the restaurant dataset, we find a large number of names of recipes and meals, composed of two, three or even more words. For example, *"hanger steak au poivre"* or *"thin crusted pizza"* are labelled as single aspect terms. In the laptop domain, multiword terms are also very important, due to a number of technical expressions (i.e. hardware components like *"RAM memory"*, software versions like *"Windows 7"* and product brands like *"Samsung screen"*). These aspect terms cannot be present in our automatically acquired aspect term list because we limit the multiword

length up to two words. The SemEval evaluation method looks for complete labels, which means that a mistake labelling an aspect term, like labelling *"processor"* instead of *"Intel i7 processor"*, would decrease both the precision (because *"processor"* alone is not a labelled aspect term) and the recall (because the *"Intel i7 processor"* aspect term has not been labelled). There are also errors coming from typos and variations in the word spelling (e.g. *ambience* and *ambiance*) that our system does not handle.

4.4 Conclusions

Aspect term extraction (also known as *features* or *opinion targets*) is an important first step to perform fine-grained automatic opinion mining. There are many approaches in the literature aiming to automatically generate aspect terms for different domains. Bootstrapping domain aspect terms and opinion words provides a useful insight for every application domain, that can lead to a better domain adaptation or domain customisation of other sentiment analysis approaches.

In this chapter, we have described an almost unsupervised method to bootstrap and rank a list of domain aspect terms using a set of unlabelled domain texts. We use a double-propagation approach based on syntactic rules, and we model the obtained terms and their relations as a graph. Then we use PageRank algorithm to score the obtained terms. We have also applied some simple heuristics to capture multiword terms. We have evaluated the approach participating in the SemEval 2014 Task 4 and our unsupervised system performs better than the supervised baseline provided by the competition organisers.

Despite being unsupervised in the sense of not being trained on a manually labelled set of examples, the described approach still makes use of language dependent resources like Part-of-Speech tagging and syntactic dependency parsing. Therefore, this approach could not be applied in a language for which there is not a reliable syntactic dependency parsing available. We want to reduce these requirements event more, so the resulting approach could be easily used for different languages and domains with the least possible adaptation effort.

In the next chapters, we will explore additional approaches towards this objective.

CHAPTER 5

Unsupervised domain sentiment lexicon generation

In this chapter, we describe an approach to calculate a sentiment polarity value for the words of a domain using just a unlabelled corpus and two seed words. We also propose an approach to separate opinion words (i.e. words that actually express a sentiment) from the rest, just by adding an additional seed word. The chapter is structured as follows. First, section 5.1 introduces the work presented in the chapter. Then, section 5.2 defines the concept of sentiment lexicon and introduces continuous word embeddings. Section 5.3 describes the proposed approach to calculate sentiment polarity values using continuous word embeddings. After that, section 5.4 shows the polarity calculation experiments and results. Next, section 5.5 describes the opinion word separation approach and its evaluation. Finally, section 5.6 presents some concluding remarks.

5.1 Introduction

A key point in Sentiment Analysis, as its name suggests, is to determine the polarity of the sentiment implied by certain words or expressions (Taboada et al., 2011). Knowing an apriori sentiment polarity value for words over a vocabulary is a valuable information to perform sentiment analysis in a text.

A basic Sentiment Analysis system can use this sentiment polarity value of words, accounted and weighted in different ways, to calculate a degree of positivity/negativity for a piece of text like, for example, a customer review. In more sophisticated systems, word polarities can be used as additional features for machine learning algorithms. Sentiment polarity can be a categorical value (e.g. positive/neutral/negative) or a real value within a certain range (e.g. from -1.0 to +1.0). That value can be plugged in supervised classification algorithms together with other lexical and semantic features to help to discriminate the overall polarity of an expression or a sentence.

Currently, words are often modelled as continuous dense vectors, known as word embeddings, which seem to encode interesting semantic information about words. Word vectors are usually computed using very large corpora of texts, like the English Wikipedia. One of the most popular methods to obtain a dense continuous representation of words is Word2Vec (Mikolov et al., 2013b). But Word2Vec is not the only one, and in fact, there are already many variants and alternatives (Le and Mikolov, 2014; Iacobacci et al., 2015; Ji et al., 2015; Hill et al., 2014; Schwartz et al., 2015). With regard to Sentiment Analysis, word embeddings are used as features to more complex supervised classification systems when training sentiment classifiers (Tang et al., 2014b; Socher et al., 2013).

In this chapter, we compare a set of existing static sentiment lexicons and dynamic sentiment lexicon generation techniques. By *static* we refer to sentiment lexicons that consist of a fixed list of general words and polarities, obtained manually or by some semi-automatic method. Such static lexicons are not specifically adapted to any application domain and do not contain any specific vocabulary or sentiment polarity values. On the contrary, by *dynamic* lexicons we refer to those lexicons that are dynamically bootstrapped from a corpus using automatic or semi-automatic processes. A dynamic lexicon can be generated from a domain corpus to obtain a domain-specific vocabulary and sentiment polarity values.

We show a simple but competitive technique to calculate a word polarity value for each word of a particular domain using continuous word embeddings. Our objective is to assess if word embeddings calculated on an in-domain corpus can be directly used to obtain a useful polarity measure of the domain vocabulary with no additional supervision. Further, we want to see to which extent word embeddings calculated on an in-domain corpus improve the ones calculated on a general domain corpus and analyse pros

and cons of each compared method.

In addition, we propose an approach to separate words that bear polarity, i.e. opinion words, from the rest, just by adding a single extra seed word. The reason is that not all the words express a sentiment, and it would be incorrect to assign them a sentiment polarity value. For example, in the sentence *"the soup was tasty"*, only the word *tasty* is actually expressing a sentiment (positive in this case). The word *soup* is not intrinsically positive nor negative, and should not have associated any apriori sentiment polarity value. We want to automatically obtain a separate list of opinion words, like *tasty* in the example, to compose the resulting sentiment lexicon only with those words.

5.2 Sentiment lexicons

A collection of words and their respective sentiment polarity value is called a *sentiment lexicon*. Sentiment lexicons can be constructed manually, by human experts that estimate the corresponding sentiment polarity value to each word of interest. Obviously, this approach is usually too time-consuming in order to obtain a good vocabulary coverage, and difficult to maintain when the vocabulary evolves or the lexicon has to be ported to a new language or domain. Therefore, it is necessary to devise a method to automate the process as much as possible.

Some systems make use of existing lexical resources like WordNet (Fellbaum, 1998) to bootstrap a list of positive and negative words using different methods. In (Esuli and Sebastiani, 2006) the authors employ the glosses from each WordNet synset to perform a semi-supervised synset classification. A WordNet synset is a set of synonym words grouped together that denote the same concept. The result consists of three scores per synset: positivity, negativity and objectivity. In (Baccianella et al., 2010) version 3.0 of SentiWordNet is introduced with improvements like a random walk approach in the WordNet graph to calculate the sentiment polarity of the synsets. Agerri and Garcia (2009) introduce another system, Q-WordNet, which expands the polarities of WordNet synsets using lexical relations like synonymy. Guerini et al. (2013) propose and compare different approaches based on SentiWordNet to improve the estimated polarity values of the synsets.

Other authors try additional bootstrapping approaches that include random walks over the WordNet graph and evaluate them on WordNets of several languages (Maks et al., 2014; San Vicente et al., 2014). A problem with the approaches based on resources like WordNet is that they rely on the availability and quality of those resources for every language. In addition, WordNet is a general resource, so it often fails to capture domain dependent semantic orientations when they differ from the general use. Like other approaches using general dictionaries, WordNet-based methods do not take into account the shifts between domains (Paulo-Santos et al., 2011).

Other methods calculate the sentiment polarity of the words directly from the text. Hatzivassiloglou and McKeown (1997) model the corpus as a graph of adjectives joined by conjunctions. Then, they generate partitions on the graph based on some intuitions like that two adjectives joined by *"and"* will tend to share the same orientation while two adjectives joined by *"but"* will have opposite orientations.

Moreover, Turney (2002) obtains the sentiment polarity by calculating the Pointwise Mutual Information (PMI) between each word and a very positive word (like *"excellent"*) and a very negative word (like *"poor"*) in a corpus. The result is a continuous numeric value between -1 and +1.

These ideas of bootstrapping sentiment polarity from a corpus have been further explored and sophisticated in more recent works (Popescu and Etzioni, 2005; Brody and Elhadad, 2010; Qiu et al., 2011).

5.2.1 Continuous word representations

Continuous word representations (also vector representations or word embeddings) represent words as n-dimensional vectors. These vectors are capable of encoding semantic information which depends on the corpus used and the process applied. One of the best-known techniques for deriving vector representations of words and documents are Latent Semantic Indexing (LSI) (Dumais et al., 1995) and Latent Semantic Analysis (LSA) (Dumais, 2004).

Currently, it is becoming very common in the literature to use Neural Networks to compute word embeddings (Bengio et al., 2003; Turian et al., 2010a; Huang et al., 2012; Mikolov et al., 2013b). Figure 5.1 illustrates what a continuous word embedding function is (see section 2.6 for further details).

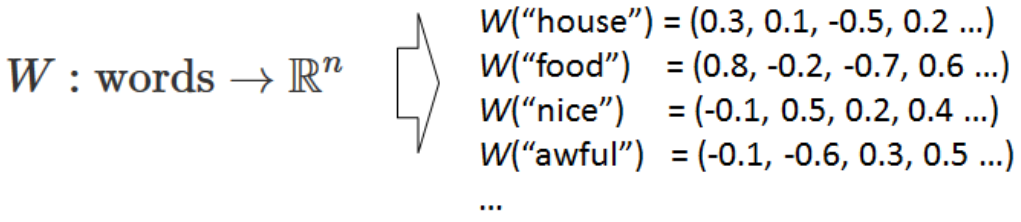


Figure 5.1: Continuous word embedding consists of calculating a function (W in the image) that maps words to N-dimensional dense vectors of real numbers.

Word embeddings offer really interesting syntactic and semantic properties. They encode information that can be exploited by performing algebraic operations over the resulting word vectors. For example, word embeddings capture a sense of *semantic distance* or *word relatedness* that can be measured by calculating the cosine of the angle of two word vectors. This semantic similarity or relatedness metric can be used for many tasks. Figure 5.2, borrowed from Collobert et al. (2011), shows an example of related words calculated using continuous word embeddings over a large corpus.

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Figure 5.2: Example of similar or related words obtained using word embeddings. Image borrowed from Collobert et al. (2011).

Word embeddings can also be used to uncover interesting word analogies. Just by simple algebraic operations, subtracting two word vectors and adding the result to a third word vector, word embeddings show semantically meaningful results. The best known real example is the operation:

$$W(\text{"}king\text{"}) - W(\text{"}man\text{"}) + W(\text{"}woman\text{"}) \approx W(\text{"}queen\text{"})$$

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Figure 5.3: Example of word analogies or relationships derived from word embeddings. Image borrowed from Mikolov et al. (2013c).

Figure 5.3, borrowed from Mikolov et al. (2013c), shows more analogies and word relationships derived from word embeddings. In the figure, the first column contains a relation between two words vectors, and each row contains the resulting pair of words when that relation is applied to a different word vector. For example, $W(\text{"}France\text{"}) - W(\text{"}Paris\text{"}) + W(\text{"}Italy\text{"}) \approx W(\text{"}Rome\text{"})$.

Word embeddings are also explored in additional tasks such as deriving adjectival scales (Kim and de Marneffe, 2013) or measuring word concreteness/abstraction (Rothe et al., 2016). It is clear that word embeddings enclose valuable information that the former *"one-hot encoding"* did not. Due to that, and because vectors of continuous values are a more natural input for many machine-learning algorithms, word embeddings are used as part of many modern sentiment analysis approaches (Socher et al., 2013; Tang et al., 2014a; Pavlopoulos and Androutsopoulos, 2014; Kim, 2014; Qian et al., 2017).

5.3 Word embedding based sentiment lexicon

We have applied Word2Vec (Mikolov et al., 2013a) and the Stanford GloVe system (Pennington et al., 2014) to calculate word embeddings. These two

Restaurant dataset word similarities		
excellent	***horrible***	***slow***
outstanding	terrible	spotty
fantastic	awful	inattentive
amazing	sucked	uncaring
exceptional	horrid	painfully
awesome	poor	neglectful
top notch	sucks	lax
great	atrocious	slower
superb	lousy	inconsistent
incredible	horrific	uneven
wonderful	yuck	iffy

Table 5.1: Most similar words in the word embedding space computed on restaurants reviews dataset, according to the cosine similarity, for words *excellent*, *horrible* and *slow*.

word-embedding models are well-known and widely used, and they are completely unsupervised (i.e. they do not need any external resource apart from a text corpus).

For the experiments presented in this chapter we have computed three word embedding models for each system: one in a restaurant reviews dataset, another in a laptop reviews dataset and a third one in a larger general domain dataset (consisting of the first billion characters from the English Wikipedia[1]).

Notice that the general domain dataset is much larger than the domain-based datasets. General domain dataset is a 700MB raw text file after cleaning it, while restaurants and laptop dataset only weight 28 and 40 MB respectively. General domain datasets, like the whole Wikipedia data or News dataset from online newspapers, capture very well general syntactic and semantic regularities. However, to capture in-domain word polarities smaller domain focused dataset might work better (García-Pablos et al., 2015b).

In table 5.1 and table 5.2 it can be observed how the word embeddings computed for restaurants and laptops domain seem to capture polarity quite accurately just by using word similarity. This is because the employed datasets are customer reviews of each domain, and the kind of content present in customer reviews helps to model the meaning and polarity of the words

[1]Obtained from `http://mattmahoney.net/dc/enwik9.zip`

Laptops dataset word similarities		
excellent	***horrible***	***slow***
outstanding	terrible	counterintuitive
exceptional	deplorable	painfully
awesome	awful	unstable
incredible	abysmal	sluggish
excelent	poor	choppy
amazing	horrid	fast
excellant	lousy	buggy
fantastic	whining	slows
terrific	horrendous	frustratingly
superb	unprofessional	flaky

Table 5.2: Most similar words in the word embedding space computed on laptops reviews dataset, according to the cosine similarity, for words *excellent*, *horrible* and *slow*.

(adjectives in this case). The tables show top similarities according to the cosine distance between word vectors computed by each model. Words like *excellent* and *horrible* are domain independent, and the most similar words are quite equivalent for both domains. But for the third word, *slow*, the differences between both domains are more evident. The word *slow* in the context of restaurants is usually used to describe the service quality (when judging waiters and waitresses serving speed and skills), while in the context of laptops it refers to the performance of hardware and/or software.

Another advantage compared to a general domain computed model is that domain-based models will contain any domain jargon words, slangs or even commonly misspelt words (as long as they appear frequently enough in the corresponding domain corpus). A general domain corpus is less likely to cover all the vocabulary present for any possible domain, limiting itself to more general words.

In order to build a domain based polarity lexicon, we have used a simple formula to assign a polarity to the words in the vocabulary, using a single positive seed word and a single negative seed word. The formula is the following:

$$\text{pol}(w) = \text{sim}(w, POS) - \text{sim}(w, NEG) \tag{5.1}$$

In the formula, *sim* stands for the cosine distance between word vectors, POS is the vector representation of the positive seed word, and analogously, NEG is the vector representation of the negative seed word. In the experiments, we have used domain independent seed words with a very clear and context- and domain-independent polarity. In particular *excellent* and *horrible* as positive and negative seeds respectively. These words are likely to be valid for any domain, since *excellent* expresses a very positive sentiment in any context, while *horrible* also expresses a negative sentiment regardless of the context.

Note that this simple formula provides a real number, that in a sense gives a continuous value for the polarity. This value can be normalised to be in [-1,+1] range. The fact of obtaining a continuous value for the polarity could be an interesting property to measure the strength of the sentiment, but to simplify the experiments we treat the polarity value as a binary label: positive if the value is greater or equal to zero, and negative otherwise. This makes the comparison with other examined lexicons easier.

5.3.1 Compared lexicons and methods

Our aim is to compare the proposed polarity calculation method, based on continuous word embeddings, to other existing sentiment lexicons and methods to find out if continuous word embeddings can be used to compute accurate sentiment polarity over the words of a domain. For that purpose, a set of well-known sentiment lexicons and sentiment lexicon bootstrapping approaches are evaluated and compared in several domains. Next, these compared lexicons and approaches are briefly described.

5.3.1.1 General lexicons

The General Inquirer (GI) (Stone et al., 1968) is a very well-known and widely used manually crafted lexicon that includes the polarity of many English words. GI contains about 2000 positive and negative words.

In addition, we have also used the Bing Liu's sentiment lexicon (Hu and Liu, 2004). According to the web page [2] it has been compiled and incremented

[2] `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`

over many years. It contains around 6800 words with an assigned categorical polarity (positive or negative).

5.3.1.2 WordNet based lexicons

SentiWordnet assigns scores to each WordNet synset[3] (Esuli and Sebastiani, 2006). SentiWordNet polarity consists of three scores per synset: positivity, negativity and objectivity. Baccianella et al. (2010) introduce the version 3.0 of SentiWordNet with improvements like a random walk approach in the WordNet graph.

We have also used the Q-WordNet as Personalized PageRanking Vector (QWN-PPV) which propagates and ranks polarity values on the WordNet graph starting from few seed words and using PageRank algorithm to weight the resulting polarities (San Vicente et al., 2014).

5.3.1.3 PMI based lexicons

Following the work at (Turney, 2002), we also have derived some polarity lexicons from a domain corpus using Point-wise Mutual Information (PMI). In few words, PMI is used as a measure of relatedness between two events, in this case, the co-occurrence of words with known positive contexts. PMI calculation is shown in equation 5.2, where $p(w1)$ and $p(w2)$ are the number of times word $w1$ and word $w2$ occur individually, and $p(w1, w2)$ is the number of times the words $w1$ and $w2$ co-occur together.

In the original Turney's work, the value of co-occurrence was measured counting hits in a web search (the extinct Altavista) between words and the seed word *"excellent"* (for positives) and the seed word *"poor"*. Using these counts the semantic orientation of the words (SO) was calculated as shown in equation 5.3

$$\mathrm{PMI}(w1, w2) = \log \frac{p(w1, w2)}{p(w1) \times p(w2)} \tag{5.2}$$

$$\mathrm{SO}(w) = \mathrm{PMI}(w, POS) - \mathrm{PMI}(w, NEG) \tag{5.3}$$

[3] A WordNet synset in a set of synonym words that denote the same concept

Firstly, we have borrowed the lexicon generated in (Kiritchenko et al., 2014) (named NRC_CANADA in the experiment tables), which was generated computing the PMI between each word and positive reviews (4 or 5 stars in a 5-star rating) and negative reviews (1 or 2 stars), for both restaurants and laptops review datasets. Note that because this approach uses the customer ratings to make its calculations, the approach is supervised.

As a counterpart, we have calculated another PMI based lexicon in the domain datasets described in 5.4.1. To calculate the PMI-based sentiment score for this sentiment lexicon we use the co-occurrence of words within a five-word window, with the word *excellent* as the positive seed and the word *horrible* as the negative seed. We name it PMI_WINDOW_5. This is potentially less accurate but requires no supervised information apart from the two seed words.

5.4 Polarity calculation experiments

We evaluate the word embeddings based polarity calculation by comparing it against other existing sentiment lexicons and sentiment lexicon generation approaches.

We perform two type of experiments with the sentiment polarity calculation. First, we evaluate the accuracy of sentiment polarity assignment against a list of domain words with a manually assigned polarity. Then, an additional evaluation is performed against the sentences of two datasets with manually labelled sentiment polarity. In both cases, the evaluated domains are customer restaurant reviews and laptops reviews.

5.4.1 Datasets and resources

In order to generate the lexicons with the methods that require an in-domain corpus (i.e. the PMI based sentiment lexicon or the in-domain Word2Vec and the GloVe computation) we have used text corpora from two different domains: customer restaurants reviews and customer laptop reviews.

5.4.1.1 Unlabelled domain corpora

The first corpus consists of customer reviews about restaurants. It is a 100k review subset about restaurants obtained from the Yelp dataset[4] (henceforth Yelp-restaurants). We also have used a second corpus of customer reviews about laptops. This corpus contains a subset of about 100k reviews from the Amazon electronic device review dataset from the Stanford Network Analysis Project (SNAP)[5] after selecting reviews that contain the word *"laptop"* (henceforth Amazon-laptops). The corpora have been lower-cased and tokenised.

Both corpora have been used to perform several processes, obtaining their respective domain-aware results. In the case of the PMI based lexicon, the corpora have been used to compute the PMI based sentiment polarity score. In the case of Word2Vec and GloVe, the corpora have been used to obtain word vector representations for each domain. In the case of Word2Vec, we have used the implementation contained in the Apache Spark Mllib library[6]. This Word2Vec implementation is based on the Word2Vec Skip-gram architecture, and we have let the default hyper-parameters and configuration[7].

5.4.1.2 Evaluation resources

In addition, for evaluation purposes we have used again these domain corpora (Yelp-restaurants and Amazon-laptops) to automatically obtain a list of domain adjectives ranked by frequency, to be used for evaluation purposes. From that list we have manually selected the first 200 adjectives with context-independent positive or negative polarity for each domain. With context-independent polarity we refer to those adjectives with unambiguous polarity not depending on the domain aspect term they are modifying (e.g. *superb* is likely to be always positive, while *small* could be positive or negative depending on the context). Then we have manually assigned a polarity label (positive or negative) to each of the selected adjectives. From now on we will refer to these annotated adjectives *restaurant-adjectives-test-set*

[4] `http://www.yelp.com/dataset_challenge`

[5] `http://snap.stanford.edu/data/web-Amazon.html`

[6] `http://spark.apache.org/mllib/`

[7] Please, refer to the Apache Spark Mllib Word2Vec documentation to see which the default parameters are

and *laptop-adjectives-test-set* respectively. The *restaurant-adjectives-test-set*[8] contains 119 positive adjectives and 81 negatives adjectives, while *laptops-adjectives-test-set*[9] contains 127 positives and 73 negatives.

We have also used the SemEval 2015 task 12 datasets[10] for evaluation. The first dataset contains 254 annotated reviews about restaurants (a total of 1,315 sentences). The second dataset contains 277 annotated reviews about laptops (a total of 1,739 sentences).

5.4.2 Polarity calculation results and evaluation

First, we examine the content of the resulting polarity word lists for several domain and languages to assess if the approach is working as expected.

Table 5.3 shows several lists of words sorted by the polarity obtained using the word embeddings based approach described in the chapter, in this particular case using Word2Vec. The lists contain the top positive and negative opinion words for customer reviews of several domains: restaurants, laptops and hotels. The positive polarity seed word used is *excellent* and the negative is *horrible*. The word lists include some misspelt words and some slang because they are taken directly from customer reviews, which are informal texts. This is a remarkable detail because a general dictionary is less likely to contain informal vocabulary or misspelt variants of a word.

Table 5.4 shows additional polarity word lists, but in this case for Spanish customer reviews of several domains: restaurants and hotels. The customer review corpora for these domains in Spanish have been obtained from a popular customer reviews website using a program to automatically download approximately 10k reviews. Since the only resources that have to be adapted are the polarity seed words, the approach can be directly applied to other languages and domains very easily. The seed words in the case of Spanish texts are: *excelente* as the positive seed, and *horrible* as the negative seed. Again, the words contain misspellings and typos, as they appear in the original texts.

[8] `https://github.com/aitor-garcia-p/resources/blob/master/restaur_adjs_test.txt`

[9] `https://github.com/aitor-garcia-p/resources/blob/master/laptops_adjs_test.txt`

[10] `http://alt.qcri.org/semeval2015/task12/`

English					
Restaurants		**Laptops**		**Hotels**	
Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
excellent	horrible	excellent	horrible	highlights	horrible
delicious	unacceptable	incredible	abysmal	convenient	terrible
fantastic	unfriendly	excelent	annoy	excellant	appalling
tasty	unhelpful	exceptional	awful	meticulously	awful
inventive	unprofessional	ingenious	bothers	terrific	nightmare
affordable	awful	perfect	clueless	easy	deafening
outstanding	disgusting	amazing	compelled	accomodating	disaster
amazing	filthy	ideal	downright	talented	disgrace
creative	gross	ample	terrible	awsome	disgusting
awesome	horrendous	awesome	forewarned	unbeatable	unacceptable
plentiful	bleh	excellant	horrendous	breathtaking	foul
innovative	horribly	outstanding	unable	exemplary	gross
superb	horrid	invaluable	unsure	ideal	grossed
interesting	tasteless	avid	plagued	flawless	horrendous
exceptional	terrible	extremely	terribly	eager	horrific
sublime	microwaved	great	rave	annual	horrified
great	nasty	extraordinarily	remedied	pros	joke
splendid	poor	good	sorely	impeccable	leaky
extensive	rude	exceptionally	stunned	painless	miserable
dependable	shitty	unbeatable	suprised	friendly	liar

Table 5.3: Examples of top positive and negative words obtained for English customer reviews of several domains using Word2Vec. The words are exactly as they appeared in the texts, including misspellings.

Spanish			
Restaurants		**Hotels**	
Top positives	**Top negatives**	**Top positives**	**Top negatives**
excelente	horrible	excelente	horrible
autenticidad	agobiados	inmejorable	terrible
céntrica	asquerosa	privilegiada	oscuro
imaginativa	prepotentes	magnifica	caliente
magnifica	refritas	perfecta	cerraba
evolución	decepcion	centrico	diminuto
dedicada	decepcionante	excepcional	extraño
honesta	desordenada	céntrica	feo
continua	desorganizado	simpatico	fria
gastrobar	diminutos	destaco	rotas
turistico	espantoso	exelente	incomodo
escueta	hervido	buenisima	olor
positivamente	indescriptible	magnífica	insoportable
primerísima	infame	ampliamente	interminables
tipica	insulsa	accesible	masificado
exelente	lento	genial	ruidosa
selecta	lentos	fantástica	vieja
actual	lentísimo	conveniente	minúscula
conveniente	repugnante	estupenda	minúsculo
fabulosa	nefasta	inigualable	molesto

Table 5.4: Examples of some top positive and negative words obtained for Spanish customer reviews of several domains. The words include misspellings and typos.

5.4.2.1 Manual gold-lexicon based experiments

We have performed two different formal evaluations. The first one uses the two domain adjective lists *restaurants-adjectives-test-set* and *laptops-adjectives-test-set* as a ground truth for polarity evaluation.

On the restaurant-adjectives-test-set and laptop-adjectives-test-set, we measure the polarity accuracy (when a lexicon assigns the correct polarity) and the coverage (when a lexicon contains a polarity for the requested word).

Tables 5.5 and 5.6 show the results for restaurants and laptops respectively. The lexicon names that appear in the tables are:

- General Inquirer (GI)
- BingLiu lexicon (BingLiu)
- SentiWordNet based lexicon (SentiWordNet)
- Q-WordNet as Personalized PageRanking Vector (QWN-PPV)
- NRC_CANADA lexicon (NRC_CANADA)
- PMI with a context window of five words (PMI_WINDOW_5)
- Word2Vec polarities calculated on a domain corpus (W2V_DOMAIN)
- Word2Vec polarities calculated on a general corpus (W2V_GENERAL)
- GloVe polarities calculated on a domain corpus (GloVe_DOMAIN)
- GloVe polarities calculated on a general corpus (GloVe_GENERAL)

In the tables, the accuracy measures how many word polarities have been correctly tagged from the ones present in each lexicon (i.e. out-of-vocabulary words are not taken as errors). The coverage measures how many words were present in each lexicon regardless of the tagged polarity.

The experiment shows that the static lexicons like GI and BingLiu's assign polarities with a very high precision, but they suffer from lower coverage. A similar behaviour can be observed for polarities based on WordNet. On the contrary, the lexicons calculated directly on the domain datasets are less accurate, but they have much higher coverage. NRC_CANADA lexicon achieves a very good result, but it must be noted that it employs supervised information. The PMI_WINDOW_5, based on windows achieve a quite good result despite its simplicity, but it does not cover all the words (i.e. some words do not co-occur in the same context).

RESTAURANTS 200 ADJECTIVES GOLD LEXICON				
Lexicon name		**Positives**	**Negatives**	**Overall**
General Inquirer	Accuracy	**0.935**	0.944	0.939
	Coverage	0.521	0.444	0.490
BingLiu	Accuracy	**0.935**	**0.979**	**0.952**
	Coverage	0.647	0.580	0.620
SentiWordNet	Accuracy	0.725	0.746	0.733
	Coverage	0.857	0.778	0.825
QWN-PPV	Accuracy	0.821	0.609	0.746
	Coverage	0.706	0.568	0.650
NRC_CANADA	Accuracy	0.933	0.753	0.860
	Coverage	**1.000**	**1.000**	**1.000**
PMI_WINDOW_5	Accuracy	0.917	0.655	0.821
	Coverage	0.807	0.679	0.755
W2V_DOMAIN	Accuracy	0.849	0.827	0.840
	Coverage	**1.000**	**1.000**	**1.000**
W2V_GENERAL	Accuracy	0.491	0.400	0.454
	Coverage	0.958	0.988	0.970
GloVe_DOMAIN	Accuracy	0.866	0.802	0.840
	Coverage	**1.000**	**1.000**	**1.000**
GloVe_GENERAL	Accuracy	0.754	0.588	0.686
	Coverage	0.958	0.988	0.970

Table 5.5: Restaurants 200 adjectives lexicon results.

The lexicons based on word embeddings calculated on the domain achieve a 100% coverage, because they are modelling the whole vocabulary, and offer a reasonable precision. Word embeddings (both Word2Vec and GloVe) calculated on general domain corpus still cover a lot of the adjectives since they have been trained on very large corpora, but they show a lower accuracy capturing the polarity of the words for both restaurants and laptops domains. According to these results, calculating word embeddings on a corpus of the target domains provides several advantages compared to using

LAPTOPS 200 ADJECTIVES GOLD LEXICON				
Lexicon name		**Positives**	**Negatives**	**Overall**
General Inquirer	Accuracy	0.965	0.971	0.967
	Coverage	0.677	0.479	0.605
BingLiu	Accuracy	**0.971**	**0.984**	**0.976**
	Coverage	0.827	0.863	0.840
SentiWordNet	Accuracy	0.795	0.833	0.809
	Coverage	0.921	0.904	0.915
QWN-PPV	Accuracy	0.895	0.661	0.814
	Coverage	0.829	0.767	0.805
NRC_CANADA	Accuracy	0.890	0.712	0.825
	Coverage	**1.000**	**1.000**	**1.000**
PMI_WINDOW_5	Accuracy	0.850	0.395	0.720
	Coverage	0.843	0.589	0.750
W2V_DOMAIN	Accuracy	0.874	0.740	0.825
	Coverage	**1.000**	**1.000**	**1.000**
W2V_GENERAL	Accuracy	0.540	0.575	0.553
	Coverage	0.992	**1.000**	0.995
GloVe_DOMAIN	Accuracy	0.890	0.740	0.835
	Coverage	**1.000**	**1.000**	**1.000**
GloVe_GENERAL	Accuracy	0.849	0.589	0.754
	Coverage	0.992	**1.000**	0.995

Table 5.6: Laptops 200 adjectives lexicon results.

word embeddings from a general source.

5.4.2.2 SemEval 2015 datasets based experiments

In addition, we have performed another sentiment polarity evaluation on the SemEval 2015 task 12 datasets, composed by customer reviews about restaurants and laptops.

SemEval 2015 datasets consist of quintuples of aspect-term, entity-attribute, polarity, and starting and ending position of the aspect-term. For this evaluation, we are only interested in the polarity slots, which refer to the polarity of a particular aspect of each sentence (not to the overall sentence polarity). We have used the different lexicons to calculate the polarity of each sentence, and then we have compared them to the gold annotations that come with the datasets.

The process of assigning a polarity to each sentence using the different polarity lexicons is the following:

- Only adjectives and verbs (e.g. *hate*, *recommend*) are taken into account to calculate polarity (auxiliary verbs like *be* and *have* are omitted)
- Negation words are taken into account to reverse the polarity of the subsequent word, in particular: *no, neither, nothing, not, n't, none, any, never, without*
- The number of positive and negative words according to each lexicon is counted. If the positives count is greater or equal to negatives count, the polarity of all polarity slots of the sentence is assigned as positive; and negative otherwise.

Note that this is a very naive polarity annotation process. It is not intended to obtain good results but for comparing the lexicons against real sentences using the same setting. That is why in general the results are lower than in the experiment with the plain adjective lists. This naive polarity annotation process is repeated for every polarity lexicon so the different lexicons and methods can be compared under the same conditions in test sets containing actual customer reviews.

Table 5.7 shows the results for restaurants dataset while table 5.8 shows the results for laptops dataset. These results have been calculated using the evaluation script provided by the SemEval 2015 organisers during the competition [11].

The results show that the best performing lexicon varies depending on the domain. It must be noted that in this case what is being annotated are whole sentences of actual reviews, so there are other facts involved apart from

[11] Available at `http://alt.qcri.org/semeval2015/task12/index.php?id=data-and-tools`

RESTAURANTS (SEMEVAL 2015 DATASET)					
Lexicon name		**Precision**	**Recall**	**F1**	**Accuracy**
General Inquirer	positives	0.783	0.937	0.853	0.760
	negatives	0.610	0.335	0.432	
BingLiu	positives	0.810	**0.958**	**0.878**	**0.799**
	negatives	**0.731**	0.431	0.540	
SentiWordNet	positives	0.790	0.896	0.840	0.745
	negatives	0.539	0.394	0.455	
QWN-PPV	positives	0.751	0.954	0.841	0.733
	negatives	0.522	0.171	0.257	
NRC_CANADA	positives	0.816	0.927	0.868	0.786
	negatives	0.648	0.471	0.546	
PMI_WINDOW_5	positives	0.811	0.842	0.826	0.732
	negatives	0.493	0.503	0.498	
W2V_DOMAIN	positives	**0.848**	0.874	0.861	0.781
	negatives	0.582	**0.605**	**0.593**	
W2V_GENERAL	positives	0.708	0.467	0.563	0.457
	negatives	0.228	0.488	0.311	
GloVe_DOMAIN	positives	0.792	0.940	0.860	0.770
	negatives	0.633	0.364	0.463	
GloVe_GENERAL	positives	0.747	0.900	0.816	0.703
	negatives	0.404	0.210	0.277	

Table 5.7: SemEval 2015 restaurants dataset sentiment polarity results.

LAPTOPS (SEMEVAL 2015 DATASET)					
Lexicon name		**Precision**	**Recall**	**F1**	**Accuracy**
General Inquirer	positives	0.631	0.939	0.755	0.651
	negatives	0.751	0.328	0.456	
BingLiu	positives	0.640	0.960	0.768	0.669
	negatives	**0.821**	0.343	0.484	
SentiWordNet	positives	0.630	0.903	0.742	0.638
	negatives	0.671	0.345	0.456	
QWN-PPV	positives	0.605	0.9411	0.736	0.614
	negatives	0.675	0.228	0.341	
NRC_CANADA	positives	0.653	0.922	0.764	0.673
	negatives	0.750	0.409	0.529	
PMI_WINDOW_5	positives	0.622	0.841	0.715	0.611
	negatives	0.580	0.366	0.449	
W2V_DOMAIN	positives	**0.728**	0.825	**0.774**	**0.708**
	negatives	0.673	**0.636**	**0.654**	
W2V_GENERAL	positives	0.533	0.443	0.484	0.441
	negatives	0.362	0.500	0.420	
GloVe_DOMAIN	positives	0.590	**0.971**	0.734	0.604
	negatives	0.762	0.159	0.263	
GloVe_GENERAL	positives	0.571	0.932	0.708	0.567
	negatives	0.528	0.120	0.196	

Table 5.8: SemEval 2015 laptops dataset sentiment polarity results.

the mere polarity of individual words. Some lexicons show a better precision capturing positive words, while others perform better for negative words. BingLiu lexicon obtains the best overall accuracy for restaurant reviews, however, it shows a much lower recall for negatives than, for example, the Word2Vec based lexicon computed in-domain. For laptop reviews Word2Vec calculated on the domain corpus show the best overall performance. It may indicate that for laptop reviews domain there is more technical and specific vocabulary, that is less likely to appear in a general lexicon. Also in this case, the domain-based word embeddings capture better the polarity than their general-domain counterparts.

If we take into account the minimal effort and linguistic resources required to compute a sentiment lexicon using word embeddings (just two seed words and a domain corpus), we can conclude that word-embedding-based sentiment lexicons show the best performance/cost ratio in these experiments.

5.5 Opinion words separation

A method that calculates a sentiment polarity value for every word in a text corpus has to deal with the fact that not all words express a sentiment. The words that bear a sentiment polarity are which are called opinion words (or opinion expressions in the case of expressions involving several words). The rest of the words play different roles, like being opinion targets, but they do not influence the sentiment of a sentence on their own. For example, the following sentences:

"The design is a disaster".

"The waitress was attentive".

The only words that are actually expressing some degree of sentiment are *disaster* and *attentive*, a positive and a negative sentiment respectively. However, if a polarity value is blindly assigned to every word (after discarding stopwords), then *design* and *waitress* would receive a polarity value, when they do not carry any sentiment polarity on their own.

In order to address this problem maintaining the low supervision of the overall sentiment lexicon generation approach, we propose a method to automatically separate opinion words from the rest. This way the resulting sentiment lexicon should be composed only of words that are likely actual

opinion words. In this section we describe and evaluate a method that will be further used in chapter 6 as part of a weakly-supervised aspect-based sentiment analysis system. The method is the following.

Let OW be a set containing opinion word seeds. For example, the two polarity words (*excellent* and *horrible*) used to calculate the polarity can be reused as opinion word seeds. Let AT be a set containing aspect term seeds consisting of a few possible aspect terms for the target domain, for example *food* or *service* for a dataset about restaurants (even a single aspect term seed may suffice). For every occurrence of a word $w_i \in OW$ in a domain text dataset, extract a training instance composed by its context words and labelled as an opinion word.

$$w_i^{OW} : (w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2})$$

Do the same with every occurrence of a word $w_j \in AT$ in the dataset, labelling them as aspect terms.

$$w_j^{AT} : (w_{j-2}, w_{j-1}, w_{j+1}, w_{j+2})$$

The idea is to use the extracted instances to train a supervised classifier that learns a simple model to estimate the probability of an unseen word being an aspect term or an opinion word. In order to improve the generalisation of the resulting model the context words that serve as features are transformed into a more abstract representation. For that, we try several unsupervised clustering algorithms and substitute each word by its corresponding cluster for each of the training instances. In this context, a cluster is just a number to represent a group of related words. Which words fall under the same cluster is determined by each particular clustering algorithm. We experiment with a classic K-means clustering of Word2Vec vectors and with Brown clusters[12] (Brown et al., 1992).

Let be $u_{w_i}^c$ the word cluster corresponding to the word w_i under the clustering type $c \in \{word2vec, brown\}$. Two extra symbols are used to cover special cases: *PAD* and *UNK*. The symbol *PAD* (meaning *"padding"*) is used to indicate that a certain context position is beyond the sentence boundary. The symbol *UNK* (meaning *"unknown"*) is used when a context word has no

[12]We have used the implementation available at `https://github.com/koendeschacht/brown-cluster`

corresponding cluster because that word was not present in the corpus used to generate the word clustering (e.g. people names, misspelt words, etc.).

After replacing each context word by its corresponding cluster, the training instances become:

$$w_i^{label \in \{AT,OW\}} : (u^c_{w_{i-2}}, u^c_{w_{i-1}}, u^c_{w_{i+1}}, u^c_{w_{i+2}})$$

We use a simple logistic regression (MaxEnt) classifier to train a prediction model over the resulting instances. The resulting classification model is used to classify all the word occurrences in the corpus, obtaining the estimated probability of being an aspect term or an opinion word for all of them. Figure 5.4 illustrates the described process using a couple of simple sentences as an example. In the figure just a single seed for domain aspect terms (e.g. *service*) and a single seed for opinion words (i.e. *excellent*) are used. The resulting training instances are derived from the contexts of the seed word occurrences in a domain corpus by replacing each particular word by its corresponding cluster identifier.

The resulting classifier can be used to classify and sort the words of a domain corpus by their probability of being opinion words. The top words of that ranked list will have a high confidence of being opinion words expressing a sentiment polarity and good candidates to be part of a sentiment lexicon.

5.5.1 Opinion word separation evaluation

In order to evaluate the opinion word separation, we measure the results given by the classifier over several specific lists of words. These lists of words act as the ground truth. One of these word lists is the BingLiu sentiment lexicon (see 5.3.1). A sentiment lexicon contains words that express some degree of sentiment polarity, so we make the assumption that the words contained in this sentiment lexicon are opinion words. We use them as a gold list of opinion words. We also use the *restaurant-adjectives-test-set* and *laptop-adjectives-test-set* (see section 5.4.1) as an additional source of gold opinion words for the same purpose.

In addition, as the aspect term counterpart, we extract the labelled aspect terms from the SemEval 2015 restaurant dataset (e.g. *burgers*, *pastries*, *owner*, *decor*, etc.). SemEval 2015 laptops dataset does not contain manually

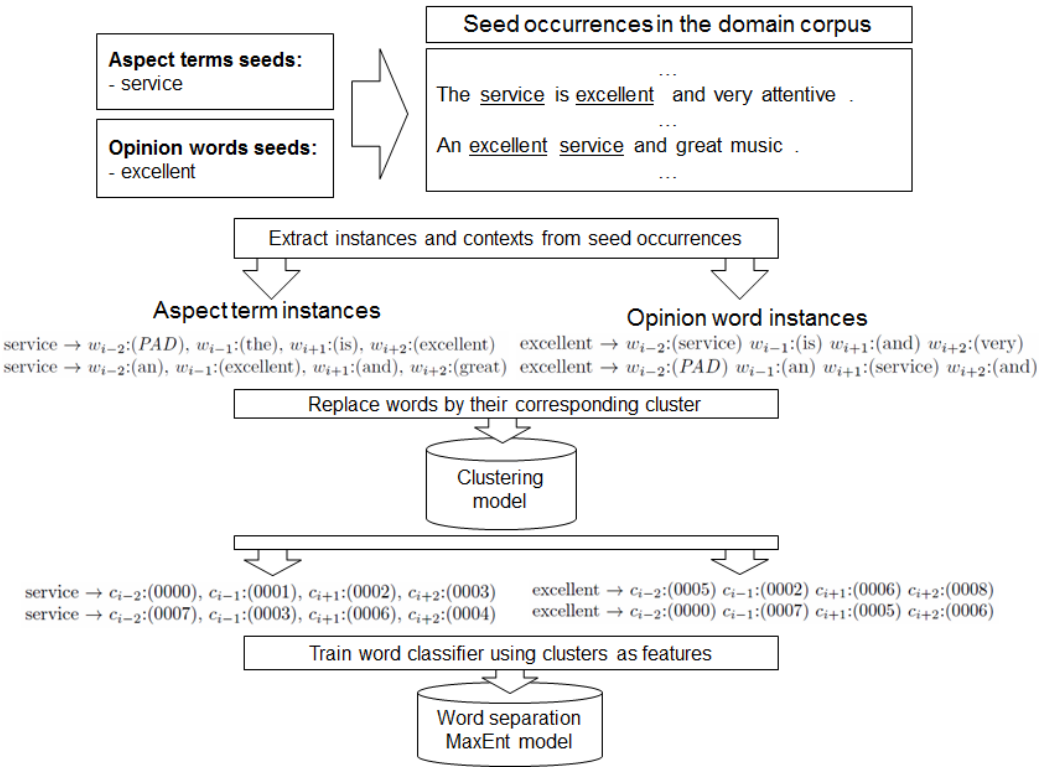


Figure 5.4: Extracting the word separation training instances from the occurrences of seed words.

labelled aspect terms, so we manually extract a list of 40 aspect terms from the laptops dataset (e.g. *laptop*, *processor*, *battery*, *keyboard*, etc.). We use these two lists as gold aspect terms.

We use these lists (gold opinion words and gold aspect terms) to evaluate the performance of the proposed word separation classifier. The word separation classifier divides words in two classes, *aspect_term* or *opinion_word*, assigning a probability to each of them for each classified word. The hypothesis is that if the classifier is working as expected, the words from the gold opinion words list will be classified under the class *opinion_word* with high probability (i.e. the classifier will be confident of its decision). The opposite will happen with the words contained in the gold aspect term lists.

The datasets used for these experiments are Yelp-restaurants and Amazon-laptops (see section 5.4.1) for their respective domains. The seed opinion

Restaurants - Opinion words averaged probability			
Gold words list	**Brown clusters**	**W2V clusters**	**Baseline**
BingLiu lexicon	**0.746**	0.708	0.500
Restaurant adjectives	**0.812**	0.767	0.500

Table 5.9: Estimated probability of ground-truth opinion words being opinion words in the restaurants dataset. The higher the better. Brown clusters used as features for the classifier outperforms the Word2Vec based K-means clusters. The baseline is the random classification of the words.

words to bootstrap and train the word separation classifier are *excellent* and *horrible*, reused from the polarity calculation process. The domain aspect term seed words are *restaurant* and *performance* for restaurants and laptops domain respectively.

Laptops - Opinion words averaged probability			
Gold words list	**Brown clusters**	**W2V clusters**	**Baseline**
BingLiu lexicon	0.621	**0.631**	0.500
Laptop adjectives	**0.747**	0.720	0.500

Table 5.10: Estimated probability of ground-truth opinion words being opinion words in the laptops dataset. The higher the better. The baseline is the random classification of the words.

For each word contained in a gold opinion word lists, each occurrence in the corresponding domain dataset is classified using the trained word separation classifier. The probability for each class is averaged. Table 5.9 and table 5.10 show this averaged probability over all the gold opinion words for restaurants and laptops domain respectively. In the tables, Brown clusters are also compared to Word2Vec based clusters. The results show a strong bias towards the correct class, *opinion_word* in this case. Brown clusters show a better performance than Word2Vec based clusters.

To ensure that the classifier is really separating the words in two classes and that it is not biased towards classifying everything as being an opinion word, we repeat the experiment with the gold aspect terms. In this case, the expected bias should go towards the *aspect_term* class. Table 5.11 shows the probability for the *aspect_term* class of the gold restaurant aspect terms

Gold aspect terms averaged probability			
Gold words list	**Brown clusters**	**W2V clusters**	**Baseline**
Restaurant aspect terms	**0.729**	0.660	0.500
Laptops aspect terms	**0.783**	0.692	0.500

Table 5.11: Estimated probability of ground-truth aspect terms being aspect terms. The higher the better. Brown clusters used as features for the classifier outperform the Word2Vec based K-means clusters. The baseline is the random classification of the words.

and the gold laptops aspect terms. In this case, the classifier has assigned more probability mass to the *aspect_term* class, showing that the classifier correctly separates the two classes. According to the results, Brown clusters work better than Word2Vec based clusters again.

5.6 Conclusions

In this chapter, we have described an approach to quickly obtain a sentiment lexicon value only with unlabelled texts. The proposed approach uses continuous word representations, and the only supervised information comes from the two seeds words, positive and negative, required by the approach. The word vector representations of the seed words are used to compute a polarity value for every word in the vocabulary.

We show results for several domains (restaurants, hotels, laptops) and languages (English, Spanish). We have compared the described approach against other existing lexicons and methods to obtain a polarity value for words in a particular domain. We have also shown that the similarity of sentiment-bearing words is better modelled using a smaller in-domain dataset rather than a bigger general dataset. We have observed a similar behaviour for other languages such as Spanish. An obvious advantage is that providing enough unlabelled domain data, word embeddings and polarity scores can be easily obtained for any language and domain.

In addition, we have described a simple approach to separate potential opinion words from the rest. The motivation is that not every word in a vocabulary is meant to carry a sentiment polarity. Words like *waiter*, *hotel*

or *processor* do not express any sentiment polarity on their own, and therefore should be excluded from a sentiment lexicon. Just by the addition of an extra seed word, a seed word of an aspect term, we propose a method to train a small Maximum Entropy classifier that separates the two types of words.

The resulting classifier outputs a probability value for each word occurrence based on its context words. This probability can be use to rank words by the confidence of being opinion words and keep only the words with the higher confidence as part of a sentiment lexicon. We have evaluated this method using several lists of known opinion words and aspect terms as the ground truth. The experiments show that the resulting classifier is effectively distinguishing both types of words, being able to correctly separate the opinion words.

The approaches described in this chapter to calculate a sentiment polarity value and to separate opinions from aspect terms, will be further explored and combined in chapter 6. The final objective is to use these weakly-supervised approaches in combination with other unsupervised techniques to build a weakly supervised aspect-based sentiment analysis system, that not only provides a value for the sentiment polarity, but also detects the domain aspects being mentioned in each sentence.

CHAPTER 6

Weakly supervised ABSA

In this chapter, we describe a system that combines some of the methods introduced in the previous chapter, like continuous word embedding based similarity or aspect term and opinion word separation, within a topic modelling approach. The result is a weakly-supervised Aspect Based Sentiment Analysis (ABSA) system that classifies texts by domain aspect and sentiment polarity that can be easily applied to corpora in different languages and domains. The chapter is structured as follows. Section 6.1 introduces and motivates the work presented in the chapter. Section 6.2 presents a brief description of related work. Section 6.3 describes the proposed almost unsupervised ABSA system. Section 6.4 shows the experimental results and evaluation. Finally, section 6.5 presents some concluding remarks.

6.1 Introduction

The vast amount of digital content, generated every day in countless websites and social networks, keeps growing and requires automated ways to be handled and classified. In the previous chapter, we have described a method to calculate a sentiment polarity value for words contained at customer reviews of a given domain, building a sentiment lexicon. This was done using just a seed word represention a positive opinion, and another word representing a

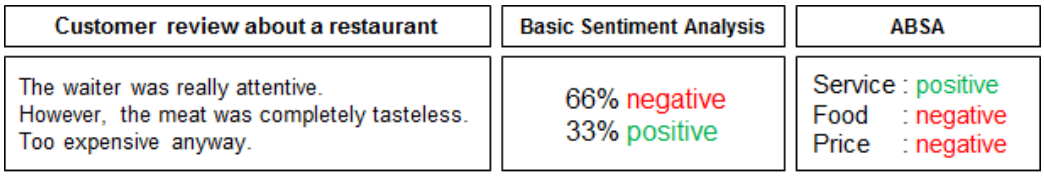


Figure 6.1: Example of classical Sentiment Analysis vs. Aspect Based Sentiment Analysis.

negative one, so the method could be readily applied to any language and domain with a minimal adaptation effort. The next step is to classify customer reviews into domain aspects, which combined with the polarity classification, results in an Aspect Based Sentiment Analysis system.

Aspect Based Sentiment Analysis (ABSA) (Liu, 2012) refers to the systems that determine the opinions or sentiments expressed on different features or aspects of the products/services under evaluation (e.g. *battery* or *performance* for a laptop). An ABSA system should be capable of classifying each opinion according to the aspects relevant for each domain in addition to classifying its sentiment polarity (usually positive, negative or neutral), as depicted in figure 6.1. The figure shows an example of a possible customer review about a restaurant, composed of three sentences containing opinions about different domain aspects: *service*, *food*, *price*. A basic, document-level, sentiment analysis would only reveal the overall sentiment. Whilst this information is useful, it still leaves unanswered questions such as *"what is causing the dissatisfaction?"* or *"what do customers value most?"*. An ABSA system would help to disclose that, in this case, the dissatisfaction comes from the quality/price relation of the offered food, while the service is doing its job (i.e. should the restaurant owner raise the salary of the waitstaff and fire the chef, maybe?).

Many ABSA systems make use of manually labelled data and language specific resources for training on a particular domain and for a particular language (Pontiki et al., 2014, 2015, 2016). This is the case of deep-learning-based systems, that provide very good performance but require a significant amount of labelled data for training (Chen et al., 2017; Araque et al., 2017).

Besides, weakly-supervised systems do not require labelled data for training, but they usually need some language specific resources, such as carefully curated lists of seed words or language dependent tools to preprocess the in-

put, (Lin et al., 2011; Jo and Oh, 2011; Kim et al., 2013). In addition, most of these works only report results for English or on a single domain.

In this chapter, we describe an almost unsupervised system for multilingual and multidomain ABSA. The system works leveraging large quantities of unlabelled textual data with a minimal set of initial seed words. We refer to it as W2VLDA. Figure 6.2 shows a schema of how a customer review corpus gets modelled by W2VLDA.

Imagine the following scenario. The owners of a famous restaurant want to monitor the opinion of their costumers with respect to a set of domain aspects. In particular, they want to know the opinion about its food, service, price, ambience, location, etc. The input of W2VLDA is a large corpus of customer reviews and a seed aspect term per domain aspect they want to monitor (for instance, *chicken* for the aspect *food*, *service* for the aspect *service*, etc.). By default W2VLDA also uses a single positive and a negative word as polarity seeds (for instance, *excellent* and *horrible*). With this input for every selected domain aspect, W2VLDA produces two main outputs. First, a weighted list of aspect terms per domain aspect (for instance, *chicken*, *salad*, *burger*, etc. for the aspect *food*), a weighted list of positive words (*tasty*, *yummy*, *homemade*, etc.) and a weighted list of negative words (*soggy*, *tasteless*, *burnt*, etc.). Thus, the proposed system performs at a word level three subtasks simultaneously: aspect classification, aspect-term/opinion-word separation, and sentiment polarity classification. Second, W2VLDA also produces a weighted list of sentences for every selected domain aspect and sentiment polarity.

The system is based on a topic modelling approach combined with continuous word embeddings and a Maximum Entropy (MaxEnt) classifier. It runs over an unlabelled corpus of the target language and domain just by defining the desired aspects with a single seed-word per domain aspect. We show results for different domains (restaurants, hotels, electronic devices) and languages (English, Spanish, French and Dutch). We compare its performance with other topic modelling based approaches, and we evaluate the performance of this approach on the SemEval2016 task 5 dataset, which provides a manually labelled set of restaurant reviews for several languages, including English, Spanish, French and Dutch.

The contributions of the work presented in this chapter are the minimal need of supervision (just one seed word per aspect/polarity) to perform ABSA over any unlabelled corpus of customer reviews. The lack of language

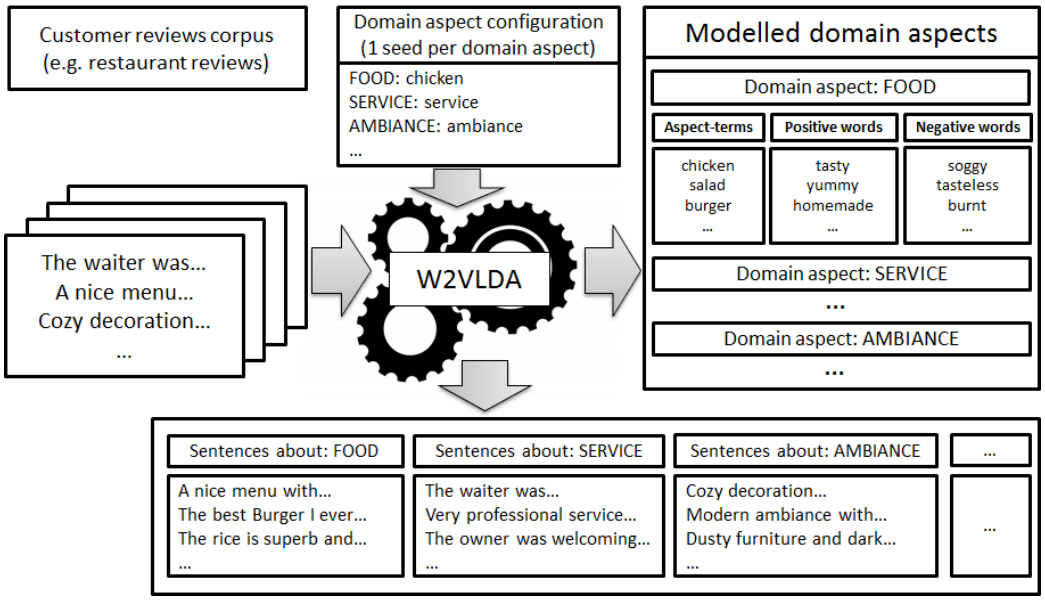


Figure 6.2: An schema of W2VLDA, with an unlabelled corpus as input and the modelled domain aspects and sentences as output.

or domain specific requirements allows the system to be readily used for other languages and domains. Another contribution is the automatic separation of the words into aspect terms, positive words and negative words per domain aspect allowing to characterise each domain aspect easily.

Note that in this chapter we sometimes use the expressions *domain topics* and *domain aspects* to refer to the same concept. Using the usual topic modelling terminology to explain the process, the outcomes are called *topics*, but those resulting topics in our case are the domain aspects being modelled according to the provided configuration.

6.2 Related approaches

During the last decade the research community has addressed the problem of analysing user opinions, particularly focusing on online customer reviews (Liu et al., 2012; Chen et al., 2014). The problem of customer opinion analysis can be divided into several subtasks, such as detecting the aspect (aspect

classification) and detecting the opinion about the aspect of the product being evaluated.

A common approach in the literature is to identify frequent nouns, lexical patterns, dependency relations applying supervised machine learning approaches (Hu and Liu, 2004; Popescu and Etzioni, 2005; Blair-Goldensohn et al., 2008; Wu et al., 2009; Qiu et al., 2011). Some works focus on automatically deriving the most likely polarity for words, constructing a so-called sentiment lexicon (Mostafa, 2013). The typical approaches use different variants of bootstrapping or polarity propagation leveraging some base dictionaries and pre-existing linguistic resources (Rao and Ravichandran, 2009; Jijkoun et al., 2010; Huang et al., 2014).

A well-known unsupervised method for text modelling documents is Latent Dirichlet Allocation (LDA) (see section 2.7). LDA is a generative model introduced by (Blei et al., 2003) that quickly gained popularity because it is an unsupervised, flexible and extensible technique to model documents. LDA models documents as multinomial distributions of so-called topics. Topics are multinomial distributions of words over a fixed vocabulary. Topics can be interpreted as the categories from which each document is built-up. They can be used for several kinds of tasks, like dimensionality reduction or unsupervised clustering. Due to its flexibility, LDA has been extended and combined with other approaches, obtaining topic models that improve the resulting topics or that model additional information (Mcauliffe and Blei, 2008; Ramage et al., 2009).

Topic models have been applied to Sentiment Analysis to jointly model topics and sentiment of words (Lin et al., 2009, 2011; Jo and Oh, 2011; Lu et al., 2011; Kim et al., 2013; Alam et al., 2016). A usual way to guide a topic modelling process towards a particular objective is to bias the LDA hyperparameters using certain apriori information. In the case of modelling the polarity of the documents, it usually means using a carefully selected set of seed words. Our method follows this idea, but replaces the need for a carefully crafted list of language or domain polarity words by only a single domain independent positive word (e.g. *excellent* for English) and a single domain independent negative word (e.g. *horrible* for English).

In general, topics coming from a topic modelling approach are anonymous word distributions, requiring an additional step to map them to a meaningful domain category. This task requires a manual inspection by an expert or a mapping calculation to an existing resource (Bhatia et al., 2016). Our

approach relies on a minimal topic configuration to define the topics for the target domain the user wants to monitor. Thus, the resulting topics match the ones defined initially by the user. This is done by leveraging semantic word similarities to guide the topic modelling towards the defined topics. This semantic word similarity is obtained using continuous word embeddings over the domain words. Continuous word embeddings are known for capturing semantic regularities of words (Mikolov et al., 2013a; Collobert and Weston, 2008). Some works have made use of this fact to improve the resulting topics (Das et al., 2015; Nguyen et al., 2015; Qiang et al., 2016), but their objective is to improve the unsupervised modelling of a corpus instead of guiding the model towards a predefined set of topics. There are works that exploit word embeddings in a supervised machine learning setting to perform sentiment analysis (Tang et al., 2014b; Giatsoglou et al., 2017).

Some authors have also attempted an automatic aspect-term/opinion-word separation within the topic modelling process (Zhao et al., 2010a; Mukherjee and Liu, 2012). Aspect terms are the words that are used to speak about the aspect being evaluated (e.g. *waiter* or *waitstaff* when speaking about the *service* of a restaurant). Moreover, opinion words express the sentiment about an aspect, such as *attentive* or *terrible*. The separation of these two kinds of words might be useful because it eases the interpretation of the resulting topics, and the sentiment classification can be focused on the opinion-words which are more likely to bear sentiment information. Zhao et al. (2010a) attempted this separation training a supervised classifier on a small manually labelled dataset and using Part-of-Speech tagging. Mukherjee and Liu (2012) elaborated on this idea trying a similar approach but substituting the manually labelled dataset with an existing lexicon of opinion words for English. Instead, we apply Brown clustering (Brown et al., 1992) to a set of training instances from an unlabelled corpus in order to train an aspect-term/opinion-word classifier that is later integrated into the topic modelling process. Following this approach no additional language-dependent resources are required, and the full process could be applied to any language and domain.

In summary, combining topic modelling, continuous word embeddings and a minimal domain aspects definition, our proposed approach can model customer reviews in different languages and domains performing three subtasks at the same time: aspect classification, sentiment classification and aspect-terms and opinion-words separation.

Aspect or Polarity	Seeds (English)	Seeds (Spanish)	Seeds (French)
food	chicken	pollo	nourriture
service	service	servicio	employés
ambience	ambience	ambiente	ambiance
drinks	drinks	bebidas	boissons
location	location	ubicación	emplacement
positives	excellent	excelente	excellent
negatives	horrible	horrible	épouvantable

Table 6.1: Example of seed words (one per aspect) used to monitor certain domain aspects of restaurant reviews in several languages, including the general polarity seeds.

6.3 System description

The main objective of W2VLDA system is to perform the three main tasks (to classify domain aspects, opinions and their polarity) of Aspect Based Sentiment Analysis at the same time. That is, to classify pieces of text into a predefined set of domain aspects and classify their sentiment polarity as positive or negative. In addition, our system separates opinion words from aspect term without requiring additional resources or supervision. The system at its core consists of an LDA-based topic model extended with additional variables, with biased topic modelling hyperparameters based on continuous word embeddings, and combined with unsupervised pre-trained classification model for aspect-term/opinion-word separation.

6.3.1 Aspects and sentiment polarity configuration

W2VLDA only requires a minimal domain aspects configuration per language and domain. This configuration consists of a single seed word to define each desired domain aspect, plus a single general positive seed word and a single general negative seed word valid for all domain aspects. This simple configuration is the only language and domain dependent information required by

W2VLDA [1]. Therefore, a quick dictionary translation of the few involved seeds should suffice to make the system work for another language or domain. Table 6.1 shows an example of domain aspects configuration for the restaurant domain in several languages.

6.3.2 Aspect-term and opinion-word separation

Part of the outcome of the system consists of the aspect-term/opinion-word separation into differentiated word classes. In order to achieve this separation without adding any language dependent tool or resource, the system uses Brown clusters (Brown et al., 1992) to model examples of aspect-terms and opinion-words and train a MaxEnt-based classification model (seed section 5.5 for more information). Brown clusters have been used as unsupervised features with good results in supervised Part-of-Speech tagging (Turian et al., 2010b) and Named Entity Recognition (Agerri and Rigau, 2016). Brown clusters are computed[2] from the unlabelled domain corpus with no additional supervision. The clusters are used as the features for the two-word context window [-2,+2] of each training example. The training instances are built using the occurrences of the aspects and sentiment seed words from the user-provided configuration, assuming that topic words are aspect-terms and polarity-words are opinion-words.

Figure 6.3 describes the process to obtain the classification model. First topic seed words and polarity seed words are used as gold aspect-terms and gold opinion-words respectively. Then the occurrences of these words are bootstrapped from the domain corpus and they are modelled according to their context window. Next, context words are replaced by their corresponding Brown cluster to build each training instance. Finally, a MaxEnt model is trained using these generated training instances.

We have experimented with a different number of Brown clusters (100, 200, 500, 1000 and 2000) but the impact of this parameter was almost negligible. The reported results have been obtained using 200 clusters.

A drawback of this approach is that every word in the vocabulary will be classified as aspect-term or as opinion-word. Obviously, there are words that

[1] A list of general stopwords for each target language is also necessary in order to obtain better results. We use the stopword lists from Apache Lucene.

[2] We use the Brown clustering implementation available at `https://github.com/koendeschacht/brown-cluster`

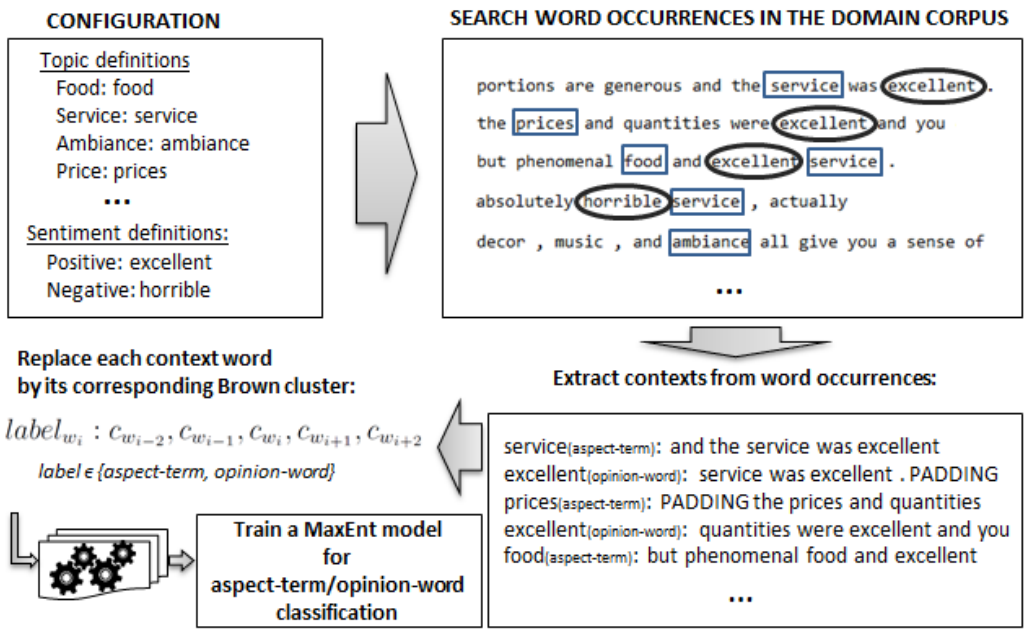
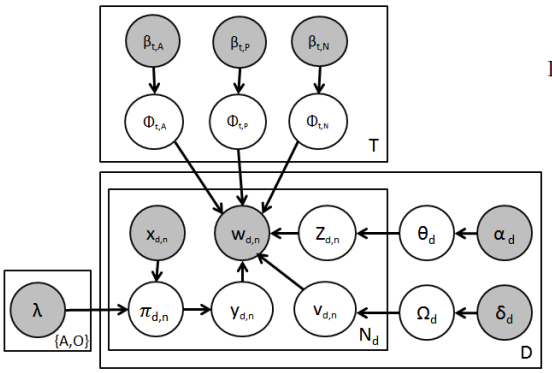


Figure 6.3: Process to train a MaxEnt model for aspect-term/opinion-word separation reusing the aspect and sentiment configuration.

do not belong to any of these categories. It would be interesting to have a third class (e.g. *"other"*), but it would require labelling training instances for that additional class, introducing a manual supervision that we want to keep to a minimum. We expect that the words that are not clearly aspect-terms or opinion-words will be spread across both classes, losing relevance during the topic modelling process.

6.3.3 Combining everything inside a topic model

The core of the proposed system consists of an LDA-based topic model, extended to include the aspect-term/opinion-word separation and the positive/negative separation for each topic. In addition, the aspect-term/opinion-word separation is guided by a pre-trained MaxEnt classifier as explained at section 6.3.2, while the domain aspects and polarity modelling are guided by biasing certain hyper-parameters according to the given domain aspects configuration. During this explanation we will refer to the configured domain



For each topic $t \in \{1..T\}$:
  sample $\phi_t^A \sim \text{Dirichlet}(\beta_t^A)$
  sample $\phi_t^P \sim \text{Dirichlet}(\beta_t^P)$
  sample $\phi_t^N \sim \text{Dirichlet}(\beta_t^N)$
For each document $d \in \{d_1..d_M\}$:
  sample $\theta_d \sim \text{Dirichlet}(\alpha_d)$
  sample $\Omega_d \sim \text{Dirichlet}(\delta_d)$
  For each word $w \in \{w_{d,1}..w_{d,N}\}$:
    sample $\pi_{d,n} \sim \text{MaxEnt}(\lambda, x_{w_{d,n}})$
    draw $z_{d,n} \sim \text{Multinomial}(\theta_d)$
    draw $y_{d,n} \sim \text{Bernoulli}(\pi_{d,n})$
    if $y_{d,n} = A$:
      sample $w_{d,n} \sim \text{Multinomial}(\phi^A_{z_{d,n}})$
    else if $y_{d,n} = O$:
      sample $v_{d,n} \sim \text{Bernoulli}(\Omega_d)$
      if $v_{d,n} = P$:
        sample $w_{d,n} \sim \text{Multinomial}(\phi^P_{z_{d,n}})$
      else if $v_{d,n} = N$:
        sample $w_{d,n} \sim \text{Multinomial}(\phi^N_{z_{d,n}})$

Figure 6.4: Proposed model in plate notation and its generative process algorithm.

aspects as *topics*, to follow a more conventional naming to explain the topic modelling process.

Figure 6.4 shows the proposed model in plate notation and the generative story modelled by the algorithm.

The generative hypothesis described by the model is the following. For each document d a distribution of topics, θ_d, is sampled from a Dirichlet distribution with parameter α_d, which is a vector with asymmetric topic priors for that document. Note that in this context each *document* correspond to individual sentences instead of full texts. Then for each word n in document d a topic value is drawn: $z_{d,n} \sim Multi(\theta_d)$, $z \in \{1..T\}$. Then a aspect-term/opinion switch variable is sampled: $y_{d,n} \sim Bernoulli(\pi_{d,n})$, $y \in \{A, O\}$. Depending on $y_{d,n}$, the word $w_{d,n}$ is emitted from the topic aspect terms distribution ($\phi_{z_{d,n},A}$) or else, a polarity value $v_{d,n}$ is sampled from Ω_d to choose if the word has to be drawn from $\phi_{z_{d,n},P}$ or $\phi_{z_{d,n},N}$ (positive and negative words respectively).

The model guides the topic and polarity modelling towards the desired values by biasing the hyper-parameters that govern the Dirichlet distributions from which the topics and words are sampled. In a standard LDA setting those hyper-parameters (commonly named α and β) are symmetric because no apriori information about the topic and word distributions is assumed. In

our model, these hyper-parameters are biased using a similarity calculation among the words of the domain corpus and the topic seed words of the initial configuration. This similarity measure is based on the cosine distance between the dense vector representation of the topic defining seeds and each word of the vocabulary. Such a dense vector representation of the words over a particular vocabulary, commonly referred as word embeddings, could be obtained using any distributional semantics approach, but in this work, we stick to the well-known word2vec (Mikolov et al., 2013a). Word embeddings are a very popular way of representing words as the input for a variety of machine learning techniques and are known for encoding interesting syntactic and semantic properties (Mikolov et al., 2013c). In this case, we exploit the semantic similarity among words that can be calculated using the cosine distance of the resulting words vectors (see sections 2.6 and 5.3 for more details). The similarity, sim, is the value between a word and a set of words (e.g. some topic defining seeds) and is calculated using 6.1.

$$\text{sim}(w, t) = \arg\max_{v \in t} \text{sim}(w, v) \tag{6.1}$$

Where w is any word found in the domain corpus, v is any of the seed words chosen to define topic t, and sim stands for the cosine distance between two word-vectors.

The α hyper-parameters control the topic probability distribution for each document as in the original LDA. But instead of having a single symmetric α value, each document has a biased α for each topic, based on semantic word similarity, as described in 6.2.

$$\alpha_{d,t} = \frac{\sum_{i}^{N_d} \text{sim}(w_{d,i}, t)}{\sum_{t'}^{T} \sum_{i}^{N_d} \text{sim}(w_{d,i}, t')} * \alpha_{base} \tag{6.2}$$

Moreover, the β hyper-parameters, which control the distribution of words for each topic, are calculated in a similar way, as shown in 6.3 and 6.4.

$$\beta_{t,w} = \text{sim}(w, t) * \beta_{base} \tag{6.3}$$

$$\beta_{q,w} = \text{sim}(w, q) * \beta_{base} \quad q \in \{P, N\} \tag{6.4}$$

Finally, the δ hyper-parameters control the polarity distribution for each document, and they are calculated for each document as shown in 6.5.

$$\delta_{d,q} = \frac{\sum_{i}^{N_d} \text{sim}(w_{d,i}, q)}{\sum_{q' \epsilon \{P,N\}} \sum_{i}^{N_d} \text{sim}(w_{d,i}, q')} * \delta_{base} \tag{6.5}$$

In the formulas $w_{d,i}$ is the *i-th* word of the document d, N_d is the number of words in that document, t is a topic from the set of defined topics T. Similarly q is a pre-defined polarity words set, P for positives and N for negatives (in our experiments P only contains *excellent* and N only contains *horrible* for English, or their equivalents for other languages).

α_{base}, β_{base} and δ_{base} are configurable hyper-parameters, analogous to the symmetric α and β in the original LDA model.

In addition to the bias of these hyper-parameters, the distribution π that governs each binary aspect-term/opinion-word switching variable, y, is set from the pre-trained aspect-term/opinion-word classifier probabilities applied to each word and its context features as described in section 6.3.2.

The posterior inference of the model is obtained via Gibbs sampling (Griffiths and Steyvers, 2004). Let $w_{d,n}$ be the n-th word of the d-th document, given the assignment of all other variables, its topic assignment $z_{d,n}$ is sampled using (6.6). Analogously, the aspect-term/opinion-word assignment $y_{d,n}$ and the polarity of the opinion-words, $v_{d,n}$ are sampled using (6.7) and (6.8) respectively.

$$p(z_{d,n} = t | z_{-d,n}, y_{-d,n}, v_{-d,n}, \cdot) \propto$$

$$\frac{n^{t,A}_{w_{d,n}} + \beta^{t,A}_{w_{d,n}}}{\sum_{v}^{V} n^{t,A}_{v} + \beta^{t,A}_{v}} \times \frac{n^{t,P}_{w_{d,n}} + \beta^{t,P}_{w_{d,n}}}{\sum_{v}^{V} n^{t,P}_{v} + \beta^{t,P}_{v}} \times \frac{n^{t,N}_{w_{d,n}} + \beta^{t,N}_{w_{d,n}}}{\sum_{v}^{V} n^{t,N}_{v} + \beta^{t,N}_{v}} \times (n_{d,t} + \alpha_{d,t}) \tag{6.6}$$

$$p(y_{d,n} = u | z_{d,n} = t, \cdot) \propto \frac{n^{t,u}_{w_{d,n}} + \beta^{t,u}_{w_{d,n}}}{\sum_{v}^{V} n^{t,u}_{v} + \beta^{t,u}_{v}} \times \frac{exp(\lambda_u \times x_{d,n})}{\sum_{u' \in \{A,O\}} exp(\lambda_{u'} * x_{d,n})} \tag{6.7}$$

$$p(v_{d,n} = q | z_{d,n} = t, \cdot) \propto \frac{n^{t,q}_{w_{d,n}} + \beta^{t,q}_{w_{d,n}}}{\sum_{v'}^{V} n^{t,q}_{v'_{d,n}} + \beta^{t,q}_{v'_{d,n}}} \times (n_{d,q} + \delta_{d,q}) \tag{6.8}$$

In these formulas, $n^{t,u}_{w_{d,n}}$ is the number of times the vocabulary term corresponding to $w_{d,n}$ has been assigned to topic t and word-type $u \in \{A, O\}$ (i.e. Aspect-terms or Opinion-words), $n_{d,t}$ is the number of words in the document d assigned to topic t, λ_u are the pre-trained aspect-term/opinion-word classifier model weights for word-type u and $x_{d,n}$ is the feature vector for $w_{d,n}$, composed by the Brown clusters of the context words. Analogously, $n^{t,q}_{w_{d,n}}$ is the number of times $w_{d,n}$ has been assigned to topic t and polarity $q \in \{P, N\}$ and $n_{d,q}$ is the number of words in the document d assigned to polarity q.

6.4 Evaluation and results

We evaluate W2VLDA for the three different subtasks that it performs: domain aspect classification, sentiment classification, and aspect-term/opinion-word separation. First, we compare it with other LDA-based methods. Then, we also evaluate the system in a multilingual ABSA dataset comparing its performance classifying domain aspects and sentiment with some supervised machine learning approaches trained on labelled data.

We show results for several datasets, demonstrating how the system works for different languages and domains just by changing the domain aspect configuration, composed of a single seed word per each desired domain aspect, language and domain.

For instance, table 6.2 shows some of the resulting words for several domains (restaurants and electronic devices) and domain aspects (food, service, ambience for restaurants, and warranty, design and price for electronic devices) for English customer reviews, including the automatic separation of aspect terms from positive and negative words per domain aspect. Table 6.3 shows the equivalent information for restaurants and hotel reviews in Spanish and French.

Likewise, table 6.4 shows examples of sentences classified under different domain aspects (food, service, ambience, location, price) for English restau-

Language:Domain	Domain aspect	Aspect-terms	Positive words	Negative words
English: restaurant reviews	Food	chicken, beef, pork, tuna, egg, onions, shrimp, curry	moist, goat, smoked, seared, roasted, red, crispy, tender	undercooked, dry, drenched, overcooked, soggy, chewy
	Service	staff, workers, employees, chefs, hostess, manager, owner	helpful, polite, knowledgeable, efficient, prompt, attentive	inattentive, rude, unfriendly, wearing, making, packed
	Ambiance	lighting, wall, interior, vibe, concept, ceilings, setting, decor	modern, beautiful, chic, nice, trendy, cozy, elegant, cool	bad, loud, uninspired, expensive, big, noisy, dark, cramped
English: electronic devices reviews	Warranty	warranty, support, repair, service, answer, center, policy	worked, lucky, owned, big, exchange, extended, longer	called, contact, broken, faulty, defective, expired, worthless
	Design	plastic, wheel, style, handle, pocket, design, exterior, wheels	adjustable, clean, good, versatile, attractive, lightweight, stylish	ugly, odd, awkward, tight, felt, weird, cute, stupid, flimsy
	Price	money, store, item, bucks, price, regret, deal, gift	paying, reasonable, penny, worth, delivered, stars, inexpensive,	disappointed, paid, cheaper, skeptical, pricey, overpriced

Table 6.2: Resulting domain aspect words distributions for English in two different domains. The domain aspects are automatically split into three different word distributions: aspect terms, positive words and negative words.

rant reviews. These examples sentences are the ones with higher a-posteriori probability for each domain aspect (i.e. its corresponding topic from the topic model estimated by the W2VLDA). Table 6.5 shows some examples of sentences in Spanish and French, after modelling a dataset of restaurant reviews and hotel reviews respectively.

6.4.1 Resources and experimental setting

In order to evaluate W2VLDA, we use the following resources. For domain aspect classification we use the dataset from Ganu et al. (2009) which contains restaurant reviews labelled with domain aspects (e.g. "food", "staff", "ambiance") for English. For sentiment classification, we use the Laptops

Language:Domain	Domain aspect	Aspect-terms	Positive words	Negative words
Spanish: restaurant reviews	Food	crema, tartar, ensaladas, sopa, brasa, patatas, salsas, alcachofas	caprese, sublime, destacar, casera, tierna, trufada, ahumada	aguada, mojar, congeladas, quemadas, fritos, rancias, reseco
	Service	camareros, camarero, maitre, dueño, encargado, metre	eficiente, eficaz, atentos, correcta, cercano, diligente	lento, pésimo, desagradable, prepotente, maleducado
	Ambiance	toques, atmósfera, material, mobiliario, bancos, modernidad	tranquilo, relajado, cálido, buena, amplio, luminoso, precioso	cutre, insoportable, pequeño, tanta, oscuro, poca, normalita
French: hotel reviews	Food	nourriture, sauce, produits, pâte, bouffe, saveur, risotto	raisonnable, michelin, excellents, merveilleuse, veritable, superbe	correcte, cuit, idem, passable, excessif, moleculaire, difficile
	Staff	personnel,ècoute, staff, gentillesse, concierge, membres	sympathique, attentionné, efficace, compètent, professionnel	dèplorable, antipathique, dèbordè, distant, constamment
	Ambiance	impression, couloirs, odeurs, personnages, hiver, escaliers	vieillissant, grand, rènovè, boone, typiquement, cosy, agrèablement	froide, vètuste, forte, incendie, bruyants, inexistante, complète

Table 6.3: Resulting domain aspect words distributions for two languages, Spanish and French, and for different domains. The domain aspects are automatically split into three different word distributions: aspect terms, positive words and negative words.

and DIGITAL-SLR dataset (Jo and Oh, 2011), consisting of English reviews of electronic products with their corresponding 5-star rating.

Additional multilingual experiments have been performed using the SemEval-2016 task 5 datasets (Pontiki et al., 2016). In particular, the restaurant reviews datasets which are labelled with domain-related categories and polarity for six languages.

In order to compute the topic model and the word embeddings, we have automatically gathered additional customer reviews about restaurants from some popular customer review websites. These unlabelled domain corpora consist of a few thousand restaurant reviews in English, Spanish, French and Dutch. The precise composition of all these datasets is explained later.

	Top sentences per domain aspect
Food	i ordered miso eggplant , stuffed chicken in mushroom , rock shrimp tempura hand roll , albacore sashimi with fried onion , spicy yellowtail cut roll .
	5 salad choices : asian shrimp salad , eggplant pepper salad , crab salad , asian chicken salad and thai cucumber salad !!
	we ordered baked mussels , fried soft shell crab , seafood salad and edamame .
	some must trys include the yang chow fried rice , beef satay , fried tofu stuffed with ground meat and salt and pepper ribs .
Service	we stayed for hours at our little patio table and the wait staff was attentive , friendly and helpful the entire time .
	the wait staff is top notch , very attentive to your needs and the hostess does a fantastic job of getting us seated in a timely fashion .
	it started with the hostess with a flat affect and moved on to the wait staff who needs to be trained to be wait staff .
	coffee keeps flowing as the friendly wait staff (my waitress kept calling me darling) dotes over you with a smile .
Ambience	clean interior that is well lit and cooled with ceiling fans during the hot summer months , nice subtle decor , comfy booths .
	it was nice to see horses tied to the fence , bikers , couples and family ' s enjoying a wonderful outdoor atmosphere .
	a nice cozy and comfortable environment along with amazing food and very nice servers
	the atmosphere is really laid back and fun - just a nice relaxing place to catch a buzz .
Location	tucked behind a parking garage and a plaza across from the performing arts center in scottsdale , this place was worth the hunt to find .
	while visiting the chandler area , i decided to check out grimaldi ' s off ray road .
	located in old town scottsdale , stumbled on this place after a san francisco giants game
	the scottsdale location has since closed down and there is now one in glendale .
Price	the selection was good and worth the price tag of $ 8 . 99 + $ 1 . 00 service fee to use a charge card .
	if you strategically go at 1015 , you can pay the breakfast price of $ 14 . 99 , eat a bit and then get lunch .
	but this is las vegas so the buffet price rather expensive low $ 20 for breakfast and lunch and high $ 20 for dinner and brunch
	if you lost all your money gambling and want to save a few bucks ... come to lunch between 10 : 30 am and 11am .

Table 6.4: Sentences sorted by domain aspect after W2VLDA modelling of a restaurant reviews corpus (i.e. domain aspect/topic with higher a-posteriori probability), using only a single seed word per configured aspect, *chicken*, *service*, *ambience*, *location* and *price* respectively. Sentences are lower-cased and white-space tokenised.

Lang: Domain	Domain aspect	Top sentences per domain aspect
Spanish: restaurant reviews	Food	probamos las croquetas melosas de jamón , milhoja de tomate y mozzarella con salsa de miel . paté de perdiz , tartar de bonito , steak tartar, paté de cabracho , brocheta de pollo y postres
	Service	el servicio a los clientes deja bastante que desear el trato es magnífico , camareros muy simpáticos y amables , un trato educado y exquisito
	Ambiance	cena agradable en un lugar de ambiente tranquilo , cosmopolita , con buena música el local es feo decorado como un bar de carretera en eeuu o un autobús
French: hotel reviews	Staff	service de qualitè et personnel extrêmement agreable , aux petits soins , disponible et serviable ! le personnel est rèactif , serviable , disponible , toujours prêt à rèpondre aux attentes des clients .
	Ambiance	l ' hotel est une attraction en soi , il y a un adventure park a l ' interieur , on se croirait a disneyland . le bâtiment a un certain charme , certaines tapisseries sont dèfraîchies , se sent londonien
	Location	a 5 minutes à pied de buckingham palace et saint james park , 10 à 15 minutes de big ben . hotel à 15 min de la gare à pied , à 15 min d ' oxford street , à 40 min du centre ville à pied .

Table 6.5: Examples of sentences with the highest posterior probability for several domains, domain aspects and languages other than English (Spanish and French). Sentences are lower-cased and white-space tokenised.

We use Word2Vec to compute the word embeddings that are used for the word similarity calculation. In particular, we use the Apache Spark MLlib [3] implementation with default parameters to compute the domain-based word embeddings.

Table 6.1 shows the domain aspect configuration used in the experiments for the domain of restaurants, just one word per domain aspect. Unless stated otherwise, the polarity seeds for every domain are *excellent* and *horrible* or their equivalents in other languages.

[3] `http://spark.apache.org/mllib/`

Method	Domain aspects											
	Staff			Food			Ambiance			Overall		
	Prec.	**Rec.**	**F-1**	**Prec.**	**Rec.**	**F-1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
LocLDA	**0.80**	0.59	0.68	0.90	0.65	0.75	0.60	0.68	0.64	0.77	0.64	0.69
ME-LDA	0.78	0.54	0.64	0.87	**0.79**	**0.83**	**0.77**	0.56	**0.65**	**0.81**	0.63	0.70
W2VLDA	0.61	**0.86**	**0.71**	**0.96**	0.69	0.81	0.55	**0.75**	0.63	0.70	**0.77**	**0.72**

Table 6.6: Comparison against other LDA-based approaches on restaurant domain.

The values for α_{base}, β_{base} and δ_{base} mentioned in 6.3.3, which play a similar role to α and β in the original LDA, are set to the values commonly recommended in the literature (Griffiths and Steyvers, 2004): 50/T for α_{base} and δ_{base} being T the number of topics, and 0.01 for β_{base}. The topic modelling process runs for 500 iterations in every experiment with a burn-in period of 100 iterations and a sampling lag of 10 iterations.

6.4.2 Comparison with other LDA based approaches

First, we evaluate W2VLDA in a domain aspect classification setting using the restaurant reviews dataset from Ganu et al. (2009). This dataset contains few thousand reviews from restaurants, classified into several categories but the authors report results only for the three main categories: *food*, *ambience* and *staff*. We compare W2VLDA against the results reported in (Zhao et al., 2010a) for two LDA-based approaches, LocLDA (Brody and Elhadad, 2010) and ME-LDA (Zhao et al., 2010a).

LocLDA and ME-LDA are LDA-based approaches, and thus, unsupervised. But the results reported in the experiment involved some supervision as described by Zhao et al. (2010a). First, the authors computed a topic model of 14 topics. Then the authors examine each topic and manually set a domain aspect label according to their judgement. W2VLDA provides already named topics at the end of the process using the configured domain aspects, so no manual topic inspection is required to label them. In order to assign a domain aspect label to a particular sentence, we use the resulting topic distribution for that sentence (θ_d) selecting the domain aspect label of the topic with highest posterior probability.

Table 6.6 shows the results of the experiment and the comparison with the other systems. Unlike the other two systems, despite not requiring a

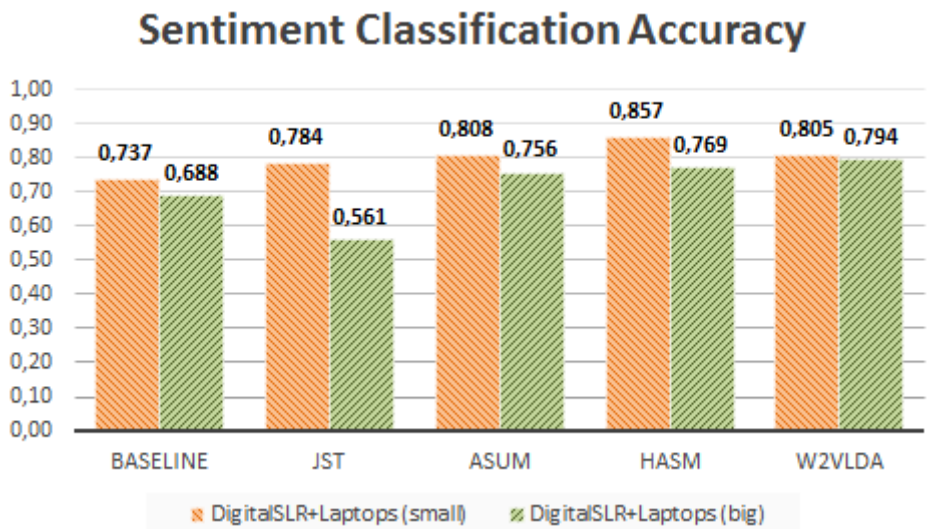


Figure 6.5: Sentiment classification accuracy comparison with other LDA based approaches in a electronic devices reviews dataset.

human labelling of the obtained topics, W2VLDA obtains slightly better overall results.

We also evaluate the ability of W2VLDA to assign correct polarities to customer reviews. We use the estimated polarity distribution of a sentence (Ω_d) to assign to a review the polarity with the highest probability. We compare our polarity classification results with respect to those from JST (Lin et al., 2011), ASUM (Jo and Oh, 2011) and HASM (Kim et al., 2013). The evaluation runs over the *Laptops and DIGITAL-SLRs* subset selected from the *Amazon Electronics dataset*[4]. As explained by Kim et al. (2013) two datasets are used, a *small* dataset containing 1000 reviews with 1 star rating (strong negative) and 1000 5 stars (strong positive), and a *large* dataset with additional 1000 reviews of 2 stars (negative) as well as 1000 reviews of 4 stars (positive). The baseline consists of a simple polarity seed word count, using the polarity seed words from Turney and Littman (2003), assigning to the sentence the polarity with the greatest proportion. As stated in previous sections, W2VLDA uses just a single polarity seed for each sentiment polarity, *excellent* and *horrible* respectively.

Figure 6.5 shows the result of this comparison. W2VLDA obtains com-

[4]Available at `http://uilab.kaist.ac.kr/research/WSDM11/`

```
<sentence id="1055910:1">
    <text>THe perfect spot.</text>
    <Opinions>
        <Opinion target="spot" category="RESTAURANT#GENERAL" polarity="positive" fr
    </Opinions>
</sentence>
<sentence id="1055910:2">
    <text>Food-awesome.</text>
    <Opinions>
        <Opinion target="Food" category="FOOD#QUALITY" polarity="positive" from="0"
    </Opinions>
</sentence>
<sentence id="1055910:3">
    <text>Service- friendly and attentive.</text>
    <Opinions>
        <Opinion target="Service" category="SERVICE#GENERAL" polarity="positive" fr
    </Opinions>
</sentence>
<sentence id="1055910:4">
    <text>Ambiance- relaxed and stylish.</text>
    <Opinions>
        <Opinion target="Ambiance" category="AMBIENCE#GENERAL" polarity="positive"
    </Opinions>
</sentence>
```

Figure 6.6: SemEval 2016 task 5 restaurants dataset example (for English).

parable results for the small dataset and better results for the big dataset despite using only a single seed word to define each polarity.

6.4.3 Multilingual evaluation on SemEval2016 dataset

We use the SemEval 2016 task 5 datasets (Pontiki et al., 2016) in order to perform a multilingual evaluation of W2VLDA. SemEval 2016 datasets consist of restaurant reviews in several languages. The reviews are split by sentence and labelled with the explicit aspect term mentions, the coarse-grained category they belong to, and the polarity for that category.

SemEval 2016 restaurants datasets are annotated for six coarse-grained categories or domain aspects: food, service, ambience, drinks, location, and restaurant. The last category, *restaurant* acts as a miscellaneous category that is used when the sentence does not refer to any other specific category but to the restaurant as a whole. Such an abstract concept cannot be represented by a seed word, so we omit this category from the evaluation. In order to avoid ambiguities and simplify the classification of each sentence, we only keep sentences with a single category label. Finally, since the domain aspects

	EN	ES	FR	NL
Food	486	364	370	374
Service	328	233	290	350
Ambience	110	145	98	117
Total	924	742	758	841

Table 6.7: SemEval 2016 dataset domain aspect distribution after filtering unwanted categories and sentence with more than one annotation.

drinks and *location* have very little representation in the datasets (less than the 5% of the instances), we keep only the three main domain aspects: *food*, *service* and *ambience*.

Table 6.7 and table 6.8 show the distribution of categories and polarities respectively for the resulting datasets, for four languages: English, Spanish, French and Dutch.

	EN	ES	FR	NL
Positive	551	417	300	405
Negative	326	273	413	369
Total	877	690	713	774

Table 6.8: SemEval 2016 dataset polarity distribution after filtering unwanted categories and sentence with more than one annotation.

Since W2VLDA is a topic modelling, it needs a reasonable amount of domain documents to build the statistical model. To cope with this requirement, we have implemented a script to automatically extract restaurant reviews of the required languages from an online customer reviews website. Due to copyright permissions, we cannot share these reviews, but table 6.9 shows the number of reviews used to feed the algorithm. The polarity mentioned on the table is based on the number of the stars from the 5-star rating (as usual, 1-2 stars meaning negative and 4-5 starts meaning positive). As it can be observed in the table, for some languages the script have not found an equal number of positive and negative reviews. We tried to compensate this fact with oversampling, to pair the number of positive and negative reviews before running the algorithm. Note that these polarities are just to get an

insight of the polarity distribution of the datasets, but they are not used for any sort of supervised training.

Restaurant customer reviews downloaded from a website				
	EN	**ES**	**FR**	**NL**
Positives (4 or 5 stars)	10000	10000	10000	10000
Negatives (1 or 2 stars)	10000	8400	5500	830
Total reviews	20000	18400	15500	10830

Table 6.9: Downloaded reviews distribution per language and polarity (using 5-star rate).

The evaluation experiment is done as follows. For each language, we use the downloaded reviews to run the algorithm. It includes calculating the domain word embeddings, Brown clusters and the topic model estimation. The domain aspect and polarity distributions, θ and Ω, are estimated for each of the sentences of the evaluation set using the generated model for each language. The domain aspect with the highest probability in the estimated domain aspect distribution for that sentence is assigned as the category label. Analogously, the polarity with the highest probability in the estimated polarity distribution for that sentence is assigned as the polarity label. Each assigned label is compared to its corresponding gold label, and the accuracy (ratio of correctly labelled examples) is calculated. The same process is followed to calculate the polarity classification accuracy.

The obtained accuracy is compared to several baselines. First, two supervised baselines are used. One is a Naive-Bayes classifier (NB), trained using the labelled sentences. The sentences are transformed to bag-of-words vectors with a vocabulary size of 80k words and normalised using tf-idf weights. The other supervised baseline is a Multilayer Perceptron algorithm (MLP), with two hidden layers, and the same tf-idf vector as input. Another baseline is the majority baseline, that shows the accuracy that can be obtained in the case of choosing the most frequent class. This is only to ensure that the datasets are not excessively unbalanced and the algorithms are really learning relevant information. Finally, the last baseline (W2VLDA_NO) is the same W2VLDA but removing the word-embeddings similarity mechanism to bias the topic modelling hyper-priors. Instead of using the word-embedding similarity to calculate a bias for every word, only the seed words receive a strong bias for their corresponding domain aspect or polarity.

Domain aspect classification				
	EN	**ES**	**FR**	**NL**
NB	0.610	0.570	0.575	0.61
MLP	0.657	0.617	0.588	0.590
Majority baseline	0.525	0.490	0.488	0.444
W2VLDA_NO	0.486	0.467	0.411	0.402
W2VLDA	**0.719**	**0.674**	**0.645**	**0.638**

Table 6.10: Comparison of domain aspect classification results against several baselines.

Polarity classification				
	EN	**ES**	**FR**	**NL**
NB	0.635	0.611	0.621	0.626
MLP	0.717	**0.680**	0.624	**0.639**
Majority baseline	0.628	0.604	0.579	0.523
W2VLDA_NO	0.558	0.587	0.528	0.515
W2VLDA	**0.736**	**0.680**	**0.643**	0.601

Table 6.11: Comparison of sentiment polarity classification results against several baselines.

Table 6.10 shows the evaluation results for the domain aspect classification (food, service, ambience, drinks and location). The scores for the supervised baselines are obtained using the average accuracy applying a 10-fold cross validation. The unsupervised methods (W2VLDA and W2VLDA_NO) are evaluated using all the labelled instances since they do not need any labelled data for training. W2VLDA obtains good results and outperforms the proposed baselines for all of the evaluated languages.

Table 6.11 shows the evaluation results for the polarity classification (positive and negative). W2VLDA obtains competitive results outperforming the proposed baselines except for Dutch. A possible explanation is that for Dutch there are not enough negative sentences in the downloaded reviews to obtain robust word-embeddings and a robust topic model with regard to polarity (see table 6.9). Studying which are the lower bounds of the required amount of data would be an interesting issue that we let for future research.

	Gold **opinion-words** probability mass		Gold **aspect-terms** probability mass assignment	
Clusters	Aspect-terms distrib.	Opinion-words distrib.	Aspect-terms distrib.	Opinion-words distrib.
None	0.231	0.210	0.326	0.333
Brown	**0.030**	0.487	**0.480**	0.120
Clark	0.131	0.359	0.380	0.250
Word2Vec	0.051	0.447	0.473	0.138
All clusters	0.037	**0.529**	0.456	**0.110**

Table 6.12: Gold aspect terms and opinion words probability mass distribution using different word clusters, in particular, Brown clusters (Brown et al., 1992), Clark clusters (Clark, 2000) and Word2Vec K-Means based clusters.

6.4.4 Aspect-term/Opinion-word separation evaluation

Finally we experiment with the aspect term and opinion word separation. As described in section 6.3.2, W2VLDA models the domain words into separated word distributions: aspect terms or opinion words.

Tables 6.2 and 6.3 show a visual result of what is the outcome of this separation. In order to obtain a quantitative measure of how this word separation is performing, we use the well-known Bing Liu's sentiment lexicon[5] as a gold-standard for opinion words assuming that words in the sentiment lexicon should be opinion words. From the 6,790 words contained in the lexicon, a 88% of them are present in our model vocabulary for restaurants. Analogously, we use the SemEval 2014 task 4 restaurants dataset[6] to obtain the list of manually annotated gold aspect terms for this domain. From the 1,212 manually annotated unique aspect terms, only 448 are entries in the model vocabulary (729 of the 1,212 gold aspect terms are composed terms like *selection of meats and seafoods*, and our model is dealing only with unigram terms for now).

The experiment assesses if after the topic-modelling process the gold opinion-words and the gold aspect-terms are assigned to their corresponding word distributions, opinion-word distributions and aspect-terms distribution respectively. The W2VLDA topic-modelling process is run as in the previous experiments, and the summed probability for gold-opinion-words and gold-aspect-terms in the resulting word distribution is accounted and averaged across all domain aspects.

[5] `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`
[6] `http://alt.qcri.org/semeval2014/task4/`

If the resulting topic model is correctly separating opinion words and aspect terms the accumulated probability mass of the gold-opinion-words in the opinion-words distribution should be noticeably higher than in the aspect terms distribution and vice-versa. Under an ideal and perfect separation the probability of a gold-opinion-word in the aspect terms distribution should be zero, and the probability of gold-aspect-terms in the opinion word distribution should also be zero. Besides, the probability of gold-opinion-words in the opinion words distribution would be high, and so would be the probability of gold-aspect-terms in the aspect terms distribution.

This evaluation provides an objective and repeatable measure of how the model is building word distributions that represent these two word-types, effectively separating aspect terms from opinion words.

Table 6.12 shows the evaluation results for the aspect-term/opinion-word separation. The table also contains results using alternative unsupervised word-clusters as features for the MaxEnt classifier that helps separating words. Apart from Brown clusters, the table shows results using Clark clusters[7] (Clark, 2000), and using K-Means clusters over Word2Vec word vectors. The same number of clusters is used for every variant.

The numbers on the table refer to the amount of accumulated probability in the multinomial distributions of words over the vocabulary of the model, averaged among all the topics. The entry *None* is a baseline, equivalent to removing the MaxEnt classifier from the process to check that effectively without the MaxEnt classification no useful separation happens.

According to these results, Brown clusters achieve the best performance in general. The combination of all cluster types provides a slight improvement. However, this improvement is not big enough to justify the increase in the computational cost derived using the three clusters types during the topic-modelling process instead of just one. In any case, the results show a strong bias to the appropriate distributions, indicating that the approach using seed words and unsupervised word-clusters to train a minimal classifier actually helps separating aspect terms from opinion words with no further resources.

Finally, another potentially relevant parameter is the number of clusters (Brown clusters in this case) used when modelling a text corpus. We repeat

[7]We have used the implementation from `https://github.com/ninjin/clark_pos_induction`

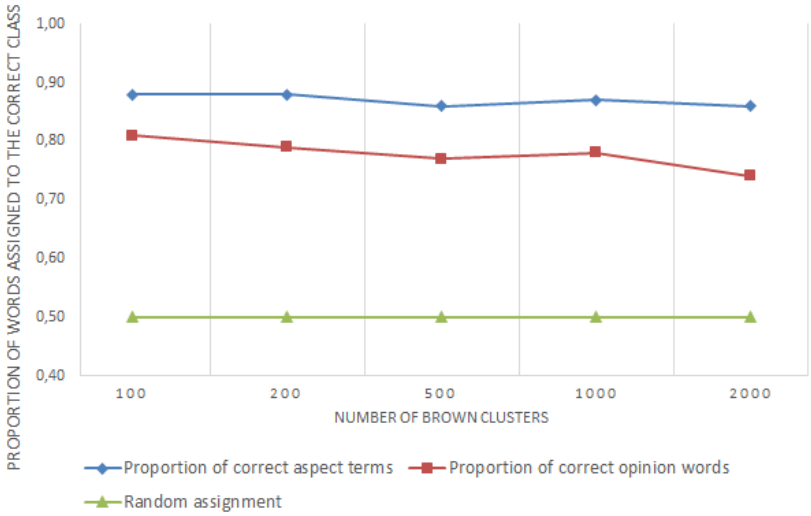


Figure 6.7: Result of aspect term and opinion word separation for English with each point indicating the proportion of aspect terms or opinion words that have been correctly classified.

the previous experiment with the gold-opinion words and gold-aspect terms, varying the number of Brown clusters.

We perform several experiments varying the number of Brown clusters involved in the process (see section 6.3.2) to evaluate if it has a noticeable impact on the word separation. Figure 6.7 shows the resulting proportions of correctly assigned aspect terms and opinion words for English. In general the correct proportions are high compared to a random assignment, which indicates that the aspect-term/opinion-word separation works correctly most of the times. Interestingly, aspect terms are better distinguished than opinion words.

6.5 Conclusions and future work

In this chapter, we have presented W2VLDA, a system that performs aspect and sentiment classification with almost no supervision and without the need of language or domain specific resources[8]. With the approach proposed in

[8]Implemented source code available at `https://bitbucket.org/aitor-garcia-p/w2vlda-last`

this chapter, we reuse the word embeddings based similarity within a more complex process, which results not only in a sentiment value estimation for each customer review but also a domain aspect classification.

More precisely the proposed system combines different unsupervised approaches, like word embeddings or Latent Dirichlet Allocation (LDA), to bootstrap information from a domain corpus. The only supervision required by the user is a single seed word per desired aspect and polarity. Because of that, the system can be applied to datasets of different languages and domains with almost no adaptation. The resulting topics and polarities are directly paired with the aspect names selected by the user at the beginning, so the output can be used to perform Aspect Based Sentiment Analysis. In addition, the system tries to separate automatically aspect terms and opinion words, providing more clear information and insight to the resulting domain aspects vocabulary. We evaluate W2VLDA for domain aspect and sentiment polarity classification using customer reviews of several domains and compare it against other LDA-based approaches. We also evaluate its performance using a subset of the multilingual SemEval 2016 task 5 ABSA dataset.

The results show that the proposed system works, and that even with this minimal supervision the performance is competitive with other existing methods that require more supervision or resources. In some cases the proposed system beats supervised baselines that make use of manually labelled data. The proposed system is evaluated for several languages and domains just by translating a few seed words (one per domain aspect and one per polarity), showing that it can be truly multilingual and multidomain with almost no adaptation effort.

However, there are quite a few possible improvements and further research that can be done based on these results. Our aim was to explore a set of methods to perform ABSA with the lowest possible dependency on language and domain based resources. We have obtained a system that fulfils this objective to a reasonable extent.

As future work, it would we interesting to include an automated way to deal with stop-words and other words that do not carry information for the ABSA task. A better-integrated handling of multi-word and negation expressions could also improve the results and the readability of the resulting aspect terms and opinion words.

Besides, the are more specialised word embeddings related to sentiment analysis (Rothe et al., 2016), and it would be interesting to study if different word embeddings bring improvements to the method keeping a minimal supervision. Even if some of those methods require some extra language dependent resources (e.g. lists of words) there may be a worthy trade-off in order to obtain a better performance or more fine-grained and precise analysis capabilities. This trade-off between the requirements and ease of adaptation of a system and the performance it obtains is an interesting subject to explore.

CONCLUSIONS AND FURTHER WORK

CHAPTER 7

Conclusions and further work

This last chapter presents a summary (section 7.1) of the objectives, research and conclusions reached in this thesis about sentiment analysis using weakly-supervised approaches. Section 7.2 lists the research papers that we have published in the process. Finally, section 7.3 outlines some possible future research lines and additional ideas for further exploration.

7.1 Summary

Sentiment analysis (also known as opinion mining (Pang and Lee, 2008)) deals with detecting and extracting subjective information, such as opinion and attitudes (Balahur, 2011). Sentiment analysis is one of the fastest growing research areas in computer science due to the high volume of on-line subjective content generation, with more than 5000 papers published in the last ten years (Mäntylä et al., 2016). Since the appearance of the Web 2.0 and social networks, this enormous amount of user reactions, feelings and opinions need to be constantly measured. This cannot be achieved but with the help of automated tools and algorithms.

Sentiment analysis can be applied to any type of human communication, such as speech, text, signs, etc. In this thesis we have described sentiment

analysis from the text analysis point of view, in particular, to understand and measure the sentiment polarity of on-line customer reviews.

Within sentiment analysis, we can make a clear distinction of some type of methods and approaches depending on their level of granularity. Systems and approaches that try to classify the sentiment polarity at aspect level are said to be performing Aspect Based Sentiment Analysis (ABSA) (Liu, 2012).

A domain aspect is a feature or facet of the entity being evaluated. For example, in the restaurants domain typical domain aspects are *"food"* or *"service"*. Domain aspects are important because they are relevant facets that can help aggregating content and discover the level of satisfaction towards the different facets of the evaluated entity. For example, if a hotel receives a report informing that customers are satisfied with the rooms but dissatisfied with the breakfast, there is a clear point of action. This situation is much more informative than simply receiving an overall positive/negative indicator that mixes all the potential satisfaction/dissatisfaction causes together.

Aspect terms are words (or multiwords) that refer to a particular aspect, for example *"burger"* for the aspect *"food"*, or *"waiter"* for the aspect *"service"*. In other words, domain aspects are coarse grained themes about the evaluated domain, and aspect terms are more fine-grained items that explicitly represent those themes in the analysed texts.

Another relevant element are the opinion words. Opinion words are those words (or composed expressions) that bear and express a sentiment, e.g. *excellent*, *expensive*, *dirty*, *wonderful*, *noisy*, etc. Detecting and assigning them a correct sentiment polarity value is critical to perform an accurate sentiment analysis.

This thesis starts with a comprehensive analysis of the state of the art with regard to sentiment analysis. This analysis includes the context and motivation that justify the interest that sentiment analysis has received from the industry and the academia during the last years. It also includes the most relevant approaches, challenges and trends in the automatic analysis of online opinions. Along the rest of the chapters of this thesis we have proposed and evaluated several weakly-supervised methods that deal with different tasks related to sentiment analysis.

We have described an approach aimed at automatically bootstrapping lists of aspect terms and opinion words for each analysed domain. This proposed approach is based on syntactic rules to build a graph and rank the

bootstrapped words, and we evaluated it participating in the SemEval2014 task 4 shared task.

We have also described a method to obtain a sentiment polarity value for domain opinion words. The proposed method is based on word embedding based similarity requiring only two seed words to work. We have evaluated and compared it against other well-known sentiment lexicons and sentiment lexicon generation approaches.

Finally, we have combined this into a system that performs Aspect Based Sentiment Analysis (ABSA). The resulting system is based on an extended Latent Dirichlet Allocation (LDA) model. This extended LDA model does not only separate aspect terms and opinion words but also estimates the domain aspect for each sentence from a pre-defined inventory of domain aspects, together with the sentiment polarity value.

The overall objective of the explored methods was to avoid, when possible, all language dependent data and tools. The described methods run over a unlabelled text corpus of the target domain, and just with the help of few seed words and a combination of unsupervised approaches, perform their task. Since no language dependent resources or manually labelled datasets are used, the resulting approaches are easily applied to other languages, and also to other domains.

We have evaluated the resulting approaches in datasets of several languages and domains and we have compared them with other existing methods. The results are competitive, taking into account the self-imposed restrictions about not using language dependent resources. In general, a fully supervised method making use of labelled data and language specific resources will obtain a better result for that particular language/domain, but it would not be able to work for texts written in a different language or from a domain for which there is not available labelled data. An interesting point of research would be to study to which extent the inclusion of language dependent resources is a good trade-off to improve the results or the analysis capabilities without sacrificing too much language and domain portability.

7.2 Publications

Below, we present in chronological order the list of publications related to the research described in this document:

- García-Pablos A., Cuadros M. and Rigau G. (2014). *V3: Unsupervised Generation of Domain Aspect Terms for Aspect Based Sentiment Analysis*. 8th International Workshop on Semantic Evaluation (SemEval-2014). Dublin, Ireland.

  The contributions of the previous publication are described in Chapter 4

- García-Pablos A., Cuadros M. and Rigau G. (2014). *Unsupervised Acquisition of Domain Aspect Terms for Aspect Based Opinion Mining*. Procesamiento del Lenguaje Natural 53, 121-128.

  The contributions of the previous publication are described in Chapter 4

- García-Pablos A., Cuadros M. and Rigau G. (2015). *V3: Unsupervised Aspect Based Sentiment Analysis for SemEval-2015 Task 12*. 9th International Workshop on Semantic Evaluation (SemEval-2015). Denver, Colorado.

  The contributions of the previous publication are described in Chapter 5

- García-Pablos A., Cuadros M. and Rigau G. (2015). *Unsupervised Word Polarity Tagging by Exploiting Continuous Word Representations*. Procesamiento del Lenguaje Natural 55, 127-134.

  The contributions of the previous publication are described in Chapter 5

- García-Pablos A., Cuadros M. and Rigau G. (2016). *A Comparison of Domain-based Word Polarity Estimation using different Word Embeddings*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portoroz (Slovenia).

  The contributions of the previous publication are described in Chapter 5

- García-Pablos A., Cuadros M. and Rigau G. *W2VLDA: Almost Unsupervised System for Aspect Based Sentiment Analysis*. Journal of Expert Systems with Applications. (Submitted)
  arXiv preprint arXiv:1705.07687

The contributions of the previous publication are described in Chapter 6

The following publications are not covered but are closely related to the topics described in this thesis:

- García-Pablos A., Lo Duca A., Cuadros M. Linaza M., and Marchetti A. (2016). *Correlating Languages and Sentiment Analysis on the basis of Text-based Reviews*. Information and Communication Technologies in Tourism 2016, 565-577.
- García-Pablos A., Cuadros M. and Linaza M. (2016). *Automatic Analysis of Textual Hotel Reviews*. Information Technology and Tourism 16 (1), 45-69.
- García-Pablos A., Cuadros M. and Linaza M. (2015). *OpeNER: Open Tools to Perform Natural Language Processing on Accommodation Reviews*. Information and Communication Technologies in Tourism 2015, 125-137.
- García-Pablos A., Cuadros M., Gaines S. and Rigau G. (2014). *OpeNER demo: Open Polarity Enhanced Named Entity Recognition*. Come Hack with OpeNER! Workshop Programme. Reykjavik, Iceland.
- García-Pablos A., Cuadros M., Rigau G. (2013). *OpeNER demo: Open Polarity Enhanced Named Entity recognition*. Proceeding of the 6th Language and Technology Conference (LTC) - demos. Poznań, Poland.
- García-Pablos A., Gaines, S., and Linaza, M. T. (2012). *A lexicon based sentiment analysis retrieval system for tourism domain*. e-Review of Tourism Research (eRTR), Vol. 10, No. 2.

7.3 Future work

Sentiment analysis is a very active and challenging research area (Strapparava, 2016). In this thesis, we have focused our attention on a very narrow set of approaches because our objective was to explore weakly supervised methods, aiming at achieving a high language and domain portability. This

self-imposed restriction has been very demanding, in the sense that many interesting and valuable tools and resources were not used.

As future work, we would like to continue exploring approaches that keep, to the possible extent, a high language and domain portability, but relaxing the restrictions about the used tools and resources.

In this thesis, we have used word embeddings. In particular, Word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), mainly because they are some of the most well-known word embedding generation approaches and they are completely unsupervised (in the sense that they are computed using only unlabelled data). Currently, there are many novel continuous word embedding generation or word embedding specialisation approaches (Rothe et al., 2016; Tang et al., 2014b), that might be worth trying. Those specialised word embeddings require some extra resources to be computed like labelled data or hand-crafted word lists. But if the use of more specific word embeddings improves the results such extra requirements could be a good trade-off.

In addition, with regard to the topic modelling approaches, there is a world of possibilities and variations to experiment with. LDA-based approaches offer a high flexibility to explore hypotheses about how a text corpus is composed and provides mechanisms to inject various kinds of information in the model estimation process. In fact, the topic model proposed in this thesis is just a first step that has required several attempts and reimplementations before it worked, and further extensions and improvements will probably be carried out.

Finally, the new trend in sentiment analysis (and basically in any machine-learning related area) is *deep learning* (Socher et al., 2013; Kim, 2014; Shin et al., 2016; Qian et al., 2017). Deep learning is, in essence, a buzzword to name the heavy use of big and stacked architectures of artificial neural networks, that leverage the computational power of modern machines. We have not explored deep learning in this thesis because deep learning approaches are supervised algorithms that, to our knowledge, require a large amount of labelled data to be trained. Nevertheless, with the use of deep learning based techniques, the research community is reaching new standards and obtaining very promising results. We would like to be part of it and explore these interesting techniques to find out if we can combine them somehow with other approaches to continue in the track of low-supervised methods.

Bibliography

Agarwal, S. and Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442.

Agerri, R. and Garcia, A. (2009). Q-WordNet : Extracting Polarity from WordNet Senses. *7th International Conference on Language Resources and Evaluation (LREC)*, pages 2300–2305.

Agerri, R. and Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63 – 82.

Alam, M. H., Ryu, W. J., and Lee, S. K. (2016). Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Anastasiou, D. (2010). *Idiom treatment experiments in machine translation*. Cambridge Scholars Publishing.

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. *Proceedings of the 26th International Conference on Machine Learning*, 29(26):997–1003.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2204.

Balahur, A. (2011). *Methods and resources for sentiment analysis in multilingual documents of different text types*. PhD thesis, Universidad de Alicante.

Balahur, A. and Jacquet, G. (2015). Sentiment analysis meets social media - challenges and solutions of the field in view of the current information sharing context. *Information Processing & Management*, 51(4):428 – 432.

Balahur, A. and Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management*, 51(4):547–556.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.

Bhatia, S., Lau, J. H., and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. *arXiv preprint arXiv:1612.05340*.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pages 339–348.

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

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, pages 1–8.

Bouchard, M., Joffres, K., and Frank, R. (2014). Preliminary analytical considerations in designing a terrorism and extremism online network extractor. In *Computational models of complex systems*, pages 171–184.

Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.

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.

Cardoso, P. M. D. and Roy, A. (2016). Sentiment lexicon creation using continuous latent space and neural networks. In *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 37–42.

Ceron, A. and Negri, F. (2015). Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle. *Rivista Italiana di Politiche Pubbliche*, 10(3):309–338.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, volume 31, pages 1–8.

Chen, T., Xu, R., He, Y., and Wang, X. (2015). Improving distributed representation of word sense via wordnet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 15–20.

Chen, T., Xu, R., He, Y., and Wang, X. (2017). Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230.

Chen, Z., Mukherjee, A., and Liu, B. (2014). Aspect extraction with automated prior knowledge learning. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 347–358.

Choi, D., Ko, B., Kim, H., and Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, 38(1):16–21.

Clark, A. (2000). Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, volume 7, pages 91–94.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Cresci, S., D'Errico, A., Gazzé, D., Duca, A. L., Marchetti, A., and Tesconi, M. (2014). Tourpedia: a web application for sentiment visualization in tourism domain. In *Proceedings of The OpeNER Workshop in The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 18–21.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian lda for topic models with word embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, pages 795–804.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Deng, L. and Wiebe, J. (2015). Mpqa 3.0: An entity/event-level sentiment corpus. In *Human Language Technologies - North American Chapter of the Association for Computational Linguistics*, pages 1323–1328.

Dennis III, S. Y. (1991). On the hyper-dirichlet type 1 and hyper-liouville distributions. *Communications in Statistics-Theory and Methods*, 20(12):4069–4081.

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., Deerwester, S., et al. (1995). Latent semantic indexing. *NIST Special Publication SP*, pages 219–219.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of 5th International Conference on Language Resources and Evaluation*, pages 417–422.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110.

Fei, G., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R. (2012). A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In *24th international conference on computational linguistics*, volume 2, pages 309–318.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., and Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. In *International Conference on Social Informatics*, pages 22–39.

Finlayson, M. A. and Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24.

Ganapathibhotla, M. and Liu, B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 241–248.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content. *WebDB*, pages 1–6.

García-Pablos, A., Cuadros, M., Gaines, S., and Rigau, G. (2013). Opener demo: Open polarity enhanced named entity recognition. In *Come Hack with OpeNER! Workshop Programme*, volume 501, page 12.

García-Pablos, A., Cuadros, M., and Linaza, M. T. (2015a). Opener: Open tools to perform natural language processing on accommodation reviews. In *Information and Communication Technologies in Tourism 2015*, pages 125–137. Springer.

García-Pablos, A., Cuadros, M., and Rigau, G. (2015b). Unsupervised word polarity tagging by exploiting continuous word representations. *Procesamiento del Lenguaje Natural*, 55:127–134.

Ghajar-Khosravi, S., Kwantes, P., Derbentseva, N., and Huey, L. (2016). Quantifying salient concepts discussed in social media content: An analysis of tweets posted by isis fangirls. *Journal of Terrorism Research*, 7(2):79–90.

Ghosh, A. and Veale, T. (2016). Fracking sarcasm using neural network. In *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 161–169.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Go, A., Huang, L., and Bhayani, R. (2009). Twitter sentiment analysis. *Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group*, 17:252.

Goikoetxea, J., Soroa, A., Agirre, E., and Donostia, B. C. (2015). Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1434–1439.

Griffiths, D. and Tenenbaum, M. (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Guerini, M., Gatti, L., and Turchi, M. (2013). Sentiment Analysis : How to Derive Prior Polarities from SentiWordNet. *Empirical Methods for Natural Language Processing*, pages 1259–1269.

Hai, Z., Chang, K., and Cong, G. (2012). One seed to find them all: mining opinion features via association. *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 255–264.

Hai, Z., Chang, K., and Kim, J.-j. (2011). Implicit feature identification via co-occurrence association rule mining. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 393–404.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages:181.

Hill, F., Cho, K., Jean, S., Devin, C., and Bengio, Y. (2014). Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448*.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Höpken, W., Fuchs, M., Menner, T., and Lexhagen, M. (2017). Sensing the online social sphere using a sentiment analytical approach. In *Analytics in Smart Tourism Design*, pages 129–146.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Association for the Advancement of Artificial Intelligence*, volume 4, pages 755–760.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882.

Huang, S., Niu, Z., and Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, 56:191–200.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1:95–105.

Inyang, I. F., Ozuomba, S., and Ezenkwu, C. P. (2017). Comparative analysis of mechanisms for categorization and moderation of user generated text contents on a social e-governance forum. *Mathematical and Software Engineering*, 3(1):78–86.

Jabreel, M., Moreno, A., and Huertas, A. (2017). Do local residents and visitors express the same sentiments on destinations through social media? In *Information and Communication Technologies in Tourism 2017*, pages 655–668.

Jakob, N. and Gurevych, I. (2010). Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045.

Ji, S., Yun, H., Yanardag, P., Matsushima, S., and Vishwanathan, S. (2015). Wordrank: Learning word embeddings via robust ranking. *arXiv preprint arXiv:1506.02761*.

Jijkoun, V., de Rijke, M., and Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594.

Jiménez-Zafra, S. M., Cámara, E. M., Valdivia, M. T. M., and González, M. D. M. (2015). Tratamiento de la negación en el análisis de opiniones en español. *Procesamiento de Lenguaje Natural*, 54:37–44.

Jiménez-Zafra, S. M., Martın-Valdivia, M. T., Molina-Gonzalez, M. D., and Urena-López, L. A. (2016a). Domain adaptation of polarity lexicon combining term frequency and bootstrapping. In *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 137–146.

Jiménez-Zafra, S. M., Martın-Valdivia, M. T., Urena-López, L. A., Martı, M. A., and Taulé, M. (2016b). Problematic Cases in the Annotation of Negation in Spanish. *ExProM 2016*, page 42.

Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international Association Computational Machinery - Special Interest Group on Information Retrieval conference (ACM-SIGIR)*, pages 244–251.

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.

Kanjo, E., Al-Husain, L., and Chamberlain, A. (2015). Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 19(7):1197–1212.

Kim, J.-K. and de Marneffe, M.-C. (2013). Deriving adjectival scales from continuous space word representations. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1625–1630.

Kim, S., Zhang, J., Chen, Z., Oh, A., and Liu, S. (2013). A Hierarchical Aspect-Sentiment Model for Online Reviews. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 526–533.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1746–1751.

Kiritchenko, S. and Mohammad, S. M. (2016). The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 43–52.

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. (2014). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *International Conference on Web and Social Media*, 11:538–541.

Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Lak, P. and Turetken, O. (2014). Star ratings versus sentiment analysis–a comparison of explicit and implicit measures of opinions. In *System*

Sciences (HICSS), 2014 47th Hawaii International Conference on, pages 796–805.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Li, Z., Zhang, M., Ma, S., Zhou, B., and Sun, Y. (2009). Automatic extraction for product feature words from comments on the web. In *Asia Information Retrieval Symposium*, pages 112–123.

Lin, C., He, Y., Everson, R., and Rüger, S. (2011). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24:1134–1145.

Lin, C., Road, N. P., and Ex, E. (2009). Joint Sentiment / Topic Model for Sentiment Analysis. *Cikm*, pages 375–384.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, K., Xu, L., and Zhao, J. (2012). Opinion target extraction using word-based translation model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1346–1356.

Long, C., Zhang, J., and Zhut, X. (2010). A review selection approach for accurate feature rating estimation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 766–774.

Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 81–88.

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Azpeitia, A., and Vossen, P. (2014). Generating polarity lexicons with wordnet propagation in five languages. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik*, pages 1155–1161.

Maks, I. and Vossen, P. (2013). Sentiment analysis of reviews: Should we analyze writer intentions or reader perceptions? In *Recent Advances in Natural Language Processing*, pages 415–419.

Manning, C. D., Schütze, H., et al. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Mäntylä, M. V., Graziotin, D., and Kuutila, M. (2016). The evolution of sentiment analysis-a review of research topics, venues, and top cited papers. *arXiv preprint arXiv:1612.01556*.

Martínez-Cámara, E., Cruz, F. L., Molina-González, M. D., Martín-Valdivia, M. T., Ortega, F. J., and Ureña-López, L. A. (2015). Improving Spanish Polarity Classification Combining Different Linguistic Resources. In *International Conference on Applications of Natural Language to Information Systems*, pages 234–245.

Mata, F. L. C., Ureña, L. A., and Sanchís, E. (2012). Extracción de opiniones sobre características: Un enfoque práctico adaptable al dominio. *Procesamiento del Lenguaje Natural*, 41:73–80.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, pages 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 746–751.

Moghaddam, S. and Ester, M. (2013). Opinion mining in online reviews: Recent trends. *Tutorial at WWW2013*.

Mohammad, S. M. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of North American chapter of Association for Computational Linguistics - Human Language Technologies*, pages 174–179.

Mohammad, S. M. and Bravo-Marquez, F. (2017). WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, pages 1–10.

Moilanen, K. and Pulman, S. (2007). Sentiment composition. In *Proceedings of Recent Advances of Natural Language Processing*, volume 7, pages 378–382.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M.-T., and Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *arXiv:1605.02019*, page 8.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

Mukherjee, A. and Liu, B. (2012). Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, pages 339–348.

Munezero, M., Montero, C. S., Kakkonen, T., Sutinen, E., Mozgovoy, M., and Klyuev, V. (2014). Automatic detection of antisocial behaviour in texts. *Informatica*, 38(1):3.

Narayanan, R., Liu, B., and Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 180–189.

Neidhardt, J., Rümmele, N., and Werthner, H. (2017). Predicting happiness: user interactions and sentiment analysis in an online travel forum. *Information Technology & Tourism*, pages 1–19.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134.

Niu, Z.-Y., Ji, D.-H., and Tan, C.-L. (2007). I2r: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 177–182.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Paulo-Santos, A., Ramos, C., and Marques, N. C. (2011). Determining the polarity of words through a common online dictionary. In *Portuguese Conference on Artificial Intelligence*, pages 649–663.

Pavlopoulos, I. (2014). *Aspect based sentiment analysis*. PhD thesis, Athens University of Economics and Business.

Pavlopoulos, J. and Androutsopoulos, I. (2014). Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. *Proceedings of Language Analysis and Social Media - European chapter of Association for Computational Linguistics*, pages 44–52.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP2014*, pages 1532–1543.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. *Computing attitude and affect in text: Theory and Applications*, 20:1–10.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste,

V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado*, pages 486–495.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pages 27–35.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346.

Popescu, A.-M. and Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28.

Qian, Q., Huang, M., and Zhu, X. (2017). Linguistically Regularized LSTMs for Sentiment Classification. *Association for the Advancement of Artificial Intelligence*, pages 1–9.

Qiang, J., Chen, P., Wang, T., and Wu, X. (2016). Topic Modeling over Short Texts by Incorporating Word Embeddings. *arXiv preprint arXiv: 1609.08496v1*, page 10.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *International Joint Conference on Artificial Intelligence*, volume 9, pages 1199–1204.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Quan, C. and Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272:16–28.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256.

Rana, T. A. and Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, 46(4):459–483.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Reed, C. (2012). Latent Dirichlet Allocation : Towards a Deeper Understanding. *Tutorial*, pages 1–13.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Rothe, S., Ebert, S., and Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

San Vicente, I., Agerri, R., and Rigau, G. (2014). Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 88–97.

San Vicente, I., Saralegi, X., and Agerri, R. (2015). Elixa: A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752.

Sarlan, A., Nadam, C., and Basri, S. (2014). Twitter sentiment analysis. In *Information Technology and Multimedia (ICIMU)*, pages 212–216.

Scanlon, J. R. and Gerber, M. S. (2014). Automatic detection of cyber-recruitment by violent extremists. *Security Informatics*, 3(1):1–10.

Schouten, K. and Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Schwartz, R., Reichart, R., and Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *CoNLL*, volume 2015, pages 258–267.

Shin, B., Lee, T., and Choi, J. D. (2016). Lexicon integrated cnn models with attention for sentiment analysis. *arXiv preprint arXiv:1610.06272*.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1631–1642.

Stone, P., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1):113–116.

Strapparava, C. (2016). Emotions and nlp: Future directions. In *WASSA 2016-Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis@ NAACL-HLT*, page 180.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(September 2010):267–307.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the Association for Computational Linguistics*, pages 1556–1566.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014a). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014b). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

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.

Tsur, O., Davidov, D., and Rappoport, A. (2010). Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*, pages 162–169.

Turian, J., Ratinov, L., and Bengio, Y. (2010a). Word Representations: A Simple and General Method for Semi-supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Turian, J., Ratinov, L., and Bengio, Y. (2010b). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Valitutti, A., Strapparava, C., and Stock, O. (2004). Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83.

van Son, C., van Erp, M., Fokkens, A., and Vossen, P. (2014). Hope and fear: interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference*, pages 26–31.

Wallace, B. C. and Kertz, L. (2014). Can cognitive scientists help computers recognize irony? In *The Annual Meeting of the Cognitive Science Society*, pages 49–50.

Wang, C., Medaglia, R., and Sæbø, Ø. (2016). Learning from e-government: an agenda for social media research in is. In *Pacific Asia Conference on Information Systems*, page 190.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1533–1541.

Xiong, S. (2016). Improving Twitter Sentiment Classification via Multi-Level Sentiment-Enriched Word Embeddings. *arXiv preprint arXiv:1611.00126*.

Yu, J., Zha, Z.-J., Wang, M., and Chua, T.-S. (2011). Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1496–1505.

Zhang, L. and Liu, B. (2014). Aspect and entity extraction for opinion mining. In *Data mining and knowledge discovery for big data*, pages 1–40.

Zhang, L., Liu, B., Lim, S. H., and O'Brien-Strain, E. (2010). Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 1462–1470.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010a). Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. *Computational Linguistics*, 16(October):56–65.

Zhao, Y., Qin, B., Hu, S., and Liu, T. (2010b). Generalizing syntactic structures for product attribute candidate extraction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 377–380.

Zhuang, L., Jing, F., and Zhu, X. (2006). Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., and Tolmie, P. (2015). Towards detecting rumours in social media. *arXiv preprint arXiv:1504.04712*.