Università
Ca'Foscari
Venezia

Master's Degree programme – Second Cycle
(*D.M. 270/2004*)
in International Management

Final Thesis

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Approaches and methods of Sentiment Analysis applied on Dainese S.p.A.

Supervisors
Ch. Prof. Claudio Silvestri
Ch. Prof. Andrea Albarelli

Graduand
Chiara Zaggia
Matriculation Number 838714

Academic Year
2015 / 2016

A mia nonna Marcellina,
che mi ha insegnato
cosa significhi resilienza.

INDEX

INTRODUCTION **4**

CHAPTER I - THE IMPORTANCE OF BIG DATA: WHY WE NEED SENTIMENT ANALYSIS

1.1 BIG DATA 8
1.2 BIG DATA SOURCES 10
1.3 HOW TO STORE, HANDLE AND ANALYSE BIG DATA 12
1.4 WHY BIG DATA ARE USEFUL? 13
1.5 ROAD TO SENTIMENT ANALYSIS 14

CHAPTER II - HOW TO COLLECT DATA

2.1 SENTIMENT CLASSIFICATION ON DOCUMENT 16

2.1.1 METHOD FOR SENTIMENT CLASSIFICATION: SUPERVISED LEARNING 17

2.1.2 METHOD FOR SENTIMENT CLASSIFICATION: UNSUPERVISED LEARNING 19

2.1.3 CROSS-DOMAIN SENTIMENT CLASSIFICATION 21

2.2 SENTENCE SUBJECTIVITY AND SENTIMENT CLASSIFICATION 24

2.2.1 SUBJECTIVITY CLASSIFICATION 24

2.2.2 SENTIMENT CLASSIFICATION ON SENTENCE 25

CHAPTER III - HOW TO SELECT DATA

3.1 SENTIMENT CLASSIFICATION ON ASPECTS 28

3.1.1 SENTIMENT CLASSIFICATION: ASPECTS 29

3.1.2 OPINION AND COMPOSITIONAL SEMANTIC: BASIC RULES 30

3.1.3 ASPECT EXTRACTION 34

3.1.4 GATHERING ASPECTS INTO CATEGORIES 37

3.1.5 ENTITY, OPINION HOLDER AND TIME EXTRACTION 37

3.1.6 COREFERENCE RESOLUTION AND WORD SENSE DISAMBIGUATION 39

3.2 SENTIMENT LEXICON GENERATION 41

3.2.1 DICTIONARY-BASED APPROACH 41

3.2.2 DESIRABLE AND UNDESIRABLE FACTS 42

CHAPTER IV - HOW TO ANALYSE DATA

4.1 OPINION SUMMARIZATION 45
4.1.1 ASPECT-BASED OPINION SUMMARIZATION 45
4.1.2 CONTRASTIVE VIEW SUMMARIZATION 47
4.2 SENTIMENT LEXICON GENERATION 48
4.2.1 PROBLEM DEFINITION 48
4.2.2 IDENTIFYING COMPARATIVE SENTENCES 51
4.2.3 IDENTIFYING PREFERRED ENTITIES 51
4.2.4 QUALITY OF REVIEWS 53

CHAPTER V - CASE STUDY: DAINESE S.P.A

5.1 HISTORY 56
5.1.1 DAINESE IN NUMBERS 58
5.1.2 DAINESE MEANS INNOVATION 59
5.1.3 WHY SENTIMENT ANALYSIS 61
5.2 REVZILLA 61
5.2.1 REVZILLA.COM'S STRUCTURE 63
5.3 MAIN COMPETITORS 64
5.3.1 ALPINESTARS 64
5.3.2 REV'IT! 64

CHAPTER VI - A MODEL FOR DAINESE S.P.A.

6.1 HOW TO COLLECT DATA 66
6.2 HOW TO SELECT DATA 70
6.2.1 THE PROCESS 70
6.2.2 THE DATA 73
6.3 HOW TO ANALYSE DATA 75
6.4 RESULTS 75
6.4.1 DAINESE PRODUCTS 75
6.4.2 ALPINESTARS PRODUCTS 78
6.4.3 REV'IT! PRODUCTS 81
6.4.4 DAINESE VERSUS COMPETITORS 82
CONCLUSIONS **85**
BIBLIOGRAPHY **87**
WEBLIOGRAPHY **87**

INTRODUCTION

In recent years, internet faced deep transformations thanks to its diffusion. Web 2.0 is the current state of online technology as it compares to the early days of the Web, characterized by greater user interactivity and collaboration, more pervasive network connectivity and enhanced communication channels[1]. Web 2.0 leads to a myriad of technologies and applications that were not available in the past. Like many other ground-breaking innovations, Web 2.0 brings not only new opportunities but also new challenges. Before the World Wide Web, there was no easy access to information: if a person wanted to buy an item, they first asked friends and family in order to have different opinions. One of the most important elements of Web 2.0 is the huge amount of information that spontaneously users share through blogs, social networks, forums and supplier websites: this is called *User Generated Content* (UGC).

An increasing amount of customers are now often sharing opinions, experiences and feelings referred to a particular product. This extensive information has not only a crucial value for companies as feedback, but also for other customers who will be most likely influenced by those reviews. We can say that Web 2.0 improves the word of mouth: few years ago, each person was able to collect only a small number of opinions, but now, the online world opens up to an improved concept of information sharing. Now the customers, in order to make up their mind whether or not buying a product, they first look it up on the internet.

For a company, it may no longer be necessary to conduct surveys, organize focus groups or employ external consultants in order to get consumers' opinions on products and competitors. Those could be considered outdated models, because the users-generated content (UGC) can provide them with such knowledge in a cheaper and faster way. For these reasons, more and more companies prefer investing on computational analysis, thus they will be able to better monitor the customer perception of their brands and articles. Companies always strive hard to satisfy customers' needs: the usage of Web 2.0 through a variety of tools is supporting companies in understanding what consumers think about their products, how they perceive them and what its target needs.

[1] Source: http://whatis.techtarget.com/definition/Web-20-or-Web-2

However, finding all the sources where the customers´ reviews are shared and monitoring all of them can be a challenging task. There is a large amount of websites where a costumer can share their review, and each website consists on a wide range of opinions. Furthermore, opinions are frequently hidden in longer texts in blogs and social media, often eluding the main focus. It is a challenging task to handle the whole process, due to the huge amount of data that needs to be processed, the so-called Big Data. Big Data can be defined as a database that can be built by a company collecting a huge number of information in terms of volume and variety. In order to use in an effective way such database, companies need automated systems and methods. In this way enterprises will be able to recognize relevant sources, collects all the significant sentences, read and summarize them, organize and analyse consumers' feedback. The Sentiment analysis, also known as opinion mining, was created to tackle this issue.

A crucial piece of data used in Sentiment analysis is textual information. Textual information is the complex amount of documents that users share online in blog, forum and social media. Online textual information can be broadly categorized into two main groups: facts and opinions. Facts are objective expressions about entities (for example products, politicians, brands, etc.) and their properties (all the aspects that constitutes entities; e.g. the entity is a jacket, its properties are the colour, collar, etc.). Opinions are usually subjective expressions describing people's feelings towards entities, events and their properties. Much of the existing research on textual information has been focused on mining and retrieval of facts, e.g., information retrieval, Web search, text classification, text clustering and many other text mining and natural language processing tasks (for definition see below). Classifying automatically a text document can be very challenging because of the complexity of the syntax: words have many aspects and meanings and they could be misleading. Opinions are playing a critical role in the decision making process, both for individuals and for organizations. Thanks to internet, opinions are also expressed in online reviews and blogs; this can allow companies to read them. Only recently researchers are investing in understand them through the development of new methods and tools.

Thanks to the fact that Big Data analysis is widely developed in several fields, sentiment analysis has flourished in the last sixteen years. Managing this complex source of reviews leads to another issue, the Natural Language Processing (NLP) challenge. We can describe the Natural Language Processing as an automated process

that reads all the information expressed by natural languages through a calculator. A natural language is what people spontaneously learn, like their mother tongue or other commonly used languages. Currently we do not have an advanced technology that allows us to process this wide variety of information in a fully automated way. Therefore, Sentiment analysis is effective only when the algorithms are monitored by people who control and correct the mistakes.

In order to understand when and how the human contribution is needed, a general idea of an algorithm structure is needed. At the starting point, based on a large amount of data the object that functions as a target needs to be defined (for example: how consumers perceive Dainese Keira Jacket). Afterwards, entity categories (i.e. a unique entity, in this case *jacket*) and entity expressions (i.e. all the words used to describe the entity, in our case *Keira, Dainese, Dai, etc*.) are determined. Other important facets to be defined are aspect categories, a unique aspect of an entity (e.g. the jacket collar) and aspect expressions (i.e. all the words used to describe the aspect). It also need to be taken into consideration the time when the review was shared and the profile of the user, since those are important aspects to understand how significant are the reviews. These various steps, from the extraction of entities to the extraction of time, need to be automated through an algorithm: this is the only way to manage Big Data. At the end of the process, human efforts need to be made. They have to leave out spam or fake reviews, which need to be excluded since they do not represent the real target of these studies.

Sentiment analysis offers methods and tools to, for example, plan a marketing strategy from the beginning, to understand who are the company's customers, what is their perception of the brand and how company can penetrate better the market.

This thesis focuses on Sentiment Analysis and its application on the case study Dainese S.p.a. We build a method able to collect and classify reviews from the site Revzilla.com. This platform is widely used as a benchmark by motorbike companies, since a great number of drivers contribute reporting their experience. The core of this thesis will be the summarization and the analysis of what the algorithm extract.

Chapter 1 helps to understand better the meaning and the importance of Big Data. Then, the need for Sentiment Analysis in order to handle those data will be explained.

Chapter 2 gives an overview of how data from documents and sentences can be collected, describing different methods and approaches reported in literature.

Chapter 3 describes various methods to select data, in order to decrease the chances of errors. This section includes a focus on which words are worth to be selected and how to select them.

Chapter 4 describes available approaches to summarize opinions, how to understand comparative opinions (comparative opinions are sentences that compare two entities based on their shared characteristics) and how to measure reviews quality.

Chapter 5 presents Dainese S.p.A. We will briefly introduce its history, evolution, market position and the role of competitors in its niche. Then, we introduce Revzilla.com and how its analysis can improve Dainese business.

In chapter 6 the method of choice for sentiment analysis is described: how it was developed and its application. In the end, the data collected is summarized and analyzed in order to find the key words linked to Dainese, through the classification of customer reviews. This will lead to a better understanding of our target group, thus improving the quality of information as starting point for setting up Dainese's new marketing campaign.

Finally, we describe the conclusion and the future scenarios that the analysis could have anticipated.

CHAPTER I

THE IMPORTANCE OF BIG DATA: WHY WE NEED SENTIMENT ANALYSIS

1.1 BIG DATA

Nowadays our lives are deeply influenced by a new phenomenon, called Web 2.0. Web 2.0 is the state of online technology as, characterized by greater user interactivity and collaboration, more pervasive network connectivity and enhanced communication channels[2]. In these years, we have created a new lexicon in order to define all the new aspects born with Web 2.0. This new lexicon comprehends words like *big data, social network, cloud, etc.* Our language has developed and has changed because of the new technologies like smartphones, smartwatches and tablets. These powerful and user friendly technologies make possible an interaction between people.

This interaction between people, produces a lot of information that we call Big Data. The Big Data is a huge and complex amount of information coming from different sources. Data can be founded in many resources and they may be numbers, document text or "likes". Social networks like Facebook and Twitter are the most important sources of information. But an user can give opinions through blogs and forums. Thus, we have to develop different methods in order to collect them.

By the end of 2016, global internet traffic will reach 88.4 exabytes (approximately one billion gygabytes) per month and this volume is growing rapidly[3]. Our society is changing faster than ever and after the era of digitalization now we are discovering the era of "datafication". Is called digitalization the conversion of analogy information–text, photographs, voice, etc.- to digital form with suitable electronic devices, such as a scanner or specialised computer chips. In this way information can be processed, stored, transmitted through digital circuits, equipments and networks[4]. The concept of datafication is connected to what Mayer-Schonberger wrote[5]: internet users produce a

[2] Source: http://whatis.techtarget.com/definition/Web-20-or-Web-2
[3] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html
[4] http://www.businessdictionary.com/definition/digitization.html
[5] Viktor Mayer-Schonberger "Big Data, a revolution that will transform how we live, work and think", 2013.

large amount of data and companies have to handle them through information technologies and human resources.

Big data are characterized also by other two aspects:

- The garbage: among the useful information there are some of them that must be deleted, in order to have a useful set of data to analyse. It is needed to connect information systems that are able to "purify" those data that provoke "rumour" and "redundancy"
- The analysis: in order to take decisions, considering the importance of those data, Big Data analysis is essential. Thus, it is necessary to develop technologies and abilities able to understand the meaning of data sets.

Viktor Mayer-Schonberger (Big Data, 2013), professor of Internet Governance at Oxford University, describes Big Data with three words: *more, messy* and *correlation*.

More: with this word, Mayer-Schonberger indicates the data availability. Nowadays, we can collect more complex data about more or less each topic.

Messy: Big data comprehend different kinds of information. Some of them could be misleading or incorrect.

Correlations: Big Data are useful to know how a person is influenced in the decision making process. It is possible to use big data to find statistical correlations among information connected by users' opinions and expressions; in fact, customers read the reviews before buying a product and these can influence them. After purchasing it, they leave a review where they link to other reviews.

Usually, literature refers to another definition of Big Data[6]. Big Data is described as "4Vs model": three of them derived from Doug Laney's model[7] and they are *Volume, Velocity* and *Variety*. The fourth is *Value*: this aspect got always more importance only in these years.

Volume: the possibility to collect data is unlimited and we need new databases in order to gather and summarize them. Big Data are so huge that new units of measurement like Terabyte (10^{12}byte), Exabyte (10^{18}byte) and Zettabyte (10^{21}byte) have been invented.

[6] Huang T., Lan L., Fang X., An P., Min J. "Promises and challenges of Big Data computing in Health Science", 2015.
[7] Laney D. "3D Data management: Controlling Data Value, Velocity and Variety", 2001.

In order to understand the dimension of one Zettabyte, we can think about the measure of all the visible Universe, which is approximately 880 zettabyte.

Velocity: data are shared each second and companies need to analyse them in real time if they want to have consistent results. Velocity is an important aspect during the analysing process, especially when data become rapidly obsolete. For example, if we analyse financial sector, the importance of time is extremely crucial because they can change suddenly.

Variety: variety refers to the multiplicity of data that can be collected, saved and analysed. Few years ago, we were not able to handle similar data, but now with Big Data and other advanced technologies we are able to achieve better performance. In fact, now we can handle data from different sources and with different structures. For instance, we can store in the same set structured and unstructured data. Structured data are those data that can be organized in tables or databases. Unstructured data, as photos, video, tweets, likes, are becoming more important: researchers estimates that in 2015, online consumers produced 5.38 Exabyte of unstructured data[8].

Value: Big Data comprehends a huge variety of information useful not only to understand consumers' opinion and behaviour, but also to improve production processes. In order to take right decisions, data must be consistent with real world and must be readable and kept separated from fake reviews.

1.2 BIG DATA SOURCES

A Big Data source is the place where data are extracted from. There are several kinds of sources and all of them have different characteristics. Each source can be structured (organized in a pre-defined model and standardized how they relate to one another) or unstructured, any kind of data that does not have a specific format like audio, videos and pictures. Sources can have different combination of volume, velocity and variety.
An important Software company (Kapow Software) studied Big Data sources and tried to make a list of the most important ones.

The main Big Data sources[9] are:

[8] Nair, Narayanan " Benefitting from Big Data- Leveraging Unstructured Data Capabilities for Competitive Advantage",
[9] https://datafloq.com/read/understanding-sources-big-data-infographic/338

- *Archives*: insurances, scanned documents, statements, paper archives, customer correspondence and files that are collected and digitalized by companies. They are unstructured data with low volume, velocity and variety.
- *Documents*: it comprehends documents in PDF, PPT, CSV, Word documents, etc. They are enough structured, an average variety, low velocity and volume.
- *Media*: images, video, photos and podcasts. They are structured, with an average variety and high velocity and volume.
- *Data storage*: all the systems that store data like SQL, NoSQL and file systems. They are structured data, with high volume.
- *Business Apps*: tools that companies uses in order to organize their work like CRM systems, Google Docs, portals and intranet. Variety, volume and velocity are on average.
- *Public Web*: web sites that collect data and information like IMdB, Wikipedia, etc. They have high volumes.
- *Social media*: Facebook, Twitter, Google+, LikedIn, SlideShare, You Tube and Instagram. Their main characteristic is the high velocity and high volumes. They are not structured.
- *Machine log Data*: all the file produced by machines like server systems, mobile app usages, applications logs, IRP, etc. High volume variety and velocity.
- *Sensor data*: medical devices, car sensors, road cameras, satellites and many other tools store a lot of information that we can call sensor data. High volume, velocity and variety characterize those data.

In Figure 1 the infographic presented helps us to understand the sources of Big Data. Each section represents a big data source. The rectangles have a different meaning depending on the colour: velocity is light blue, variety is green and volume is blue. These three variables have different levels, and are represented like this: low only one rectangle, medium two rectangles and high three rectangles. The small grey rectangle on the external circumference represents whether the source is structured or unstructured.

Figure 1 Big Data sources

Sources: https://dataflog.com/read/understanding-sources-big-data-infographic/338

1.3 HOW TO STORE, HANDLE AND ANALYSE BIG DATA

In a recent study from Zhejiang University of Hangzhou are summarized the steps useful to analyse Big Data[10].

1) *Problem definition*. In order to analyse Big Data, objective target to be achieved has to be settled. Usually, the goal is expressed by a question and the Big Data analysis' aim is to find the answer.
2) *Sources and methods*. Big Data can help us to find an answer to the problem definition, but we have to define the most useful sources where information have to be extracted from.
3) *Storage*. When we know the sources and the methods to extract Big Data, we have to collect them in a specific database.

[10] Huang T., Lan L., Fang X., An P., Min J. "Promises and Challenges of Big Data Computing in Health Sciences", 2015, ScienceDirect.

4) *Accuracy*. When we have enough data, we must check their quality and their precision in order to have consistent results.
5) *Statistics*. After clearing data, we can compute our statistical operations in order to have a readable result, useful to solve the problem that we have formulated in the first step.
6) *Data visualization*. Here researchers or data analysts build a report in order to present the results. The final report has to be clear and easily interpreted in order to allow a huge public to understand it.
7) *Assessment*. This final step helps people to evaluate the quality of analysis, considering if the process leads to useful results or not.

1.4 WHY BIG DATA ARE USEFUL?

Big Data is a new phenomenon which is not completely explored and it is difficult to forecast its evolutions. The wide application of Big Data varies from financial sector to consulting, from companies to telecommunications, from retail to political elections. Big Data can be collected from several contexts and the computational operations can be infinite.

Nowadays, Big Data are stored mainly for customer-centric strategies. The data are collected directly inside companies' systems, for instances, turnover, most visited selling points and top sold articles. However, recently, companies are interested to acquire information from the Web, especially from social networks. A research -called "Business opportunity: Big Data"[11]- presented by European Commission, reported which kind of opportunities IBM might have approaching to Big Data. Below the results:

- Improvement of *e-commerce*: it helps the company to build a stronger customer relationship.
- Big Data outsourcing and hosting: Big Data can be sold and analysis can be useful for other companies. As explained in one of the following chapters, Revzilla.com have a large amount of data that could be useful for Dainese.
- Big Data consulting. In order to manage and handle such Big Data, companies need specialists that analyse all the information in the web. The analysis lead to

[11] European Commission, "Business opportunity: Big Data", 2012.

create job opportunities and a new market, where consulting agencies can sell their report.

- Public Administration. Big Data can be applied to improve the life within the society: this information can lead to know better the segmentation of population in order to create customized services for everyone.
- Remote Monitoring of products in real time. Through internet we can have access to all the information about products even without seeing it. For example, if you send a gift to another Country, you can track your pack and know exactly where it is. With other technologies, you can know if a food is used before its expired date or how much gas all the people consume.

Big Data are crucial in several fields of our society, but some limits have to be overtaken. For instance, users share in social networks sensitive data that can be covered by privacy policies. This can obstacle Big Data analysis because law may not allow companies to handle those data. In other cases, management may forbid the access to some data in order to protect information. Furthermore, companies do not have qualified employees able to manage Big Data, or they do not have enough money to develop and implement tools.

1.5 ROAD TO SENTIMENT ANALYSIS

As we said before, Big Data have myriad of applications, but in this paper, we will focus on Big Data applied on Sentiment Analysis.

The aim is to understand better Sentiment Analysis, because it represents the most important phenomenon in terms of volume of information. Sentiment Analysis concerns the extraction of opinions from a document text available on the internet. This field is a hot topic for companies[12] and for this reason we chose to study in deep this particular application of Big Data.

Almost all companies have tools and information technologies that can elaborate non-textual data, but they do not have useful systems to store, summarize and analyse textual data. This data are increasingly being shared especially because of the growing selling trend of smartphones and because more and more people use social networks.

[12] http://www.forbes.com/sites/louiscolumbus/2015/03/22/56-of-enterprises-will-increase-their-investment-in-big-data-over-the-next-three-years/#1314b89b88b1

The next challenges for most of the companies consists of extracting this large amount of data that everyday online users produce. For instance, Unilever can use Big Data in order to know how consumers perceive its products, or insurance companies can discover which are the services that lead people to talk about in online forums, blogs and social networks. In general, every sector could take advantage of Sentiment Analysis[13].

[13] http://www.ninjamarketing.it/2016/05/06/web-marketing-cose-e/

CHAPTER II

HOW TO COLLECT DATA

2.1 SENTIMENT CLASSIFICATION ON DOCUMENTS

Sentiment Analysis is a discipline applied to online text documents. In order to understand how it works, we have to delineate the target of Sentiment Analysis as the following[14]: given an opinion document *d* that evaluates an entity, the aim of Sentiment Analysis is to extract the following five variables

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

Where e_i refers to the entity name, a_{ij} is the entity aspect, s_{ijkl} is the feeling connected to the aspect, h_k is the opinion holder and t_l is the time when an opinion holder shares their opinion[15].

There are two different problems depending on the type of feeling *s*:

- when the feeling *s* is expressed in a text document, the problem consists in categorizing it;
- when the feeling is expressed by a score or by a number – the so-called regression- the problem consists in measuring it.

The first case is difficult to manage because an opinion can be written in many different ways, using different words. In addition to that, is not unusual for a customer to use sarcasm. The problem is, therefore, how to recognize which words express negative feelings and which ones positive.

In the second case, we have to define which entity is rated and then understand how to interpret the results.

Since the second case is only partially related with Sentiment Analysis. In fact, an opinion can be classified as positive or negative, just making an average of all the scores. In this chapter, we will focus on how to find opinions in a document text.

[14] Bing Liu, "Sentiment Analysis and Opinion Mining", 2012.
[15] Bing Liu, "Sentiment Analysis and Opinion Mining", 2012.

In order to make the analysis easier, we assume that a document expresses one single opinion for a unique entity, without reporting other opinion holders' thoughts. This assumption allows us to explain how to extract opinions from documents, facing different type of expressions and sayings.

First of all, we will describe the most important methods for extracting opinions from documents: supervised learning method (that uses datasets recognizable thanks to the tag addition- this data are called labelled data) and unsupervised learning method (that has not a set of labelled data, but it identify the sentence structure). Then, we will try to understand how contexts can change meanings. In the end, we will give an overview of methods that extract opinions from sentences.

2.1.1 METHOD FOR SENTIMENT CLASSIFICATION: SUPERVISED LEARNING

We now describe the methods and approaches used to classify customer reviews based on the texts they shared.

Classifying an object (an item, a word, a document, etc.) means trying to cluster it in a group. This concept has as well to be used on text, allowing us to classify it. This classification helps us to separate positive and negative text. In each document, words are of course key features, especially the ones which express positive or negative trends. In "*Thumbs up?-Sentiment classification using machine learning techniques"*[16] was presented for the first time a method that analyses movie reviews. The researchers selected a vast list of positive and negative words and set an algorithm on these two groups of words. Thus, the algorithm is able to identify those words within reviews. The recognized words are then count. When a review has more negative words than positive ones, it is classified as a negative review.

Researchers who wrote *"Thumbs up?-Sentiment classification using machine learning techniques"* used a rudimental version of what we now call *supervised learning method*, deeply evolved in the future by other researchers.

[16] Pang, Bo, Lillian Lee, Vaithyanathan "Thumbs up? –Sentiment classification using machine learning techniques", 2002.

The aim of supervised learning method was to build a set of effective structures, like those listed by Liu (Liu, 2012)[17]:

- *Terms and their frequencies.* This feature was set to find a specific word and count how many times is written. In some cases, the position of the word can be taken into consideration because the position can change its meaning; for example, in the sentence "I do not want to buy this yogurt anymore", the meaning of *not* is different from the meaning of the same word in the sentence "My Nokia phones break up as often as not"
- *Part of speech.* The same word in different part of speech can have a different meaning: that is why is needed to add to each word a *part-of-speech tags* (POS). A POS tag must be codified if we want to classify words and determine their importance in the text. That is why we use *Penn Treebank POS tags*, as shown in Table 1. If every word is codified through a POS tag, we can easily recognize which words are more likely to be opinion words (e.g. adjectives) and which expressions describe product features (usually nouns). The POS tags are manually added to each word and after that, the researcher add also the orientation (positive, negative, neutral).

Table 1 Penn Treebank Part-Of-Speech (POS) tags

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote

Source: Bing Liu, Sentiment Analysis and Opinion Mining, 2012

[17] Bing Liu, "Sentiment Analysis and opinion mining", 2012.

- *Sentiment words and phrases.* In sentiment analysis, we should consider that all the words can be connected to a positive or negative feelings. For example, *good* and *wonderful* are positive sentiment words, instead *awful* and *bad* express negative sentiment opinions. There are also *sentiment phrases* and *idioms*.
- *Sentiment shifter.* There are some expressions used to change the opinion of a phrase, from positive to negative and vice versa. Negation words are the most important class of sentiment shifter (not, worse, worst, etc.). Not all the words that are considered shifter are used always as sentiment shifters: for example the "*not*" in the construction "*not only… but also*".
- *Syntactic dependency.* Researchers study words dependency (i.e. the relation that links two words frequently) based on features recognized by some tools. One of this tool is the parsing, a computational method that analyses a set of data in order to determine a structure within them.

One of the most important features of supervised learning method is the connection between words and tags; each words must have a tag in order to be recognizable. When the algorithm has codified all the words of our document text, there must be a person that assigns an orientation to the opinion words. Now the algorithm is able to recognize all the words and it can define if the document is positive or negative. The algorithm can be applied to other documents in order to analyse them. There must be always a person who control that every word has a tag.

This approach requires an important effort in terms of time, but it is effective and the outputs are reliable.

2.1.2 METHOD FOR SENTIMENT CLASSIFICATION: UNSUPERVISED LEARNING

One of the critical point of opinion classification is represented by words that express feelings. The classification of a text could be made not only by supervised learning method, but also by unsupervised learning method. The method developed by Turney in "*Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of review"* performs classification based on some fixed syntactic patterns

that are more likely to be used to express opinions. The syntactic patterns are composed of part-of-speech (POS) tags. Three steps[18] lead to the final algorithm:

- **Step one.** Two consecutive words are extracted only if they fit to one of the patterns present in Table 2

Table 2 Patters of POS tags for extracting two-words phrases

	FIRST WORD	**SECOND WORD**	**THIRD WORD (NOT EXTRACTED)**
1	ADJECTIVE	NOUN (SINGULAR, MASS OR PLURAL)	Anything
2	ADVERB (also comparative or superlative)	ADJECTIVE	Not noun (singular, mass or plural)
3	ADJECTIVE	ADJECTIVE	Not noun (singular, mass or plural)
4	NOUN (SINGULAR, MASS OR PLURAL)	ADJECTIVE	Not noun (singular, mass or plural)
5	ADVERB (also comparative or superlative)	VERB (base form, past, past participle, gerund)	Anything

Source: Bing Liu, Sentiment Analysis and Opinion Mining, 2012

Obviously, we use this table because adjectives, adverbs, comparative adverbs and superlative adverbs often express opinions. We cannot know what opinion they express only codifying their framework, but this is another step forward in order to analyze opinion.

- **Step 2.** It estimates the sentiment orientation (SO) of the extracted phrases using the *pointwise mutual information* (PMI) measure:

$$PMI\ (term_1, term_2) = \log_2\left(\frac{\Pr\ (term_1\ \Lambda\ term_2)}{\Pr(term_1)\ \Pr(term_2\ )}\right)$$

PMI measures the degree of statistical dependence between two terms. Here, $\Pr\ (term_1\ \Lambda\ term_2)$ is the actual co-occurrence probability of $term_1$ and $term_2$ and $\Pr(term_1)\ \Pr(term_2\ )$ is the co-occurrence probability of the two terms if they are statistically independent. The sentiment orientation (SO) of a

[18] Turney, "Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of review. In Proceedings of annual meeting of the Association of computational linguistic", 2002

phrase is computed based on its association with the positive reference word "excellent" and the negative reference word "poor":

$$SO(phrase) = PMI(phrase, \text{"excellent"}) - PMI(phrase, \text{"poor"})$$

The probabilities are calculated by issuing queries to a search engine and collecting the number of *hits*.

- **Step 3.** Given a review, the algorithm computes the average SO of all phrases in the review and classifies the review as positive if the average of SO is higher than zero, and negative if lower.

This method does not find specific words inside a review and does not recognize every single word with a POS tag (like supervised learning method), but it understands the syntactic form. This can be a reliable method in some context, for example, automotive[19], where the dependency between words is easy to understand. In another, it can be difficult to apply because language constructs are different depending on the domain: the same word could have a positive orientation in a context and a negative in another one.

2.1.3 CROSS- DOMAIN SENTIMENT CLASSIFICATION

In order to explain cross-domain sentiment classification, we need to introduce the definition of labeled data and of sentiment classifier. Labeled data are those data that we select as reference for our extraction in reviews. For example, in the supervised learning approach, as first step we define two sets of words (positive and negative); these two groups are labeled data.

A sentiment classifier is a tool that links a set of features to specific labeled information, in order to recognize the features inside text. For example, supervised learning and unsupervised learning are two classifiers because in the first case, it extracts words from a text recognized by the group of words previously settled, and in the second case, the classifier sets specific structure and finds it in the document texts.

Usually, when we export a sentiment classifier from one domain to another, it performs poorly because lexicon changes depending on the context. It is not easy to understand

[19] Bing Liu "Sentiment Analysis and Opinion Mining", 2012.

how to transfer sentiment classifier from one context to another, but some researchers tried to solve the problem. Here we present two approaches:

- The first method was developed by Aue and Gamon[20] and it uses a small amount of labeled data for the new domain;
- The second needs no labeled data for a new domain (Blitzer et al., 2007)

Aue and Gamon proposed the transfer of sentiment classifiers to new domains in absence of large amount of labeled data. They experiment it with four strategies[21]:

1. *All data.* This approach consists of applying the same classifier in different domains, combining labeled data from different domain. Note that the number of labeled data from each domain must be the same. Then, researchers pick one text where there is one target domain (the target domain is the topic that the document talks about) and applied all the labeled data to it. The results show that this method is less reliable than a specific classifier for each domain.
2. *Limit approach.* Researchers applied a small modification to "All data" method. Here labeled data are not examined for the target domain, but only for the outside-domain labeled data that cannot be applied in this domain. In other words, the algorithm recognizes that the target domain has some words that are not in the outside-domains of labeled data, and it automatically links these words to the target domain. Obviously, if one word have different meanings in different domains, the results can be less reliable.
3. *Ensemble of classifiers.* Researchers tried to mesh up different classifiers, with different structure and target domains to create a general classification, applicable to many target domains. Each classifier, after being adapted, can be applied to a target domain. Each classifier gives its result on each domain topic, this results can be weighted and averaged in order to analyze it. This method is useful especially for those words that have different connotations in different domain.
4. *Using in-domain labeled data.* This approach tries to combine in a same domain a small amount of labeled data with a large amount of unlabeled

[20] Aue, Gamon "Customized sentiment classifiers to new domains: a case study", Proceeding of recent advances in Natural Language Processing, 2005.
[21] Aue and Gamon, "Customizing sentiment classifiers to new domains: a case study", 2005

data. Firstly, the classifier is used only for the labeled data to define their probability distribution. Then, the classifier is applied to the unlabeled data, and the probability distribution is calculated. The comparison between the probability distribution of labeled data and the probability distribution of unlabeled ones shows us how to adjust the model parameters. In this way, to the aim is to have the probability distribution of unlabeled data equals the probability distribution of labeled data.

SVM (Support Vector Machine) was used for strategies number 1, 2 and 3; for the fourth strategy, they applied EM (Expectation Maximization).

SVM (Support Vector Machine) is a complex classifier that uses a supervised learning machine to classify words. Given a set of words and phrases to take as example, all labeled and belonging to different tags, SVM builds an algorithm that finds in texts other words that can be synonyms or antonyms of the first words and phrase taken as example(we call it *training data*). In this way, the algorithm assigns words in one category or another. The process continues until no words can be found.

EM (Expectation Maximization) is an algorithm based on estimations and probabilities. The EM algorithm is used to find maximum likelihood parameters of statistical model when the equation cannot do directly and there are missing or hidden data[22]. Instead of using a method that labels all the words, this algorithm uses a set of labeled data and other variables to estimate if a single word is positive or negative, basing on probabilities and estimations.

In "*Domain adaptation for sentiment classification*" Blitzer, Dredze and Pereira[23] analyzed in deep cross-domain with no labeled data. The most important characteristic of this method is to find similarities in different domains. The researchers have a set of texts where topic is known and recognizable by tags, and other texts where the argument is not known. The algorithm proposed finds a set of words that occurs both in the labeled (documents that have a known topic) and in unlabeled data (texts whose topic is not known). Then, all these data are divided in two groups: positive and negative.

[22] Borman, "The expectation-maximization algorithm", 2004.
[23] Blitzer, Dredze, Pereira "Domain adaptation for sentiment classification",2007.

2.2 SENTENCE SUBJECTIVITY AND SENTIMENT CLASSIFICATION

The analysis used to classify the whole text focuses on sentence subjectivity. It will be applied all the same methods as before with the only assumption that a sentence contains only one opinion. We can summarize the aim of sentence subjectivity in one phrase: given a sentence *x*, determine if *x* is positive, negative or neutral.

The sentence classification is an intermediate and useful step. It allows us to understand the entities involved in the document, thus we recognize positive or negative opinions.

We can do sentence classification in two ways:

- three-class classification
- two separate classification

The three-class classification problem divides sentences is in the following groups: objective sentences, subjective positive sentences and subjective negative sentence.

The second way, two separate classification, is easier to use and more intuitive. We split the problem into two sub-problem: we first classify if a sentence expresses an opinion or not (*subjectivity classification*), then, if the sentence is subjective, we can classify the opinion as positive or negative (*sentence sentiment classification*).

2.2.1 SUBJECTIVITY CLASSIFICATION

Subjectivity classification is the analysis that divides sentences in two groups: subjective and objective ones. An objective sentence expresses the features of an entity and a subjective sentence expresses personal point of views and opinions regarding the features of a product. Wiebe in *"Recognizing contextual polarity in phrase-level sentiment analysis"* (2005) proposed a method for subjectivity classification. This method compares a list of given opinion words with a sentence. In this way the set of words will be analysed and divided into positive and negative. The list used by Wiebe cannot be exhaustive because there is a huge amount of words that expresses positive and negative opinions. Moreover, one word can express both positive and negative opinions, depending on the context. This method is therefore be less accurate than others, but it is very flexible and can be applied to many different contexts.

Hatzivassiloglou and McKeown (1997) found a way to improve Wiebe's method: the concept of gradability. *Gradability* measures how a word can intensify or diminish an expression. (Hatzivassiloglou and Wiebe (2000)). In this way they managed to separate the subjective expressions from the objective ones.

In order to solve the problem of knowing the orientation of a sentence, several supervised learning method have been developed. Those methods are more accurate but more difficult to manage, due to the variability of the opinions. In order to simplify the issues of supervised learning method, Riloff and Wiebe in their paper called *"Learning extraction patterns for subjective expressions"* (2003) tried to build a program that automatically labels data. This algorithm first identifies subjective and objective sentences comparing series of words. Then it classifies a sentence counting the numbers of words: the sentence is a positive if two or more good words are founded, a sentence is negative if the algorithm finds at least two negative words. The so-classified sentences are enhance the sets of labelled data in order to have more labels for the next comparison. In this way, the number of labelled data is always increasing. The process goes on every time a new sentence is analysed and more the method is used more it becomes efficient.

Riloff and Wiebe (2005) used this mechanism to develop a rule-based method that produces labelled data. Instead of labelling subjective sentences, researchers built an algorithm that finds out objective sentences. A sentence is considered objective when it does not have any strong subjective words and when there are expressions that are more likely present in such sentences. For example, in the Wall Street Journal, words like "price" or "profits" are in objective sentences (i.e. " iPhone price is 799$), even though they could be also in a subjective sentence (i.e. the price of Samsung is very cheap)[24]. In this way, the method divides subjective and objective sentences. It takes into account also the information that are not opinion, but factual data that can help customers whether purchasing or not the product.

2.2.2 SENTIMENT CLASSIFICATION ON SENTENCES

As we said before, for sentiment classification there will be the following assumption: a single sentence expresses a single opinion.

[24] Wiebe, Riloff "Creating subjective and objective sentence classifiers from unannotated texts", 2005.

The method proposed by Yu and Htzivassiglou in *"Towards answering opinion question: separating facts from opinions and identifying the polarity of opinion sentences"* (2003) improves the method of Turney (see Paragraph 2.1.2). For his approach, Turney used two group of adjectives (positive and negative) and the *pointwise mutual information* (i.e. a correlation measure of two events *x* and *y*: it expresses the probability that two words are in the same sentence) in order to understand the structure of a text and through this understand the text orientation. The method proposed by Yu and Hatzivassiglou does not divide words in positive and negative, but uses a large group of not labelled words. They developed an equation that determines positive or negative orientation of the sentence analysing its structure. This equation assigns an average equation scores to words present in the sentence. This score is then compared with the two fixed threshold in order to know if the sentence is positive, negative or neutral.

In "Mining and summarizing customer reviews" (2004), Hu and Liu proposed an algorithm that consists of three steps[25]:

1. Pulling out product features that have been commented by customers
2. Identifying sentences that express opinions in each review and deciding whether each opinion sentence is positive or negative
3. Summarizing the results

In order to achieve these steps, there are some sub-steps to cover.

First of all, the algorithm downloads all the reviews and it creates a database. Each word is recognized by a POS tagging and the most frequent features that many reviewers have rated are recognized using WordNet. WordNet is a lexical-semantic database for English vocabulary that defines the relations between words. This method is effective because the words used to describe a product, are usually the same; if other words have been used, it is unlikely that they are product features[26]. In order to remove all the words that are incorrectly recognized as frequent features, Hu and Liu adopted two systems of pruning:

- *Compactness pruning.* In order to understand compactness pruning, we have to assume that words that respect a specific order are more likely to be meaningful.

[25] Hu, Liu "Mining and summarizing customer reviews", 2004.
[26] Hu, Liu "Mining and summarizing customer reviews", 2004.

If some words have a specific order, we can say that they are frequent features. Compactness pruning aims to recognize the words' order and eliminates those words that does not respect it.

- *Redundancy pruning.* This step removes all the single words that are superfluous and does not represent a feature.

After identifying product features, the researchers identified opinion words using WordNet synonyms and antonyms. They analysed the orientation of a small group of words, and predicted other words' orientation comparing them with synonyms or antonyms.

Hu and Liu took into account that there are product features that cannot be categorized as frequent features only because they are cited only few times. However, these aspects of a product could influence potential customers and for this reason it is worth to analyse them. Researchers solve the problem using opinion words: when the algorithm finds an opinion words but not a frequent feature, it assigns the tag of feature to the nearest noun or noun phrase.

The algorithm decides the orientation of the sentences, which is determined by the dominant orientation of the opinion words. If in the sentence there are more negative opinion words than positive ones, then the sentence is negative (and vice versa).

CHAPTER III

HOW TO SELECT DATA

3.1 SENTIMENT ON ASPECTS

As explained in the previous chapter, in order to start a Sentiment Analysis is necessary to classifying opinions in text and sentences. Moreover, a proper analysis needs to assign a positive or a negative orientation. If we assume that each text refers to a single entity, we cannot simplifying the concept considering a single document as totally positive (or totally negative). An opinion holder can share a review about a single entity, in which various aspects of it are expressed. An aspect is a feature of the entity examined. For example, if the entity is a Dainese Jacket, one aspect could be the protection, another one could be the style.

The structure of this chapter is divided into two sections: the first part focuses on aspects of a certain entity, the second one on opinions shared by customer. In the first part, we give an overview of the methods used to extract aspects from a text and, especially, from a sentence. We define some basic rules in order to gather aspects into categories and to understand the semantic group of words that constitutes an entity. We briefly present some methods to extract information about opinion holders and the time of the review. In the end, we underline what are the most important obstacles in aspect extraction.

The second part presents two methods that can extract opinion words in two different ways: *dictionary based approach* and *corpus based approach*. The last part of the chapter is dedicated to opinions that are expressed with words that does not have a defined opinion trend: *desirable and undesirable facts*.

The aim is to discover the five variables $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ in a given document *d*, as already presented in Sentiment Analysis: on document level and on sentence level. To achieve this, we have to cover two important steps[27]:

1. **Aspect sentiment classification.** The opinions related to different aspects need to be defined. Thus, an opinion can be classified as positive, negative or neutral.

[27] Bing Liu, "Sentiment Analysis and Opinion Mining", 2012.

2. **Aspect extraction.** The step in which a specific aspect is retrieved from a text. It is important to underline that an aspect is always connected to an entity.

3.1.1 SENTIMENT CLASSIFICATION: ASPECTS

To determine the orientation of the feelings expressed on each aspect, there are two approaches:

- Supervised learning approach
- Lexicon-based approach

The supervised learning approach is similar to what we described for document texts and sentences: a method that classifies document text labelling each word. After the analysis, we have to determine the opinion trend of each expression. In the norm, this goal is achieved by studying the structure of the sentence and recognizing the opinion words. Usually a given algorithm finds an opinion word and it searches the nearest noun or noun phrase and connects them. The noun (or noun phrase) is the aspect and the opinion words represent the feeling associate to it. This method does not take into account the context and it could lead to misleading results because the association feeling-entity could be wrong. In fact, it is not always true that an opinion word is connected to the nearest noun. For example, the sentence "Coca-Cola is more delicious than Pepsi" expresses a positive opinion about Coca-Cola, but the method would recognize a positive opinion about the other brand.

In order to solve this problem, Jiang, Yu, Zhou, Liu and Zhao (2011) invented a target-dependent and context-aware approach. Specifically, target-dependent approach is achieved recognizing the structure of the sentence that incorporates the syntactic of features. In the case of the context-aware, the method is developed to study the review into its context. This method uses labelled data and it cannot be applied to a large amount of data, because the time requested to label all the words is too wide.

The lexicon-based approach can be an alternative to the supervised learning method and it can be efficient in several domains, sources where different documents refer to the same topic. It uses a sentiment lexicon (a list of words, idioms, phrases, composite expressions) and possibly the sentence parse tree (a graphical representation of the syntactic structure of a sentence) to determine the orientation of the review on each aspect of a sentence. A lexicon based method and its application will be described. Ding

et al. (2008) presented the following method in four steps. Here we assume that aspects and entities are known[28]:

1. **Mark sentiment words and phrases:** A computer program will underline all the words that represent a feeling in the sentence. +1 will be assigned for each positive word, and -1 for each negative word.
2. **Apply sentiment shifters:** Sentiment shifters are noun or phrases that can change the orientation of a review. For instance, negation words like *not, never, nobody, none, nowhere, neither and cannot*; negative words, etc.
3. **Handle but-clauses:** The sentences that contain the word "but" usually are divided in two parts by this word. The part before "but" usually expresses the opposite orientations of the part after "but". With this assumption, we can affirm that if we are not able to understand the orientation of a but-clause, there are two opposite orientation and we connected each of them to the nearest noun.
4. **Aggregate opinions:** The last steps finds the opinion on the overall text We can do this only if previously we assign scores to opinion words and phrases (see step 1). In order to aggregate opinion, Liu (2012) developed an equation as here described: the sentence *s*, contains a set of aspects $\{a_1, \dots, a_m\}$ and a set of opinion words or phrases $\{s_{w1}, \dots, s_{wn}\}$ with their scores obtained in the first step. The sentiment of each aspect a_i in a sentence *s* is determined by this aggregation function:

$$score(a_i, s) = \sum_{sw_j \in s} \frac{sw_j.so}{dist(sw_j, a_i)},$$

Where sw_j indicates a sentiment word or phrase in *s*, $dist(sw_j, a_i)$ is the distance between the aspect a_i and the sentiment word sw_j in *s*. The distance is used to give lower weights to sentiment words, that are far away from a_i. If the final score is higher than zero, the opinion orientation is positive, and vice versa[29].

3.1.2 OPINION AND COMPOSITIONAL SEMANTIC: BASIC RULES

In order to define words' orientation, we need to settle some rules useful to create a specific procedure. In this way there is no misunderstanding and the method can be

[28] Ding, Xiaowen, Bing Liu, "Resolving object and attributing conference in Opinion Mining", 2008.
[29] Bing Liu "Sentiment Analysis and Opinion Mining", 2012.

easily replicated. In this case, we define opinion rules, useful for understanding orientation of words: these rules explain how to identify and classify each opinion, thus we can labelled each opinion as positive or negative.

An opinion rule is essential to classify opinions. This rule had to take into consideration common sense and the specific context: without these elements, the rules do not have sense. This section describes some of these rules.

The rules are represented using a method similar to the Backus-Naur Form (BNF). BNF is a precise and unambiguous way to describe syntactic in information technologies. We adopt this method in the description of language syntactic in order to define clearly all rules. In this sense, Liu (2010)[30] presented the following rules:

1. POSITIVE::= P
2. | PO
3. | sentiment_shifter N
4. | sentiment_shifter NE
5. NEGATIVE::= N
6. | NE
7. | sentiment_shifter P
8. | sentiment_shifter PO

The non-terminals P and PO represent two types of *positive sentiment expressions*. P indicates a word or a sentence that express a positive opinion, while PO is a positive opinion composed by several positive expressions. The non-terminal N and NE express the same concept but in the negative aspect. "Sentiment shifter N" and "sentiment shifter NE" represent the negation of N and NE, so these are positive expressions. Both "sentiment shifter P" and "sentiment shifter PO" are the negation of positive expressions. These rules are in the same time intuitive and interchangeable, because they have the same simple structure and all the rules refer to the concepts of positive and negative[31].

[30] Liu, Feifan, Dong Wang, Bin Li and Yang Liu, " Improving blog polarity classification via topic analysis and adaptive methods", 2010.
[31] Liu, Feifan, Dong Wang, Bin Li and Yang Liu, "Improving blog polarity classification via topic analysis and adaptive methods", 2010.

Sentiment shifters include different categories: *negation words* like *no, nobody, never, etc.* are the most common ones and *modal auxiliary verbs* (e.g. *would, should, might, etc.*) are another type.

N, NE, P and PO with any sentiment shifter can be divided into six different categories[32].

1. *Sentiment word or phrase.* This is the most used and the simplest category. A single word or phrase, P or N, can express positive or negative sentiment related to an aspect.

   9. P ::= a_positive_sentiment_word_or_phrase
   10 N ::= a_negative_sentiment_word_or_phrase

2. *Decreased and increased quantity of an opinion word* (N and P). This set of rules expresses the "intensity" of an opinion. A phrase (noun) can often change the orientation with these rules. The concept of decreasing also includes the possibility of *removal* (the opinion is totally changed) and *disappearance* (the opinion is totally annulled).

   11. PO ::= less_or_decreased N
   12. | more_or_increased P
   13. NE ::= less_or_decreased P
   14. | more_or_decreased N

   We can see that rule 12 and 14 cannot change the sentiment orientation, but they change the intensity of the opinion. Usually, these kind of words can be before or after N or P.

3. *High, low, increased and decreased of a positive or negative potential item*. For some item, a small value or quantity is negative and a large value/quantity is positive. For example "*the battery life is short*" and "*the battery life is long*". Such items are called *positive potential items* (PPI). For some other aspect, a small quantity/value of them is positive and a large quantity/value is negative- i.e. *the negative potential item* (NPI). Both PPI and NPI express any sentiment, but when there are some decreasing or increasing items they inforce an opinion.

[32] Bing Liu, "Sentiment Analysis and Opinion Mining", 2012

15. PO ::= no_low_less_or_decreased_quantity_of NPI
16. | large_larger_or_increased_qunatity_of PPI
17. NE ::= no_low_less_or_decreased_quantity_of PPI
18. | large_larger_or_increased_quantity_of NPI
19. NPI ::= a_negative_potential_item
20. PPI ::= a_positive_potential_item

4. *Desirable or undesirable fact*: The rules above all contain some subjective expressions, but they can also express desirable and undesirable facts. Such sentences usually do not use sentiment words, but they use words that cannot be recognize as positive or negative, but in a particular context, they acquire an opinion trend. For example, the word *valley* does not express an opinion, but in the sentence "After two weeks, my mattress has a valley in the middle" the word expresses a negative opinion.
   21. P ::= desirable_fact
   22. N ::= undesirable_fact

5. *Deviation from the norm or a desired value range*: in some domains, the value of an item has a desired range or norm. If the value deviates from the normal range, it is negative. For instance, in this sentence "After taking the drug, my blood pressure went to 210", we can say that the drug has a negative effect in our body). Such sentences are often objective sentences as well.
   23. P ::= within_the_desired_value_range
   24. N ::= deviate_from_the_desired_value_range

6. *Produce and consume resource and waste*: we can affirm that an entity that produces a large quantity of physical resources (i.e. water, energy, etc.) has a positive opinion, when an entity consume a large quantity of resource recall a negative opinion. When an entity consume a large quantity of waste is positive, when it produces a large quantity of waste is negative. For example, *water* is a resource, and if I say "My mother uses a large amount of water when she showers" I express a negative opinion because in this case my mother waste a resource; when I say "This year windmill blades produced more energy than the previous year" I express a positive opinion. Below the rules:

25. P ::= produce a_large_quantity_of_or_more resource
26. | produce no,_little_or_less waste
27. | consume no,_little_or_less resource
28. | consume a_large_quantity_of_or_more waste
29. N ::= produce no,_little_or_less resource
30. | produce some_or_more waste
31. | consume a_large_quantity_of_or_more resource
32. | consume no,_little_or_less waste

3.1.3 ASPECT EXTRACTION

Aspect is a part of the entity reviewed; the aspect extraction is comparable to entity extraction, because we have to select the words that express features and words that express opinions; furthermore, we can use similar methods and techniques. The key characteristic of aspect extraction is that in a sentence, an opinion is linked to features that are the main topic that reviews refer to.

In the following pages we will describe how the four methods have been applied to online reviews to extract aspects. The online reviews are usually divided into two formats[33]:

Format 1- **Pros, Cons and the detailed review:** the reviewer first describes pros and cons of an entity and then they give a detailed review.

Format 2- **Free format:** the reviewer writes without a specific pattern and there is not a list of pros and cons.

Extracting pros and cons from a review with format 1 is relatively easy. Liu et al. (2005) used an automated method to that assigns to all pros words a positive score and to all cons a negative one. Usually Pros and Cons are very concise and short sentences or phrases that contain only one aspect (a part of the entity reviewed) and can be separated by commas, periods, *but*, *etc*.

In order to retrieve aspects from a sentence, there are four main approaches[34]:

[33] Bing Liu, "Opinion Mining and Sentiment Analysis", 2012.
[34] Bing Liu, "Opinion Mining and Sentiment Analysis", 2012.

1. Extraction based on frequent noun phrases and noun
2. Extraction by exploiting opinion and target relations
3. Extraction using supervised learning
4. Extraction using topic modelling

FINDING FREQUENT NOUNS AND NOUN PHRASES

The aim of this method is to find *explicit aspect expressions*, nouns that indicate in various ways the same aspect of an entity. The *explicit aspect expressions* are characterised by a wide variability, because people use different slangs and words to communicate. Hu and Liu (2004) built a data mining algorithm in order to find frequent explicit aspect expressions. This algorithm uses a part-of-speech (POS) tagger able to identify this kind of words in a given review. After finding them, the algorithm counts how many times these specific words have been used. The words, which have a high frequency, are kept for further analysis.
Depending on the aim of a specific company, it is possible to set the frequency target. For instance, in some researches a good frequency is three, which could be considered a low in others.

Hu and Liu assumed that people use the same vocabulary to describe the same aspect or the same entity. The non-frequent nouns or aspects are usually not important.

USING TARGET AND OPINION RELATION

Since opinions refer always to an entity (or part of it), we can use this relationship in order to extract opinion aspects. Hu and Liu (2004) assume that the same word can be used to describe or modify different aspects. This extraction method works as following: if a sentence does not have a frequent aspect but has some "sentiment" words, the noun or noun phrase right before or after a sentiment word is considered as an aspect. For example, in the sentence "My tablet is amazing", the word *amazing* is recognized as a word that expresses a feeling and then *tablet* is extracted. This method consists in an approximation because is not always true that an opinion word and its entity or aspect are always close. The researchers did not built a structured and well defined method like the supervised learning method. However, the results are reliable

and this method can be applied without any support. In fact, it does not need any another method that comprehends labelled data, in order to work.

USING SUPERVISED LEARNING

Many of the algorithms proposed in these years to extract information were based on supervised learning[35]. There are many kinds of methods that used supervised learning: the most important one is called *sequential learning* (or *sequential labelling*). As we saw in the previous chapter, since this is a supervised method, we need to manually label data: doing this, we create a labelled set that are useful to recognize words in a document text. In this way, a separation between non-aspects and aspects separated takes place.

As we have seen previously, data are matched with a list of labels helping researchers to recognize what is an aspect. A list of labels for a Dainese jacket could put together words like: collar, colour, waist adjustment, protections, etc.

USING TOPIC MODELS

In recent years, topic models have emerged as a principal way to retrieve specific subjects from a large collection of text document. Topic modelling is an unsupervised learning method that takes the following into consideration: in each document there is a large amount of topics and the sensible words connected to a topic have a high frequency within the document text. For example, when we talk about jacket, it is more likely that there are words like "collar", "colour" and "fabric" instead of "dog" and "meal". A mathematical algorithm captures this probability and finds the most likely topic of the document. The output of a topic model is a set of word called clusters. Each cluster comprehends a topic.

There are two basic models:

- pLSA (Probabilistic Latent Semantic Analysis)[36]. The goal of pLSA is to map data in order to find relations between different aspects; this is possible thanks to

[35] Hobbs, Jerry R., Riloff "Information extraction", 2010.

[36] Hofmann "Probabilistic latent semantic indexing", 1999.

Expectation Maximization (EM) algorithm (i.e. EM algorithm estimates the likelihood of words presence).

- LDA (Latent Dirichlet allocation)[37] is a probabilistic model that is used to collect data in order to know the main topics of a given documents and the words linked to them. It is a three-level hierarchical Bayesan model. A Bayesan model is a method to extract words from a given document in a probabilistic way. The theorem of Bayes estimates the likely causes for a certain situation: in the same way, we can affirm that the Bayesan model estimates what are the main topic of a given document.

Even though they are mainly used to extract and describe topics extracted from text, these two methods can be extended to analyse many other types of information. Topic models can contain both aspect expressions and sentiment words, and to be applied to sentiment analysis they need to be separated. This separation is made by extending the basic model (e.g. LDA) to join both aspects and sentiments.

3.1.4 GATHERING ASPECTS INTO CATEGORIES

After extracting aspects, we need to gather them together into categories. Each category is formed by synonymous of the same aspect. This process is quite hard even for sentiment analysis: even if many dictionaries can help to some extent, they are far from sufficient because many words are domain dependent. For example, “movie” and “picture” are synonyms in movie reviews, but not in a camera review, where “picture” is more likely to “photo” and “movie” to “video”. Furthermore, many aspect expressions are formed by a sentence and a dictionary cannot recognize them. The majority of expressions linked to the same aspect are not synonyms. For example, “expensive” and “cheap” can both indicate the aspect “price”, but they are not synonyms of each other or synonyms of price.

3.1.5 ENTITY, OPINION HOLDER AND TIME EXTRACTION

Extracting Entity, opinion holder, and time is the classic problem of Named Entity Recognition (NER). NER has been studied in several fields: text mining, data mining,

[37] Blei, David M., Andrew Y.ng and Michael I. Jordan “Latent dirichlet allocation”, 2003.
Griffiths, Steyvers “Prediction and semantic association”, 2003.
Griffiths, Steyvers “Integrating topics and syntax”, 2005.

machine learning and natural language processing, under the name of information extraction (Hobbs and Riloff, 2010; Mooney and Bunescu, 2005; Sarawagi, 2008).The principal goal of NER is to classify words in specific categories like location name, person name, date, time, etc[38]. The main approaches are statistical or based on mathematical rules. Statistical methods are mainly based on Hidden Markov Models (HMM) (Rabiner, 1989; Jin and Ho, 2009) and Conditional Random Fields (CRF) (Lafferty et al., 2010). Both HMM and CRF are supervised method.

HMM is a generative model, such models use a joint probability distribution (i.e. this model takes into consideration the probability that two events, X and Y, happens together). The model assigns joint probability to opinions and labelled data in order to connect each other. The model is then optimized, maximizing the likelihood between opinions and labelled data. This method works well especially when opinions are represented by words and are near to entities[39].

Conditional Random Field is a conditional model where the opinion are fixed and have a particular sequence. Given that opinion sequence, we can know the context where the opinion is expressed: with this information we can know the aspects reviewed. This method considers possible that the context can influence the way people write entities.

Note that especially in European languages, finding person names, location names is easier because of the capital letter, that almost every time identifies a proper name.

Usually, it is easy to extract the user, the date and the time from a text. These information have a specific place in a review and are always written outside the corpus. Extracting entities could be more articulate. For example, a same concept might be written in different ways, some reviewers may write “Dainese” some others “Dai”. For entity extraction is important to recognize which are the words that refer to the same entity, extract them from the sources and cluster them in the same group.

Li et al. (2010) formulated the problem as following[40]: Given a set of Q of seed entities (entities settled to be recognize inside a document text) of a particular class C, and a set D of candidate entities, we wish to determine which of entities in D belong to C. In this

[38] Morwal, Jahan, Chopra “Named Entity Recognition using Hidden Markov Model”, International Journal on Natural Language Computing, 2012.
[39] Morwal, Jahan, Chopra “Named Entity Recognition using Hidden Markov Model”, International Journal on Natural Language Computing, 2012.
[40] Li, Fangtao, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang and Hao Yu, “Structure aware review mining and summarization”, 2010.

way, the class C will be completed by a set of seed examples Q. This problem is often solved as a ranking problem: ranking the entities in D based on their likelihoods of belonging to C.

The most used method to solve this problem in Natural Language Processing is based on distributional similarity (Lee, 1999; Pantel et al., 2009). The method compares the similarity of the words that surround each entity with those of the seed entities (the first words that the model recognizes as entities). The candidate entities will be then ranked based on the similarity values.

3.1.6 COREFERENCE RESOLUTION AND WORD SENSE DISAMBIGUATION

In this section, we would like to highlight all the Natural Language Processing (NPL) issues and problems in the sentiment analysis context. Only a few researchers studied NLP in opinion mining. NPL community focuses especially on the co-reference resolution. These studies aim to determine which expressions in a sentence or document refer to the same object. For example, in "I bought an iPhone two days ago. It looks very nice. I made many calls in the past two days. They were great". "It" in the second sentence refers to iPhone, that is an entity, and "They" in the fourth sentence refers to the calls, which is an aspect. Recognizing these differences is very important for sentiment analysis. If the opinion in each sentence is considered alone or not connected to any aspect or entity, there is the possibility to lose that opinion. In fact, even if we know that the second and the fourth sentence are related to a specific opinion, we cannot identify the target those refer to. In order to clarify the relationship between the two elements we refer to Ding and Liu (2010). In their paper, they tried to solve the problem of entity and aspect co-reference resolution. They developed a supervised learning approach, assuming that aspects and entities were previously extracted. Firstly, they scan all the document with a POS tagger, then every phrase and word are tagged and every entity and aspect are match to all the opinions. Thus, the researchers will constitute the labelled data set. These data are compared with other texts in order to find some match. The method is based on sentiment analysis of regular and comparative sentences, based on the idea of *sentiment consistency* (i.e. the coherence between opinions expression in the same review, that have implicit aspects or entities to refer to). The following example will be used as explanation: "*The Nokia phone is better than*

this Motorola phone. It is cheaper". "It" refers to "Nokia phone" because both the first part referring the Nokia phones and the second sentence are positive. In order to connect "Nokia phone" with "it", the system must have the possibility to determine positive and negative opinions expressed in regular and comparative sentences[41].

The second aspect considers the modifiers of positive and negative words (or phrases). If we consider the phrase "I bought a Nokia phone two days ago. The sound quality is not bad. It is cheap too", "the sound quality" refers to Nokia phone, because a sound quality cannot be cheap. But the system, in this case, needs to know which sentiment words are associated with which entities or aspects. In this case, the word "bad" is the word that expresses opinion, but the modifier "not" changes its connotation and the result is a positive opinion. These two features are semantic features that current general co-reference resolution methods do not consider.

To solve the problem of sentiment consistency with modifier words, Akkaya et al. (2009) studied *subjectivity word sense disambiguation (SWSD).* The purpose of this research is to find a system that automatically recognizes which word have a subjective senses and which word examples have an objective senses. The method is a mix between a dictionary classification and a contextual interpretation. The researchers used a sort of POS tag in order to recognize if a sentence is subjective or objective. The subjective sentences are then analyse in the context in order to extract data.

There is another factor to take into account: many words have both an objective and a subjective sense, depending on the context. For example, the word *cheap* has an objective sense if it is used for a jacket, but it has a subjective meaning if we refer to a person. Disambiguation represents a huge source of errors. For this reason, the authors built a supervised SWSD model to categorize if subjectivity words have a subjective meaning or an objective one. The method takes into account not only the definitions from dictionaries but also the context, in order to understand if a word expresses opinions or factual data. Researchers uses various classifiers, depending on which of them have labelled data. They list a set of words as subjectivity lexicon, because they proved that with these words there are opinions[42]. Those words that match into this are subjective, otherwise they are objective one.

[41] Ding, Liu "Resolving Object and Attribute Coreference in Opinion Mining", 2010.
[42] Akkaya, Wiebe, Mihalcea "Subjectivity word sense disambiguation", 2009.

3.2 SENTIMENT LEXICON GENERATION

This paragraph helps us to understand how to analyse data in order to understand which are the feelings connected not only to words, but also to phrases and idioms. In this paragraph, we will assume that sentiment words are both words and sentences.

We can divide sentiment words in two types:

- *base type*
- *comparative type*.

Base type are words like *awful, poor* and *amazing* and they express a regular opinion on an entity. Regular opinion expresses a judgment without considering other entities. Comparative ones are words like *best, worse, better, etc.* and they express opinions on more than one entity.

In order to treat the base sentiment words there are two different approaches[43]:

- *dictionary-based approach*
- *corpus-based approach*.

In the following paragraphs, we will describe these methods.

3.2.1 DICTIONARY-BASED APPROACH

The corpus-based approach has been applied in two different cases:

1) When we know the orientation of a set of words and then we find other opinion words in the same document.
2) When we want to adapt a general lexicon to a new domain, in order to use sentiment analysis.

Setting a list of words and recognizing a specific lexicon related to the domain is not the proper solution. In fact, a word might have a positive orientation in a context and in another one might have a negative. Even if the corpus-based approach is used to the same purpose of the dictionary-based approach, the second one is more effective because is based on a complete list of words.

[43] Bing Liu, "Sentiment Analysis and Opinion Mining",2012.

One of the most important and early ideas that built the basis of this method was proposed by Hazivassiloglou and McKeown (1997). The authors used a document and some seed sentiment words to find other sentiment adjectives in the corpus. The method considers different language rules or conventions on connectives (a word that connect phrases, other words, clauses, etc.) to identify more adjective sentiment words and their orientations in the corpus. For example, in this sentence "This car is beautiful and spacious", if *beautiful* is known to be positive, thanks to connection *and*, we can say that also *spacious* is positive. The method works with other connectives (but, or, either-or, etc.). This idea is called *sentiment consistency*. In practice, this rule could be misleading, because reviews do not always have this structure.

The next step is to determine if two connected adjectives have the same or different orientations. First, the adjectives with the same-and different-orientation are drawn in a graph that measures the links between adjectives. Then a cluster of positive words and a group of negative ones are built.

Kanayama and Nasukawa (2006) extended the application of this method by introducing the intra-sentential (within a sentence) and inter-sentential (between neighbouring sentences) sentiment consistency, the so-called *coherency*[44]). The method is similar, but instead of words, it is applied to sentences.

3.2.2 DESIRABLE AND UNDESIRABLE FACTS

In order to express opinion, people can use some words that are not opinion words, but in specific context, they have opinion trends. For example, the sentence "I bought on the internet this jacket but when I wear it, a wrinkle appears near my neck: it seems like there is a surplus of material" does not uses opinion words, but it express opinion because the word "wrinkle", in this specific context, has a negative connotation. For these sort of sentences is difficult to set an automatic algorithm able to recognize them.

Zhang and Liu (2011) proposed a method to find words and expressions that are aspects in some situations and express sentiment in other context. For example, the noun "valley" does not have orientation if considered alone, but, when we talk about mattress, it has a negative connotation. If I write something like "I bought a mattress two months

[44] Kanayama, Nasukawa "Fully automatic lexicon expansion for domain-oriented sentiment analysis", 2006.

ago and there is already a valley in the middle", I express a negative opinion. Identifying these opinions is very challenging. The risk of errors is very high and the ability to recognize them is difficult to automate. The algorithm of Zhang and Liu[45] is based on the following concept: even if these objective words are alone and they are not followed by a subjective expression, in some occasions the reviewer add some adjectives that have the same orientation of the objective word. To distinguish such words, we have to do an observation[46]:

> **Observation:** normal aspects that cannot have a specific orientation, they can assume both positive and negative orientations. (e.g. the noun phrase *voice quality* has not an orientation; it express an opinion only with an adjective "good voice quality" or "bad voice quality"). In a different way, for words that express desirable or undesirable facts they can have only one orientation, but not both (for example, the noun *valley* in a sentence where the entity is a mattress cannot be positive).

With this observation in mind, we now list the steps in order to recognize words that express desirable and undesirable facts automatically.

1. *Candidate identification.* If the sentence is positive, the noun or noun phrases are more likely to be positive. If the context is negative, it is more likely that an undesirable fact (a factual data that reviewers reposted as prove of their bad opinion) happened. In this way, we can list a set of aspects with positive opinion and a set of aspects with a negative orientation.
2. *Pruning:* this step let us prune the two lists made in the previous steps. When an aspect is opinionated in both positive and negative way by adjectives, we can say that these are not opinionated aspects: these noun aspects are cut off. Two types of relations were used:

   - Type 1: when opinion depends on an aspect through the adjective. E.g. "*this tv has <u>a good</u> picture quality"* in this case "a good picture quality" influences the opinion;

   - Type 2: when sentiment word depends on both aspect and entity through dependency relations. E.g. *"The springs of the mattress are <u>bad</u>"* here the modification sentiment word decide where the review is clustered.

[45] Zhang, Liu "Identifying noun product features that imply opinions", 2011.
[46] Bing Liu, "Sentiment Analysis and Opinion Mining", 2012.

This approach is only a draft and it needs to be more complete, because now its precision is not high. Probably, in the future, more researchers have to invest this field[47].

[47] Bing Liu, "Sentiment analysis and opinion Mining", 2012.

CHAPTER IV

HOW TO ANALYZE DATA

4.1 OPINION SUMMARIZATION

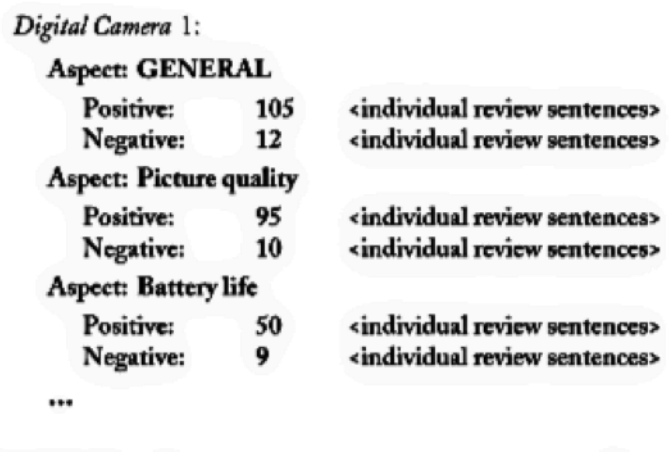
Sentiment analysis is effective if a large number of opinions is taken into consideration. Analysing is just the first step, but then, in order to take some conclusions, we need to summarize the results. The summary can be done in a structured form (graphs, numbers, etc.) and in an unstructured form, as a short text document.

In general, all the opinion summarizations can be seen as a sort of *multi-document text summarization*. Multi-document text summarization is an automatic procedure that extract information from multiple texts written about the same topic. Many researches in Natural Language Processing (NLP) field studied and applied text summarization. This kind of summary is not a traditional one: in this case, aspects, entities and their related opinions have to be described both quantitatively and qualitatively. A traditional text-document summary consists of a short text, explaining the most important concepts taken from the original review. In an aspect based opinion summarization (a text that comprehends entities, target and their related feelings), there are some rules that have to be followed.

4.1.1 ASPECT-BASED OPINION SUMMARIZATION

Aspect-based opinion summarization has two main characteristics. First, it captures the essence of opinions: opinion targets (entities and their aspects) and feelings related to them. Second, it expresses opinions in a quantitative way. It provides the number or percentage of positive and negative opinions about a product or a service. Since opinions are subjective, it is quite important to summarize them in a quantitative way in order to measure the trends. The final result of summarization is a structured document with the number of positive and negative sentences of all the reviews, divided per different aspects. The summary is like the following one, about a digital camera:

FIgure 2 An aspect-based opinion summary

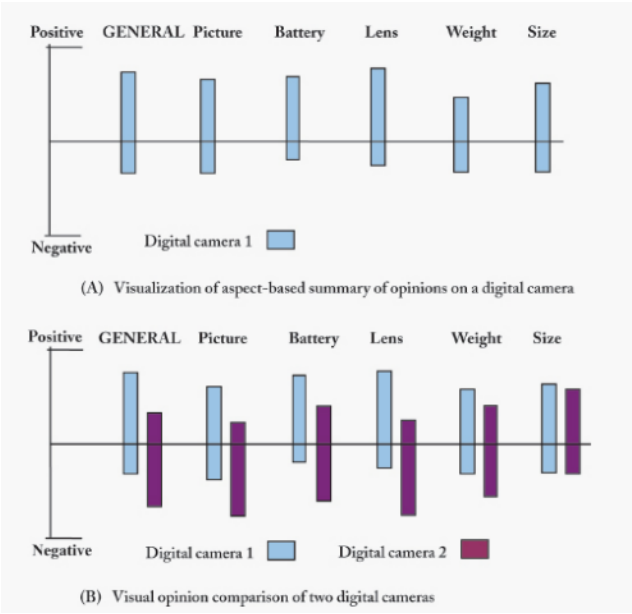


Source: Bing Liu, "Sentiment Analysis and Opinion Mining",2012.

The aspect GENERAL refers to customer opinions given to the whole camera. In this case, for instance, 105 people have a positive overall opinion, 12 a negative one.

In Figure 3 (A) another representation of the summary related to the camera: a summary can be represented in different ways, but it is important to express clearly positive and negative opinions. The bars in the upper part of the graph express positive opinion related to the aspect given at the top. Bars under the X-axis represent negative ones for the same aspect.

Figure 3 Visualization of aspect-based summaries of opinions



Source: Bing Liu, "Sentiment Analysis and Opinion Mining",2012.

Thanks to the digitalization and the "datafication" (i.e. a modern technological trend that turns many aspects of our life into computerised data[48] and transforms this information into new forms of value[49]) the amount of information in our possession is huge. Handling data increases the quality of structured summaries. For example, data from opinion holders help researchers to make comparisons in different time: if a product is improved during a period and the same customer write about this change, the company could know if the improvement is well welcomed by customers. Furthermore, if we focus on the time of a review, we will know how many times an aspect is mentioned in a specific period.

4.1.2 CONTRASTIVE VIEW SUMMARIZATION

Many researchers tried to find an answer to the problem of summarizing contrastive point of view in reviews as the following. An opinion holder writes "*The voice quality of my Nexus is really good*" and another one expresses "*The voice quality of my Nexus is very lousy*". A pair of opinionated sentence that focuses on the same aspect but giving opposite opinions, is called *contrastive sentence pair* [50].

The researchers Kim and Zhai (2009) studied and tried to solve the problem of summarizing contrastive sentence pair. First, they classify all the opinions in the reviews in two sets: a positive and a negative one. In each group, sentences are divided per topic and for each topic they extract one sentence, which is a good representative of each group. These two sentences constitute the set of *k contrastive sentence pairs* sentence from the sets. I.e. a pair of opinionated sentences (*x*,*y*) is called a contrastive sentence pair if *x* and sentence *y* are about the same topic aspect, but have opposite sentiment polarities[51]. Each contrastive sentence pair is compared with the sentence that constitute the data to be analysed.

Another pull of experts tried to solve the problem. Paul et al. (2010) set up an articulate method that gives two types of summary: a macro multi-view summary and a micro multi-view summary. The macro multi-view summary comprehends different set of

[48] Cukier, Mayer-Schoenberger, Viktor, "The rise of the Big Data", *Foreign Affairs*, 2013.
[49] O'Neil, Schutt, "*Doing Data Science*", O'Reilly Media, 2013.
[50] Kim, Hyun Duk and Zhai "Generating comparative summaries of contradictory opinions in text", 2009.
[51] Kim, Zhai "Generating comparative summaries of contradictory opinions in text", 2009.

sentences, both positive and negative. The micro multi-view group consists of a set of pairs of contrastive sentences. The algorithm they built, works in two steps[52]:

- **Step 1.** The algorithm uses a topic modelling approach in order to represent and mine aspects and sentiments. A topic modelling approach is a machine learning that uses natural language process to capture the main topic of the text and forecast a list of words that probably compare in the text because they are related to the company.
- **Step 2.** A method scores sentences and pairs of sentences from opposite viewpoints: the assessment is based on both their representativeness and their contrast with each other.

4.2 ANALYSIS OF COMPARATIVE OPINIONS

There are many different ways to express opinions: not only describing an aspect or an entity, but also comparing similar entities. *Comparative opinions*[53] are a particular form of regular opinions. In one hand, they can have an opinion or not as the regular opinion. On the other hand, they have different meanings and syntactic forms.

We can divide comparative opinions in two groups: *comparative opinions* and *superlative opinions*. The main difference consists of using comparative and superlative adjectives respectively.

As comparative opinion is a particular form of regular opinions, different techniques and approaches have to be applied as presented in the following paragraphs.

4.2.1 PROBLEM DEFINITIONS

A comparative sentence describes a relation between two or more entities, based on differences or similarities. We can compare entities through *gradable* or *non-gradable comparisons*[54], as described below.

[52] Paul, Michael J., ChengXiang Zhai and Roxana Girju "Summarize contrastive viewpoints in opinionated text",2010.

[53] Jindal, Niting and Liu "Identifying comparative sentences in text documents", 2006a.
Jindal, Niting and Liu "Mining comparative sentences and relations", 2006b.

[54] Jindal, Niting and Liu "Identifying comparative sentences in text documents", 2006a.
Kennedy "Comparatives, Semantics of" 2005.

Gradable comparisons: A comparison that contains an adjective that expresses preferences indicating "intensity". It has three sub-types:

1. *Non-equal gradable comparison*: when the customer expresses their favourite entity. This sub-type describes an unbalanced relation: an entity that refers to the same aspect is *greater* or *less than* another entity, e.g., "*Coke tastes better than Pepsi*". Non-equal gradable comparisons also comprehend preferences like "*I prefer Dainese to Alpinestars*".
2. *Equative comparison*: when the customer expresses two equal entities. e.g., "*Sprite and Schweppes taste the same*".
3. *Superlative comparisons*: these sentences express the best or the worst entities of the same aspect among all others, e.g., "*Ceres taste the best among all the other beers*".

Non-gradable comparison: this comparison describes a relation between two or more objects without ranking them. There are three main sub-types[55]:

1. Entity A is similar to or different from Entity B based on some of their shared aspects, e.g., "*Coke tastes differently from Pepsi*".
2. Entity A has aspect *a1* and entity B has aspect *a2* (*a1* and *a2* are changeable), e.g., "*Desktop PCs use external speakers, laptops use internal speakers*".
3. Entity A has aspect *a*, but entity B does not have, e.g., "*Nokia phones come with earphones, but iPhones do not*".

Only the gradable comparisons will be analysed in deep.

We will focus our analyse taking in consideration the English grammar, where comparatives are composed by the suffix *–er*, whereas superlative words are characterized by the suffix *–est* (we call them *Type 1 comparatives and superlatives*). There are, of course, exceptions: words composed by two or more syllables and not ending in y are preceded by *more, most, less, least, etc.* (the so-called *Type 2 comparatives and superlatives*). Both Type 1 and Type 2 can be considered *regular comparative and superlatives*.

Words like more, most, least, less, better, best, worse, worst, further/farther, and furthest/farthest that are not preceded by any adjectives do not follow the previous rules

[55] Jindal, Niting and Liu "Identifying comparative sentences in text documents", 2006a.
Kennedy "Comparatives, Semantics of" 2005.

and they are defined *irregular comparatives and superlatives*. Note that they are built in a different manner, but they respect the rules of Type 1, so we can group them under that type.

There are many ways to express comparatives opinions, with other words and phrases: Jindal and Liu (2006a) made a partial list. This list and the above standard comparatives and superlatives are called *comparative keywords*.

Comparative keywords used in non-equal gradable comparisons can be divided into two categories[56]:

- *Increasing comparative*: a comparative that expresses increased quantity (e.g. *more* and *longer*)
- *Decreasing comparative*: a comparative that expresses decreased quantity (e.g. *less* and *fewer*)

Objective of mining comparative opinions[57]: given an opinion document *d*, discover all the six comparatives opinion of the form:

(E_1, E_2, A, PE, h, t)

Where E_1 , E_2 are the entities set being compared on their shared aspect. A, PE ($\in\{E_1, E_2\}$) is the preferred entity set of the opinion holder *h*. *T* is the time when the opinion is expressed. For a superlative comparison, if one entity set is implicit, we can use the special set U to denote it. For the equal comparison, we can use the special symbol EQUAL as the value for PE.

Mining regular opinions is the same process of extracting comparative opinions, entities, aspects, opinion holders and time are in both cases known. We can also affirm that comparative opinions have a determined structure[58] and we can easily recognize entities, comparative keywords and aspects. Not all the sentences that have comparative keywords are comparative opinions. Moreover, not all the comparative sentences contain comparative keywords.

[56] Jindal, Niting and Liu "Identifying comparative sentences in text documents", 2006a.
[57] Bing Liu "Sentiment Analysis and Opinion Mining", 2010.
[58] Liu "Sentiment Analysis and Opinion Mining", 2010.

Now we only focus on studying two specific problems of comparative opinion sentiment analysis: identifying comparative sentences and determining the preferred entity set.

4.2.2 IDENTIFYING COMPARATIVE SENTENCES

As we said before, some phrases contain comparative keywords, but they are not comparative sentences, e.g. "*I cannot agree with you more*".

Jindal and Liu (2006) found out that almost every comparative sentence has a keyword indicating comparison. Using a list of keywords, 98% of comparative sentences were identified. The keywords are:

1. Comparative adjectives (JJR) and comparative adverbs (RBR) and words ending with –er.
2. Superlative adjectives (JJS) and superlative adverbs (RBS) and words ending with –est.
3. Other non-standard indicative words and phrases such as favour, beat, win, exceed and outperform.

Their algorithm first identifies comparative sentences, then it classifies them into four types: *non-equal gradable*, *equative*, *superlative* and *non-gradable*. Jindal and Liu showed that keywords and key phrases are sufficient in order to recognize the sentences that are analysed by Support Vector Machine (SVM).

4.2.3 IDENTIFYING PREFERRED ENTITIES

There are some differences between regular and comparative opinions: for example, comparative opinions do not express a clear sentiment about an aspect or an entity. For this reason, we cannot analyse them in the sentiment classification field. The main characteristic of comparative sentences is the ability to classify two entities that have some aspects in common.

Most comparative sentences make an evaluation between two sets of entity. In this case applying the sentiment analysis means to find out which is the preferred entity for a customer.

Below we describe how to identify the preferred entity set in a comparative sentence.

Studies made for regular sentences can be applied also for comparative ones: the structure is different, but they express an opinion, so many discoveries can be applied in both sentences. Many researchers like Ding, Ganapathibhotla and Liu found out the method to do this transfer[59].

In order to do that, we identify two types of comparative words[60]:

1. *General-purpose comparative sentiment words: Type 1 comparatives and superlatives words* (this category includes words like *better, worse, etc.*, e.g. iPhone battery is better than Samsung one), which often express positive or negative feelings independently from the context; if a sentence has these words, is easy to recognize in which entity set it belongs to. In the case of Type 2 comparatives, formed by adding *more, less, most* or *least* before adjectives/adverbs, the preferred entity set are determined by both words. The following rules are applied[61]:

   Comparative Negative::= increasing_comparative N
   | decreasing_comparative P
   Comparative Positive::= increasing_comparative P
   | decreasing_comparative N

   In these rules, N (respectively, P) indicates a negative (positive) opinion regarding the aspect that sentences analysed contained. The first rule define the following: the combination of an increasing comparative and a negative sentiment word produces a *negative comparative opinion*.

2. *Context-dependent comparative sentiment words:* in Type 1 comparatives ad superlatives, words like *higher and lower* can be included in this type of comparative words. For example, when we say "*Nokia phones have longer battery life than Motorola phones*" a positive opinion about Nokia phones and a negative opinion about Motorola phones are expressed. However, without knowing the topic, it is hard to know if "longer" is positive or negative. It is the same problem for regular opinion. In this case, "battery life" is a *positive*

[59] Ding, Xiaowen, Biung Liu, Lei Zhang "Entity discovery and assignment for opinion mining applications", 2009.
Ganapathibhotla, Liu "Mining opinions in comparative sentences", 2008.
[60] Bing Liu "Opinion Mining and sentiment analysis", 2012.
[61] Bing Liu "Opinion Mining and sentiment analysis", 2012.

potential item (PPI). In the case of Type 2 the situation is similar, but in this case comparative words and aspects are all important in determining the favourite entity.

As we said before, the context is important and crucial to understand the opinion orientations. The *pair (aspect, context_sentiment_word)* – where *context_sentiment_word* is an opinion word related to an aspect- represents an opinion context. To determine whether a pair is positive or negative, the algorithm built by Ganapathibhotla and Liu (2008) uses a large amount of external data from different contexts. From this huge set of data, they lists Pros and Cons of each entity reviewed. The purpose is to determine if the association between aspect and context_sentiment_word are more likely with Pros or with Cons. If the association between aspect and context_sentiment_word is more likely with Cons, the opinion is positive; otherwise, the opinion is negative. Note that if an adjective or an adverb is positive (respectively, negative), also its comparatives and superlative are positive (negative). The comparatives are divided into increasing or decreasing comparatives[62]. To do such lists, they applied this rule[63]: in a comparative sentence, without negation words (i.e. *not, no, ecc*.), if the comparative word in the sentence is positive (or negative), then the entities before (or after) *than* has a positive connotation. Otherwise, the entities after (or before) *than* are positive.

4.2.4 QUALITY OF REVIEWS

The aim of this session is to whether a review is helpful and useful for other users. In our analysis we will not focus on fake reviews, because they might be well crafted and difficult to recognize and easy to be mistaken with real review with poor quality. There are several ways to determine if a review is useful or not. For example, on Amazon.com, users give their opinion on other reviews by answering to the question "*Was the review helpful to you?*" and they can assign a score to each review.

[62] Bing Liu "Opinion Mining and Sentiment Analysis", 2012.
[63] Ganapathibhotla, Liu "Mining opinions in comparative sentences", 2008.

QUALITY AS REGRESSION PROBLEM

Determining the quality of reviews is a helpful method to enhance the trust of customers. This helps companies to understand the consistency of the reviews. There are several ways to ask users whether a review posted by another customer was useful or not. One way is asking directly to people and let them assign a score to each review. This information can be used in review classification or analysis. Although, there are many other indirect approaches to get the same information.

Kim et al. (2006) used Support Vector Machine (SVM) to solve the problem: they assess helpfulness basing their method on some features as presented below.

The features comprehend[64]:

> *Structure features*: review length, number of sentence, percentage of question marks (these are common when people are sarcastic) and exclamations, and the number of HTML bold tags <b> and line breaks
;
>
> *Syntactic features*: percentage of tokens (i.e. a symbol representing a word) that are of open-class (they refer to a general group composed by nouns, verbs, adjectives and adverbs), percentage of tokens that are nouns, percentage of tokens that are verbs, percentage of tokens that are verb conjugated in the first person, percentage of tokens of adverbs or adjectives;
>
> *Semantic features*: product aspects and sentiment words;
>
> *Meta-data features*: review rating (number of stars).

The researchers extract all the reviews using Amazon Web Services Application Programming Interface (API; i.e. a set of procedures that are able to achieve a goal within a program). They select reviews that have more than five feedbacks. Using Support Vector Machine, the algorithm select and divide all the feedbacks in helpful and unhelpful. They connect this data to another tool that links feedbacks with reviews, in order to understand why these reviews are helpful or unhelpful.

[64] Kim, So-Min, Patrick Pantel, Tim Chklovski, Marco Pennacchiotti "Automatically assessing review helpfulness", 2006.

Lu et al. (2010) tried to define the quality of a review, from a different point of view. They thought that the social context of reviewers can influence the language: important because it may help to understand the accuracy of the results. Specifically, these are the main hypothesis of their approach[65]:

> *Author consistency hypothesis:* reviews from the same opinion holder are the same in terms of quality: the authors assume that if a reviewer writes an helpful reviews, the following ones will be helpful (and vice versa).
>
> *Trust consistency hypothesis:* A link from a reviewer r_1 to a reviewer r_2 is an explicit or implicit statement of trust. Reviewer r_1 trusts reviewer r_2 only if the quality of or review r_2 is at least as high as that of review r_1.
>
> *Co-citation consistency hypothesis:* People trust reviews believed by other people. So if two reviewers, r_1 e r_2, are trusted by the same reviewer r_3, then their quality should be similar.
>
> *Link consistency hypothesis:* if two people are linked through social networks, their reviewers should be similar.

In order to prove the efficiency of the method, they applied it to www.ciao.co.uk. It is an e-commerce site where you can find any kind of products, which is also an enormous database of reviews. This site gives the possibility to leave reviews and to rank reviews from other users. Furthermore, user can create a trust circle composed by people that leave helpful reviews. Unfortunately, this method cannot be applied to websites, which have any social network in it.

[65] Lu, Yue, Hiuzhong Duan, Wang, Zhai, "Exploiting Structured Ontology to Organized Scattered Online Opinions", 2010

CHAPTER V

CASE STUDY: DAINESE S.P.A.

5.1 HISTORY

In 1968, Lino Dainese, the founder, at his 20s travelled to London with his Vespa. When he arrived, he noticed “Ton up bikes” (what we would call cafè racers) whose riders wore the first sets of leathers. When he came back to Italy, inspired by this trip, he decided to found a company that produced protections for motor bikers.

Figure 4 First logo draft

Source:www.dainese.com/it_en/timeline/#time_2

In 1971, the idea of producing motorbike clothing started to come true and the logo became a red devil, that represents rebellion and dynamism.

In 1972, Dainese was founded in Molvena (Vicenza) and the first article produced was a motocross pants.

In 1974, Dainese introduced elasticised inserts in order to improve the comfort and the wearability of its products. In these years the company started to invest in sponsorships and the collaborations with motorbike champions became an important part: Dieter Braun became the first Dainese official racer in the World Championship.

In 1978, the first concept of composite protector was developed. The idea was to produce a soft support for rigid parts that guarantees the absorption and the dispersion of the impact. During the year, the first back protector was realised thanks to the collaborations with Barry Sheen, a motorbike champion during 1970s, and the designer Marc Sadler, a designer famous for winning many Compasso d'Oro Awards.

In 1980 the way of driving a motorbike started to change. In fact, the professional racers would start to slither their knees on the ground. Dainese tested the first *knee sliders* with Kenny Roberts. This product was called *"Istrice"* because of its cylinder form that comes out from a base applied at the knee of the suit.
In this year, Dainese started to produce gloves. During the following years, Dainese specialized on different protections. The new mission of the company was to protect motorbikers from head to toe.

In 2001 the first prototype of D-Air was produced: this totally changed the concept of security for motobikers. D-Air is a new air bag technology that protect the rider because its inflation begins before the trigger.

Safety plays a fundamental role in the business of the company. To Dainese it means protection, comfort and reliability, factors that, when properly balanced, lead to the creation of highly effective products. The features of the products meet the requirements of a wide range of customer needs. In fact, Dainese products are not only used by professional riders but also by common ones. Riders during official races need a different balance between protection and comfort than, for instance, scooter riders. That is why engineers at D-Tec (Dainese Technology Center) work constantly to transfer the experience gained from the extreme conditions of competitions, to create a collection of products suitable for all types of motorbike riders.
Research, development and competition are, therefore, the essence of a journey that began more than forty years ago. From Giacomo Agostini to Valentino Rossi, Dainese became part of motorcycling with the greatest riders. This could be possible thanks to their inventions like back protectors, knee sliders or the evolution of existing products with the introduction of innovative technologies and materials. It was Dainese who

conceived the idea of "head-to-toe" protection which combines various clothing components with rider protection, and optimises performance and weight. It was also Dainese who created D-air®, the intelligent protection based on air bag technology for bike riders.
In 44 years of history, Dainese has become leaders on motorbike clothing and the group has acquired two important companies, leaders in producing helmets: POC and AGV.

Taking inspiration from the geometry of medieval armoury and from nature itself, Dainese made technological innovation for protection its true mission, especially in the market of official competitions.

5.1.1 DAINESE IN NUMBERS

Dainese S.P.A. headquarter is located in Molvena, while in Vicenza there are the market & sales departments and the warehouse. Subsidiaries are also located in California (USA) and in Hong Kong. The total number of employees in the group was 560 in 2014[66]. Many of them are young, under thirty years old. Lino Dainese thought that new generations could give something more to the company because they have new competencies useful to exploit business potential.

In 2014 the annual turnover was 130 million of Euros[67] and divided among different countries as in the figure below.

[66] http://www.pambianconews.com/2014/07/17/dainese-cedera-il-70-80-decido-dopo-lestate-149977/
[67] http://www.lastampa.it/2014/03/10/economia/tuttosoldi/dainese-tute-per-bikers-e-astronauti-lavoriamo-per-la-sicurezza-mhFlS3hfYq3Ov8jckDzQXK/pagina.html

Figure 5 Dainese's Turnover percentage in different Countries

Source: Dainese S.p.A.

Dainese S.p.a. is famous especially for leather suits and leather jackets: its products are a status symbol all over the world.

5.1.2 DAINESE MEANS INNOVATION

Dainese's history goes along with the evolution of motorbike protective clothing. In forty years, Dainese S.p.a. changed the life and the protection of motorbikers, thanks to a myriad of innovations. From the back protector called "Lobster", for its similarity with lobster's exoskeleton (this protection was introduced in 1978), to the aerodynamic hump in order to protect cervical vertebras, introduced in 1988. The last important innovation, born in 2007, consists of a special airbag called D|Air. This complex airbag is integrated within leather suits and it has some sensors that can activate the airbag's explosion before the driver falls. This technology is now available not only for racers but also for every day motor bikers. Furthermore, Dainese developed also the D|Armor: an open platform project that takes the airbag system from the D|Air Racing suits and puts it in an undersuit that can be worn with leather suits made by other companies. All these important innovations were applied in different fields, like snow, bicycle and equestrian sports.

A very significant project in terms of innovation is the collaboration between Dainese S.p.a. and MIT (Massachusetts Institute of Technology) of Boston to develop a spacesuit, called Biosuit. The project aims to improve the comfort for astronauts: usually suits were made of rigid materials and these caused traumas on astronauts' body. Dainese and MIT worked together in order to develop a special-suit with sensors able to pressurize the human body without limiting the movements. The pressurization is necessary to increase the air pressure, thus human being can live in space. This pressurization is not made by the usual bubble of air that creates bulky spacesuit, but it is possible thanks to special filaments that cover all the human body.

Figure 6 Dainese Biosuit for astronauts

Source: http://www.myconfinedspace.com/2013/12/18/space-lady-3/

All these innovation were inspired by Lino Dainese, the founder, who has a challenging goal in mind: save people in unusual ways, allowing them to practise sports and to drive motorbikes.

5.1.3 WHY SENTIMENT ANALYSIS?

Dainese S.p.A. produces technical products especially for motor bikers. The customer opinions are critical in this case, because Dainese is not only selling a clothing, it is offering a protection, a way to feel safer when you ride your bike.
Before purchasing such complex products, people want to have all the information in order to make their choice. Thanks to the internet, opinions are shared through blogs, forums and specialized website and reviews are available to customers. In the sector of motorbiker's clothing there are many specialised website that collect reviews, make video to describe products and have a section for the e-commerce. In particular, one of them is up-to-date and constitutes a point of reference for motor bikers: Revzilla.com. For these reasons, we decide to take this site for our Sentiment Analysis, since there are a lot of information to extract data. The results should lead us to make important considerations in order to improve Dainese products: knowing what costumers need and want will help Dainese to develop a product in line with their demands.

5.2 REVZILLA

Revzilla was founded in 2007 by three IT savvies with the passion of motorcycles: Anthony Bucci, Matt Kull and Nick Auger. Their aim was to bring the best possible shopping experience to all motor bikers addicted. This lead them to build a website that turned to be the crucial platform for all the motorbike enthusiasts.

The sector of motorbikes clothing is very competitive and Revzilla tries to catalogue a large and rider-specific product selection, providing customers with a set of products as complete as possible. Furthermore, they not only have an e-commerce section, but they make videos comparing products and underlining their differences. They rank products and collect a great number of reviews made by online users, who are not necessarily customers of Revzilla.com. Its You Tube page has 68 million visualizations and more

than 200.000 subscribers[68], which is a great number for this industry[69]. YouTube represent a great source for Bucci and co.: creating well-done and detailed videos about each product, is fundamental to spread information through all the motorcycle enthusiasts.

Revzilla literally reinvented the shopping experience for motorcycle enthusiasts. Before Revzilla.com, people bought their suits, jackets, trousers and accessories inside specialized stores and the amount of products proposed by websites was small[70]. Revzilla presents a wide gamma of articles, from several brands and categories. In this way, bikers can find a product that fits perfectly to their needs. Furthermore, Revzilla experts express their reviews on products: the goal of Revzilla is to provide a complete service, in order to allow customers to choose what is best for them. This service is profitable also because it is built on good network of suppliers. The partnerships between Revzilla and brands like Dainese, Alpinestars and Rev'it! are based on trust and reliability. There is also a sophisticated logistic structure behind the company that helps the customers to get their products delivered within few days.
In 2015 the company reached a revenue of 75 million dollars and it has 162 employees. The headquarter is located in Philadelphia, Pennsylvania.
The figures are important and this is one of the reasons that leads us to analyse Revzilla's reviews. Revzilla is also the perfect candidate for the Sentiment Analysis we are going to present. In fact, the products of Dainese's main competitors, such as Alpinestars and Rev'it!, are sold through the website. We will analyse the products of the competitors in order to compare the results of Sentiment Analysis. In this way we are focusing not only on Dainese, but also on the perception that customers have of Alpinestars and Rev'it!. The complete framework of the market that we are going to picture, represents a more reliable image of what customers need and want. The Dainese slogan is "Inspired by humans" because usually protections are developed studying the human body. This paper wants to pursue the same goal, listening customers' thoughts and needs, analysing their assessments and understanding what are the main important features for them.

[68] https://www.youtube.com/user/RevZillaTV/about
[69] http://www.bloomberg.com/news/articles/2016-04-21/revzilla-turns-employees-into-motorcycle-gear-geeks
[70] http://www.bloomberg.com/news/articles/2016-04-21/revzilla-turns-employees-into-motorcycle-gear-geeks

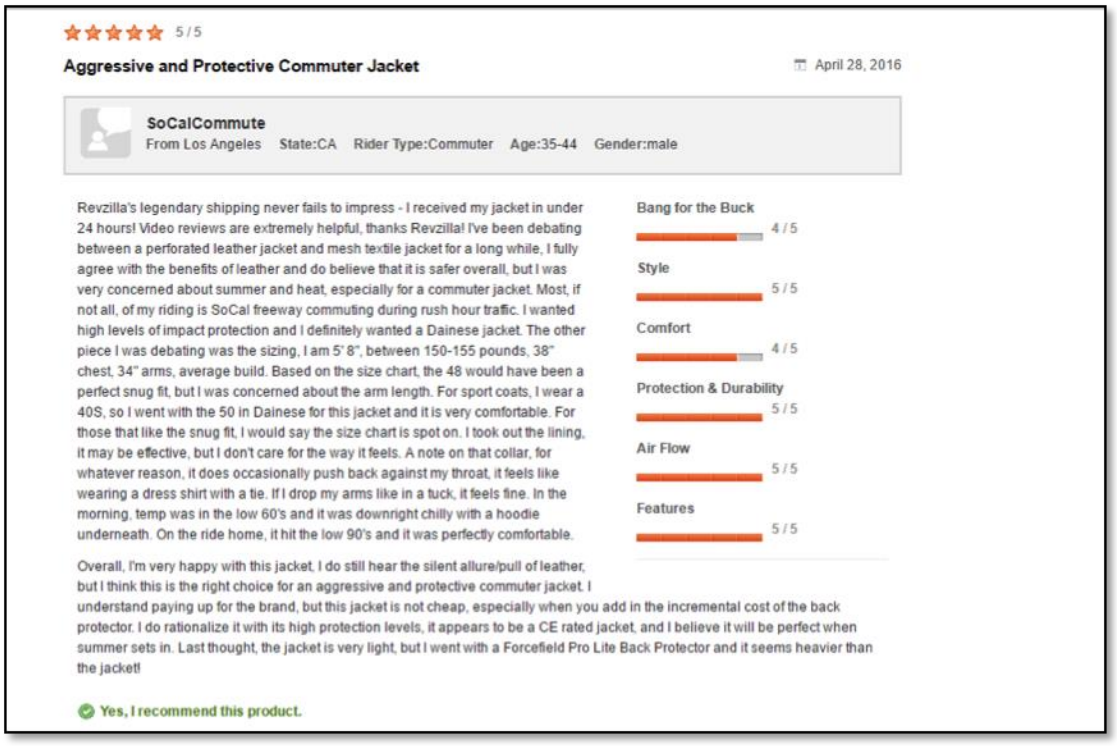
5.2.1 REVZILLA.COM'S STRUCTURE

Revzilla is a site where people can buy motorbike clothing, see video where products are reviewed and comment their purchases. The focus on people's reviews are the inputs of Sentiment Analysis.

Only customers who have an account in Revzilla.com can write a review, post photos with their products and share their experiences with other users. A person can log in directly creating an account or through Facebook or Google+.

In each review, there is the nickname for each user, and the information about where they come from, the type of riders they are, the gender, and the age. The reviewer gives a score for the overall product and for specific features as its value for money, the comfort, the protection grade and durability and also the fitting. In the end, the user can recommend or not the product. In order to share a review, the user has to create a summary with a title. When those elements are set then the user's opinion, that must be at least 50 characters long, can be posted. To complete the review, users can upload their photos, link a video and include related products. Before posting it, you can express an opinion about Revzilla using scores and giving a motivation. Once the review is posted on the internet, the visualization is similar to the Figure 7.

Figure 7 An example of review in Revzilla.com



Source:http://www.revzilla.com/motorcycle/dainese-super-speed-textile-jacket?login=returning#reviews_tab

5.3 MAIN COMPETITORS

5.3.1 ALPINESTARS

Sante Mazzarolo, a leather craft man who produced motocross boots, founded Alpinestars in 1963. The headquarter is in Asolo (Treviso), Italy. In 1960s motocross was a new sport that was enhanced in all over Europe and especially Italy. The boots produced by Mazzarolo became a standard for motocross bikers.

The motocross champion Roger Decostes noticed these boots and he wanted a pair of them during his official racings. He became one of the first brand ambassador and sponsored athlete of Alpinestars.

Alpinestars' know-how was declined in all the most important road racing, covering all the motorsports discipline: from Formula 1 to motocross.

In 2014 the annual turnover was 127 million of euros and the net profit was 2 million of euros. The annual turnover is quite similar to Dainese's one.

Alpinestars' research and development developed in the last fifteen years a new technology called Tech Air, the wearable airbag. This technology comprehends a set of sensors that are able to predict the fall (or the crash accident): this activate the process of inflation and the airbag is activated.

5.3.2 REV'IT!

Rev'it! was founded in 1995 by Ivan Vos, an entrepreneur who wanted to create a company of motorcycle clothing that could cover the lack of the market: in term of price, all the motorcycle clothing companies were at the top of the market or at the bottom of it. He founded a mid-scale oriented brand, without losing quality.

In 2000, REV'IT! introduced some innovations like the Engineering skin® design concept, that enables to re-allocate stitches to areas at a lower risk of impact[71].

In 2002 the first product was sold in USA: the market was very competitive, but the positive feedbacks lead company to found a branch in New York in 2006.

In 2008 REV'IT! arrived in MotoGP thanks to Bautista and Petrucci.

In 2009 started the collaboration with GORE-TEX, a company well-known for the Gore-Tex material, a waterproof but also breathable membrane. The partnerships are also made with motorcycle companies like Ducati and Yamaha. Customers who are

[71] https://www.revitsport.com/en/revit-en/about-us/

particularly addicted to motorcycle brands, can wear it without losing all the characteristics of a Rev'it! Jacket, that are mid-price and specifically for customers that take long trips.

In 2011, REV'IT! acquired Tryonic, an Italian company specialised in protection. This could expand the possibilities for REV'IT! because they work together in order to create high performance motorcycle clothing.

CHAPTER VI

A MODEL FOR DAINESE S.P.A.

Applying a method for Sentiment Analysis to Revzilla.com might be extremely useful for Dainese S.p.A.. Revzilla is a website where customers express their opinion not only about Revzilla's services (shipping costs, time delivery, etc.) but also about the quality of motorbike products, the way to use them and sometimes they write tips for other riders. This platform gives Dainese S.p.A. the opportunity to get opinions about their and competitors' products. Taking into consideration this site allows us to compare products from different brands and this is useful for a complete sentiment analysis.

Sentiment analysis is important especially in this sector, in which motorbike clothes are studied mainly to protect people in several situations. Thus, riders want to know if a specific product is suitable for their needs and their trips. For example, a touring rider has completely different needs from a racing one. These two type of customers use different products and different protections, so they rely on other customers' opinion in order to take a decision.

In order to apply a Sentiment Analysis to Dainese's case it was necessary to create., a partnership with a student of computer science. His work helped to achieve the results, especially in data collection and data selection.

This application has been developed with the use of open source software. Many free programs are useful to extract data from reviews. Companies and organizations specialised in sentiment analysis use paid tools for a more detailed analysis.

The chapter is divided into three parts: the first one explain how we collect data to constitute the database, the second part focuses on how we select data. In the end, we analyse the results and explains how Dainese could take advantage from this study.

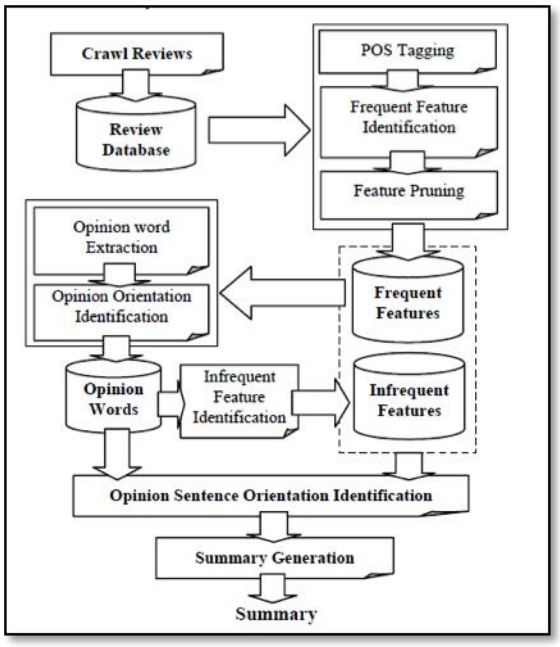
6.1 HOW TO COLLECT DATA

In this application, data collection was not fully applied due to some problems occurred during the extraction of reviews from Revzilla.com. The site protects its reviews in a very secure way, and it is impossible to automate the entire process. Revzilla.com

gathers a huge amount of data that have a great value for companies that produce motorbike clothing. Knowing that, the owners of the website, have built a complicate method in order to protect those data: in this way, they could sell the information or eventually keep it only for Revzilla's analysis.

According to the literature, the problem can be solved using the method developed by Hu and Liu (2004). In their paper "Mining and summarizing customer reviews", the process was represented as in the following figure

Figure 8 Feature-based opinion summarization



Source: Hu, Liu "Mining and summarizing customer reviews", 2004.

This method describes the sentiment analysis process, from the inputs, the customer reviews, to the outputs, the number of positive and negative results for each product. This method comprehends three main steps[72]:

1. Finding product aspects that are commented by customers
2. Identifying opinion sentences of each review and dividing them in positive and negative
3. Summarizing the results

These main steps are divided into sub-steps that are mentioned in Figure 8.

[72] Hu, Liu "Mining and summarizing customer reviews", 2004.

We now describe briefly all the process. Each step is then deeply described in order to illustrate how this method has been adjusted to our case study.

The inputs (in this case are URL pages) are downloaded (in Figure 7 the word “crawls” means “download”) and put in a review database, that is the place where all the reviews are stored. The POS tagging method (described in paragraph 2.1.1) is then applied in order to recognize frequent features (aspects and words that are written many times). Unfortunately, not all the frequent words are frequent features, so the method uses two types of pruning (see paragraph 2.2.2) in order to eliminate what are infrequent features. After that, opinion words are extracted using semantic orientations with WordNet. Knowing frequent opinion words, we can delete all the infrequent opinion words (features, aspects of a product that only a small number of people talk about) because the algorithm do not recognize these words as frequent, since the count of them is low.

DOWLOAD REVIEWS

In order to implement the method of Hu and Liu (2004), we need to identify a huge amount of reviews. In our case, the reviews are online, therefore we have to download them from Revzilla.com. From this site, we picked up a list of web addresses (or Uniform Resource Locator or URL) that have been chosen in our work. In the Table 2 stored the URL indicating all the following characteristics of the referred product:

- Brands: Dainese, Alpinestars or Rev’it!
- Part number of each item
- Extended name for each item
- Price
- URL
- Number of pages for each item
- Number of reviews for each item

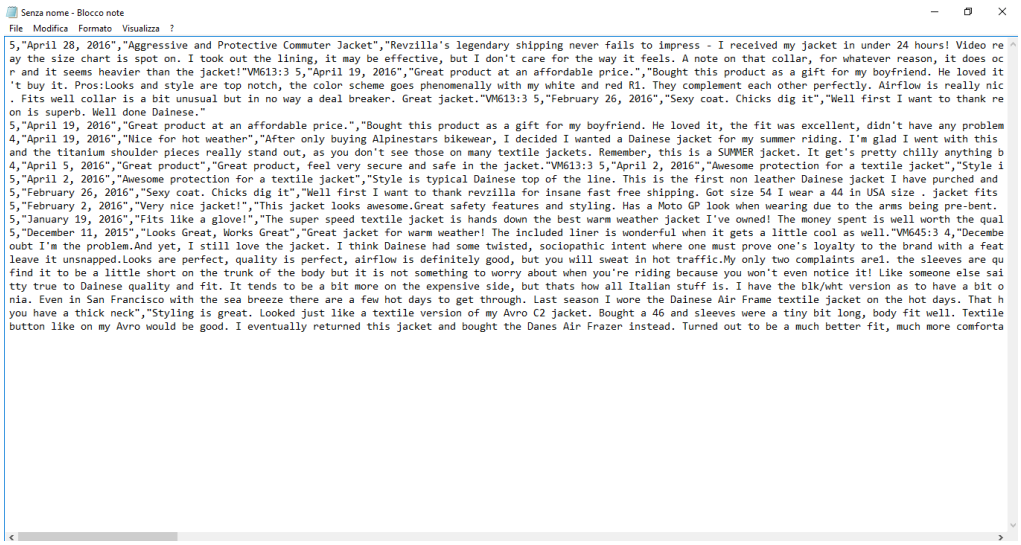
Table 2 URLs

BRAND	CODE	ARTICLE NAME	PRICE	URL	REV. PAGES	REV. NUMBERS
Dainese	1735143	SUPER SPEED TEX	€ 299,95	http://www.revzilla.com/motorcycle/dainese-super-speed-textile-jacket#reviews_tab	17	131
Dainese	1735114	AIR FRAME TEX	€ 199,95	http://www.revzilla.com/motorcycle/dainese-air-frame-textile-jacket#reviews_tab	15	112
Dainese	1815850	4 STROKE EVO	€ 179,95	http://www.revzilla.com/motorcycle/dainese-4-stroke-evo-gloves	12	92
Dainese	1795155	TRQ-TOUR GORE-TEX	€ 249,95	http://www.revzilla.com/motorcycle/dainese-trq-tour-gore-tex-boots	19	148
Alpinestars	3302716	VIPER AIR TEXTILE JACKET	€ 179,95	http://www.revzilla.com/motorcycle/alpinestars-viper-air-jacket	19	147
Alpinestars	3105015	CELER LEATHER JACKET	€ 549,95	http://www.revzilla.com/motorcycle/alpinestars-celer-jacket	8	58
Alpinestars	3556713	GP PRO GLOVE	€ 219,95	http://www.revzilla.com/motorcycle/alpinestars-gp-pro-gloves	14	105
Alpinestars	2224012	SMX-1 BOOTS	€ 159,95	http://www.revzilla.com/motorcycle/alpinestars-s-mx-1-riding-shoes	38	299
REV'IT!	FJT180	GT-R AIR TEXTILE	€ 189,99	http://www.revzilla.com/motorcycle/revit-gt-r-air-textile-jacket	19	150
REV'IT!	FJT150	SAND 2	€ 399,99	http://www.revzilla.com/motorcycle/revit-sand-2-jacket#reviews_tab	12	94

Each item has many reviews that are divided into different pages. The application we build is not able to download automatically all the pages in one time, because the site is well protected. In our case, each reviews had to be manually downloaded.

With the help of Console, a Google tool, each URL has been opened and analysed. Using programming language we are able to extract the data as in the following figure:

Figure 9 Data extraction



For each product, we collect all the reviews in a Word document. The download of reviews was possible thanks to a Google tool called Console. This application gives us the possibility to create some Java scripts that allow us to download the reviews.

The data extracted from the reviews are:

- Score (from 1 to 5)
- Date of the review
- Title
- Review's text

In the Word document, a comma delimits each information. Unfortunately, we are not able to collect all the scores about bang for the buck, style, comfort, protection and durability, airflow and features because this information are well protected.

All the reviews for each product have been collected in an excel sheet (fig. 10).

Figure 10 Database of Dainese Super Speed Tex Jacket

SCORE	DATE	TITLE	REVIEW
5	May 6, 2016	Best jacket I have ever owned for riding!!!!	Recently purchased the white black and red one. Not only is the jacket extremely comfortable when riding but the styling of it is unbelievably attractive and attention grabbing with other riders who know the Dianese accessory line. This is the 5th riding jacket I have owned and I can honestly say that I will be ordering at least two more for future use. I am that confident in this being the perfect jacket. I am extremely confident that it will be a long time before I find another Jacket for the price of this one that can offer the detail and style and comfort that the Super Speed Textile Jacket does. Absolutely you should consider this jacket and with the multiple color options it is offered in I am sure you will find one that will match or go with your current gear / bike colors. Just keep in mind it is a little smaller sized than most. Usually I wear small at 5'9 150 lbs but in this I am a medium which fits really snug!"
5	May 3, 2016	Quality Jacket!	The care and attention to detail that Dainese put into this jacket is astounding. Shoulder, elbow, and forearm protectors come standard, while I snagged the back protector to round out the coverage that this jacket offers. The cut is aggressive, but following the sizing guide on Revzilla's product page made sure this thing fit like a glove. I padded my chest size by 2 inches since I knew the back protector was going to snug things up. Massive style points, incredible fit, and supreme airflow make this a jacket that i'm going to be very happy with for a long, long time.
5	April 28, 2016	Aggressive and Protective Commuter Jacket	Revzilla's legendary shipping never fails to impress - I received my jacket in under 24 hours! Video reviews are extremely helpful, thanks Revzilla! I've been debating between a perforated leather jacket and mesh textile jacket for a long while, I fully agree with the benefits of leather and do believe that it is safer overall, but I was very concerned about summer and heat, especially for a commuter jacket. Most, if not all, of my riding is SoCal freeway commuting during rush hour traffic. I wanted high levels of impact protection and I definitely wanted a Dainese jacket. The other piece I was debating was the sizing, I am 5' 8
5	April 19, 2016	Great product at an affordable price.	Bought this product as a gift for my boyfriend. He loved it, the fit was excellent, didn't have any problems with payment or shipping.
4	April 19, 2016	Nice for hot weather	After only buying Alpinestars bikewear, I decided I wanted a Dainese jacket for my summer riding. I'm glad I went with this one, as the protection feels sturdy and airflow keeps you cool on warmer days. I'm 5'11 150 lbs and size 50 fits me well. Sleeves could be a bit longer
4	April 5, 2016	Great product	Great product, feel very secure and safe in the jacket.

6.2 HOW TO SELECT DATA

6.2.1 THE PROCESS

The sources of our data have already been determined and settled. Now it is necessary to find a tool that divided the useful information and detect the feelings within the reviews.

The open source software called RapidMiner and one of its extension (i.e. Aylien) was used in order to select all the data. RapidMiner is a platform developed for data mining, sentiment analysis and business analytics. It has many applications, in different fields: business, education, training and research. This platform allows the user to aggregate and analyse data. RapidMiner is intuitive and user-friendly: that is why we decide to apply it to our project.

This opens source software is suitable for our project because its sub-tools are able to do all the steps of the Hu and Liu's method. As we said before, after the collection of data, there is the selection of them; the first step in the Hu and Liu method is the POS tagging.

POS TAGGING

In the method proposed by Hu and Liu, the first step to select data is the classification of each word with a *part-of-speech* label (the process where each word is classified depending on its nature: adjective, noun, etc.). In RapidMiner there is a tool that filters

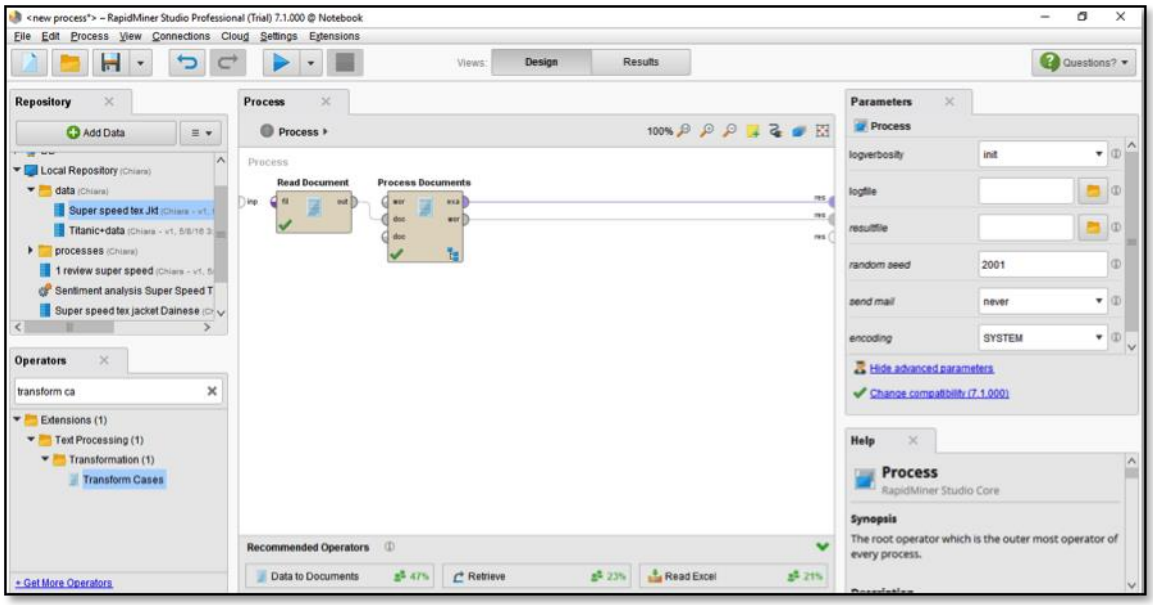
the whole document through POS tagging, in order to label every word: this filter is already built in the tool that makes sentiment analysis.

FREQUENT FEATURES

After labelling each word with POS tags, frequent features have to be highlighted, to find which are the entities taken into account. Words that express entities or aspects are recognized as frequent features if they appear in more than 1% of the reviewed sentences[73]. Hu and Liu, in their model, add an a-priori algorithm that uses a "bottom-up" approach that is based on one assumption: if an item is frequent, also its sub-items and extensions are frequent, but if an item is not frequent, so neither its sub-items and extensions are. For example, if the word "Jacket" is frequent (infrequent), also "Jackets", "Jkt" are frequent (infrequent). RapidMiner does not work in the same way, but we can create a process that is able to extract the same information.

In fact, with RapidMiner, we can build a process that can count the words. The process is represented in the following picture.

Figure 11 Process in RapidMiner to find Frequent Features



Since we have to read all the reviews and extract from them the frequent words, we need two programs: one of them that read the document where all the data are stored (in

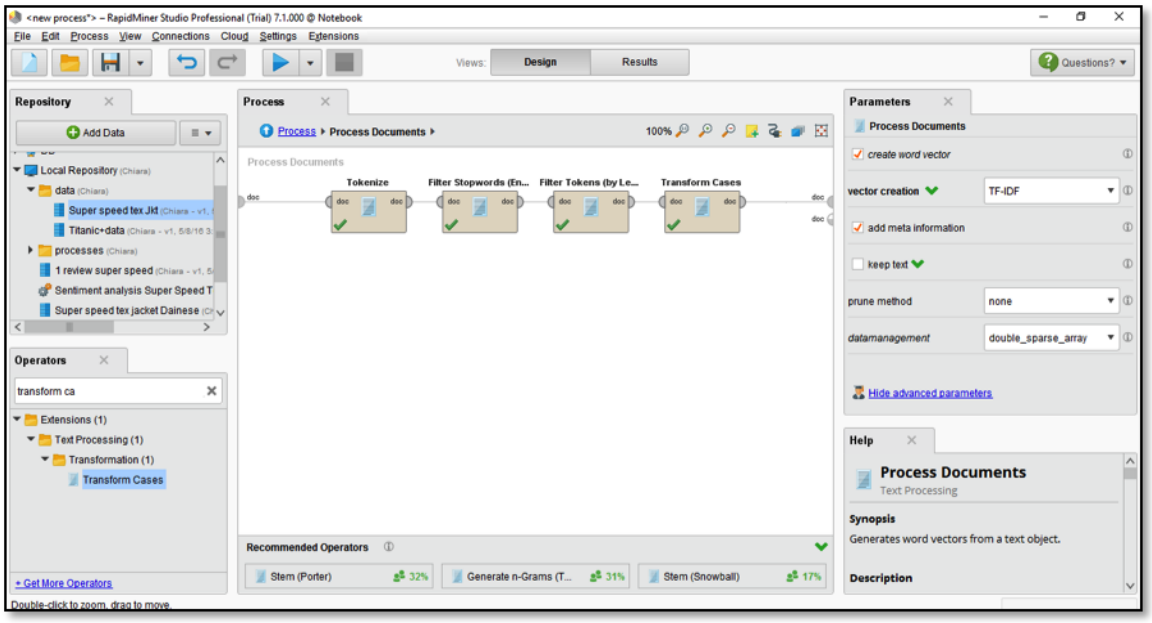
[73] Hu, Liu "Mining and summarizing customer reviews", 2004.

our case an excel file) and the second one that count the words and ranks the words by frequency.

The function of the program, called Read Document, will be used to read the document where all the reviews are.

After that, we need to use the tool *Process document* (fig. 12).

Figure 12 Process document



Inside the Process tool, four different sub-tools are merged:

- *Tokenize.* This tool is useful in order to separate all the words and count how many times each of them appears in the document
- *Filter Stopwords.* This program eliminates what in the model is called infrequent features identification. In fact, this sub-tool recognizes the aspects and the entities that are not frequent and eliminate them. In order to explain what the infrequent features are, we report an example. Considering these two sentences:

  "The voice quality of my iPhone is amazing"

  "The software of my iPhone is amazing"

  The same opinion words are used to express feelings about different aspects or entities[74] of the iPhone.

[74] Hu, Liu "Mining and summarizing customer reviews", 2004.

- *Filter Tokens by length.* This tool deletes all the words that are present in every review because they are connections, pronouns, etc. These words are composed by three or two letters most of the time, so a filter by length can cut them off.
- *Transform Cases.* The process counts in different ways words that are in lowercase instead of uppercase: this tool is able to aggregate the frequency of the words independently from the way they are written.

OPINION WORDS EXTRACTION & ORIENTATION IDENTIFICATION

Statistically, when in a sentence there is an adjective, there also is an opinion[75]. Hu and Liu built an algorithm that is able to recognize and store an adjective when it is connect to entities or aspects. Then, using WordNet (a word database where adjectives are organized in bipolar clusters: positive and negative), they extracted the orientation of those words. This is not possible with RapidMiner because the tools are not developed enough, but we can highlight the frequency of a word and decide its orientation manually.

6.2.2 THE DATA

Below the table that summarize the results in terms of frequent features and opinion words for each product. These are the words that appear frequently in reviews for each product: the distinction between frequent features and opinion words is not automatic, but manual.

[75] Wiebe, "Learning subjective adjective from corpora", 2000.

Table 3 Frequent features and opinion words

PRODUCT	FREQUENT FEATURES	OPINION WORDS
DAINESE SUPER SPEED TEX JACKET	Jacket, Dainese, Protection, Textile	Great, Good, Comfortable
DAINESE AIR FRAME JACKET	Jacket, Liner, Summer, Dainese, Protection	Great, Nice, Good
DAINESE 4 STROKE EVO GLOVES	Gloves, Protection, Hands, Dainese	Great, Good, Comfortable
DAINESE TRQ TOUR GORE TEX BOOTS	Boots, Protection	Comfortable, Grate, Good, Narrow
ALPINESTARS VIPER AIR JACKET	Jacket, Textile, Protection, Summer	Great, Good, Comfortable
ALPINESTARS CELER JACKET	Jacket, Size, Leather, Protection	Great, Comfortable
ALPINESTARS GP PRO GLOVES	Gloves, Protection, Size	Great, Good, Comfortable
ALPINESTARS SMX-1 BOOTS	Boots, Size, Protection	Great, Good, Comfortable
REV'IT! GT-R AIR JACKET	Jacket, Summer, Fit, Protection, Mesh	Great, Good, Comfortable
REV'IT! SAND JACKET	Jacket, Rain, Liner	Great, Good, Comfortable

As we can notice, the same frequent features and opinion words are connected to all the products. However, for Alpinestars and Rev'it!, the brand does not appear as frequent features. Instead, for Dainese, people have clear in mind what is the brand awareness, i.e. they know the brand. The company can take advantage from this piece of information, making the logo more visible in the products. In fact the brand is recognised as a symbol of quality and reliability. The customer knows the brand and wants to be a brand ambassador and transfer the brand awareness to other potential customers.

6.3 HOW TO ANALYSE DATA

In order to aggregate all this information, we need to summarize them and do sentiment analysis. RapidMiner has a function able to analyse words that express feelings. A feature of this function is also provides us the level of consistency of the result.

The model developed by Hu and Liu completes the sentiment analysis with two steps:

- *Predicting the Orientation of Opinion Sentence.* The orientation of opinion sentence is the dominant orientation of the words that composed the sentence; i.e. if there are more positive words than negative ones, the dominant orientation is positive, otherwise is negative. If there are more positive (negative) opinion words, then the orientation of the opinion sentence is positive (negative).
- *Summary Generation.* All the entities and aspects linked to opinion words are divided in positive and negative, according to the orientation of the opinion words. A tool is computed in order to understand if there are more positive opinions than negative ones (and vice versa). All the entities and the aspects are ranked accordingly to the frequency of appearance in the customers reviews.

These two steps are included in the RapidMiner tool called Analyse Sentiment. The conclusions exposed in the following paragraph are the result of the analysis made by this tool.

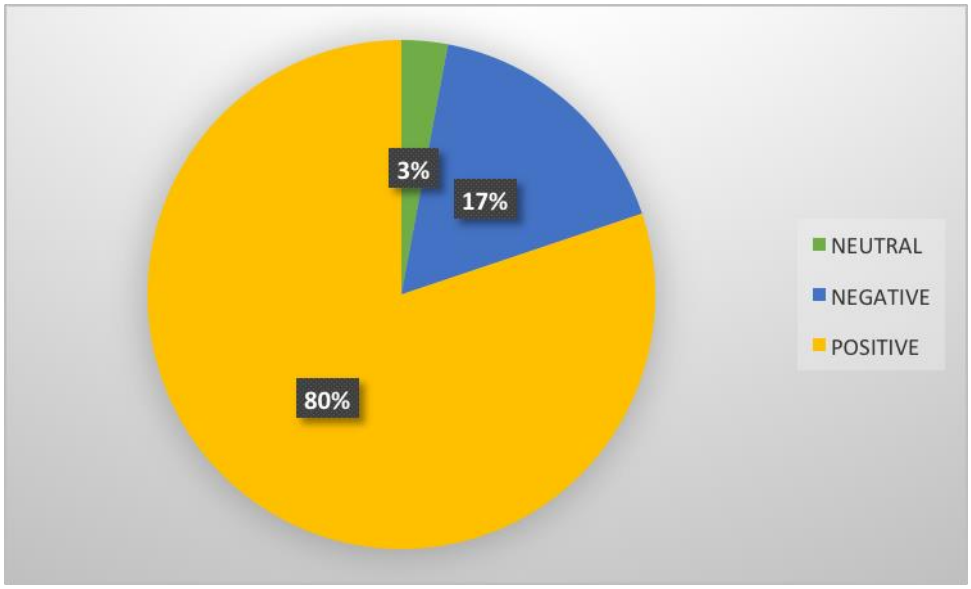
6.4 RESULTS

6.4.1 DAINESE PRODUCTS

SUPER SPEED TEX JACKET

Dainese Super Speed Tex jacket has 136 reviews divided in this way: 4 reviews are recognized as neutral, 23 as negative and 109 as positive. RapidMiner provides some graphical representations, useful to have a clear perception of the overall sentiment analysis. Below we report the pie where the results are expressed in terms of percentage (fig. 13).

Figure 13 Pie chart with percentage of positive, negative and neutral reviews for Super Speed Tex Jacket



Source RapidMiner

These results have an average polarity confidence (i.e. the reliability of the result) of 0.951 out of 1, so we can say the result is reliable.

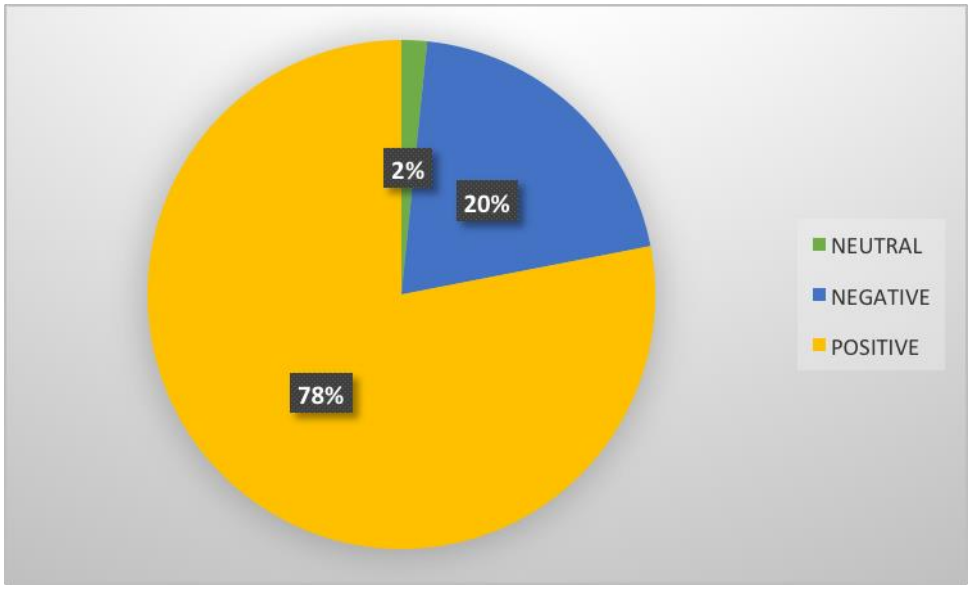
RapidMiner is able also to determine the average score taking into account every evaluation that each reviewer did. Each costumer can give from one to five stars, where five is the highest score. For Super Speed Tex jacket, the average score is 4.173 out of 5.

The average of scores and the 80% of positive reviews communicate the same result: there are more positive reviews than negative ones.

DAINESE AIR FRAME JACKET

Dainese Air Frame jackets has 123 reviews: 96 positive, 2 neutral and 25 negative (fig. 14).

Figure 14 Pie chart with percentage of positive, negative and neutral reviews for Air Frame Jacket



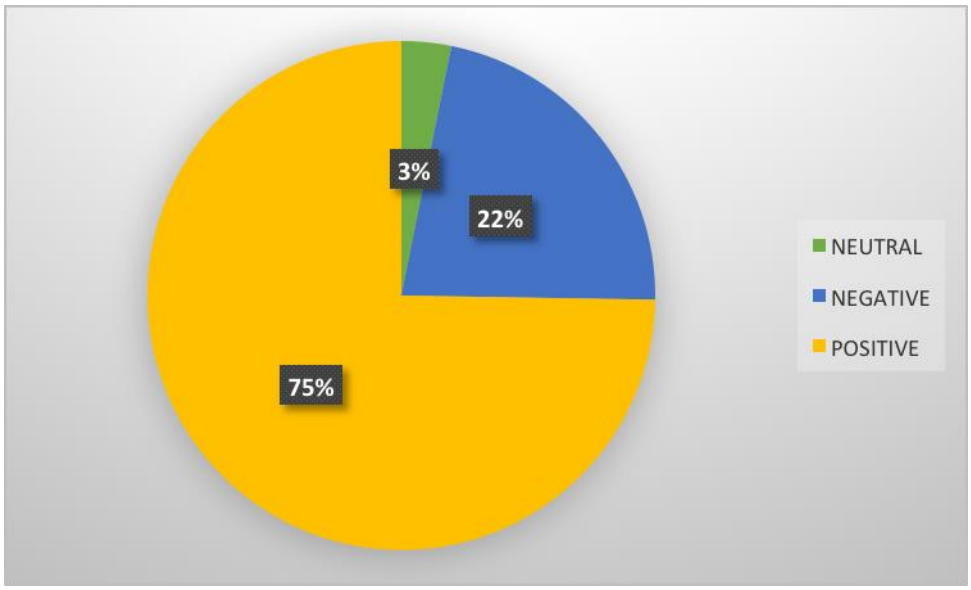
Source: RapidMiner

The average polarity confidence is 0.967 out of 1. The average score is 4.175 out of 5. The percentage of positive reviews and the average score tell us that most of the reviews are positive.

DAINESE 4 STROKE EVO GLOVES

Dainese Stroke Evo Gloves collected 94 reviews: 71 positive, 3 neutral and 20 negative ones (fig. 15).

Figure 15 Pie chart with percentage of positive, negative and neutral reviews for 4 Stroke Evo Gloves



Source: RapidMiner

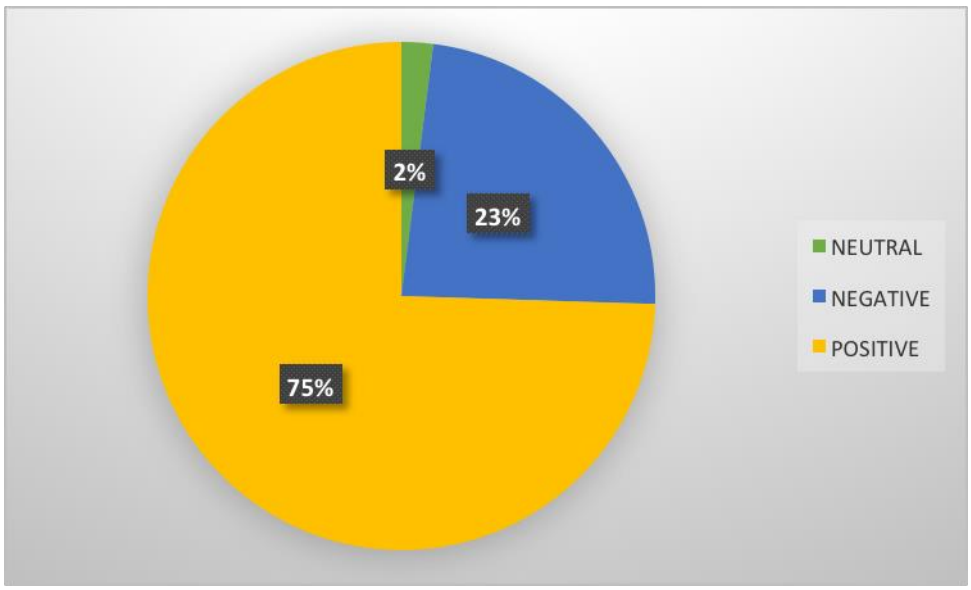
The results are reliable because the average polarity confidence is 0.950 out of 1.

The average score is 4.660 out of 5. This information confirms the majority of positive reviews.

DAINESE TRQ-TOUR GORE-TEX BOOTS

Dainese TRQ-Tour Gore-Tex Boots has 149 reviews split in this way: 111 positive, 35 negative and 3 neutral (fig. 16).

Figure 16 Pie chart with percentage of positive, negative and neutral reviews for TRQ Tour boots



Source: RapidMiner

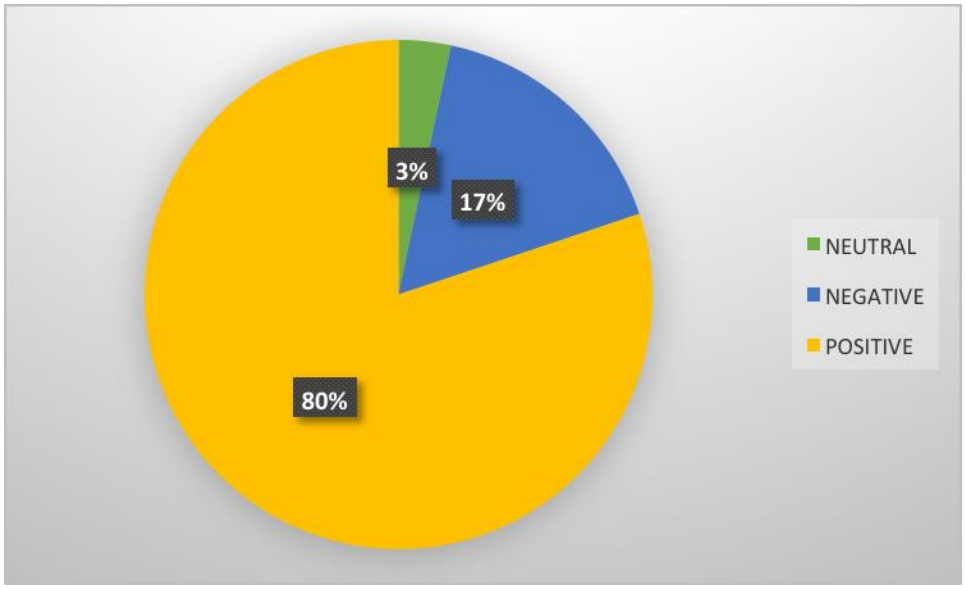
The average polarity confidence is 0.972 out of 1. The average score is 4.503 out of 5.

6.4.2 ALPINESTARS PRODUCTS

ALPINESTARS VIPER AIR JACKET

The reviews for Alpinestars Viper Air jacket are 154: 151 of them are positive, 25 negative and 5 neutral (fig. 17).

Figure 17 Pie chart with percentage of positive, negative and neutral reviews for Viper Air jacket



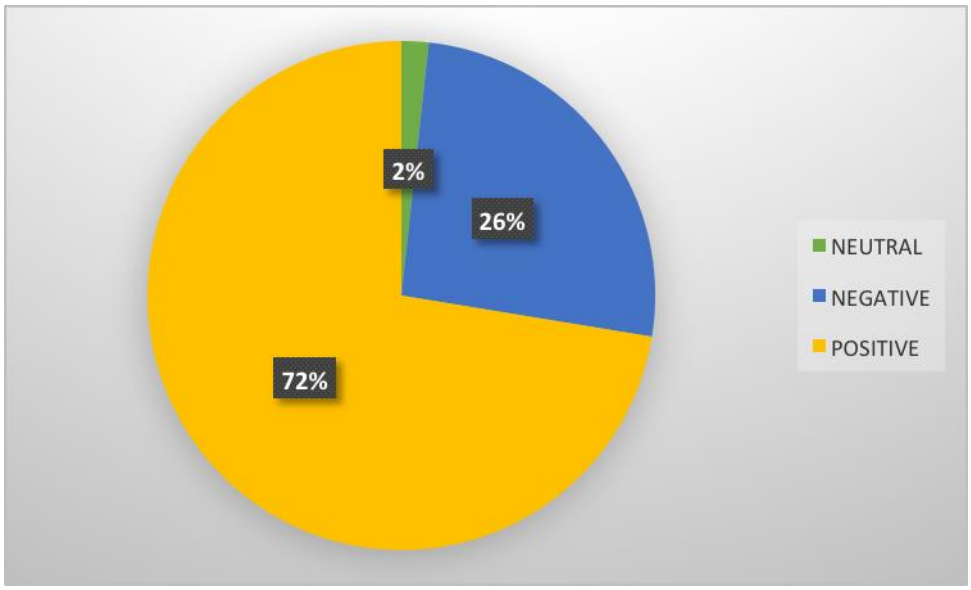
Source: RapidMiner

The average polarity confidence is 0.956 out of one. The average score is 4.481 out of 5.

ALPINESTARS CELER JACKET

There are 58 reviews in Revzilla.com for Alpinestars Celer Jacket: 42 are positive, 1 is neutral and 15 are negative (fig 18).

Figure 18 Pie chart with percentage of positive, negative and neutral reviews for Celer jacket



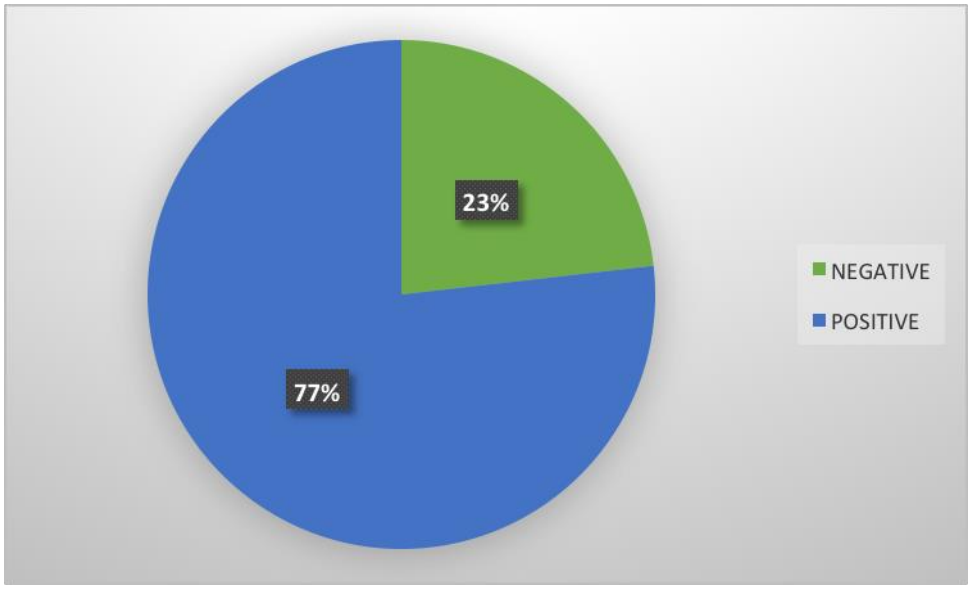
Source: RapidMiner

The average polarity confidence is 0.957 out of one. The results are also reliable because the average score is high: 4.81 out of 5.

ALPINESTARS GP PRO GLOVE

Alpinestars GP Pro gloves has 125 reviews: 96 positive and 29 negative (fig. 19)

Figure 19 Pie chart with percentage of positive, negative and neutral reviews for GP Pro gloves



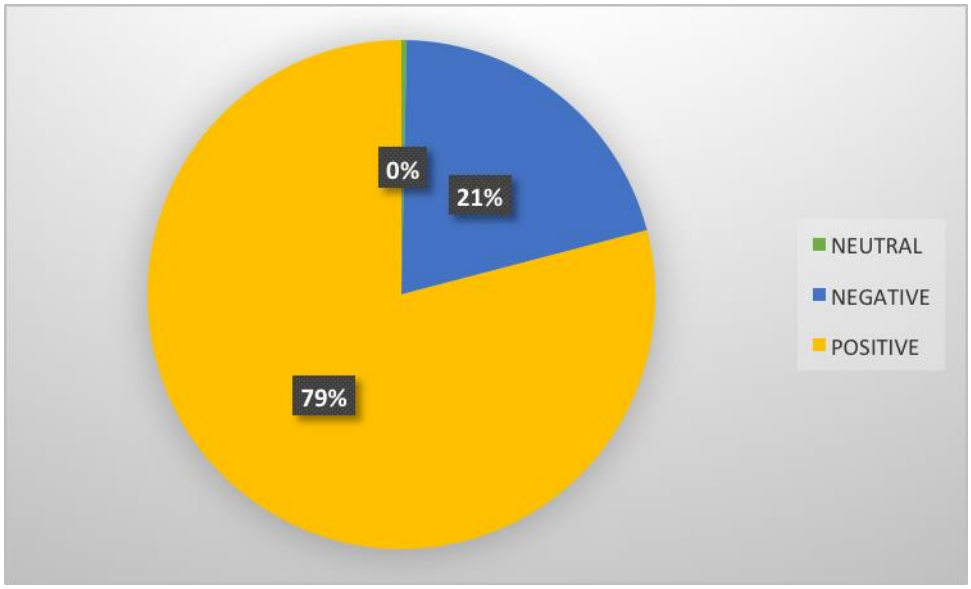
Source: RapidMiner

The average polarity confidence is 0.963 out of one. The average score is 4.552 out of 5.

ALPINESTARS SMX-1 SHOES

Alpinestars SMX-1 shoes has 301 reviews: 238 positive, 62 negative and one neutral (fig. 20)

Figure 20 Pie chart with percentage of positive, negative and neutral reviews for SMX-1 shoes



Source: RapidMiner

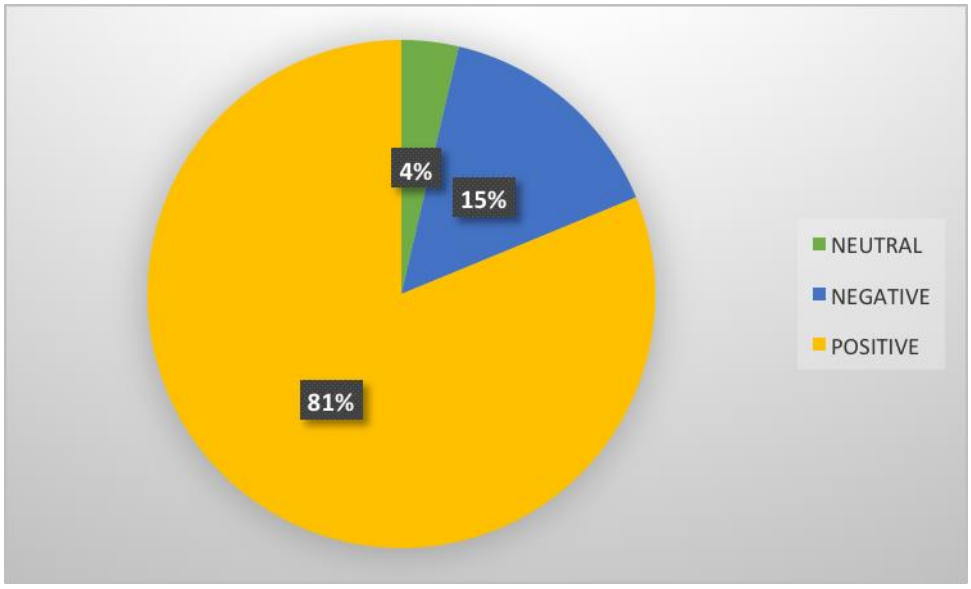
The average polarity confidence is 0.959 out of one. The average score is 4.684 out of 5.

6.4.3 REV'IT PRODUCTS

REV'IT! GT-AIR JACKET

Rev'it! GT-air jacket has 165 reviews: 134 positive, 25 negative and 6 neutral (fig. 21)

Figure 21 Pie chart with percentage of positive, negative and neutral reviews for GT-air jacket



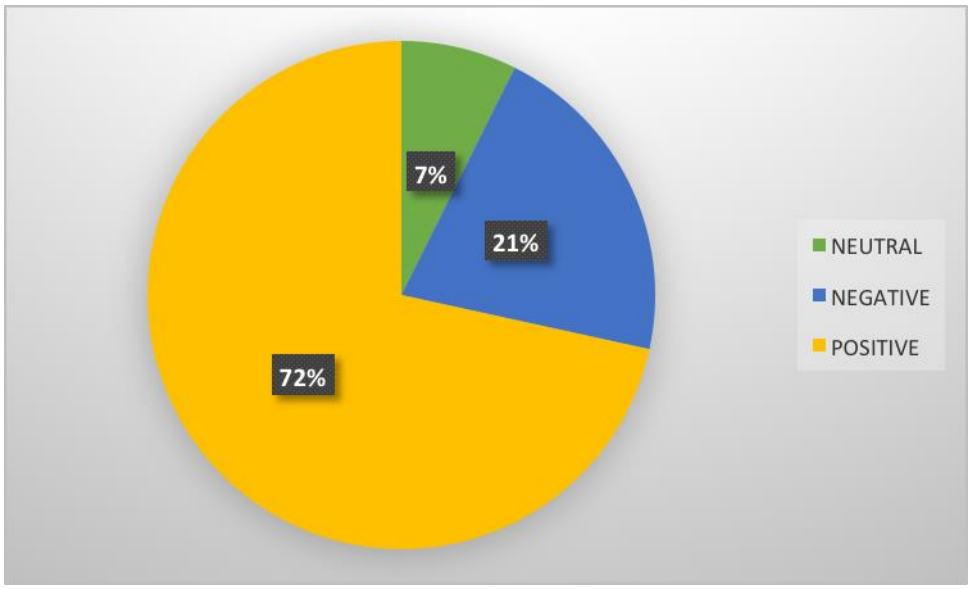
Source: RapidMiner

The average polarity confidence is 0.965 out of one. The average score is 4.711 out of five.

REV'IT! SAND JACKET

Rv'it! Sand jacket has 95 reviews: 68 positive, 20 negative and 7 neutral (fig. 22)

Figure 22 Pie chart with percentage of positive, negative and neutral reviews for Sand jacket



Source: RapidMiner

The average polarity confidence is 0.949 out of one. The average score is 4.642 out of five.

6.4.4 DAINESE VERSUS COMPETITORS

The pie charts show that most reviews are positive: this is an important information because we can assume that there are more satisfied people willing to leave a review, than unsatisfied ones.

78% of the reviews are positive for Dainese, Alpinestars and Rev'it!. The negative reviews are 21% for Dainese and Alpinestars and 17% for Rev'it!. Even though the numbers of reviews are different, all brands have almost the same percentage of positive and negative reviews.

This analysis has a great value for Dainese. In fact, knowing that the customer appreciates the quality of products (78% of reviews are positive) and its brand, we can assume that Dainese is on the right way: in fact, the annual turnover is growing year by year and there are more satisfied customers than unsatisfied ones.

On the other hand, the figures show that customers are satisfied also from other competitors. Probably these companies reached a level of perception of quality good enough that customers are satisfied in the same way. As we notice on the POS tagging, the brand awareness that Dainese has, it is not recognizable neither in Alpinestars or Rev'it!. Dainese can exploit this competitive advantage to promote its products, especially during events like official racings. We can say that professional racers broaden the target audience and they can attract customers. Moreover, people that know Dainese only for its leather suits are more willing to buy jackets and pants for everyday ride.

Before Sentiment Analysis, companies used to ask customers opinions and perceptions through surveys and focus group. These methods push people to give their opinion in a partial way, in fact, they are influenced by the context. There are disadvantages for both focus group and survey: focus groups cannot have statistic relevance, the quality of information depends on the ability of moderator[76]. Surveys can be distorted by people who participate (i.e. they can declare something that are not real), the questions must be standardized in order to avoid different interpretation and the approach does not analyse the casual correlation between choices or facts. Survey describe only what people think, overlooking the reason why they think in that way. Sentiment Analysis is more effective than the others methods, because the sample is greater and the reviews are spontaneous, and people are free to express what they think[77].

Until now Dainese has consulted extreme users, like MotoGP racers and people who usually make long trips by motorbike. Moreover, product managers travel around the world in order to talk with shop owners, who are in contact with customers. All this information influences new products. Now, with Sentiment Analysis, Dainese S.p.A. can analyse the point of view of different people, who may not be extreme users or dealers, but have some important information that the company does not have.

[76] http://www.freniricerchedimarketing.com/wordpress/wp-content/uploads/2013/04/focus.pdf

[77] Wright, Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services, Journal of computer-mediated communication, 2006.

The method proposed in this thesis represents only a part of what we could discover analysing the reviews. In fact, other applications can be implemented to understand why customers are not satisfied and what Dainese can do to meet their needs.

Sentiment Analysis can bring benefits in many different ways, using the information to:

- Improve customer care: the analysis of feelings and expectations could give a framework of opinions, brand awareness, expectations for the future and comparison with competitors.
- Enhance the brand identity: using Sentiment Analysis, companies can know who is interested in their brand, what these people notice and what they consider connected to the brand. This method can extract useful information to develop better and more effective marketing campaigns.
- Take advantage against competitors: Sentiment Analysis can give us a benchmark of competitors and this is important especially during the market researches.

Dainese S.p.A. can take the opportunity to improve customer care, enhance the brand identity and gain more market share on competitors only implementing Sentiment Analysis. The company could get some information that cannot emerge from numbers and usual analysis: this information could be the key to improve the products.

CONCLUSIONS

Internet transformed our life in many different ways: the possibility of connection is unlimited and everyone can contact other people at all times. E-commerce, social networks, forum and blogs changed the way people face a new purchasing. In fact, customers can get access to an enormous amount of data in order to take a decision on a product. They have the possibility to check feedbacks and reviews of almost every product. Obviously, positive and negative reviews influence the behaviours of a customer. Consumers are not only passive, they also want to share their opinions and get in touch with the community of other customers with the same values. Expressing an opinion, interacting and becoming actors of the process is important for them. They are not influenced by judgements or others' opinions: this is the main reason why a company should take into consideration their opinions. Listening to the customers opinions and addressing their concerns is a way to connect directly with purchasers in order to understand their needs and their expectations. The visibility that a brand has on the internet is crucial for companies. They are interested in being reviewed in a positive way in order to push the selling. Therefore, companies should monitor the feedbacks to know if their products are appreciated or not.

Sentiment Analysis offers several applications that allow companies to analyse costumers' behaviours. Sentiment Analysis is very useful when it analyses Big Data.
Big Data gathers a huge amount of information coming from different sources. Companies usually collect information about customers or reviews made by them, but Big Data may also be applied to store sensible data Obviously, among the huge amount of information stored there are some information that are not useful. New technologies have been developed to filter the valuable information.
Collecting data is important, but clearly companies have to summarize and analyse them to obtain valuable and sensible information.
Big Data have myriad of applications: in this paper we focused on Sentiment Analysis. Sentiment Analysis is usually applied to extract customers' opinions from a text, most of the time available on the internet. This discipline helps to categorize, select and analyse feelings and feedbacks in order to delineate what users think about the company.

We described different methods that can be applied for Sentiment Analysis to select the useful information within different kind of text and sentences; moreover, the summarization aspect was taken into consideration in order to organize the data and therefore adapt the companies' strategy.

The aim of this paper is to describe a method able to analyse customers' feedbacks from online reviews about Dainese S.p.A. and its competitors taken from Revzilla.com. The process was divided in two phases: the first, collection of information; second, analysis of the reviews.

The results show us that Dainese has a better brand awareness than the other competitors and the average percentage of positive feedbacks is the same for all the three brands (Dainese, Alpinestars, Rev'it!).

The brand awareness measures the ability to recognize a brand image, and also associate it with a certain product or service[78]. Dainese is a status-symbol for motorbike enthusiasts: from the analysis that has been carried we can deduce that Dainese's products are appreciated from online users. This information can be very useful for future strategies. Considering that the brand is well-known among professional races, the company might present their collections during official racings in order to involve fans.

The percentage of positive feedbacks is the same for each company. There might be two reasons: on one hand, the overall amount of positive reviews is higher than the negative ones. Probably, people who have something to claim use different ways to communicate: Revzilla.com is not the website of companies, so if there are some problems, people may try to contact the company directly. On the other hand, we might have considered the three most important competitors, which are the top gamma in terms of quality. In fact, they offer technical products, where the level of quality has to be certified for safety reason. We conclude affirming that it would be interesting to analyse negative reviews as well. Understanding which words are connected to negative reviews and why customers have given a negative feedback would be an opportunity to improve Dainese's products.

[78] http://trackmaven.com/marketing-dictionary/brand-awareness/

BIBLIOGRAPHY

- Agrawal, Imielinski, Swami, "*Mining Association Rules between Sets of Items in Large Databases"*, IBM Almaden Research Center, 1993.
- Agrawal, Srikant, "*Fast Algorithms for Mining Association Rules*", IBM Almaden Research Center, 1994.
- Akkaya, Wiebe, Mihalcea, "*Subjectivity Word Sense Disambiguation*", Conference on Empirical Methods in Natural Language Processing, 2009.
- Aue, Gamon, "*Customizing Sentiment Classifiers to New Domains: a Case Study*", Microsoft Research, 2005.
- Bing Liu, "*Opinion Mining and Sentiment Analysis",* Graeme Hirst, University of Toronto, 2012.
- Bing Liu, "*Sentiment Analysis and Subjectivity"*, Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F.J. Damerau editors, 2010.
- Bishop, "*Pattern Recognition and Machine Learning*" Jordan, Kleinberg, Schölkopf Information Science and Statistics, 2006.
- Blitzer, McDonald, Pereira, "*Domain Adaptation with Structural Correspondence learning*", Department of Computer and Information Science, University of Pennsylvania, 2006.
- Blei, David M., Andrew Y.ng and Michael I. Jordan "Latent dirichlet allocation", 2003.
- Blitzer, Dredze, Pereira "*Domain adaptation for sentiment classification*",2007.
- Borman, "*The expectation-maximization algorithm*", 2004.
- Carenini, Ng, Zwart, "*Extracting knowledge from Evaluative Text*", Computer Science Department, University of British Columbia, 2005.
- Cukier, Mayer-Schoenberger, Viktor, "*The rise of the Big Data*", *Foreign Affairs*, 2013.
- Ding, Liu, "*Resolving Object and Attribute Coreference in Opinion Mining*", 2010.
- Ding, Xiaowen, Bing Liu, Lei Zhang "*Entity discovery and assignment for opinion mining applications*", 2009.
- European Commission Research, "*Business opportunity: Big Data*", 2012.
- Hofmann "*Probabilistic latent semantic indexing*", 1999.
- Hofmann, "*Probabilistic Latent Semantic Analysis*", EECS Department, Computer Science Division, University of California, Berkley & International Computer Science Institute, Berkley, CA, 2013.
- Hu, Liu, "*Mining and Summarizing Customer Reviews*", Department of Computer Science, University of Illinois at Chicago, 2004.
- Huang T., Lan L., Fang X., An P., Min J. "*Promises and challenges of Big Data computing in Health Science*", 2015.
- Huang T., Lan L., Fang X., An P., Min J. "*Promises and Challenges of Big Data Computing in Health Sciences*", 2015, ScienceDirect. Hobbs, Jerry R., Riloff "*Information extraction*", 2010.

- Jiang, Yu, Zhou, Liu, Zhao, "*Target-dependent Twitter Sentiment Classification*", 49th Annual meeting of the Association for Computational Linguistic, 2011.
- Jindal, Niting and Liu "*Identifying comparative sentences in text documents*", 2006.
- Jindal, Niting and Liu "*Mining comparative sentences and relations*", 2006.
- Kanayama, Nasukawa "*Fully automatic lexicon expansion for domain-oriented sentiment analysis*", 2006.
- Kim, Hyun Duk and Zhai "*Generating comparative summaries of contradictory opinions in text*", 2009.
- Kim, So-Min, Patrick Pantel, Tim Chklovski, Marco Pennacchiotti "*Automatically assessing review helpfulness*", 2006.
- Ganapathibhotla, Liu "*Mining opinions in comparative sentences*", 2008.
- Griffiths, Steyvers "*Prediction and semantic association*", 2003.
- Griffiths, Steyvers "*Integrating topics and syntax*", 2005.
- Lafferty, McCallum, Pereira, "*Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*", Department of Computer & Information Science, University of Pennsylvania, 2001.
- Laney D. "*3D Data management: Controlling Data Value, Velocity and Variety*", 2001.
- Li, Fangtao, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang and Hao Yu, "*Structure aware review mining and summarization*", 2010.
- Liu, Hsu, Ma, "*Integrating Classification and Association Rule Mining"*, Department of Information Systems and Computer Science, National University of Singapore, 1998.
- Liu, Hu, Cheng, "*Opinion Observer: Analyzing and Comparing Opinions on the Web*", Department of Computer Science, University of Illinois at Chicago, 2005.
- Liu, "*Sentiment Analysis and Subjectivity*", Handbook of Natural Language Processing, Second Edition, 2010.
- Liu, Feifan, Dong Wang, Bin Li and Yang Liu, "*Improving blog polarity classification via topic analysis and adaptive methods*", 2010.
- Lu, Yue, Hiuzhong Duan, Wang, Zhai, "*Exploiting Structured Ontology to Organized Scattered Online Opinions*", 2010.
- Morwal, Jahan, Chopra, "*Named Entity Recognition using Hidden Markov Model*", International Journal of Natural Language Computing, 2012.
- Nair, NSponarayana, "*Benefitting from Big Data- Leveraginf Unstructured Data Capabilities for Competitive Advantage*", Booz&co., 2012.
- Narayanan, Liu, Choudhary, "*Sentiment Analysis of Conditional Sentences*", Conference on Empirical Methods in Natural Language Processing, 2009.
- O'Neil, Schutt, "*Doing Data Science*", O'Reilly Media, 2013.
- Pang, Bo, Lillian Lee, Vaithyanathan "*Thumbs up? –Sentiment classification using machine learning techniques*", 2002.
- Paul, Michael J., ChengXiang Zhai and Roxana Girju "*Summarize contrastive viewpoints in opinionated text*",2010.

- Sobhana N.V., Mitra, Ghosh, "*Conditional Random Field Based Named Entity Recognition in Geological Text*", International Jpurnal of Computer Applications, 2010.
- Turney, "*Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of review. In Proceedings of annual meeting of the Association of computational linguistic*", 2002
- Viktor Mayer-Schonberger "*Big Data, a revolution that will transform how we live, work and think*", 2013.
- Wiebe, Riloff, "*Creating Subjective and Objective Sentence Classifiers from Unannotated Texts*", Department of Computer Science University of Pittsburgh, 2005.
- Wright, "*Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services*", Journal of computer-mediated communication, 2006.
- Zagibalov, Carroll, "*Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese text*", 22nd International Conference on Computational Linguistics, Cooling, 2008.

WEBLIOGRAPHY

- http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf
- http://whatis.techtarget.com/definition/Web-20-or-Web-2
- http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html
- http://www.businessdictionary.com/definition/digitization.html
- https://datafloq.com/read/understanding-sources-big-data-infographic/338
- http://www.forbes.com/sites/louiscolumbus/2015/03/22/56-of-enterprises-will-increase-their-investment-in-big-data-over-the-next-three-years/#1314b89b88b1
- http://www.ninjamarketing.it/2016/05/06/web-marketing-cose-e/
- http://www.pambianconews.com/2014/07/17/dainese-cedera-il-70-80-decido-dopo-lestate-149977/
- http://www.lastampa.it/2014/03/10/economia/tuttosoldi/dainese-tute-per-bikers-e-astronauti-lavoriamo-per-la-sicurezza-mhFlS3hfYq3Ov8jckDzQXK/pagina.html
- https://www.youtube.com/user/RevZillaTV/about
- http://www.freniricerchedimarketing.com/wordpress/wp-content/uploads/2013/04/focus.pdf
- http://www.bloomberg.com/news/articles/2016-04-21/revzilla-turns-employees-into-motorcycle-gear-geeks
- http://www.bloomberg.com/news/articles/2016-04-21/revzilla-turns-employees-into-motorcycle-gear-geeks
- https://www.revitsport.com/en/revit-en/about-us/
- http://ricerca.gelocal.it/ilpiccolo/archivio/ilpiccolo/2016/03/15/nazionale-sportsystem-nei-geni-la-proiezione-globale-35.html
- http://trackmaven.com/marketing-dictionary/brand-awareness/