Benefits of Using Data Analysis for Crowd-Sourced data

September 2014

MASTER THESIS

INFORMATION TECHNOLOGIES FoR BUSINESS INTELLIGENCE

Technical University of Catalonia

Author : Murar Claudiu Ionut

Supervisor : Professor Carles Farré Tost

Abstract

Gathering data through Crowdsourcing System has been gaining a huge popularity nowadays. The aim of the thesis project is to determine how data analysis and data visualization can be efficiently used in the context of crowd-sourced data. One of the most common ways of gathering data while using the power of crowds, is through distributed human intelligence platforms such as Amazon Mechanical Turk, or through social platforms, e.g.: Blogs, Discussion Forums, Social media website, etc. When dealing with huge amounts of data coming from social platforms , an important factor that should be taken into account is analyzing and interactively visualizing all this information , in order to offer several perspectives .

This thesis concludes my Master Degree in Information Technologies for Business Intelligence. The thesis has been performed throughout my 4'th semester of studies, from February to July 2014. My supervisor on this project has been Carles Farré Tost, from the Technical University of Catalonia.

Acknowledgment

I am grateful to all the people who supported me throughout my thesis work.I would like to express my sincere gratitude to my advisor Carles Farré Tost for the continuous support of my study and research, for his patience, motivation and enthusiasm. His guidance helped me during all the time of research and writing of this thesis.
I would also like to thank the rest of my professors for their encouragement, insightful comments, and all the help they have provided me with throughout my studies. Last but not the least, I would like to thank my family: my parents Marian Murar and Titiana Murar, for supporting me spiritually throughout my life. I am deeply grateful to my sister as well, who has been advising and helping me a lot during the last 6 months.

Barcelona,September 2014
Murar Claudiu Ionut

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Crowdsourcing is a recently used technique by which technological systems can solve certain issue that are difficult to be deciphered computationally, by handing them out to human users. In the last years crowdsourcing has been a very appealing research area, which is why a lot of improvement has been done towards this direction. The concept is based on the fact that instead of relying only on developing newer and better algorithms to handle different tasks, we could look towards other solutions and try to employ human participation[13]. Crowdsourcing systems(CS) allow humans to peform a variaty of tasks. Over the past decade, numerous crowdsourcing systems have appeared on the World-Wide Web. Examples include Wikipedia, Amazon Mechanical Turk[25], Yahoo! Answers. Moreover, crowdsourcing applications have also appeared under several names, such as: community systems, human computation, user-powered systems, social media,etc.[7]

It is important to understand what is a crowdsourcing system (CS). A crowdsourcing system is defined as being a collaborative model which allows a crowd of humans to solve a problem proposed by system administrators[7]. One of the reasons for building crowdsourcing systems is to employ human participation. and later on gather the results their results.
The challenge that might appear when gathering multiple results from a group of people, is to analyze the realiability of the gathered data in an efficient manner. The data gathered from CS systems is refered to as crowd-sourced data. Data analysis over crowd-sourced data can offer new perspectives and possibilities in monitoring, transforming and modelling the data, with the overall goal of discovering useful information[9]. Depending on the type of the data that is gathered using crowdsourcing processes, several techniques for analysing the data can be applied such as: Data mining, Text mining, Natural Language Processing.

1.2 Context

Throughout the development of my thesis project , I had to take into account one of the projects of the Crowdsourcing Research Group from UPC (Technical University of Catalonia), in which my thesis supervisor is involved. Therefore one of the goals of this thesis paper is to offer a contribution to one of the previous projects done in collaboration with the research group, entitled "Linguistic problem ranking for people with Dyslexia using crowdsourcing"[4].

1.3 Motivation

Nowadays, the software industry is experiencing a continuous growth due to the explosion of data. An alternative solution to effectively gather useful data is to employ human participation, by using crowdsourcing methods[13]. Due to this fast adoption of CS systems, harnessing the human crowd with the general purpose of collecting relevant data, has become a common phenomenon. Moreover, showing the benefits of using data analysis techniques on different types of data generated by CS systems, has captivated my attention since it can solve certain challanges that appear in the crowdsourcing field.
At a personal level, my motivation for tackling this topic for my thesis project is applying the business intelligence and data analytics knowledge that I gathered during the last two years of studies. Furthermore, regarding the academic part of my research I am positive that the results of my study can provide good insights and contribute to future research directions to the Crowdsourcing Research Group, from UPC.

1.4 Problem Area

There are multiple models and frameworks for classifying CS systems, proposed by different authors. One of the reasons why there is no standard method to categorize this systems is the fast market adoption which always creates new dimensions under which CS systems can be classified[7]. Therefore, there is a need for an overview presenting all the existing frameworks and classification models of CS systems.
A different challenge is to apply data analysis and data visualization methods on the crowd-sourced data. The general purpose for performing data analysis in this context is that in most of the cases when a person is crowdsourcing a task through a CS system, he/she is also interested in monitoring and analyzing the results received from the users. Moreover,since the individual that is interested in this analysis can be either technical or non-technical there is a need of using data visualization in order to offer a better insight. In order to benefits of using data analysis and data visualization in this context, we need to take into consideration the data type generated by the CS systems. For example, the data generated by Social Media crowdsourcing platforms such as Amazon, Wikipedia, etc., can be analyzed using text processing techniques. On the other hand, when analyzing the data generated by CS platforms such as AMT or Crowdflower,other kind of data analysis can be performed, e.g.: Descriptive,Predictive or Prescriptive analysis. In order to better explain the issues that might appear ,we can mention below the following

cases.

a) Crowdsourcing through Social Media and Community QuestionAnswering platforms

If we consider social platforms the term 'crowdsourcing' can be defined as a technique to get ideas, opinions and reviews from large groups of people. Nowadays, since many people are using social media websites, these have become a convenient way to collect extra information[9].
The text data gathered through Social Media is exponentially increasing. One of the main challenges that should be taken into account is related to text processing analysis. Consider a simple example where an individual is interested to know other people's opinion about a certain product (Movie, Book, Game, etc.). By searching for product reviews on social platforms or forums, the requester can get a good insight from other web users, e.g.:if the number of positive reviews is high, the requester might be interested in purchasing the product. Due to this reason the system administrators of social platforms are highly interested in knowing the general users' opinions regarding their company's products. Performing opinion mining techniques or sentiment analysis is a challenge that has already been tackled in the previous years. Many of the already existing sentiment analysis algorithms are efficient if the input is formed by short sentences (ie. tweets), therefore using opinion mining techniques over long sentences (reviews) can be a bottleneck since the accuracy is considered to be important. Data visualization is also a challenge that should be considered since the people interested in this analysis could have a non-technical background, which is why it is important to display the gathered analysis over the data, in a visually interactive manner.

b) Crowdsourcing a task through distributed human intelligence

The problem that might appear in this case is that some of the answers received from the crowd might be incorrect or inefficient. If we consider Amazon Mechanical Turk[12] as an example, in many cases some workers are submitting fake answers, voluntarily or involuntarily, in this way obliging the requester of the task to consider analyzing each worker profile. Therefore one of the main issues is to check for existing outliers within the gathered data. When considering uni-variable outlier detection, data distribution is important since it can influence the accuracy of outlier detection.

1.5 Scope

In order to keep a centralized approach for the thesis and due to some limitations presented later on in the chapter, I have narrowed the scope of the thesis. When considering this two main concepts, data analytics and crowdsourcing, there are two intertwined approaches.

a) Performing analysis over the data gathered from the crowdsourcing systems

b) Using crowdsourcing for deploying data analytics tasks.

Extensive research has been done lately in how to use crowdsourcing for partitioning data analytics tasks . Because of the risk of having a too broad perspective over the topic I will mainly focus my thesis topic over the first approach. Although the initial literature review presented multiple models and frameworks of crowdsourcing systems, the analysis that I have performed is centered only around two different types of crowdsourcing systems.

1.6 Limitations

In order to apply analytics to a real-case scenario, I have initially considered to use the crowdsourcing platform internal database, presented in the paper "Linguistic problem ranking for people with Dyslexia using crowdsourcing"[2]. This database contains relevant information related to worker's profile , such as : age, time spent per each task, user efficiency , etc. One of the limitations that I encountered was that this database was inaccessible over the period of this thesis. Due to this reason , checking for ways to improve the data quality in a real-scenario crowdsourcing system has become out of the scope of this thesis.

In order to overcome this limitation I searched for Yahoo! Webscope Library[18], which contained a particularly interesting dataset for my project . These datasets are provided freely for research purposes. Unfortunately the information gathered in the dataset did not prove to be relevant enough for the purpose of building efficient predictive analysis , relevant to my thesis focus.

A third option that I took into account was to use the Amazon Mechanical Turk[12] crowdsourcing platform. Assigning a task to a certain number of people and later on analyzing the results received from the workers. This approach has been out of my scope since it would have implied a relatively high cost to hire a huge number of workers from whom I can obtain relevant results.

Therefore, the solution I have found was to generate my own dataset containing similar information about the worker's profile , as we can see in the crowdsourcing platform internal database .

1.7 Objectives

The main objectives of this master thesis project are :

a) Provide a theoretical overview over the main frameworks of crowdsourcing systems classification

b) Show how data analysis can be applied over the information resulted from different crowdsourcing systems

c) Present the benefits of using data visualization, over the information resulted from different crowdsourcing systems

d) Relate the contribution of this thesis to the project "Linguistic problem ranking for people with Dyslexia", tackled by the crowdsourcing UPC research group.

1.8 Structure of the Report

The rest of the report is organized as follows: Chapter 2 presents the definitions of several used concepts throughout the thesis , as well as a comprehensive overview over several crowdsourcing system classification. Chapter 3 presents the methodology and approaches I have considered in the context of social media platforms . Chapter 4 presents an example of how data analyiss and data visualization can be applied in the context of a different type of crowdsourcing system . Chapter 5 determines a series of discussions about the results and work achieved during the thesis project. Chapter 6 comprises several future improvements that can represent a continuation of my work. Furthermore a list of useful information is presented in Apendix A, B and C.

Chapter 2

Research Background

2.1 Basic Concepts

In the context of CS systems , there are two frequentely used concepts[13]:

- *Requester* : For crowdsourcing a task , the owner who submits the task to the crowdsourcing platform is called 'requester'
- *Workers*: The people who are accomplishing the task and submitting their contributions are known as 'workers'

2.2 State of the Art

This section emphasizes the relevant work that has been done in this field, to the best of my knowledge. Crowdsourcing systems have lately become widely used in several domains, reason why there are over 7 million results for the term "crowdsourcing" on a Google search. Due to the globalization and the explosion of different crowdsourcing start-ups, it has become challenging to define a unique framework for crowdsourcing systems. Although there is not a unique way to define a crowdsourcing model, one of the classifications presented in [2], take into account the following aspect:

- The type of labor performed through crowdsourcing
- Motivation for tackling crowdsourcing methods
- Functionality of the crowdsourcing application
- Issues that crowdsourcing is trying to solve

a) Crowdsourcing systems classification based on the type of labor performed

By considering the way in which the individuals from the crowd communicate with each other, we can define the following frameworks:[9]

- *Social-production crowds:* Different individuals that are using their skills and talents in order to create a product , e.g. Wikipedia, Linux
- *Averaging Crowds*:Using the crowd to provide an average judgment which could prove valuable for certain individuals, e.g.: Stock Markets
- *Networking crowds:*The information is shared over particular social networks, ie. Twitter, Facebook
- *Transactional Crowds*:This system defines a group of people that are centered on point-to-point transactions

b) Crowdsourcing systems classification based on functionality

:

- *Crowd-Creation:* A familiar principle commonly used in management, states that "No matter who you are, most of the smartest people work for someone else"[16]. A crowd-creation system, refers to the idea of using the creative capabilities of many people, in order to achieve a common goal.
- *Crowd Voting:* The crowd-voting systems have become extremely popular, with the explosion of social media web-sites. This types of CS systems allow people to express their opinion about a topic through different rating pools.
- *Crowd Funding:* This type of CS system is a popular practice for funding certain projects. It implies raising small amounts of money from a group of people, generally via internet.

c) Crowdsourcing systems classification based on the type of problem that is being solved This particular types of CS systems are centered around the question: " What are the problems that crowdsourcing is best suited to solve?" In this particular case, we can define the following model:

- *Knowledge discovery management:* A group of people is asked to find and collect certain information from the past , while later on they are meant to report their results.
- *Broadcast search:* The people from the crowd are given the task to solve certain tasks, such as scientific challenges, e.g.: Kaggle[8]
- *Distributed human intelligence tasking:* A large data problem is decomposed into several smaller tasks which require human intelligence to be solved. The most popular example in this case is Amazon Mechanical Turk,Crowdflower,Freelancer.com, etc.

Another classification framework for CS systems is introduced by Doan et.al[7]. The classification takes into account several dimensions under which crowd-sourcing systems can be classified: *nature of collaboration* and *type of target problem*. The paper also divides the classification, by using the following concepts:

a) **Explicit systems:** This type of CS systems allow users to evaluate "items" such as books,movies,products by using textual comments, numeric scores, or tags. Systems that offer sharing and evaluation benefits include Facebook, Napster, Youtube ,etc.

b) **Implicit systems:** This type of systems allow users to have an implicit collaboration for solving certain problems or tasks received from the system owners.

Bar et.al[2] classifies CS systems in four cathegories:

- *Collective Knowledge Management*: e.g.: Wikipedia
- *Collective Creativity*:e.g.: The Sheep Market application which employs the creativity of crowds
- *Collaborative Gaming*:e.g.: Online games
- *Collaborative Voting*:e.g.:MTurk

The author considers Social Networks as part of the classification , but it discusses this type of systems independently since the interaction between individuals is different than in the previous cases of CS systems.

For the purpose of this thesis I will be referring at two different types of crowdsourcing systems that could cover both the explicit and implicit CS system frameworks. Distributed Human Intelligence systems include platforms that require collaboration to solve a certain problem, therefore we could consider it part of the implicit CS systems classification. Another reason for choosing to perform the analysis over the distributed human intelligence CS platforms, is due to the fact that the system presented in the project[4], is also based on distributed human collaboration. On the other hand, based on the information presented above, another important type of CS system is based on the data gathered from Social Media. CS systems that generate crowd-sourced data are gaining a lot of popularity, reason why there is a need of research into that direction.

2.2.1 Distributed human intelligence crowdsourcing system

As mentioned before, examples of systems that rely on distributed human intelligence are: AMT,Crowdflower,etc . Some of the biggest problems encountered in DHI CS systems refer to the detection of anomalies within the gathered data. Different techniques have been used to solve the problems that appear when performing data analysis on the data gathered from DHI systems. Outlier detection analysis techniques have already been takeled in several studies. [12] and [11] Moreover, uni-variable and multi-variable outliers have been approached in [3] and [26]. When considering crowd-sourced data, the two different types of outliers can be found which makes it even more harder to perform the detection analysis.

2.2.2 Crowdsourcing through social media platforms

As previously stated, crowdsourcing can be defined as a technique for getting information from relatively large groups of people by using different social networks. A very convenient and popular way of using crowd-sourced

data is through social platforms such as Twitter, Facebook. Moreover using opinions and reviews gathered from the crowd has become another convenient way of gathering and collecting valuable information [24] Most of the crowdsourcing campaigns are using social media in order to obtain a bigger number of contributions from groups of people. These contributions are measured under different parameters such as : Views, Shares, or Feedbacks
A proof of how social media platforms are using CS systems to obtain several benefits is the fact that in the last years more and more social platforms have decided to use CS platforms to translate their services into different languages. We can exemplify here Twitter, that decided to offer web users a chance to contribute for the initiative of offering their services in several different languages[14].

Natural Language Processing Collecting data from a group a Web Users has become a popular phenomenon. However, according to the main challenge is being able to know what kind of information we are searching for. The biggest problems that appear is performing efficient NLP analysis on the gathered data. Sentiment analysis and Opinion mining is has been intensively tackled by many researchers [22]. The main two categories of sentiment analysis algorithms fall under the classification of supervised or unsupervised models.
In artificial intelligence the term “supervised learning “is used when the data that needs to be feed to the algorithm is annotated. On the other hand, “unsupervised learning “algorithms try to offer results without using an external output. There is also a middle category of weackly supervised algorithms that is situated between supervised and unsupervised models [20].

Chapter 3

Social Media Crowdsourcing platforms

Nowadays the amount of user-generated content is continuously increasing and constitutes a new source of information. In order to facilitate the information sharing and retrieval process, many systems such as blogs or discussion forums have been developed. Nowadays, social media has given a lot of power to customer, reason why customer complaints are being made public and are amplified through different social platforms[23]. Having the right sentiment analysis tools, nowadays we can measure the customers attitude and tone, which can give us a good picture of their opinion on certain products. Therefore, many efforts are nowadays being made for assessing the authenticity and quality of gathered data.

By asking customers to write positive or negative reviews and by allowing them to communicate their real opinions, we can increase the reliability of the information. One of the key influencers of customer's behavior with respect to a certain product is the opinion of other people regarding that product[18]. Therefore, customers perception of reality and the choices they make are highly influenced by the opinions of other people.

Two of the major aspects that could influence the customer's perception over a certain product are:

a) *Sentiment orientation of the reviews:* The overall sentiment of the reviews submitted by other customers

b) *Quality of the text contained in the reviews:* The quality of the text can refer to the politeness ,conciseness or informativeness level of each review. For example we can take into consideration reviews that express a positive sentiment, but contain inconcise phrases or impolite words. This sort of reviews can be negatively interepreted by other users.

3.1 Sentiment Analysis

Sentiment analysis is a field of study that analyzes people's evaluations, sentiments, and attitudes towards certain entities. This concept can be found under many names which slightly differ one from the other, e.g.: opinion extraction, opinion mining, sentiment analysis, review mining, subjectivity analysis, etc. However, all of the above mentioned concepts are generally labelled as opinion mining and sentiment analysis.

Whenever a decision has to be made, organizations want to know the opinion of others. In reality, businesses and organizations are always interested in finding the public opinions about their services and products. Moreover, individual customers are interested in knowing the opinions of other existing users that have purchased the same product/service in the past. Until recently, when an individual needed to gather certain opinions, he/she had the choice to ask friends and family. On the other hand, when an organization needed public customers opinions , the solution was to conduct opinion polls and surveys. Nowadays, if an individual wants to buy a product, he/she is not limited anymore in asking friends or family for opinions about that particular product, since there are already many reviews and public discussions on the Web regarding the product. Furthermore, for organizations or businesses there may no longer be necessary to conduct opinion polls or surveys in order to gather certain public opinions since there already is an abundance of information available

In the last years sentiment analysis algorithms have found applicability in most of the domains that deal with people opinions, attitudes or feelings.

Detecting the sentiment of the text can offer meaningful insights for the system administrator that is supervising the review information. It can be useful both for detecting outliers within the reviews , or for simply having a better overview over the opinion state of the text submitted by customers. Therefore , two different ways for performing sentiment analysis are either through a supervised or an unsupervised method.

3.1.1 Unsupervised sentiment analysis algorithms

The challenge of building efficient unsupervised sentiment analysis algorithms has been tackled in many recent studies. The main goal for building unsupervised algorithms is to overcome the domain dependency in sentiment classification and to reduce the usage of annotated training data [21].

3.1.1.1 Unsupervised learning algorithm using Pointwise Mutual Information for predicting the Semantic Orientation(SO)

A pointwise mutual information (PMI) model , uses the similarity of certain pairs of words or phrases in a given phrase, to calculate the average semantic orientation of the whole text. For computing the semantic orientation of the phrase, its similarity to a positive referential word is compared to its similarity to a negative referential word (“ Excellent” and “Poor”). The pointwise mutual comparison between word1 and word2 , is defined as we can see in the below mathematical equation [5]:

$$PMI(word1, word2) = \log_2 \frac{p(word1 \wedge word2)}{p(word1)p(word2)} \tag{3.1}$$

In this case p(word1 word2) defines the probability of the first and second word to co-occur. The words are independent then the product p(*word1*)p(*word2*) gives the probability of the two words to co-occur. The statistical

dependence degree between the words is given by the fraction between p(*word1 word2*) and p(*word1*)p(*word2*) The semantic orientation of a given sentence , is calculated as follows:

$$SemanticOrientation(Phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor") \tag{3.2}$$

Turney [27] , uses the PMI algorithm in order to calculate the semantic orientation of phrases. By computing the average semantic orientation of the phrases in a review that contains sentiment words such as adjectives or adverbs, the author is able to predict the overall sentiment orientation of the text. The semantic orientation of each phrase is computed by using the mutual information between the sentence and the word “excellent”, minus the mutual information between the sentence and the word “poor”. After calculating the average semantic orientation of all the phrases , the author classifies the review as being either positive or negative. The algorithm presented in this paper is evaluated by using reviews from four different domains, e.g.: reviews of movies, banks, automobiles and travel destinations. The accuracy of the algorithm ranges when using data from different domains but achieves and average level of 74The work performed by Turney in building an algorithm capable of detecting the semantic orientation of a text is closely related to Hatzivassiloglou and McKeown work[10]. Their algorithm performs well in predicting the semantic orientation , but it is designed only for particular categories of words (adjectives and adverbs) , rather than phrases which contain sentiment words.

3.1.1.2 Unsupervised Sentiment Analysis using Bayesian models

Natural language processing researchers have given a considerable amount of attention towards Bayesian Models. Most of the unsupervised Bayesian algorithms that exist nowadays, extend Latent Dirichlet Analysis (LDA) which is a widely spread and well-understood model.

a) **Latent Dirichlet Analysis:** The LDA concept assumes that text documents are a mixture of several topics. Due to this reason an LDA model is also called a topic model, where each word is assigned a particular topic[19]. An LDA framework has several hierarchical levels , where initially topics are associated with documents and later on words are associated with the topics. An LDA model follows several steps, such as:

   - The probability distribution, is labeled to each document.
   - According to this probability, a number of labels ’l’ is generated in each document
   - Each label is generating a word, by following a particular distribution p(w/l)
   - Parameters have been generated by previous Dirichlet distributions

   Recent studies have focused on using an LDA model to predict the sentiment polarity of a given text. Lin and He [17], focus on the conjunction between document-sentiment classification and topic detection. The model proposed in the paper, joint sentiment–topic model (JST), extends the state-of-art LDA model by adding an additional sentiment layer.The model is based on the premise that topics are depended on the

sentiment distributions, while words are being generated conditioned on the topic-sentiment pairs. The authors of the paper are evaluating the model by using both supervised and unsupervised setups, but no advantage could be demonstrated of using an additional layer.

b) **Unsupervised Sentiment Analysis using Naïve Bayes:** An alternative to using Latent Dirichlet Analysis, by building an unsupervised Naïve Bayes algorithm is predsented in paper [17] The results obtained in the paper show that this unsupervised model with a Bayesian Dirichlet distribution , can achieve a higher speed and classification accuracy comparing to LDA.

3.1.1.3 Lexicon-based unsupervised sentiment analysis algorithm

The most important indicators of sentiments are the sentiment words, which are also called opinion words. These words are used to express either positive or negative sentiments. Apart of individual sentiment words there are also some common phrases that are used to express the general feeling of the text. A sentiment lexicon can be defined as a list of such opinion words or phrases. Even though opinion words are important for performing sentiment analysis, the problem is far more complex than this. Therefore, using a lexicon-based approach in detecting the general sentiment of a text is necessary, but far from being sufficient.

a) A positive or negative opinion word can have opposite meanings and is highly dependent of the context

b) A sentence that contains opinion words doesn't necessary have to express any sentiment. This phenomenon can happen for different types of sentences. Interrogative or conditional sentences are two important types of sentences.

c) Sarcastic sentences that contain opinion words are also hard to deal with,e.g.: "Such a good car! It got damaged 10 days ago". Even though sarcasm is not that common in product reviews, it is still an issue that should be taken into account.

After looking at the above issues, we can understand why the lexicon-based approach in building an unsupervised sentiment analysis algorithm offers a great opportunity for NLP researchers to make tangible progresses towards this direction. Paper [9], presents a lexicon-based approach to extract sentiment from text. It uses several dictionaries of words annotated with their corresponding polarity, while taking into account certain sentiment shifters, such as negations.
Moreover, each sentiment word has a corresponding annotated strength, which emphasizes the degree to which a word is considered to be positive or negative. After taking into account sentiment intensifiers and sentiment shifter words, the lexicon-based approach presented in the paper presents a significant improvement in terms of accuracy.

Text Analysis on Amazon product reviews by using lexicon-based approach

As stated before it has become a common practice for individuals to search through online reviews/opinions for different purposes. For example, if a person wants to buy a product, he/she typically goes to a review site (e.g.,

amazon.com) to read more information about the product together with the associated reviews submitted by other customers who have purchased the same product. If most reviews are positive, the person is likely to decide to buy the product. On the other hand if most reviews are negative, he/she will almost certainly decide not to buy it. The above example is a commonly encountered scenario in which customers' opinions can result in significant financial gains or looses for certain businesses.
In order to perform a use case I have used a dataset provided freely by Amazon after submitting a reaust to them.The dataset contains a high number of reviews submitted by different customers, with relevant information about Automotive products reviews. The dataset comes in a semi-structured format, presenting relevant metrics such as:

Product information:

- productId
- productTitle
- productPrice

Reviewer information:

- userId
- profileName

Review information:

- reviewHelpfulness
- reviewScore
- reviewSummary
- reviewText

An extract example from the dataset, is presented below. It is important to mention that each product can have more than one review.

product/productId: B0002MA4X4

product/title: Aluminum Ramp Kit

product/price: 32.70

review/userId: AQJXV779GJNDJ

review/profileName: bob

review/helpfulness: 1/1

review/score: 4.0

review/summary: ramps

review/text : " *They did what i needed them for, i did not use the pins that keep it on the truck, i did not.."*

Feature selection

For the purpose of the experiment, the main features from the dataset have been extracted. This features represent:

a) *productTitle* : Name of the product of which the customer has submitted the review for

b) *reviewerName*: Name of the customer who submitted the review

c) *reviewHelpfulness* : The degree of helpfulness of each review. Reviews come with annotations such as "26 of 32 people found the following review helpful."

d) *reviewScore*: The user can choose to rate the product from a scale of 0 to 5.

The above features presented in the dataset have been fetched using a script written in the Python programming language . The main functionality of the script is to get the product information from the dataset and create a relational database which contains the above features placed in different columns.
The other two important features referring to the review information : *reviewText* and *reviewSummary* , have not been directly extracted to the database. Each review can be considered to be either positive , negative or neutral.
As stated above, the lexicon-based approach can help to avoid some of the issues encountered in a supervised learning context and has been proved to perform well in several domains. When dealing with unsupervised learning methods it is important to take into consideration several factors, such as: sentiment shifters, negations and other different grammatical constructs which could affect the sentiment analysis results. Below, a lexicon-based method of this approach is introduced over the reviews dataset presented earlier:

a) **Mark sentiment words:** This step marks all sentiment words in each text review. Each positive word is assigned a sentiment score of +1 , while each negative word is assigned a sentiment score of -1. If we take into consideration the sentence , "The engine power of this car is not good, but the durability is long." After this step, the sentence presented above becomes "The engine power of this car is not good [+1], but the durability is long [+1]" because "good" is a positive sentiment word . The algorithm is based on counting the total number of positive words and total number of negative words presented in each review. If the review contains a higher number of positive words than negative words, then the review is considered positive. An attempt to categorize the text as being neutral has also been introduced. The review is considered to be neutral if the

difference between the total number of positive/negative or negative/positive words has a particular minimum value. After several experiments , results have shown that using a value of 2 provides the best accuracy. The Python script for counting the total number of positive/negative words can be seen in Appendix B.

b) **Apply sentiment shifters:** Sentiment shifters , which are also called valence shifters are words that have the power to change sentiment orientations. There are several types of word shifters. Negation words like never, not or none are the most common types. After taking into account the sentiment word shifters , the sentence is turned into "The engine power of this car is not good [-1], but the durability is long [+1]" due to the negation word that is present before the opinion word. Handling negations is an important aspect in detecting performing sentiment analysis, especially when using a lexicon-based approach[24]. Moreover it is important to take into account that sentiment shifter words may not necessarily be present right before the opinion word. Therefore, it is essential to consider sentiment shifter words that could appear up to a certain distance before the opinion word.
The Python script for handling negations by considering sentiment shifters, can be found in Appendix B.

3.1.2 Supervised Learning Algorithm

The purpose of using supervised learning algorithms is to optimize the performance of the system using example data or past experience. Machine Learning algorithms provide a solution to the classification problem that involves two steps[24]: 1) Learning the model from a corpus of training data 2) Classifying the unseen data based on the trained model. The following are some of the Machine Learning approaches commonly used for Sentiment Classification

3.1.2.1 Nayve Bayes Classifier

The Nayve Bayes classifier is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting language detection and sentiment detection. Despite the naïve design and oversimplified assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems.The Naïve Bayes model involves a simplifying conditional independence assumption[17]. That is given a class (positive or negative), the words are conditionally independent of each other.
The performance of a Nayve Bayes sentiment classification is evaluated by using four different indexes : Accuracy, Precision, Recall and F1-score

- Accuracy is the portion of all true predicted results against all predicted results. An accuracy of 100 percent means that the predicted results are exactly the same as the actual instances.

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{3.3}$$

- Precision is the portion of true positive predicted instances against all positive predicted instances

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

- Recall is the portion of true positive predicted instances against all actual positive instances

$$Accuracy = \frac{TP}{TP + FN} \tag{3.5}$$

- F1 is the average of precision and recall

In the following subsections I will perform a series of tests over the Naive Bayes classifier in order to improve its accuracy in detecting the text sentiment. When dealing with a Naive Bayes classifier there are different steps that should be considered. In order to determine the initial accuracy of the classifier we first have to train and test the classifier over a given annotated dataset.

Training and testing the classifier

The movie reviews corpus used for this example contains 700 positive reviews and 700 negative reviews and has been used in the paper [6]. I will be using 3/4 of them as the training set, and the rest as the test set. The classifier training method expects to be given a list of tokens in the form of [(feats, label)] where feats is a feature dictionary and label is the classification label. In our case, feats will be of the form word: True and label will be one of 'pos' or 'neg'. For accuracy evaluation, we can use nltk.classify.util.accuracy available in Python.

The accuracy of the classifier is displayed as: ***"Train on 1050 instances, test on 350 instances: Accuracy=0.626"***

Calculating the precision , recall and F-measure of the classfier

Computing the classifiers metrics such as precision, recall and F-measure is done by using the *NLTK metrics module*.

The results obtained are displayed below:

pos precision: 0.65159

pos recall: 0.98

neg precision: 0.5612349

neg recall: 0.7539

Improving the feature extraction engine of the classifier by eliminating the stop words and Collocations

Improving feature extraction can often have a significant positive impact on classifier accuracy (and precision

and recall). I'll be evaluating the two modifications of the feature extraction method: Filter out stopwords and bigram collocations. The NLTK module in Python comes with a list of 120 stop words and collocations in 'English' language. Therefore, I will be using this list for checking the metrics of the classifier.

After applying the NLTK module we obtain the following metrics:
pos *accuracy: 0.67*
pos precision: 0.62159
pos recall: 0.91
neg precision: 0.51
neg recall: 0.72
neg F-measure: 0.4917

As we can see the accuracy if the classifier has slightly increased after eliminating the stopwords and collocations from the reviews.

3.1.2.2 Maximum Entropy Classfier

Maximum Entropy (ME) classification is another technique, which has proven effective in a number of natural language processing applications. Sometimes, it outperforms Naive Bayes at standard text classification[27]. In Maximum Entropy classification, the probability that a document belongs to a particular class given a context must maximize the entropy of the classification system.
Measuring the accuracy performance of a Maximum Entropy classfier is based on the same metrics as the Nayve Bayes classifer, e.g.: precision,recall, F-measure.

3.1.3 Support Vector Machine Classfier

Many recent studies have approached the idea of using SVM algorithms in opinion mining detection. Anothre approach as been to use the Support Vector Machines technique, on three different domain data sets. The chosen data sets have different sizes and contain diversified review information about three categories of products. During the experiments carried out by the authors of the experiemnts, is noticeable that the accuracy results are strongly influenced by the dataset corpus that is being used. Nevertheless, a maximum accuracy value of 91.5% has been achieved when using TFIDF, 10-fold cross validation and bigrams.
O'Keefe and Koprinska [1] have evaluated a range of feature selectors based on different weighting methods , using Support Vector Machine and Naïve Bayes classifiers. Feature selection is important in order to provide the most useful information to the classifier. For example, in the Pang et al. Movie dataset which has been widely used in several research studies, part of the feature information is irrelevant for the classifier. During the experiments, the authors have used a feature selector based on categorical proportional difference (PD). The categorical proportional difference metric has firstly been introduced by Simeon Hilderman []. The obtained results present an

accuracy of 87,15% after using as the PD metric as a feature selector, feature presence (FP) as a weighting method and SVM as a classifier. Moreover, the paper shows that in terms of accuracy the SVM classifier performs better than the Naive Bayes classifier.

3.1.4 Sentiment Analysis API

Sentiment analysis APIs are capable to compute the overall document-level sentiment with a high accuracy. Moreover they can compute the sentiment of a given text by taking consideration the following structures: entity-level , quotation-level and keyword-level sentiments. Due to the heavy research done into the opinion mining field, sentiment analysis APIs have been constantly improved in order to offer a better and higher accuracy[27]. The use-cases of sentiment analysis APIs range from social media monitoring to trend analysis.
Nowadays there is a huge list of sentiment analysis APIs that are available online.Several sentiment analysis APIs have been taken into account when trying to detect the overall sentiment of the reviews in our database. The APIs differ in terms of the functionalities that they offer. In the context of my thesis paper, when trying to detect the overall sentiment analysis of the reviews, I have considered using the following APIs:

a) **ML Analyzer :** This complex API can perform tasks such as: text classification , language detection, sentiment Analysis.

b) **Text processing API:** This API is particularly useful when trying to perform stemming and lemmatization over a given text. Due to grammatical reasons, text documents contain words that appear under different forms, e.g.: perform, performs and performing. The goal of using lemmatization and stemming techniques is to reduce the inflectional and derivationally forms of a word. Moreover this API is highly efficient in entity extraction, also known as named entity recognition. Entity extraction is useful when trying to classify parts of the text in pre-defined categories such as : organizations, names of persons, locations etc.

c) **AlchemyAPI:** This sentiment analysis is highly efficient in detecting the sentiment analysis of a given text . The algorithm behind the API can look for words that have positive or negative connotations and afterwards tries to compute to which person, place or thing they refer to . As mentioned in the previous sections, one of the main challenges that appear when trying to detect the polarity of a text is trying to handle the negations and the sentiment shifters. The algorithm behind the AlechmyAPI is able to efficiently compute and understand negations and sentiment shifters which is one of the major issues encountered in sentiment analysis.

The main reasons behind this choice were:

a) **High accuracy:** The algorithm is considered to be one of the most powerful and exact ways of detecting the sentiment polarity of a given document. It uses powerful text processing and machine learning techniques in order to offer accurate sentiment analysis results.

b) **Continuous improvement:**

- *Word disambiguation improvements:* One of the aspects that can affect the sentiment polarity of a review text is word disambiguation. AlchemyAPI presents continuous improvements regarding this issue, therefore an increased capacity in accurately determining the sentiment polarity of the review text
- *Quotation extraction improvements:* When processing the text of a review, certain phrases can occasionally be found under quotation symbols. Quoted phrases can affect the sentiment polarity of the text, reason why is important to classify this phrases and analyze them separately. According to the latest release of this API, a lot of improvement has been done in extracting and analyzing the quoted phrases.[27]
- *Named entity extraction improvements:* The engine that performs the entity extraction task is based on an algorithm that has been continuously improved in the API's latest release.

c) **Accessibility:** The API provides an advanced cloud-based text analysis infrastructure , which eliminates the difficulty of integrating certain natural language processing algorithms in our system. Moreover, AlchemyAPI provides a free plan of maximum 1000 calls per day which has proved to be more than enough for the purpose of detecting the sentiment polarity of each review text.

.

3.1.5 Analysis conclusions

a) **Accuracy:** In terms of accuracy , sentiment analysis APIs perform better than the previous stated supervised and unsupervised algorithms

b) **Cross-domain applicability:** One of the major drawbacks in using supervised algorithms is the need of obtaining annotated data for training and testing the algorithm, even though a lot of research has been done into improving the accuracy of the supervised learning algorithms over multiple domains[4]. On the other hand unsupervised or weakly supervised learning algorithms are domain independent.

c) **Accessibility:** Sentiment analysis APIs offer a cloud-based access to its functionalities, but the drawback is that most of the APIs are limited to a number of calls per day. In case there is a continuous need of performing the analysis, the necessity of using an API might be questioned. On the other hand building on premise supervised or unsupervised algorithm and efficiently improving its accuracy can confer a better flexibility

3.2 Quality Assessment

Text quality is important, because it can offer deeper insights for the system administrator. The quality level of the text review can have a direct influence over the user's behavior. For example, if a text review submitted by a certain user contains at least one impolite word , that review is considered to be impolite. In the thesis paper , we will focus on detecting the informativeness, conciseness and politeness of the review text. Therefore, quality assessment of

user-generated content has also attracted a lot of interest in the NLP area. Many methods have been developed to achieve this goal. State-of-the-art NLP techniques heavily rely on machine learning algorithms. There is a series of quality assessment metrics, such as:

a) *Review Length* : Checking for the informativeness level of the review is done by using a word counter . After eliminating all the words smaller than two characters in each review, the review length is given by the count of total number of words. This method has been previously used, in order to determine the informativeness of Wikipedia articles.

b) *Review level of politeness* : Politeness is the degree of respect for others' feelings and opinions. For this qualitative measure , we are using a separate list that contains offensive and injurious English words publicly avaialble. The algorithm for checking the level of politeness is based on eliminating all the prepositions and pronouns from each review, followed by the step of checking if the review contains any word contained in the list mentioned above. Therefore , for the purpose of the analysis each review can have two states:

   - Polite: The review does not contain any word from the list
   - Impolite: The review contains at least one word from the list

   The Python script for checking the politeness review level is found in Appendix B.1.1

c) *Review level of conciseness* : Determining how compact is the review's structure. One of the features of checking the level of conciseness into a text is by looking at the minimization of words repetition. Checking for words repetition , is done by using a lexical delimiter which basically calculates the number of distinct words / number of total words into a text. This step is done after previously eliminating all the prepositions and pronouns from the review text. The Python script for checking the conciseness review level is found in Appendix B.2.1
   After performing the text sentiment and qualitative assessment analysis the results are being fetched into the PostgreSQL Database.

3.3 Visual Analytics

Visual Analytics combines visualization and data analytics in the same platform and it is growing rapidly as a way for people to explore and understand the data. In this context , the usage of visual analytics is essential since there is a clear need of interactively displaying the analytical results obtained after the text analysis process. In order to justify the relevance and importance of visual analytics, the thesis paper will also take into account the quality assessment metrics of the text reviews and interactively display the results [20].

3.3.1 Using dashboards for displaying the results

A dashboard is a data visualization tool used to graphically display certain metrics, such as the features from the PostgreSQL database. Creating a customizable interface and the ability to pull real-time data from multiple sources, is one of the major reasons for using dashboards.
The idea behind building visualization dashboards , is to help technical or non-technical people to better understand the results gathered in the database. In our context, by creating visualization dashboards we can answer several questions. Each dashboard reffers to different metrics which can answer different types of questions.

a. Visualization dashboards used for displaying the sentiment analysis results

- **Question 1:** What is the overall sentiment of the text review posted by users, e.g.: positive, neutral or negative? Can we have a visualization dashboard that displays the sentiment analysis results?
- **Question 2:** How similar is the sentiment of the *review text* and the *review score* submitted by users? Can we have a visualization dashboard that displays any potential outliers within the data?

b. Visualization dashboards that display the quality assessment results

- **Question 1:** What is the politeness , conciseness and informativeness level of each review text? Can we build a visualization dashboard that displays this results under the form of pie-charts and bar-charts?

For tackling the above questions, we first need to accurately determine the sentiment of each review text. In order to gather highly accurate sentiment analysis results, I have decided to use the AlchemyAPI. The reason for using a sentiment analysis API instead of using any of the previous supervised or unsupervised learning algorithms described in the previous chapter, is due to accuracy constraints. Visualization dashboards are highly depended on the accuracy of data. Therefore the decision of using sentiment analysis APIs for detecting the sentiment polarity of a review text is to obtain accurate data that can later on be displayed by using a visualization dashboard.

PostgreSQL Database connection

After using the AlchemyAPI to detect the sentiment results of the reviews, I have added the results into the PostgreSQL database. PostgreSQL offers different connection facilities with Python. Figure 3.2 presents an example extract of the PostgreSQL database, with the features information about two different products: “Radiator”, and “Aluminum Ramp”. Moreover,it contains information about the sentiment orientation and the politeness level of each review.
In our case, the first step is to connect the database to a visualization tool that can perform real-time and in-depth analysis over the data. For this purpose we have used Tableau Visualization Tool, because of its integration capabilities with multiple data-sources and the graphical interactivity that it offers.

product_id character var	product_title character var	user_id character var	profile_name character var	review_helpfi character var	review_score character var	text_sentime character var	politeness_le character var
B0002MA4X4	Aluminum Ra	A92PYHU2388	Arnold R We	7/7	4.0	mixed	polite
B0002MA4X4	Aluminum Ra	A1YP6OZSQWH	Vincent P.	6/6	2.0	positive	polite
B0002MA4X4	Aluminum Ra	A2GTW6J6GI0	Kevin "pani	3/3	4.0	positive	polite
B0002MA4X4	Aluminum Ra	A9XWTKYBOP2	Carol M. Be	2/2	5.0	positive	polite
B0002MA4X4	Aluminum Ra	A2V0C4H1T8H	Vinster	4/5	4.0	positive	polite
B0002MA4X4	Aluminum Ra	AZL0H1HTUA9	F. Simko	1/1	4.0	positive	polite
B0002MA4X4	Aluminum Ra	A34H6RM3GMG	Mark Daniel	1/1	5.0	positive	polite
B0002MA4X4	Aluminum Ra	AYOVMBKO4FI	Scott Jorge	1/1	5.0	positive	polite
B0002MA4X4	Aluminum Ra	AQJXV779GJN	bob	1/1	4.0	mixed	polite
B0002MA4X4	Aluminum Ra	A1L5FPPZXL9	RailFan65	1/1	5.0	positive	polite
B000AAMY86	Automotiv C	A2NHIHQN85D	Jerry7111	8/16	2.0	mixed	polite
B000AAMY86	Automotiv C	A3NC6H0SQQ5	Ciro Villa	0/6	4.0	mixed	polite
B000AAMY86	Automotiv C	A2S05YVU3RO	Rick Master	1/9	2.0	mixed	polite
B000AAMY86	Automotiv C	A60FTP4V69M	captainmart	9/27	1.0	negative	polite
B000AAMY86	Automotiv C	A29246QM8Y7	Manuel Rive	0/8	4.0	mixed	polite
B000AAMY86	Automotiv C	A2UC08NHB63	M. W. Roska	5/20	2.0	mixed	polite
B000AAMY86	Automotiv C	AGHXVVN45T3	Robert M. M	0/12	5.0	positive	polite
B000AAMY86	Automotiv C	A3RW48814WL	Alamoman "A	16/57	1.0	negative	polite
B000AAMY86	Automotiv C	A3URPKYEXRP	cvrvr	16/72	4.0	mixed	polite
B000AAMY86	Automotiv C	A3U5NCYN7N9	V. Nguyen "	7/140	1.0	negative	polite
B000AAMY86	Automotiv C	A1RYYEBS3SZ	R. G. Hatch	3/4	5.0	positive	polite

Figure 3.1: *Database Extract sample*

The dashboard displayed in Appendix C.1 presents a graphical example, that allows the technical /non-technical person to interactively visualize several features , such as :

a) The overall Sentiment orientation of the text reviews

b) The score submitted by each customer

c) Average reviews score rating for each product

d) Politeness level of the reviews

Moreover the person interested in visualizing the analysis, has the flexibility in choosing to see this information for several different products. In the case presented below, the dashboard displays the analysis for three products. The dashboard is build in a visually dynamically way, so that whenever the person interested in this type of visualization, he/she can switch from one product to another while all the charts and diagrams will dynamically and automatically change.

Appendix C.2 presents different dashboards which displays the information gathered in the the database from a different perspective. For example ,if an individual is interested to see how similar is the sentiment of each review

with its corresponding review score. This can be useful in detecting certain outliers that might appear in the data and even spammers. Opinion Spamming is a concept that describes different “illegal activities” that try to deceive customers, by giving inaccurate positive or negative opinions about certain products. It can have many forms , e.g. , fake comments, fake reviews or fake social network postings.[16] As stated before one of the key features of social media is that it gives the opportunity to anyone from anywhere in the world to express their views and opinions freely, without disclosing their true identity. This fact comes with a price, because sometimes it allows certain people with hidden intentions and agendas to infiltrate into the system and post fake reviews and opinions in order to either promote or discredit different products, services, individuals or organizations[17]

For example a customer can rate a review with a score of “5.0”, but the text sentiment analysis could show that actually the text has a negative connotation.

Chapter 4

Data gathered from Distributed Humman Intelligence crowdsourcing sysmets

Nowadays, a considerable amount of data is coming from DHI platforms. One of the major sources for gathering this data are the crowdsourcing platforms (AMT , Crowdflower, etc) where a crowds of people are involvded into solving certain tasks. The individuals that are submitting the task can be called *requesters*, while the people who are involved into solving the task can be called *workers*.

The DHI platform presented in the paper "Linguistic problem ranking for people with dyslexia using Crodsourcing" is also an example of how crowdsourcing techniques can help in gathering useful data from groups of people. Another method that involves human participation in gathering information is by collecting survey and questionaire data

One of the issues that are often encountered when gathering data from groups of people is the fact that some users can submit random or irrelevant results to particular questions. This problem can be encountered in DHI crowdsourcing platforms when workers are voluntarily/involuntarily submitting wrong answers that contain exceptional values. Outliers within the data can strongly affect the quality of the analysis. These values can also be caused by incorrect measurements, including data entry errors.

Paper [12] presents a more exhaustive list of applications that utilise outlier detection. We can mention below some of the most common applications:

- Fraud detection: Detecting fraudulent usage of applications for credit cards or for banking systems
- Loan application: Detecting fraudulent potential customers
- Intrusion detection: Detecting illegal activities such as unauthorised access to a computer

Most real-world DHI crowdsourcing systems contain outliers that could have unusual values. It is important to consider that even though outliers can cause a negative effect over the analysis of data, in some scenarios they can

also provide interesting and useful information about the data.

One of the things that should be considered in outlier detection methods is the distribution of data. Research studies have previously presented several outlier labeling techniques , based on the distribution of data. Each outlier labeling method has certain different metrics , which are applied according to the data distribution.

Data Distributions

a) Normal distribution: Normal (Gaussian) distribution , such as the bell curve the data is distributed symmetrical, meaning that the right and left part of the distribution are mirror images of one another. The normal distribution can differ according to their dtandard deviations and means. Paper [12] presents several features of normal distributed data:

   - Gaussian distribution are symetric around their means
   - The mode, median and mean of a Gaussian distribution are equal
   - Gaussian disitributions are defined by using two parameters,*mean* and *standard deviation*
   - Around 95% of the Gaussian distribution area is found within two standard deviations, from the mean

b) Non-Normal Distributed Data:

   - *Skewed to the Right:* The data that is skewed to the right presents a tail that is extending to the right side. Data that is skewed to the right is considered to be positively skewed. Considering positively skewed data, the median and the mean are supposed to be greater than the mode.
   - *Skewed to the Left:* In case data is skewed to the left,the long tail is extending to the left side. Data skewed to the left is considered to be negatively skewed. In general cases, if the data is negatively skewed the mean should be less than the median.

   One way to determine if the data is skewed to the left or to the right is to calculate the Pearson's first coefficient. The Pearson's coefficient of skewness is also one of the most common metrics in detecting the assymetry of the data.

In our context, the data coming from DHI crowdsourcing systems can be encountered under both distributions, normal and non-normal.
Moreover the outliers found in crowd-sourced data can be either dependent or independent of other variables. Therefore it is important to consider both uni-variable and multi-variable outlier detection techniques.

4.1 Uni-variable Outlier Detection

One of the most common ways of detecting uni-variable outliers within the data, is Tukey's method[] who constructed a graphical boxplot in order to point out the outliers. A boxplot is a simple graphical tool used to display information about uni-variate parameters , such as the lower quartile, upper quartile, lower extreme and upper

extreme of a data set. The method that Tukey has used is considered to be efficient when the data contains extreme values.

Other methods used for uni-variable outlier detection consider the standard variance algorithms and sample mean.

Crowdsourcing data can sometimes contain extreme values that can affect the analysis. This anomalies can be caused either voluntarily or involuntarily. If we consider the project regarding gathering data from people with Dyslexia, we can see why this sort of anomalies can appear. The parameters of the analysis, presented in the project that are sensitive to outliers are:

- TimeSpent: The time spent per each task can slightly depend for one person to another. Although the results of the experiments performed during the project, showed that the *TimeSpent* value range between certain values, outliers can still appear. This thing can happen because of the fact that some people are using the application while performing other activities in the same time.
- *NumberofTasks*: This metric has been used to count how many times a person is completing certain exercises. Anomalies can also appear in this case, since some users can randomly send answers to exercises.

Regarding the metrics presented above, we notice that by applying uni-variable outlier detection methods we can detect potential red flags within the data.

As stated before, another way of gathering data though DHI crwdsourcing systems is to collect survey and questionnaire data from groups of people. In this case we can take into consideration the applicability of boxplots over data coming from survey data sets.

When answering a survey a person might be asked to answer questions referring to several parameters. Different studies have approached the challenge of finding outliers in surveys and questionnaires[12].

Boxplots are graphical tools, widely used to visualize uni variable outliers. When using boxplots, the information that fall outside the wiskers is considered to be an outlier.

- The Inter Quartile Range represents the distance between the lower and upper(Q1,Q3) quartiles
- Inner fences are situated at a distance of 1.5 IQR below Q1 and above Q3
- Outer fences are situated at a distance of 3 IQR below Q1 and above Q3
- The values between the inner and outer fences are possible outliers. On the other hand,an extreme value beyond the outer fences is considered to be probable outlier

In our context, due to the fact that the data gathered by using DHI crowdosuring platforms can can be highly skewed, in some cases using a standard boxplot the results will not be efficient enough. One of the solutions proposed by Vanderviere and Huber[15] was to adjust the standard boxlot in case the data is highly skewed. The experiments performed by the authors have shown that an adjusted boxplot can present more accurately results.

4.1.1 Adjusted Boxplot

The boxplot proposed by Tukey can be used for both skewed and symmetric data. As stated before the more skewed the data is , the more analysis observations can be detected as outliers. This is due to the fact that this method is based on robust measures such as the IQR ,lower and upper quartiles , without the skewness of the data[]. The method proposed by Vanderviere et. al[15], to adjust the standard boxplot, can be a solution for detecting outliers found in DHI crowd-sourced data. The steps followed by the author in adjusting the standard boxlpot, by considering highly skewed data, are presented bellow:

a) **Measuring the skewness of data:** Before building an adjusted boxplot , it is important to detect the distribution of data. In case the data is found under a Gaussian distribution other techniques could be applied,such as calculating the mean and standard deviations, rather than creating and adjusted boxplot. On the other hand if the data is not heavily skewed, building a standard boxplot would be enough for getting accurate results in outlier detection.

   In order to check the skewness of data, the author has proposed using the *medcouple* measure, noted as MC. The MC measure always ranges between [-1,1]. If the MC value is smaller than 0, data is skewed to the left. The opposite situation takes place in case data is skewed to the right, the value will be grater than 0. Brys et al[] have shown that using the MC measure give better results, comparing other methods based on quartile skewness.

b) **Building the adjusted boxplot**

   The following equation has been proposed by Vanderviere, in order to adjust the deviation of the standard box plot.

$$
\begin{aligned}
[L,U] &= [Q1 - 1.5 \times exp(-3.5MC) \times IQR, Q3 + 1.5 \times exp(4MC) \times IQR] if MC \geq 0 \\
[L,U] &= [Q1 - 1.5 \times exp(-4MC) \times IQR, Q3 + 1.5 \times exp(3.5MC) \times IQR] if MC \leq 0
\end{aligned}
\tag{4.1}
$$

   where L is the lower fence, and U is the upper fence of the interval. The values which fall outside the interval are considered to be outliers.

The results presented by the author in the above paper, show an innovative method of tackling highly skewed data. Regarding the thesis context, since the crowd-sourced data gathered from DHI systems can be highly skewed, creating an adjusted boxplot is a solution in tackling this issue.

A complete range of libraries for helping the developers build adjusted boxplots, have been recently included in R programming language. We can use the *adjbox* and *mc* functionality, for calculating the MC measure and creating the adjusted boxplot.

4.2 Integrating outlier detection analysis with data visualization

Different types of data analysis can be performed on DHI crowdsourcing systems. I have previously mentioned how outlier detection techniques can be applied in order to solve one of the issues that can affect the accuracy of

the analysis.
When looking at the individuals or entities interested in this analysis results, we can notice that there is a need of making the analysis as understandable as possible. Furtermore,data visualization can offer certain insights into the analysis, both for technical and non-technical people.
Nowadays most of the programming languages , such as Java, C, C++, offer graphical capabilities that can be used to display the results of the program. On the other hand standard open-source software such as R and Python, which are powerful and efficient in performing analytical operations, have only lately started to offer integration capabilities with existing visualization tools.
In this section I will refer to the interoperability of R software with two of the most popular visualization tools and platforms available on the market, *Tableau Software* and *QlikView*. I will also present an example of how outlier detection analysis can be integrated with data visualization.

4.2.1 R integration with QlickView

QlickView is considered to be one of the most popular visualization tools available on the market[]. It can provide integration capabilities with various types of data. Accoring to several comparison surveys the main benefits that QlickView provides are:

- Support for building efficient predictive and descriptive analysis
- Fast integration with data coming from multiple choices
- High performance with no restrictions and limitations on the amount of data that is beying analyzed

Recently the integration between R and QlikView can be done by using *Rattle*[]. Rattle has a graphical user interface that can generate R code which can later on be used in a QlikView application. The interoperability between R and QlikView is encountered when a user is trying to generate predictive models by using R , and display the results with QlikView.
Regarding the context of this thesis, one of the challenges that appear when gathering data from DHI systems is creating predictive models that can point out patterns within the data and certain future directions. The results gathered from the predictive model can be used later on to generate prescriptive analysis, that can have a direct impact over the functionality of the organisation that is interested in this analysis.
Building predictive models and displaying the results to organisation managers by using QlikView, can improve the reaction time to any changes that can affect the functionality of the organisation.

4.2.2 R integration with Tableau Software

Tableau is considered to be one of the leading visualization tools available on the market. When comparing Tableau with QlikView, we can notice certain advantages and disadvantages in using one instead of the other. Tackling this

differences is out of the scope of this thesis, but based on certain comparison surveys[], QlikView offers better functionality in performing descriptive predictive analysis over wider categories of data. On the other hand Tableau is more focused on offering in-depth visualization capabilities.
In the Tableau Software latest release, the integration with R has been made possible.The package used for coupling Tableau with R is called Rserve[]. In the thesis context, this interoperability can give several advantages to individuals and organisations that want to explore various types of data sources gathered from the DHI crowdsourcing systems.

As a proof of concept I have used the integration between Tableau and R , in order to display the outliers found in a survey dataset. The decision of using Tableau instead of QlikView is based on the fact that Tableau offers a complete range of visualization capabilities.
The dataset that I have used for this proof of concept, is a combined survey data gathered between 2004-2010 which contains information about retail manufacturing. The survey dataset has been previously used in other studies by Bloom et.al[] who used this dataset in order to show the best management practices across certain countries and firms.
The survey contains information regarding the following metrics:

- *Country:* The survey takes into account data coming from three different countries: Canada, USA and UK
- *Ownership:* The information regarding the ownership of certain companies found in this countries: Founder,Private individuals, Stakeholder,etc.
- *Competition:* Displays the number of competitors for each category of owners
- *Management:* The information regarding the management score for each cathegory of owners

The steps that I followed for this proof of concept, starting from data integration , outlier analysis and data visualization are detailed below:

a) In order to detect the outliers within this dataset, I have first deployed the survey data into a PostgreSQL database which can later on be connected with Tableau Software.

b) I have chosen the information found in the *Management* column to be considered for outlier detection.

c) After completing the Tableau connection with the database, I have used certain R modules within the visualization tools in order to detect the outliers

d) Adding other parameters such as *country* and ownership to the dashboard in order to confer a wider perspective over the survey data

The result of this proof of concept is a dashboard presented in Appendix C.1. By displaying in the same dashboard the *country* and *ownership* information, together with the outlier detection results we can offer a complete

overview over different types of information contained in the survey dataset. Moreover the visualization dashboard presented in Appendix C.1 offers the possibility to view the data at different levels, by giving the flexibility of filtering the *country* and *ownership* parameters.

Chapter 5

Conclusions

5.1 Crowdsourcing Systems classification

Research studies have presented different models and dimensions for classifying the crowdsourcing systems. By starting from the generic definition of a crowdsourcing system, which states that a CS is an application that allows humans to help solve a variety of tasks, we can say that all the classification frameworks are based on a common ground. It is also typical for an emerging area,such as crowdsourcing, to appear under different classification models since its applicability in many domains is growing very fast .

If we take a closer look to the classification models presented in the literature review, we can notice that many of the dimensions used when classifying the systems have the same functionalities. For example, the DHI systems are closely connected to the implicit systems presented in the paper[15], since they both rely on distributed human collaboration. On the other hand explicit crowdsourcing systems are based on the idea that users can evaluate, share and execute tasks in a collaborative manner. *Evaluating* and *Sharing* are two of the main functionalities that social media platforms are offering.

5.2 Benefits of using text analysis and data visualization techniques for Social Media crowdsourcing systems

5.2.1 Text analysis

Performing text analytics techniques on the data gathered through crowdsourcing, can have an impact over the management of any organisation who's business is dependent of the reliability of the information. Suc organisations include, but are not limited to: Amazon, Yelp, Yahoo Answers, etc. Another important issue that can affect the integrity of the organisation, is the qualitative level of the reviews.

By using text processing techniques I was able detect the degree of politeness, conciseness and informativeness of each review. Moreover the paper presented an overview over some of the existing sentiment analysis algorithms,

together with the advantages and disadvantages of using certain types of algorithms. Furthermore text analysis can be useful in dealing with opinion spammers. One of the main problems that could appear is when individuals are posting fake or inaccurate information in order to discredit their competitors. Text analysis can help in detecting the sentiment orientation of the reviews. In the use case presented in this thesis project, the sentiment orientation of each review is analyzed in order to check if it corresponds with the score submitted for the review. This technique can be used for detecting any potential red flags(outliers) within the information.

5.2.2 Visual analytics

Integrating visual analytics tools with the data analysis results, can offer deeper insights into the information. By using visual analytics non business managers from the organisations can benefit from graphically interacting with the data. As we can see in the examples performed in the thesis project, visual analytics can allow real time analysis to be performed .
Moreover , the person interested to see the analysis can have a big flexibility in visualizing the information according to certain filters. For example , both dashboards presented in the previous chapters allows the choice selection of certain parameters that can be interesting to visualize.

5.3 Benefits of using data analysis and data visualization techniques for DHI crowdsourcing systems

5.3.1 Data Analysis

Outlier detection methods can help in solving one of the issues that DHI systems are encountering nowadays. It is important to point out that the data distribution in DHI systems can highly influence the output of the analysis. Moreover, by using an adjusted boxplot[3] that can deal with highly skewed data outliers can be detected more accurately.

5.3.2 Visual Analytics

Being able to integrate R programming language with Tableau, shows that the power of data analytics software which can be combined with the interactivity of visualization tools. In this way the users who are interested in the analysis can now have a deeper overview regarding the data.

5.3.3 Contributions

a) Linguistic problem ranking for Dyslexia people:

- Data analysis: One of the objectives of this thesis has been to offer a contribution for the project presented in paper[8]. The crowdsourcing system presented in the paper is part of the DHI model. Applying outlier detection mechanisms,by taking into account the data distribution inside the system, can result in the accuracy of the analysis to highly grow.
- Data Visualization: Creating dashboards can give a certain flexibility for non-technical people to better understand the data. In the context of the project mentioned above, the developers who created the application can partially perform the data analysis directly from the visualization tools and present the analysis in a vizualy interactive manner.

Chapter 6

Recommendations for Further Work

Possible recommendations and extensions to my work could be applied to providing a wider overview over are several sentiment analysis algorithms. The unsupervised lexicon-based algorithm can also be further improved. I would mainly reffer to the following aspects:

a)Ambiguous words - For example there could be case when the same opinion word can have two different meanings. "This car is working terribly" vs. "This car has always been terribly good". In the case mentioned above the word "terrible" has a dual" opinion polarity. This issue can be tackled by using a feature extractor algorithms, in order to check the weight of each word. Moreover, there publicly available lists that contain ambiguous phrases and words. This list can be crosschecked with the opinion words found after performing the feature extraction algorithm.

b)Missed negations: In the algorithm build during the thesis project, negations and sentiment shifter words have been taken into account . Nevertheless, the distance between the negation word and the opinion word , used in the thesis project, had a particular value. There can be cases where the negation word can appear at the beginning of the sentence, for example: "I would never in my entire life think that this car can be worth buying"

c)Quoted text: Even though quoted text is not commonly encountered in social reviews, it can still be an issue worth considering in order to increase the accuracy of the algorithm. Sometimes quoted text can contain sarcastic language which is also an opinion influencer. In order to tackle this issue, the sentence can be parsed and split into chunks. Later on, each chunk can be analyzed individually.

Appendix A

Acronyms

NLP Natural Language Processing

CQA Community Question Answering

AMT Amazon Mechanical Turk

Y.A Yahoo Answers

DHI Distributed humman intelligence

Appendix B

Additional Information

language=R, backgroundcolor=, commentstyle=, keywordstyle=, numberstyle=, numberstyle=, stringstyle=, basicstyle=, breakatwhitespace=false breaklines=true, captionpos=b, keepspaces=true, numbers=left, numbersep=5pt, showspaces=false, showstringspaces=false, showtabs=false, tabsize=2, frame=shadowbox, rulesepcolor=, columns=fullflexible, deletekeywords=for, if, morekeywords=sseakomo, keywords=[2]TRUE, FALSE, keywords=[3]for, in, if, keywordstyle=[2], keywordstyle=[3], literate= 0011112213314415516617718819 91

booktabs multirow float table hyperref

[style=sedem2, caption=Matrix Function]

(i) Fast Fourier Transform:

y <- fft(x)

B.1 Python scripts

This section presents the most important algorithms developed during the thesis project . I have used Python open source programming language.

B.1.1 Checking for politeness level

```
def check_politeness_level(list_of_reviews):

    words_offensive = [’’]
    listing_politness = [’’]

        if len(words.strip()) != 0:
            words_offensive.append(word.strip())
    for name_of_user, review in list_of_reviews:
```

```
        counting_words = 0
        for word in list_of_offensive:
            if review.find(word) != -1:
                counting_words += 1
            if counting_words == 0:
            listing_politness.append([name_of_user, "polite"])
        else:
            listing_politness.append([name_of_user, "impolite"])
    return listing_politness
```

B.1.2 Checking for conciseness level in the reviews

```
def checking_for_conciseness(list_of_reviews):
    list_conciceness = []
    for name_of_user, review in list_of_reviews:
        words_1 = review.split(" ")
        words_2 = []
        for item1 in words_1:
            if len(item1.strip()) > 2:
                words_2.append(item1.strip())

        level_of_conciseness = float(len(set(words_2)))/len(words_2)

        list_conciceness.append([name_of_user, level_of_conciseness])
    return list_conciceness
```

B.1.3 Checking for the length of each review

```
def informative_reivews(list_of_reviews):
    number_of_words = ["", 0]
    for name_of_user, review in list_of_reviews:
        counting_words = len(review.split(" "))
        if counting_words > number_of_words[1]:
            number_of_words = [name_of_user, counting_words]
    return number_of_words[0]
```

B.1.4 Handling negations

```
def sentiment_analys(dictionary_review, list_of_positive, list_of_negative, list_of_offensi
    list_of_positive_cnt = 0
    list_of_negative_cnt = 0
    list_of_offensive_cnt = 0
    words_of_reviews = [item.strip(".,") for item in dictionary_review["text_of_review"].l
    for word in list_of_positive:
        if word in words_of_reviews and dictionary_review["text_of_review"].find("not " +''
            list_of_positive_cnt += 1
    for word in list_of_positive:
        if dictionary_review["text_of_review"].find("not " + word) != -1:
            list_of_negative_cnt += 1
    for word in list_of_negative:
        if word in words_of_reviews:
            list_of_negative_cnt += 1
    for word in list_of_offensive:
        if word in words_of_reviews:
            list_of_offensive_cnt += 1
    if list_of_offensive_cnt == 0:
        dictionary_review["politeness_level"] = "polite"
    else:
        dictionary_review["politeness_level"] = "impolite"
        if list_of_positive_cnt > (list_of_negative_cnt + 1):
        dictionary_review["sentiment_of_text"] = "positive"
    elif list_of_negative_cnt > (list_of_positive_cnt + 1):
        dictionary_review["sentiment_of_text"] = "negative"
    else:
        dictionary_review["sentiment_of_text"] = "neutral"
    return dictionary_review
```

B.1.5 Inserting data into the PostgreSQL database

```
def inserting_into_database(dictionary_review, database_cursor):
       try:
```

```
    database_cursor.
    ("""INSERT INTO reviews (id_of_product, id_of_user, name_of_user,
    helpfulness_of_review, score_of_review, sentiment_of_text, level_of_politeness) \
              VALUES (%(id_of_product)s, %(id_of_user)s, %(name_of_user)s, %(helpfulne
     %(score_of_review)s, %(sentiment_of_text)s,
     %(level_of_politeness)s);""", dictionary_review)
return True
```

Appendix C

Visualization Dashboards

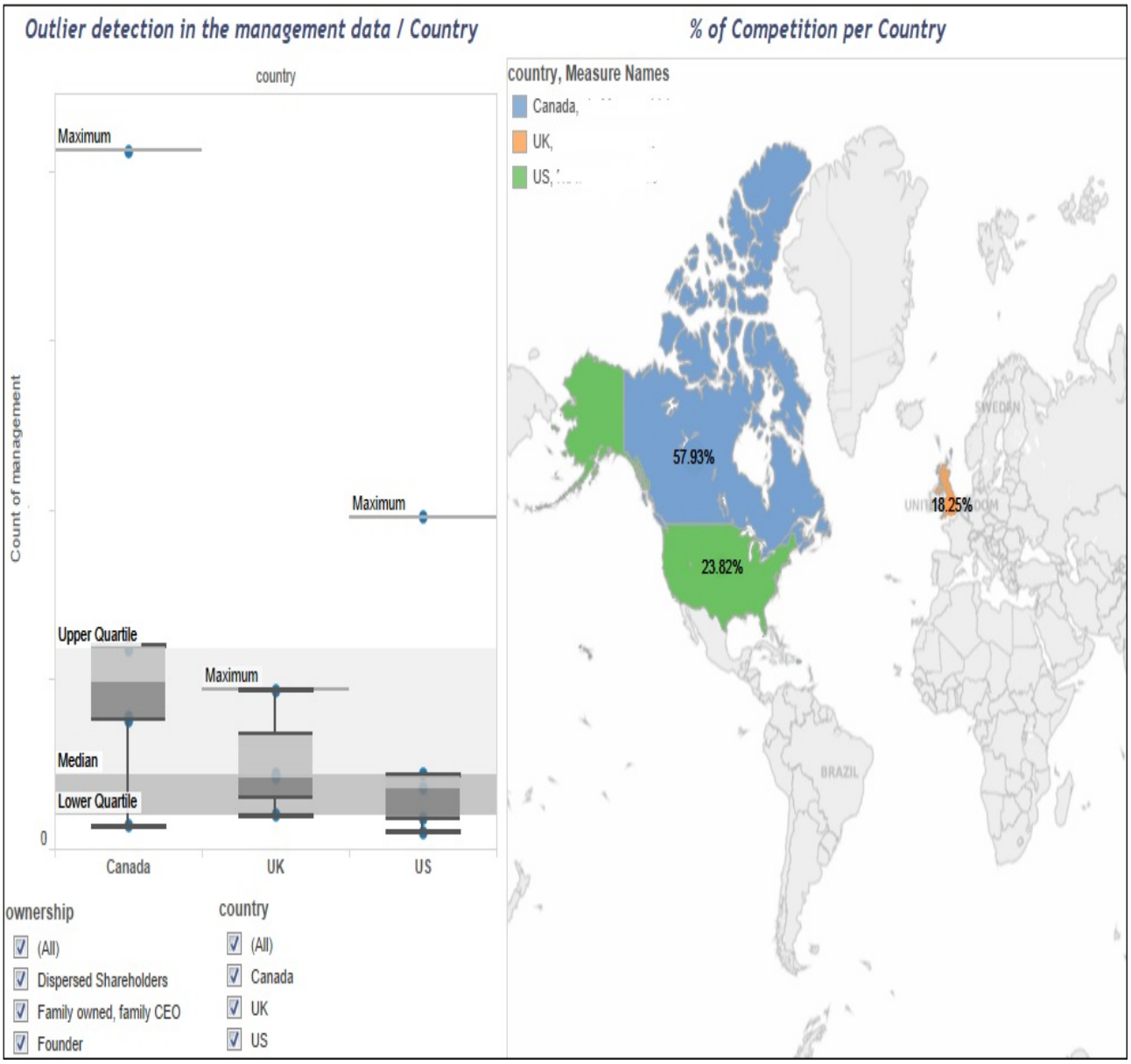

Figure C.1: *Review information dashboard*

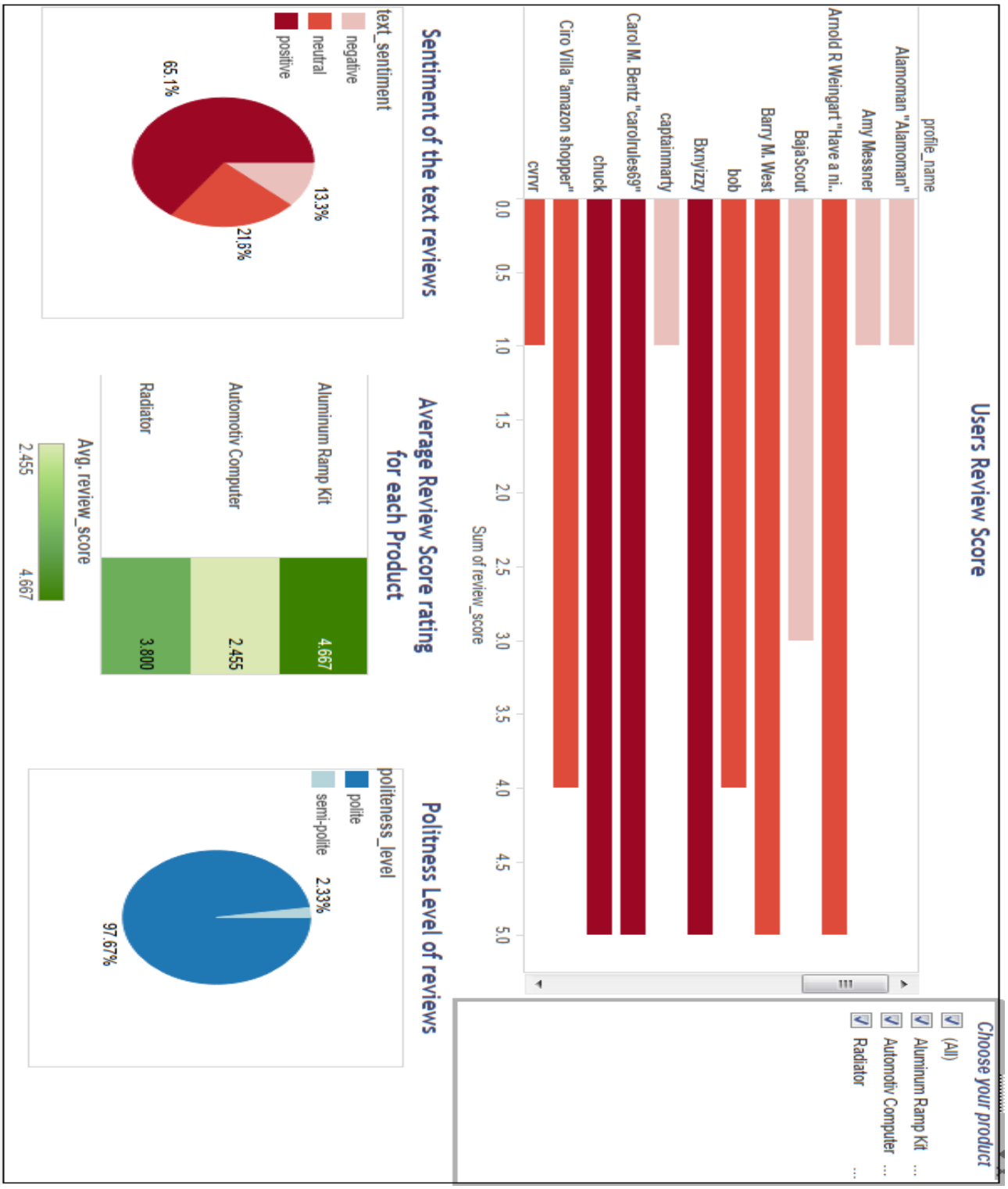

Figure C.2: *Review information dashboard*

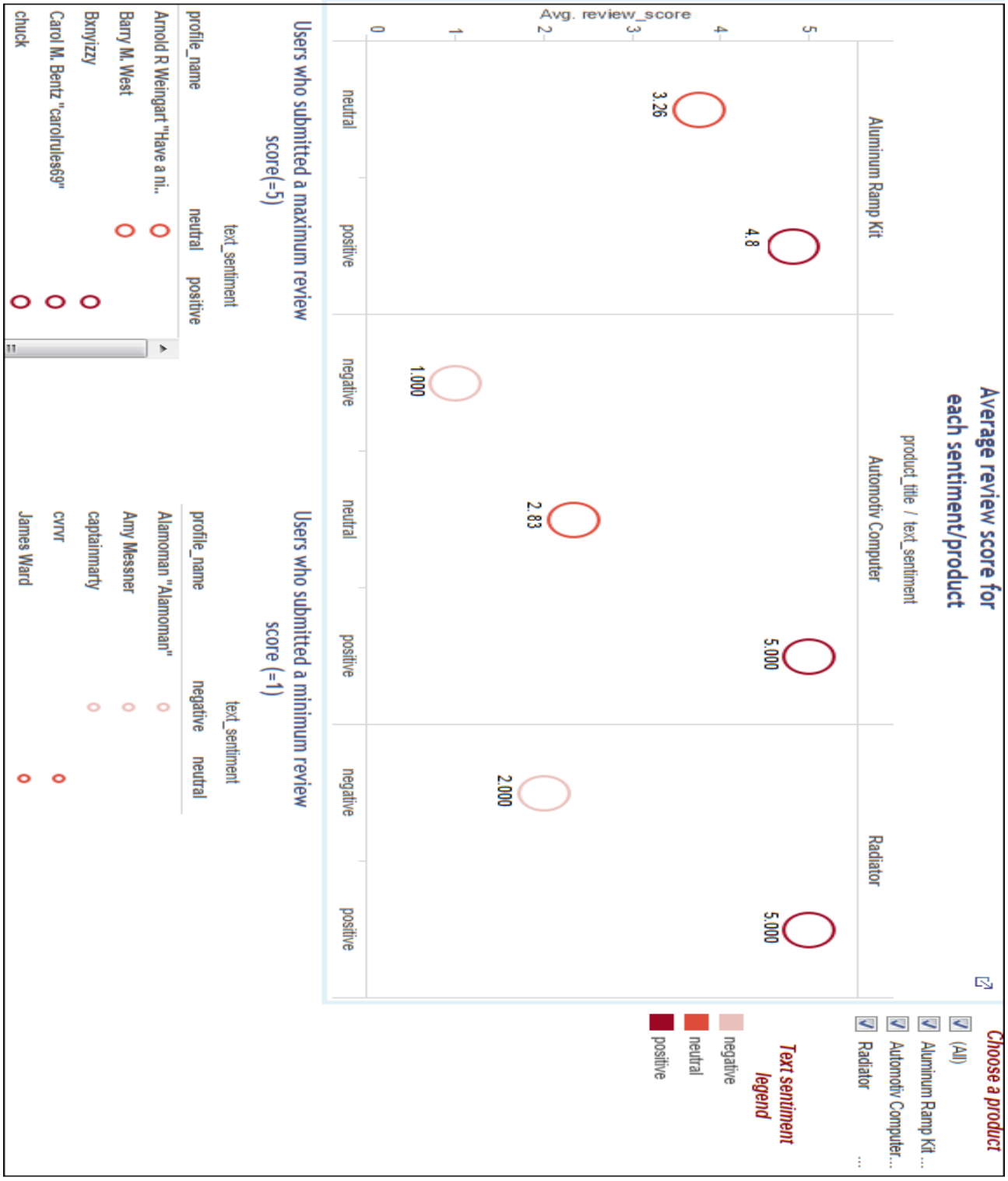

Figure C.3: *Dashboard showing the comparison between review score/review sentiment*

References

[1] Michelle Annett and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs". In: *Advances in artificial intelligence*. Springer, 2008, pp. 25–35.

[2] Amin Ranj Bar and Muthucumaru Maheswaran. "Crowdsourcing Information Systems". In: *Confidentiality and Integrity in Crowdsourcing Systems*. Springer, 2014, pp. 5–9.

[3] Vic Barnett and Toby Lewis. *Outliers in statistical data*. Vol. 3. Wiley New York, 1994.

[4] Clara Bayarri Romaña. "Linguistic problem ranking for people with Dyslexia using crowdsourcing". In: ().

[5] Gerlof Bouma. "Normalized (pointwise) mutual information in collocation extraction". In: *Proceedings of the Biennial GSCL Conference*. 2009, pp. 31–40.

[6] Pimwadee Chaovalit and Lina Zhou. "Movie review mining: A comparison between supervised and unsupervised classification approaches". In: *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE. 2005, pp. 112c–112c.

[7] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. "Crowdsourcing systems on the world-wide web". In: *Communications of the ACM* 54.4 (2011), pp. 86–96.

[8] P Dugan, W Cukierski, Y Shiu, A Rahaman, and C Clark. "Kaggle Competition". In: *Cornell Univerity, The ICML* (2013).

[9] David Geiger, Stefan Seedorf, Thimo Schulze, Robert C Nickerson, and Martin Schader. "Managing the crowd: towards a taxonomy of crowdsourcing processes". In: (2011).

[10] Vasileios Hatzivassiloglou and Kathleen R McKeown. "Predicting the semantic orientation of adjectives". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 1997, pp. 174–181.

[11] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. "Outlier detection using replicator neural networks". In: *Data Warehousing and Knowledge Discovery*. Springer, 2002, pp. 170–180.

[12] Victoria J Hodge and Jim Austin. "A survey of outlier detection methodologies". In: *Artificial Intelligence Review* 22.2 (2004), pp. 85–126.

[13] Jeff Howe. "The rise of crowdsourcing". In: *Wired magazine* 14.6 (2006), pp. 1–4.

[14] Bernardo A Huberman, Daniel M Romero, and Fang Wu. "Social networks that matter: Twitter under the microscope". In: *arXiv preprint arXiv:0812.1045* (2008).

[15] Mia Hubert and Ellen Vandervieren. "An adjusted boxplot for skewed distributions". In: *Computational statistics & data analysis* 52.12 (2008), pp. 5186–5201.

[16] Karim R Lakhani and Jill A Panetta. “The principles of distributed innovation”. In: *innovations* 2.3 (2007), pp. 97–112.

[17] Chenghua Lin, Yulan He, and Everson. “Weakly supervised joint sentiment-topic detection”. In: *Knowledge and Data Engineering, IEEE Transactions on* 24.6 (2012), pp. 1134–1145.

[18] Bing Liu. “Sentiment analysis and subjectivity”. In: *Handbook of natural language processing* 2 (2010), pp. 627–666.

[19] Jon D Mcauliffe and David M Blei. “Supervised topic models”. In: *Advances in neural information processing systems*. 2008, pp. 121–128.

[20] Vincent Ng and Claire Cardie. “Weakly supervised natural language learning without redundant views”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, pp. 94–101.

[21] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. “Cross-domain sentiment classification via spectral feature alignment”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 751–760.

[22] Bo Pang and Lillian Lee. “Opinion mining and sentiment analysis”. In: *Foundations and trends in information retrieval* 2.1-2 (2008), pp. 1–135.

[23] Mark Stuart Powlen and Michael Brandon Ferrel. *Social media platforms*. US Patent App. 13/588,368. Aug. 2012.

[24] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. “Influence and passivity in social media”. In: *Machine learning and knowledge discovery in databases*. Springer, 2011, pp. 18–33.

[25] Alexander Sorokin and David Forsyth. “Utility data annotation with amazon mechanical turk”. In: *Urbana* 51.61 (2008), p. 820.

[26] James P Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[27] Peter D Turney. “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 417–424.