ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

A Text-Mining based model to detect unethical biases in online reviews:

A Case-Study of Amazon.com

Ana Rebello de Andrade da Costa

Dissertation submitted as a partial requirement for the conferral of Master degree in Computer Science and Business Management

Supervisor:
Dr. João Guerreiro, Assistant Professor, Department of Marketing, Operation and Management, ISCTE-IUL

Co-supervisor:
Dr. Sérgio Moro, Assistant Professor, Department of Information Science and Technology, ISCTE-IUL

September, 2017

To my grandparents, for being my living examples of kindness, strength, hard work and accomplishment.

Acknowledgements

To my supervisors – Prof. João Guerreiro and Prof. Sérgio Moro, for their valuable guidance. You definitely provided me with the tools I needed to choose the right direction and helped me overcome a rather rough start. I also have to thank you for your patience and understanding for when it seemed too difficult to keep up the pace. Thank you for helping me successfully complete my dissertation.

To my parents, for being the best parents I could have asked for. Thank you for everything you provided me, especially my education. I can honestly say I would not be where I am today if it wasn’t for both of you. You pushed me through my limits and made me believe that if I put my mind to it, I could do anything.

To my sisters Teresa and Tan and my brother Kiko, for being my best friends and my unconditional support since ever. Each one of you has helped me get through this phase, whether it was by teaching me how to put things into perspective, cheering me up or just making me laugh despite all the stress.

To Avó Tê, for not resting until she had done everything in her power to help me overcome the rocky start of this dissertation, even when she had no clue what the project was about. I admire you so much.

To Quel, for being one of the most important role models I have ever had. I am so grateful that you walked into our lives more than 20 years ago, and I am proud to call you family. I have to thank you for raising me and my siblings and for setting such a great example of hard work and strength in overcoming all obstacles in our lives. You are one of the bravest people I have ever met.

To Raquel, for being one of the key factors in this achievement. You were the one who helped me the most in the writing process of this dissertation. You taught me what I needed to carry this project to a successful conclusion and helped me when I thought I could never finish it. And most of all, thank you for putting up with my snippiness.

Abstract

The rapid growth of social media in the last decades led e-commerce into a new era of value co-creation between the seller and the consumer. Since there is no contact with the product, people have to rely on the description of the seller, knowing that sometimes it may be biased and not entirely truth. Therefore, reviewing systems emerged in order to provide more trustworthy sources of information, since customer opinions may be less biased. The problem was, once sellers realized the importance of reviews and their direct impact on sales, the need to control this key factor arose. One of the methods developed was to offer customers a certain product in exchange for an honest review. However, in the light of the results of some studies, these “honest” reviews were proved to be biased and skew the overall rating of the product.

The purpose of this work is to find patterns in these incentivized reviews and create a model that may predict whether a new review is biased or not. To study this subject, besides the sentiment analysis performed on the data, some other characteristics were taken into account, such as the overall rating, helpfulness rate, review length and the timestamp when the review was written.

Results show that some of the most significant characteristics when predicting an incentivized review are the length of a review, its helpfulness rate and the overall polarity score, calculated through VADER algorithm, as the most important sentiment-related factor.

Keywords: online reviews, text mining, sentiment analysis, VADER.

Resumo

O rápido crescimento das redes sociais nas últimas décadas levaram o comércio electrónico a uma nova era de co-criação de valor entre o vendedor e o consumidor. Uma vez que não há contacto com o produto, os clientes têm de se basear na descrição do vendedor, mesmo sabendo que por vezes tal descrição pode ser tendenciosa e não totalmente verdadeira. Deste modo, surgiu um sistema de *reviews* com o propósito de disponibilizar um meio de informação de maior confiança, uma vez que se trata de partilha de informação entre clientes e por isso mais imparcial. No entanto, quando os vendedores se aperceberam da importância das *reviews* e o seu impacto direto nas vendas, surgiu a necessidade de controlar este fator chave. Uma das formas de o fazer foi através da oferta de determinados produtos em troca de *reviews* honestas. Contudo, à luz dos resultados de alguns estudos, foi demonstrado que estas *reviews* “honestas” são tendenciosas e enviesam a classificação geral do produto.

O objetivo deste estudo foi o de encontrar padrões na forma como estas *reviews* incentivadas são escritas e criar um modelo para prever se uma determinada *review* seria enviesada. Para esta análise, além da análise de sentimentos realizada sobre os dados, outras características foram tidas em conta, tal como a classificação geral, a taxa de *helpfulness*, o tamanho da *review* e a hora a que foi escrita.

Os modelos gerados mostraram que as características mais importantes na previsão de parcialidade numa *review* são o tamanho e a taxa de utilidade e como característica sentimental mais relevante a pontuação geral da *review*, calculada através do algoritmo VADER.

Palavras-chave: *online reviews, text mining*, análise de sentimentos, VADER.

Table of contents

List of tables

List of figures

Acronyms

AJAX: Asynchronous JavaScript and XML

ANEW: Affective norms for english words

AUC: Area under the curve

CRISP-DM: Cross-Industry Process for Data Mining

eWOM: Electronic Word of Mouth

GI: General inquirer

JSON: JavaScript object notation

LIWC: Linguistic inquiry word count

NLP: Natural language processing

ROC curve: Receiver operating characteristic curve

VADER: Valence Aware Dictionary for Sentiment Reasoning

WSD: Word sense disambiguation

1. Introduction

Online shopping has become a widespread form of business over the Internet. It allows consumers to buy goods or services anytime, anywhere, using devices such as desktop computers, laptops, tablets, and smartphones. Online stores allow customers to search for specific items, brands or models, paying them using online methods of payment such as credit cards or PayPal, and receiving them at their desired location. Given that there is no physical interaction with the product or seller, customers started to rely not only on the description of the product but also on the comments provided by other customers that also bought the same product (Mudambi and Schuff, 2010). This led to the creation of a review system for e-commerce sites, in which people can rate the products based on their own experience. Product reviews have become a key factor in the decision making process of purchasing an item online (Blazevic et al., 2013). Although at different scales, both positive and negative reviews have a direct impact on the decision of the customer to buy a product (Hu et al., 2008; Lee et al., 2008). Additionally, online reviews became very useful to the sellers because they provide information on their customers' opinion regarding their products. Thus, it is crucial that this form of information exchange stays reliable.

The problem addressed in this dissertation is related to the fact that it has been proved that many reviews on Amazon are biased (Hu et al., 2011). In addition to mislead customers, this also compromises the rating and reviewing system, which may lead to a general sense of mistrust and ultimately the end of the system itself.

For the purpose of this study, two types of biased reviews were considered, depending on who wrote them and why. There are “paid reviews”, which are written by a person or company who charges for that service, and “incentivized reviews”, written by real customers who acquired the product for free or at a discount. Contrary to paid reviews that were always prohibited, the latter were formerly accepted by Amazon if they included a disclaimer about their affiliation with the seller. Due to recent studies, in particular the one conducted by ReviewMeta (ReviewMeta, 2016), the overall rating of a product was influenced by incentivized reviews. Therefore, Amazon started banning such reviews (amazon.com, 2016a). However, it is naïve to assume that this practice will cease to exist. Therefore, these reviews will continue to influence the rating system, only it is now harder to identify bias and ignore it in the process of decision making. Different methods for recalculating an overall rating of a product exist. These mechanisms require the link of the product as input and recalculate the overall rating by excluding confirmed biased reviews and averaging the remaining ratings. However, there is a noticeable absence of a mechanism with the purpose of identifying bias in a new-coming review that has no disclaimer. This work aims to bridge such gap, with the main goal of helping users of e-commerce platforms to properly assess the information they base their decisions on. It also provides a more efficient approach for companies to identify dishonest sellers and manage them, making the platform more reliable in the eyes of the customers.

The main contribution of this dissertation is to create a model of association rules for Amazon's reviews using text mining techniques to find patterns in biased reviews, in order to predict the probability of a new review being biased.

The dissertation is organized as follows: "State of the art" section offers a brief literature review on online reviewing systems, biased reviews and their relevance in Amazon, and the text mining techniques that are of interest for this dissertation. In the "Methodologies" section, the datasets and methods used are presented and the results obtained are discussed. Lastly, the "Conclusion" section presents the final thoughts and findings on this work.

2. State of the art

2.1. Social commerce

The concept of electronic commerce (e-commerce) has multiple definitions, all of them based on the idea of using electronic communications and digital information processing technology in business transactions between companies and consumers (Lallana et al., 2000). This type of business model, or segment of a larger business model, operates in all four of the major market segments: business to business, business to consumer, consumer to consumer and consumer to business.

E-commerce as we know has experienced a tremendous revolution with the rise of the Web 2.0. This concept first came up in 2004 through the American company O'Reilly Media. Tim O'Reilly defined it as the business revolution in the computer industry, the network as platform delivering software as a continually-updated service that gets better the more people use it (O'Reilly and Battelle, 2009). The main focus of this new Web generation is the ability for people to collaborate and share information online (Aghaei et al., 2012). The infrastructure that has allowed the current social experience of social media and consumer creation has risen and created a situation where consumers expect a richer context (Berthon et al., 2012). Instead of simply expecting content to be fed to them, consumers wish to interact, to collaborate and to interject. Web 2.0 means participation and action – there is nothing passive about it. It is here that consumers are encouraged to be involved and personal (Fournier, 2011). Users gather in communities where they can communicate and share information and began to perform a more active role in providing and sharing the information available on the Web (Morente-Molinera et al., 2015).

The challenge in defining Web 2.0 is that it is more than just a set of technologies. It also incorporates attributes with a social dimension including new business models, user-contributed content and user-generated metadata, and increased simplicity in design and features (Wigand et al., 2008). The three anchor points are technology and architecture, community and social, and business and process. The first one consists of the infrastructure of the Web and concept of Web platforms, which includes technologies like Rich Internet Applications, AJAX and Flash, representational state transfer and Really

Simple Syndication. The second point looks at the dynamics around social networks, communities and other personal content share models, wikis and other collaborative content models. The last one refers to business models enabled by web services and include advertising and long-tail economics (Wigand et al. 2008). As technologies such as AJAX evolve and are adopted in large scale, Web 2.0 techniques are quickly becoming the expected user experience for the web. Mainstream examples of AJAX include the Google-based applications (Google Maps, Docs and Calendar) as well as Microsoft-based applications (Hotmail and Windows Live-based applications). As users start utilizing these types of applications in their everyday lives, they will come to expect the same type of functionality in every other application.

One of the most interesting elements of Web 2.0, and certainly the most relevant one for this work, is electronic word of mouth (eWOM). The advances of information technology have profoundly changed the way information is transmitted and have transcended the traditional limitations of word-of-mouth. Consumers can now easily and freely access information and exchange opinions on companies, products, and services on an unprecedented scale in real time (Duan et al., 2008). EWOM is defined by King (2014) as any positive or negative statement about a product or company, made by potential, actual or former customers, and made available to a large number of people and institutions through the Internet. EWOM allows consumers to make informed purchase decisions over digital platforms, by engaging socially with other consumers and discussing ideas (King, 2014). This works just as ordinary word of mouth would, but with the added element of the extended volume that comes with being on the Internet – anyone, anywhere, could read your opinion and make a decision based on it. What is interesting about eWOM is that consumers do not only seek it out when they are considering a purchase, but also when they have no specific desire to purchase at all (King, 2014). This means businesses must understand that interaction across social media should not only be to facilitate a current purchase, but also to aid brand loyalty and brand awareness. These customers might not be interested in purchasing now, but possibly they will later.

The dramatic growth of social media and Web 2.0 has provided a huge potential to transform e-commerce from a product-oriented environment to a customer-centered one (Wigand et al., 2008). Within this environment, customers have access to social knowledge and experiences to support them in better understanding their online purchase purposes, and in making more informed and accurate purchase decisions (Dennison et al., 2009). This has meant the most to retailers because they started being able to capture customers' behavior, which gave them insights into their shopping experiences and expectations, and how to develop successful business strategies (Constantinides and Fountain 2008). Businesses started to realize the necessity of engaging with consumers via social media, and the potential competitive advantage of a well-managed social media presence (King, 2014).

E-commerce went through a new evolution by adopting a variety of Web 2.0 features, functions and capabilities in order to enhance customer participation (Kim and Srivastava, 2007), promote customer relationships (Liang et al. 2011), and achieve greater economic value (Parise and Guinan 2008). This e-commerce evolution was the origin of social commerce and the main differences between these two are mostly in terms of business goals, customer connection and system interaction (Huang and Benyoucef, 2013).

2.2. User feedback in the decision making process

For years, retailers have ignored data reporting the significant value of user reviews, not only because the review content was difficult to manage and analyze (Mostafa, 2013) but also because they feared the impact of negative customer feedback. However, with the appearance of social commerce, this began to change since its main goal of meeting consumers' growing expectations passes by allowing their actions to have an impact on social websites.

A preliminary analysis suggests that social media impacts decision making by creating more connections to enable the access to information and opinions. In general, people tend to trust the opinions of participants in online networks in which they have also chosen to participate. Social media are rich information sources and these tools facilitate crowd behavior and increase peer pressure (Power et al., 2011).

Customer reviews and comments on products or services appeal to the very human need to know "what everybody else is doing". Since high levels of trust exist in information obtained from online networks, reviews can move shoppers from consideration to purchase (Bulmer and DiMauro, 2009). In 2010, the Nielsen Company reported that 60 percent of customers read online reviews when making a purchase and take them into consideration. When walking down a busy city street looking for a place to eat, most people will gravitate to a location with customers already inside, rather than an empty restaurant. Similarly, websites with an absence of customer community activity may soon feel “empty" compared to those that feel "alive" with activity and communication with and among their users. Social media changes decision making by challenging the notion of who is an authoritative and reliable source (Garland, 2009).

2.3. Online reviews

One of the most dominant channels to produce electronic word-of-mouth is online customer review systems (Dellarocas, 2003). As stated by Gefen in 2002, many scholars have argued that trust is a prerequisite for successful commerce because consumers are hesitant to purchase unless they feel the seller is trustworthy. Consumer trust may be even more important in electronic transactions than it is in traditional, “real world” transactions. This is because Internet transactions are blind, borderless, can occur 24 hours a day 7 days a week, and are non-instantaneous (payment may occur days or weeks before delivery is completed) and that makes consumers concerned that the seller will not adhere to its transactional obligations (Kim et al., 2008). The results of a study conducted by Utz et al. (2012) that aimed to examine the impact of online reviews on consumer trust in an online store, showed that reviews turned out as the strongest predictor of trustworthiness judgments. Store reputation had no significant effect compared to the reviews. EWOM plays an important role in consumer decision making, indicating that online consumer communities indeed empower consumers.

A study from Kim et al. (2012) also showed the importance of eWOM, but more specifically the value of a common general opinion, opposed to a single discordant opinion: negative emotional expressions in a single negative review tend to decrease the reviews' informative value and make consumers' product evaluations less negative because consumers attribute the negative emotions to the reviewer's irrational dispositions. Meanwhile, positive emotional expressions in a single positive review do not influence consumers' product evaluations significantly, even though consumers attribute the positive emotions to the product. However, when multiple convergent emotional expressions are present in multiple user reviews, both positive and negative emotional expressions increase informative value of the reviews and polarize consumers' product evaluations in the respective direction.

The customer review percentage – customers that do write a review, can vary dramatically and depends on numerous factors, like for example the number of reviews that the product currently has. According to Neil Campbell (Campbell, 2012), former Category Leader on Amazon, data shows that if a product has no reviews then the feedback rate is higher, but this decreases rapidly as the number of reviews increase. People do not see the value of adding their opinion if it is according to the majority of the existing reviews. However, this does not apply if their opinion is against the prevailing sentiment of the review, no matter how many reviews the product has, especially if theirs was a negative experience compared to a positive sentiment. Considering the results from the Nielsen Company report (Nielsen, 2010), another factor that influences the decision of reviewing a product is how controversial or high profile the item is: the higher the profile then the more feedback it will get, thus outstripping its actual sales. In March 2010, the Nielsen Company conducted a survey and polled more than 27,000 Internet users in 55 markets from the Asia-Pacific, Europe, Middle East, North America and South America to look at how consumers shop online: what they intend to buy, how they use various sites, the impact of social media and other factors that come into play when they are about to purchase a product or service. According to the research (Nielsen, 2010) presented in this report, the business areas where opinions are most important are consumer electronics, where 57% of online respondents consider reviews prior to buying, cars (45%) and software (37%). Generically, 40% of online shoppers indicated that they would not even buy electronics without consulting online reviews first.

With the popularity of social commerce, an increasing number of businesses have moved online (Cheng et al., 2013). With technology shortening distances between continents, countries and cities, people living in different parts of the world can now develop similar tastes, perceptions, and styles, benefitting from worldwide accessibility at the distance of a click. This requires a global online shopping model for e-tailers – online retailers to address a larger number of diversified customers. Website developers and e-tailers need to concentrate on the motivational attributes of the websites to attract customers, such as easy interface, effective search engines, nice layouts, updated information, multimedia contents, e-catalogs, efficient navigation scheme, simple payment procedures and easy checkout process (Akhlaq and Ahmed, 2014). Revenues of online retailers follow a power-law distribution: there are few giant-sized e-tailers like Amazon.com and Buy.com while there is a long tail of small-sized online businesses (Cheng et al., 2013).

2.4. Paid and incentivized reviews

Leading e-tailers have enabled consumers to submit product reviews for many years. Once the relevance of these reviewing systems was acknowledged, companies tried their best to use them in favor of their businesses. However, to have the advantage over the competition, a new practice emerged: paid and incentivized reviews. These types of review are commonly found on e-commerce platforms. Paid reviews are bought as a service from companies whose job is to create positive reviews about a product to increase its sales. When this action became a regular practice, businesses went even further and started to pay for negative reviews on competitors' products (Butterworth, 2016). On the other hand, incentivized reviews are written by consumers who acquired the product for free or at a deep discount from the seller in exchange for an "honest and unbiased review" – but still a form of paid review. According to Mayzlin et al. (2014), the presence of this undetectable (or difficult to identify) paid reviews have at least two harmful effects on consumer and producer surplus: on one hand the consumer is misled by the promotional reviews and could make suboptimal choices, on the other hand, the potential presence of identified biased feedback may lead consumers to mistrust the reviews. Either way, the dishonesty in these assessments has always adverse effects.

2.5. Amazon

The current dissertation is focused on the reviewing system of one of the major companies of our time. Amazon is an e-commerce and cloud computing company, based in Seattle, Washington, founded by Jeff Bezos in 1994. It is the largest Internet-based retailer in the world by total sales and market capitalization, according to a 2016 report from CNCB. Initially, Amazon began as an online bookstore, later diversifying to all kinds of products, from streaming to apparel, electronics to food. Amazon is primarily a retail site with a sales revenue model and makes its money by taking a small percentage of the sale price of each item that is sold through the website. Amazon also allows companies to advertise their products by paying to be listed as featured products. At the end of 2015, the American multinational announced 304 million active customer accounts. Amazon’s reviewing system is definitely one of the most important assets of their business. According to the community guidelines (amazon.com, 2016b), Amazon allows users to submit reviews on any product, whether it was purchased or not. In addition to the written opinion, reviewers must rate the product on a rating scale from one to five stars. Every user has a profile and all the reviews made are associated to it. If the reviewer is one of the top reviewers by popularity, Amazon provides a badging option, which gives them credibility. Customers may comment or vote on other reviews, indicating whether they found the review helpful or not. If a review is given enough "helpful" votes, it appears on the front page of the product, so customers have easier access to the most helpful information. In 2010, Amazon was reported as being the largest single source of Internet consumer reviews (Freeman, 2010). When asked why Amazon would publish negative reviews, Jeff Bezos defended the practice by claiming that Amazon was "taking a different approach... [Amazon] wants to make every book available – the good, the bad, and the ugly... to let truth loose”, which shows their wish and effort to make its review and rating system unbiased and more helpful to online shoppers.

With this in mind, Amazon’s Vice President of Customer Experience Chee Chew stated, in an update on customer reviews published on Amazon website, that Amazon has always prohibited compensation for reviews – even going so far as to sue those businesses who pay for fake reviews, as well as the individuals who write them. However, the so called incentivized reviews were actually allowed. The only condition was that those reviewers would have to disclose their affiliation with the business in question in the text of the review, writing something like “I received this product for free or at a discount in exchange for my honest, unbiased opinion”. Although these reviewers claimed to write their true opinion on the product – positive or negative – these incentivized reviews have tended to be overwhelmingly biased in favor of the product being rated, according to the study conducted by ReviewMeta, mentioned earlier, on 2016. This is due to a combination of factors, like the fact that the seller has likely reached out to those reviewers who are less critical or the fact that reviewers may think they would no longer have the opportunity to receive these sorts of offers if they said negative things (ReviewMeta, 2016). Additionally, the amount of eWOM a user has produced is also a factor for being chosen, since it may influence the way the review is written (Kim et al., 2015). This is because users asked to write reviews tend to write them using explaining language, which may carry more sentiment and provide a better interpretation of the product (Moore, 2012). These reviews, with the elevated sentiment, are more effective to consumers (Kim, 2015).

Then, in the incentivized category, there is also Amazon’s Vine Program (amazon.com, 2016c), which consists on reviews of free samples by top Amazon reviewers. However, this works differently than normal incentivized reviews. According to their own definition, explained in Amazon website, the Vine Program invites the most trusted reviewers on Amazon to post opinions about new and pre-release items to help their fellow customers make informed purchase decisions. Amazon itself invites customers to become Vine Voices based on their reviewer rank, which is a reflection of the quality and helpfulness of their reviews as judged by other Amazon customers. Amazon, not the sellers, provides Vine members with free products that have been submitted to the program by participating vendors. Vine reviews are the independent opinions of these trusted reviewers. The vendor cannot influence, modify or edit the reviews, neither does Amazon. A Vine review is properly identified with the green stripe “Customer review from the Amazon Vine Program”, so the consumer always knows the context of the review they are reading. The program was created to provide customers with more information including honest and unbiased feedback from some of Amazon's most trusted reviewers. These reviewers keep this status as long as they comply with Amazon’s posting guidelines.

Nevertheless, since almost every disclaimed review was 5-star, it became a subject of interest for some researchers to study the possibility of bias in these type of reviews. As suspected, the results from the study conducted by ReviewMeta (ReviewMeta, 2016), of over 7 million Amazon reviews indicated that the average rating for products with incentivized reviews was higher than non-incentivized ones: a 4.74 versus a 4.36 average rating, out of 5 stars. Even if 0.38 star does not seem a lot, the impact was considerable – based on the rating distribution also calculated by this group, with this difference products would rise from the 54th to the 94th percentile. Effectively, incentivized reviews could create top-rated products. The study also found that incentivized reviewers were 12 times less likely to give a 1-star rating than non-incentivized reviewers and almost 4 times less likely to leave a critical review in general. Data shows that those who participate in incentivized reviews have written an average of 232 reviews, while those who have not, only wrote an average of 31 reviews.

Once these statistics were made public, Amazon was forced to reconsider their policy. It was announced that, going forward, the only acceptable incentivized reviews will be those from the Vine Program, according to the update on customer reviews published on Amazon website. Furthermore, this update also stated that in an attempt to have a fair rating system, while most online sites will simply calculate the average (mean average of all the ratings), Amazon changed the way it manages reviews: a product's overall score is based on an algorithm that considers several factors including the age of a review, the number of helpful votes received and whether the reviews are from verified purchasers. This means that some reviews count more towards the overall score than others.

Despite all efforts to have the most transparent reviewing system, these type of reviews will always exist. Therefore, there is a need to study the text content and find patterns in these type of reviews that can differentiate true from fake, without needing a badge or disclosure. As stated by Hu et al. (2012), the rapid growth of online social media in the form of collaboratively created content presents new opportunities and challenges of information to both producers and consumers. With the large amount of data produced by various social media services, text mining provides an effective way to meet users' information needs.

2.6. Text Mining

Text mining is the field of computer science research that tries to solve the crisis of unstructured information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management (Feldman et al., 2007). In the words of Aggarwal and Zhai (2012), research in information retrieval has traditionally focused more on facilitating information access rather than analyzing information to discover patterns, which is the primary goal of text mining. Text mining can be regarded as going beyond information access to further help users analyze and digest information and facilitate decision making. These techniques and processes discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing. As claimed by

Feldman (2007), text mining implicates the preprocessing of document collections (text categorization, information extraction and term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis and association rules) and visualization of the results.

Since the most natural form of storing information is text, text mining is believed to have a commercial potential higher than that of mining information from structured data. Although it is said to be a very complex task, because it involves dealing with inherently unstructured and fuzzy data (Tan, 1999), it is a worthy task since several authors estimate that about 80% of business-relevant information originates in unstructured form, primarily text. Such estimate was firstly reported on 1999, on a study using unstructured data conducted by the Data Warehousing Institute (Linstedt, 2006), which showed that while uncertain, unstructured and semi-structured data represent definitely the majority of a company's data. That is why this 80 percent rule is still used by many contemporaneous authors (Grimes, 2013).

These techniques have already been used in several businesses, like on an e-commerce platform. In a study from Cao et al. (2011), content analysis was used to quantify the feedback text comments. Findings suggest that rich content plays an important role in building a buyer's trust in a seller. Text mining can help an organization derive potentially valuable business insights from text-based content such as posts, comments or reviews on social media (Calheiros et al., 2017; Guerreiro and Moro, 2017). In the beginning, text mining applications used basic models like bag-of-words to structure the data, categorizing them based on a few determined classes or grouping them in a natural way (Sharda et al., 2014). This model consists essentially in breaking the text up into words and considering each of them as a feature, while the order and co-occurrence of words are completely ignored (Nassirtoussi et al., 2014). Since humans have a logical speech and do not use words in a random order or structure, to fully extract the value of this type of data, new models were developed.

Input data has to go through a pre-processing stage, where the unstructured text is transformed into a representative format that is structured and can be processed by the machine (Nassirtoussi et al., 2014). In data mining in general, and specifically in text mining, the pre-processing phase has a significant impact on the overall outcomes (Uysal and Gunal, 2014). Such process is performed by natural language processing.

Nevertheless, managing unstructured data is a challenging task, because natural language text is often inconsistent. It contains ambiguities caused by syntax and semantics (Katariya et al., 2015).

2.7. Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken (Ehsani and Knodt, 1998). NLP is a component of artificial intelligence and its applications development is challenging because computers usually require humans to “speak” in a precise, unambiguous and highly structured programming language or sometimes through a limited number of clearly-enunciated voice commands. Human speech, however, is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects, double entendres, sarcasm and social context. In NLP, two aspects of language are attentively researched: semantics and syntax. Semantics deals with the meaning of words and syntax deals with their order and relative positioning or grouping (Nassirtoussi et al., 2014).

As stated by John Rehling (2011), an NLP expert at Meltwater Group, in “How Natural Language Processing Helps Uncover Social Media Sentiment”, apart from common word processor operations that treat text like a mere sequence of symbols, NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. By analyzing language for its meaning, NLP systems have successfully performed tasks such as correcting grammar, converting speech to text and automatically translating between languages (Church and Rau, 1995). By using NLP, it is possible to organize and structure knowledge to perform a multiple number of tasks such as automatic summarization, translation, named entity recognition, relationship extraction, speech recognition, part-of-speech tagging and parsing, topic segmentation, and sentiment analysis.

What are the clues we use to understand who did what to whom, or when something happened, or what is fact and what is supposition or prediction? To understand human language means to understand not only the words, but the concepts and how they’re linked together to create meaning. While nouns, verbs, adjectives and adverbs are the building blocks of meaning, it is their relationship to each other within the structure of a sentence, within a document and within the context of what we already know about the world that conveys the true meaning of a text. According to Feldman (1999), people extract meaning from text or spoken language on several levels, namely the phonetic, morphological, syntactic, semantic, discourse and pragmatic. Because each of these levels of language understanding follows definable patterns or templates, it is possible to inject some language understanding into a computer system by using those definitions. However, it becomes more difficult, the higher the level. Despite language being one of the easiest things for humans to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

A vivid area of research in NLP is called sentiment analysis (Cambria et al., 2015). The type of applications belonging to this domain addresses the interests that big commercial companies have to interpret opinions from their clients involving their products, in order to improve them or to estimate future trends. Big data methods put to work on text files are being employed successfully here. Still, a sentiment means much more than the mere and openly expression of a taste or inclination (Cristea, 2016). Historically, NLP has been used as a text mining technique to perform sentiment analysis on social media data, using text classification to distinguish between positive, negative and neutral reviews, in order to reduce cost, effort and time to manage large scale public opinion (Bollen et al., 2009).

2.8. Sentiment analysis

Sentiment analysis, also called opinion mining by Liu (2012), is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes. There is not a consensus and commonly accepted definition for this concept, but one that relates to the object of this work is the one from Mostafa (2013), saying it is an automatic knowledge discovery technique that aims to find hidden patterns in textual comments. Mostafa (2013) further adds that in order to calculate the sentiment it is necessary to compare it to a dictionary to measure the strength of the sentiment. Sentiment classification usually deals with two opposite classes – positive and negative, and the gap in between (Sharda et al., 2014) or the variance of the strength of opinion (Pang and Lee, 2008). There has been an increasing interest in this area, both for document classification and for word and phrase polarity studies (Sharda et al., 2014).

Opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality and the choices we make are conditioned upon how others see and evaluate the world (Calheiros et al., 2017). For this reason, when we need to make a decision we often seek out the opinions of others (Holleschovsky and Constantinides, 2016). This applies not only to individuals but to organizations. Sentiment analysis studies opinions and its related concepts such as sentiments, evaluations and emotions. Since early 2000, sentiment analysis has grown to be one of the most active research areas in NLP. It is also widely studied in data mining, web mining and text mining. In fact, it has spread from computer science field to management sciences and social sciences due to its relevance to business and society as a whole, as Liu (2012) points out.

Sentiment analysis systems have found their applications in almost every business and social domain. Regarding social commerce, sentiment analysis has its greatest relevance in online reviews. Most research has been focusing in product reviews with the purpose of knowing if the customer recommends such product, which is not hard to perceive since the majority of reviewing systems is associated with a rating that is generally a numerical evaluation (Paltoglou and Thelwall, 2012). So, the emphasis of this type of analysis is to understand the sentiment towards the product and its strength. Due to underlying concepts, expressions and context, it is a complex problem and for that reason there is no standard process to conduct sentiment analysis (Sharda et al., 2014).

With this work, we aim to study the patterns in sentiment polarity and strength in reviews, positive or negative, with and without disclaimer in order to predict the probability of a new review being biased. Besides plain features such as length and rating, the sentiment in a review and its strength are thought to be of great relevance in determining associating rules to predict bias. Another mechanism that may help in forecasting a biased review is one that finds relationships between reviews that use similar terms or address the same attributes.

Recently, a new algorithm called VADER (Hutto and Gilbert, 2014) was developed to perform sentiment analysis and, according to its authors, it outperforms most of the competitor tools. The reason for its great performance is related to the fact that it takes into consideration several factors usually ignored by others, like capitalization, excess of punctuation, etc. and therefore it classifies reviews with more accuracy in a sentiment intensity scale.

VADER Algorithm

VADER stands for Valence Aware Dictionary for sEntiment Reasoning. VADER is a rule-based model for general sentiment analysis (Hutto and Gilbert, 2014). To build this model, a combination of qualitative and quantitative methods was used to produce, and then empirically validate, a *gold-standard* list of lexical features, along with their associated sentiment intensity measures. This list of features is specifically attuned to sentiment in social media-like contexts (Hutto and Gilbert, 2014). Next, this list was combined with consideration for five generalizable rules that include grammatical and syntactical conventions that humans use when expressing sentiment intensity. According to the empirical validation of this study, in social media domain VADER performed as well as individual human raters at matching ground truths, and it even outperformed them in accuracy at correctly classifying the sentiment of tweets into positive, negative or neutral. This algorithm was already successfully used in other studies such the one by Araújo et al. (2016) and Ribeiro et al. (2016).

A substantial number of sentiment analysis approaches rely greatly on an underlying sentiment lexicon. A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative (Liu, 2010). Since manually creating and validating such lists is so inefficient, most studies in this field rely on preexisting lexicons. However, most of the benchmark lexicons are rather limited regarding sentiment intensity (Hutto and Gilbert, 2014). These lexicons usually categorize words as being either positive or negative, but not the strength of the feeling. So, two reviews saying "This book is *good*" and "This book is *amazing*" would be scored equally since there is a positive term in each, and yet it is obvious that the second one conveys much more enthusiasm and positiveness than the first one. Analysts and researchers need to be able to recognize variations in sentiment intensity over time in order to detect when rhetoric is heating up or cooling down (Wilson et al., 2004). Hence, it is common opinion that a general lexicon with strength valences would be beneficial (Hutto and Gilbert, 2014).

As a result of all these limitations, there was a gap to develop a new algorithm that would be based on the benchmark lexicons but would also overcome the existing flaws. Thus, Hutto and Gilbert began by constructing a list inspired by well-established sentiment banks like Linguistic Inquiry Word Count (LIWC), General Inquirer (GI) and Affective Norms for English Words (ANEW). Next, several lexical features common to sentiment expression in social media, such as emoticons, slang, acronyms and initialisms, were added to this list.

Qualitative analysis techniques were applied in order to identify properties and characteristics of the text that affect the perceived sentiment intensity and that normally would not be captured by a typical bag-of-words model. This deep analysis resulted in isolating five generalizable heuristics:

- punctuation, namely the exclamation point (!), increases the intensity of the feeling without modifying the semantic orientation;
- capitalization, mostly if a sentiment-relevant word is all-caps and the rest of the text is not. This also increases the magnitude of the sentiment without modifying the semantic orientation;
- degree modifiers, also called intensifiers, booster words or degree adverbs, which can increase or decrease the intensity of a sentiment;
- contrast words, like “but” or “however”, that usually mean a shift in sentiment polarity, giving emphasis on the text after the contrast word;
- examining the tri-gram preceding a sentiment-laden lexical feature, in 90% of the cases negation flipped the polarity of the text.

After evaluating and comparing VADER to the benchmark lexicons, the conclusion was that VADER outperformed all other lexicons and in some cases even humans (Hutto and Gilbert, 2014). Moreover, the simplicity of VADER carries several advantages such as the speed and computational economy without sacrificing accuracy, its transparent lexicon and rules that are accessible for everyone to understand, extend or modify. It also does not require an extensive set of training data to perform well in diverse domains.

The positive (*pos*), negative (*neg*) and neutral (*neu*) scores are ratios for proportions of text that fall in each category, and so they add up to be 1. These are more useful metrics to have a multidimensional measure of sentiment for a given sentence.

The *compound* score is a unidimensional measure of sentiment for a given sentence – it is a normalized weighted composite score. It is computed by summing the valence scores of each word in the lexicon and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive) (Hutto and Gilbert, 2014).

After this initial analysis where the text is structured, the next step is to use this data to feed a model that will be able to predict a certain variable through data mining processes.

2.9. Modeling

Mining knowledge from data requires training models to apprehend the hidden patterns that may be translated into such useful knowledge (Moro et al., 2017). Models can be based on supervised or unsupervised learning. Unsupervised machine learning is applied to uncategorized data, with no output variable. The goal is to model the underlying structure or distribution to learn more about the data and it is mostly based on pattern discovery techniques which are then used to group the data according to its characteristics (Chaney and Blei, 2012). Unsupervised learning problems can be interpreted as clustering and association problems (Rose, 1998). In clustering, the aim is to discover the inherent groupings in the data, such as grouping customers by purchasing behavior, while association aims to discover rules that describe large portions of the data, such as people who buy X, also tend to buy Y. Some popular algorithms of unsupervised learning are k-means and apriori.

In supervised models, there are input variables (X) and an output variable (Y) and the main goal is to use an algorithm to learn the mapping function from the input to the output: Y = f(X). The ultimate purpose is to approximate the mapping function so well that it is possible to predict the output variables for new input data. Supervised learning problems can be grouped into regression and classification. The main difference between these two is the output variable type. Regression aims to predict the output value, since it deals with continuous variables, while classification aims to group the output into a class, since it treats discrete variables (Hastie, 2009).

Decision tree models

Decision tree models allow the development of classification systems that predict or classify future observations based on a set of decision rules (Pradhan, 2013). This approach, sometimes known as rule induction, has several advantages. Firstly, the reasoning process behind the model is evident when browsing the tree, in contrast to other "black box" modeling techniques in which the internal logic can be difficult to work out. Secondly, the process will automatically include only the attributes that really matter in making a decision. Attributes that do not contribute to the accuracy of the tree are ignored (Pradhan, 2013). However, it is very frequent to have overfitting problems in this type of models. Overfitting refers to a model that models the training data too well. It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data (Larose, 2005).

In this type of model, all the fields of the dataset are examined to find the one that gives the best classification or prediction by splitting the data into subgroups. The process is applied recursively, splitting subgroups into smaller and smaller units until the tree is finished (as defined by certain stopping criteria). The target and input fields used in tree building can be continuous (numeric range) or categorical, depending on the algorithm used. Four models are typically used in the literature for such purpose:

C5.0 – this model works by splitting the sample based on the field that provides the maximum information gain (Quinlan, 2009). Each subsample is then split again, usually based on a different field, and the process repeats until the subsamples generated cannot be split any further. The lowest-level splits (leafs) are reassessed and pruned if their contribute to the value of the model is not significant.

C&R Tree – the Classification and Regression Tree method is similar to C5.0, but it splits the data based on the reduction in an impurity index (Gini), and repeats until the subsamples can no longer be split (Breiman et al., 1984).

CHAID – standing for Chi-squared Automatic Interaction Detection, this model identifies the optimal splits using chi-square statistics (Kass, 1980). Using a chi-square independence test, it examines the input field for significance and if there is more than one statistically significant relation, CHAID will select the most significant one (smallest *p* value).

QUEST – Quick, Unbiased, Efficient Statistical Tree is an attempt to simplify and reduce the processing time required by C&R Tree (Loh and Shih, 1997). It uses a sequence of rules, based on significant tests, to evaluate the input fields at a node. Splits are determined by running the quadratic discriminant analysis using the selected input on groups formed by the target categories. This method results in a speed improvement over exhaustive search to determine the optimal split.

Statistical models

Statistical models use mathematical equations to encode information extracted from the data. In some cases, statistical modeling techniques can provide adequate models very quickly. Even for problems in which more flexible machine-learning techniques (such as neural networks) can ultimately give better results, some statistical models can be used as baseline predictive models to judge the performance of more advanced techniques.

Logistic Regression – Also known as nominal regression, is a statistical technique for classifying records based on values of input fields. Logistic regression works by building a set of equations that relate the input field values to the odds associated with each of the output field categories (Cox, 1958). Once the model is generated, it can be used to estimate probabilities for new data. For each record, a probability of membership is computed for each possible output category. The target category with the highest probability is assigned as the predicted output value for that record.

Discriminant – This analysis builds a predictive model for group membership (Fisher, 1936). This model is composed by a discriminant function based on linear combinations of the predictor variables that provide the best discrimination between the groups. Discriminant analysis makes more stringent assumptions than logistic regression but can be a valuable alternative or supplement to a logistic regression analysis when those assumptions are met.

Neural Networks

This method, as opposed to decision trees, is not easy to understand. Neural networks have a remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. In case the interpretability is not important, this can be a good alternative to model the data (McCulloch and Pitts, 1943).

In order to evaluate the performance of classification models, there is a very common technique called confusion matrix, which is displayed in Table 1.

Table 1 - Confusion matrix

		PREDICTED	
		+	–
REAL	+	true positive **TP**	false negative **FN**
	–	false positive **FP**	true negative **TN**

This matrix represents a summary of the predicted results. The number of correct and incorrect predictions is summarized with count values and broken down by each class. The confusion matrix shows the ways in which the classification model is confused when making predictions (Hinton, 2015). The rows represent the real (known) values of the output variable and the columns are the classification assigned by the model. The true positive and true negative values are the cases that the model classified correctly, while the false positives and false negatives are the misclassified ones. These numbers give information not only about the errors that are being made, but more importantly the type of errors that are being made (Davis and Goadrich, 2006).

The most commonly used performance measure is accuracy – it shows the rate of correctly classified cases. The more cases it classifies correctly, the more accurate is the model:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

However, some authors defend that this is only a reliable measure when dealing with balanced datasets (Chawla, 2009; Yap *et.al.*, 2014; Liu, 2008). Other important measures that result from the output of the confusion matrix are precision, recall, specificity and F-score (Yap et.al., 2014). Precision measures how well the positive values were classified considering all predictions made, i.e. a high precision value means that most cases classified as positive were in fact positive:

$$\text{Precision} = TP / (TP + FP)$$

On the other side, recall (or sensitivity) indicates the model's ability to effectively predict the positive values. This is calculated by the rate of positive classified values, considering the real number of positives:

$$\text{Sensitivity} = TP / (TP + FN)$$

The goal is to increase sensitivity without neglecting precision. Since these two factors affect each other negatively, it is necessary to understand the problem properly, in order to be able to make a decision on whether it is worth sacrificing one over the other. There is another useful measurement that helps pondering the trade. It is called F-measure and it aims to estimate the optimal balance between precision and recall:

$$\text{F-measure} = (2 \text{ x precision x recall}) / (\text{precision} + \text{recall})$$

Finally, similar to sensitivity, there is specificity. This calculates the ratio of correctly identified negatives, considering the actual number of negative cases:

$$\text{Specificity} = TN / (TN + FP).$$

3. Hypotheses

As mentioned before, according to Moore (2012), users who are asked to write reviews tend to do it using explaining language, which carries more emotion and provides a better interpretation of the product for other users. This actually works either with positive and negative reviews. If the user had a good experience, they want to prove other users their opinion is honest and not bought, and so they tend to be detailed about the positive aspects of the product. On the other hand, if the experience was not so good and they still have to write their feedback, they will be more extra explanatory so they can excuse themselves in the eyes of the seller, who gave them something. Either way, the user feels like they have to justify their opinion. Hence, the first hypothesis states as follows:

H1a: Incentivized reviewers tend to be more extensive than regular reviewers.

H1b: Incentivized reviewers tend to be more emotional than regular reviewers.

As stated by Kim (2015), this type of reviews, with high emotional charge and sentiment, are more effective to consumers. This is the kind of feedback that remains in consumers' minds longer and is remembered when considering whether to buy a certain product or choosing an alternative. This means that these reviews weigh more than others when making a decision. Thus, another hypothesis to test is:

H2: There is a positive correlation between sentiment score and helpfulness of reviews.

4. Methodology

4.1. Overview

This empirical research adopts the CRISP-DM toward knowledge discovery. This is one of the most popular approaches to tackle problems in the field of knowledge discovery in databases, and a robust and well-proven methodology.

CRISP-DM stands for CRoss-Industry Process for Data Mining. The CRISP-DM methodology provides a structured procedure to planning a data mining project. It is the result of a study conducted by specialists in order to define the best practices and a standard process model for the systematization of knowledge discovery (Chapman, 2000).

According to CRISP-DM, the life cycle of a data mining project consists of six phases, as shown in Figure 1. Although it is presented in form of a cycle, this structure is not rigid, which means that one can – and should – go back and forth between the different phases always focusing in optimizing the results.

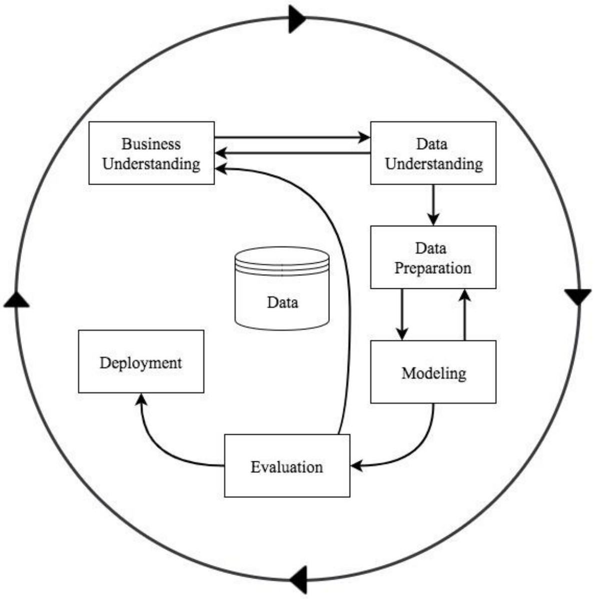


Figure 1 - CRISP-DM methodology phases (source: Chapman et al., 2000)

The initial phase of this cycle – Business Understanding – focuses on comprehending the objectives and necessities from a business perspective and translate this knowledge into a data mining problem and a plan to achieve the objectives. This leads to the Data Understanding phase, where the data is collected and there is the first contact with it. This is the moment to identify possible data quality problems or even some less obvious insights to the dataset. Then there is Data Preparation, which in a text mining problem, is one of the most important and complex phases (Hotho et al., 2005). In this phase, the data will go through multiple tasks in order to get to a final dataset that will be able to feed the modeling tool. In addition to the basic data mining preparation tasks, there is also a set of tasks specifically for unstructured data, which makes this phase much more complex for text mining. Once the dataset is prepared, in the Modeling phase various techniques are applied and the parameters calibrated to achieve optimal values. Since every model has different requirements, it is common at this stage to go back to the Data Preparation phase to make some changes and improve the modeling results. After creating the models, there is the Evaluation phase, where the performance of the models is analyzed and it is assessed if the objectives were achieved. In the end of this cycle there is the phase where the results are exposed in the form of a report, presentation or even the implementation of the created models in the organization. This final phase is called Deployment.

4.2. Business Understanding

The first step to begin an investigation is to study the context in which it will take place. This is the phase that focuses on and uncovers important factors like success criteria, business and data mining objectives and requirements, business terminologies and technical terms (Shafique and Qaiser, 2014). It is essential to acknowledge the importance of this step, since neglecting it can mean that a huge effort is put into producing the right answers to the wrong questions.

As mentioned in state of the art, the amount of unstructured data is growing exponentially. Until a few years ago, the most part of this information was disregarded since there was no efficient way to deal with it. Nowadays, with new techniques like text mining, businesses have come to realize the advantage they can obtain by knowing their customers' opinions and being able to focus their improvements. When it comes to online shopping, this subject has a special importance, not only to businesses but also to other customers. In platforms like Amazon reviews can be a determinant factor to an undecided customer.

That said, it is easy to understand why sellers take the reviewing process so seriously – and sometimes too far – to promote their products. This is the basis of this study.

Amazon has millions of products and even more product reviews. However, when a product is released, it is sometimes difficult to sell because there are no reviews to verify the quality of the product. This is one of the origin sources of incentivized reviews. Sellers would offer the product – or at least a significant discount, so people would buy it and write an honest review (ReviewMeta, 2016). If the product was actually good, this cycle would flow on its own, because good reviews generate sales and sales generate real reviews, and so on. But what started as a reasonable means to promote a product, soon became a not so legitimate way to deceive customers, once dishonest sellers started to adopt this procedure on their bad reviewed items. They would offer a product in exchange for a good review, and the customer would write it even if the product was not good. The latter is the other source of incentivized reviews, here mentioned as paid reviews.

As mentioned before, the study conducted by the team from ReviewMeta (2016) showed that, although the average star rating difference does not look very high (0.38 stars), considering the average rate on Amazon products, it could boost a product from mediocre to a top rated one, as shown in Figure 2.

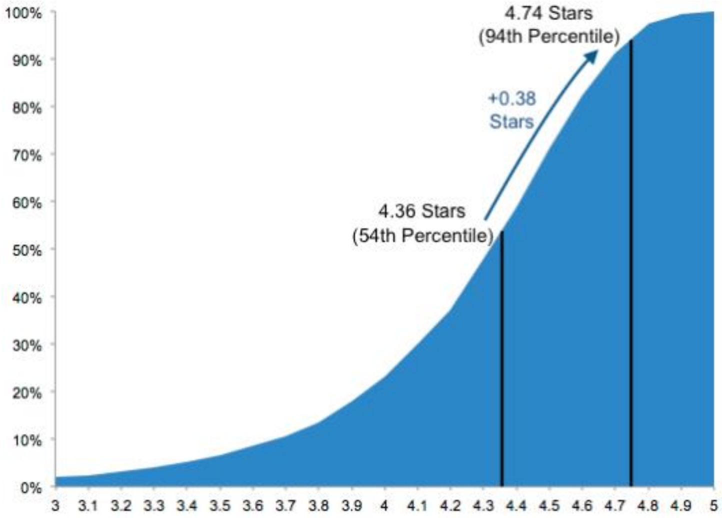


Figure 2 - Distribution of ratings of Amazon products (ReviewMeta, 2016)

In light of all this information, Amazon started to ban this practice in October 2016, in order to protect their customers and guarantee them a safe and trustworthy shopping platform.

As already mentioned before, it is naïve to assume this practice will cease to exist. The only difference is that it will be more difficult now – or even impossible – for a customer to identify these reviews. Therefore, the analysis intended in this dissertation is to find patterns in this type of reviews with the aim of recognize a possible incentivized review even if it does not have a disclaimer.

4.3. Data Understanding

According to CRISP-DM, after the business understanding phase starts the data understanding. The first task in this phase is to collect the data. Since this analysis is focused mostly on text data, at this stage it is necessary to have a tool that can process unstructured data (Sánchez et al., 2008). In order to ease the process of exploring, data should be presented in a database called *corpus* that contains all of the observations to be treated and analyzed (Feinerer et al., 2008).

The database used was retrieved from the University of California San Diego and first analyzed by McAuley (2016). It consists in more than 140 million product reviews and metadata from Amazon, spanning from May 1996 to July 2014. This dataset includes review features like ratings, text and helpfulness votes, product metadata such as descriptions, category information, price, brand and image features, and links from also viewed/also bought products. The dataset was separated by category in several subsets. There are 24 categories including Books, Electronics, Movies and TV, Video games, Beauty, Apps for Android, Baby, Musical instruments, among many others. Additionally, it was also provided a collection of k-core subsets. These are dense subsets of each category that have been reduced to extract the k-core, such that each of the remaining users and items has at least k reviews each. In this case, the author opted to create 5-core datasets. To avoid very sparse data, these were the subsets from where the samples to this work were collected. The reviews dataset had the structure presented in Table 2:

Table 2 - Description of variables

Variable	Description
reviewerID	ID of the reviewer
asin	ID of the product
reviewerName	name of the reviewer
helpful	helpfulness rating of the review
reviewText	text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime	time of the review (raw)

In order to better illustrate the dataset, following in Table 3, are some examples of data:

Table 3 - Data examples

Variable	Example 1	Example 2	Example 3
reviewerID	A2SUAM1J3GNN3B	A2GSY9DM54Y0HS	A3ORB4W3GI1YFG
asin	0000013714	0002007770	B000001OMA
reviewerName	J. McDonald	chibineko	Mike Tarrani
helpful	[2, 3]	[14, 20]	[3, 13]
overall	5.0	5.0	1.0
summary	Heavenly Highway Hymns	sooo many books, so little time!	Don't read
unixReviewTime	1252800000	1317686400	1219881600
reviewTime	09 13, 2009	10 4, 2011	08 28, 2008

For this study, two subsets were selected. The intention of using two different datasets is to see if there is a relevant difference in this behavior depending on the category of the product. The datasets selected were Books and Electronics, since they are two totally different categories and are also two of the larger subsets: 8,898,041 reviews for books and 1,689,188 for electronics. Even though these are only subsets of the major dataset, they are still too massive to work with and to be processed, so a random balanced sample was generated from each of the subsets. The extraction task was performed by a program written in Python, built specifically for this step. The samples were extracted in two steps: a first run through the data would extract all the incentivized reviews found in the 5-core subset, and a second run would extract the same number of non-incentivized reviews. A review was classified as incentivized if it contained any disclaimer in its text. After a manual exploratory analysis of hundreds of reviews, a list of the observed disclaimers was created, gathering terms such as:

- disclaimer
- discount for review
- discount to review
- free sample
- free to review
- freebie
- in exchange for a review
- product for review
- product sent for review
- review sample
- reviewing purposes
- sample for an honest review
- testing and reviewing purposes

and some other variations of those. In the first version of this list, there were nearly 70 words/expressions to identify incentivized reviews. In the first run through the data, all reviews were converted to lower case, and so were the terms in the disclaimers list, to make the searching process more effective. The program extracted all the reviews that contained at least one of the disclaimers on the list. After a quick manual verification of the extracted reviews, more than 40 terms were excluded since they were causing the program to misclassify some reviews as incentivized. This process was repeated a couple more times, until the list had only the more infallible terms. The final list had 26 expressions and can be found attached (in appendix A). After extracting the subset for incentivized reviews, the sample of non-incentivized reviews was collected, based on the assumption that it could not have any of the expressions from the previous list and also had to have some word/expression mentioning the act of an actual purchase. The expression used was “I bought”. So the program ran a second time, stopping once it extracted the same number of cases in the incentivized reviews subset. This way, the samples generated were balanced: 50% incentivized reviews and 50% non-incentivized.

The final Books dataset had a total of 105,202 reviews, being exactly half incentivized and half non-incentivized, while the Electronics dataset consisted of 5,594 reviews in total. Following is an example of an incentivized review and a non-incentivized one:

Table 4 - Examples of text reviews

Incentivized review	This case is the absolute best tablet or Fire case I have EVER used!!! Not only did it fit well and have all the cutouts placed correctly, it was nice to sit the tablet up as a mini display when in meetings so everyone could watch. I am very very very happy with the case!! The dark blue was a nice change of pace from my old hot pink case. The styluses and the charging cords all worked reliably, they were all a great bonus for the new case. The value is also superb, too! This case was sent for **reviewing purposes**, but these are my honest opinions after using this bundle for over a month. I've experienced no issues with anything, including the cords and in fact it's a great feature to add all these value extras!
Non-incentivized review	For the price you just cant beat this item. **I bought** it for a house I was selling so I could hang some tv and it worked great. Good quality for the price.

Since the original datasets were extremely large, the extraction process was designed to avoid data quality problems. For example, the program would only extract observations with no missing values, sparing the trouble of managing this issue further ahead. However, there is one exception. In the "helpful" field, the value is given in form of a fraction: *X* helpful votes in a total of *Y* votes. This means that *Y* people voted but only *X* found the review helpful. Although there is not a missing value in the dataset *per se* for any of the reviews extracted, when *Y* is 0 it can be interpreted as a lack of information, since there were no votes whatsoever. The reason for not excluding the cases with a value of [0, 0] for "helpful" was because most people do not vote on the helpfulness of a review unless they actually feel that it was clearly useful or clearly useless (usually in case the review is fake) for their decision. This leads to a considerable number of unvoted reviews that are not necessarily invalid. In order to be able to include them in the dataset to assess this issue in more detail, 3 variables were accounted: the number of total votes (*Y*), the number of helpful votes (*X*) and a normalized measure that shows the proportion of people who found the review helpful (*X/Y*). Thus, these reviews were differentiated from the reviews where the normalized helpfulness was actually zero. The number of cases accounted for the helpfulness rate ("helpful") were only the cases where "helpful_total_votes" was different than 0. So for this analysis, the cases with no total votes were considered as a missing value.

The software used to perform an overview analysis on the datasets was IBM SPSS Statistics. It provides a range of techniques, including ad-hoc analysis, hypothesis testing and reporting, to make it easier to access and manage data, and share the results. A descriptive statistics analysis of the variables for both datasets is presented in Tables 5 and 6.

Table 5 - Descriptive statistics - Electronics

Electronics – Statistics

	N		Mean	Std. Deviation	Minimum	Maximum
	Valid	Missing				
reviewerID	5594	0				
asin	5594	0				
helpful_votes	5594	0	10.03	52.196	0	1461
helpful_total_votes	3615	1979	17.31	67.947	1	1549
overall	5594	0	4.21	1.142	1	5
unixReviewTime	5594	0				
reviewTime	5594	0				

Table 6 - Descriptive statistics - Books

Books – Statistics

	N		Mean	Std. Deviation	Minimum	Maximum
	Valid	Missing				
reviewerID	105202	0				
asin	105202	0				
helpful_votes	105202	0	4.11	45.659	0	10755
helpful_total_votes	60132	45070	9.46	65.823	1	11479
overall	105202	0	4.06	1.179	1	5
unixReviewTime	105202	0				
reviewTime	105202	0				

Both in Electronics and Books datasets, the values of the mean and standard deviation calculated for the variable “helpful_votes” are not correct since all observations are being considered. For the analysis on the particular case of helpfulness rate, the cases with value of zero for “helpful_total_votes” should not be considered. Therefore, the adjusted values are presented in Tables 7 and 8.

Table 7 - Adjusted metrics - Electronics

Electronics – Statistics

	N		Mean	Std. Deviation	Minimum	Maximum
	Valid	Missing				
helpful_votes	3615	0	15.53	64.273	0	1461
helpful_total_votes	3615	0	17.31	67.947	1	1549

Table 8 - Adjusted metrics - Books

Books – Statistics

	N		Mean	Std. Deviation	Minimum	Maximum
	Valid	Missing				
helpful_votes	60132	0	7.18	60.210	0	10755
helpful_total_votes	60132	0	9.46	65.823	1	11479

Regarding the overall rating, the distribution is similar in both datasets as depicted in Figure 3 and Figure 4.

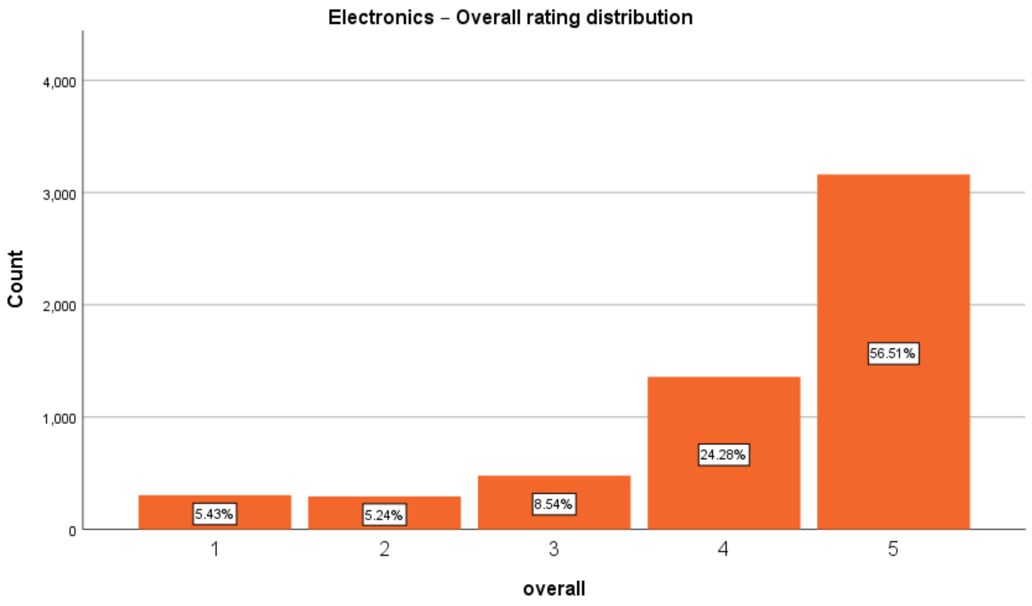


Figure 3 - Overall rating distribution - Electronics

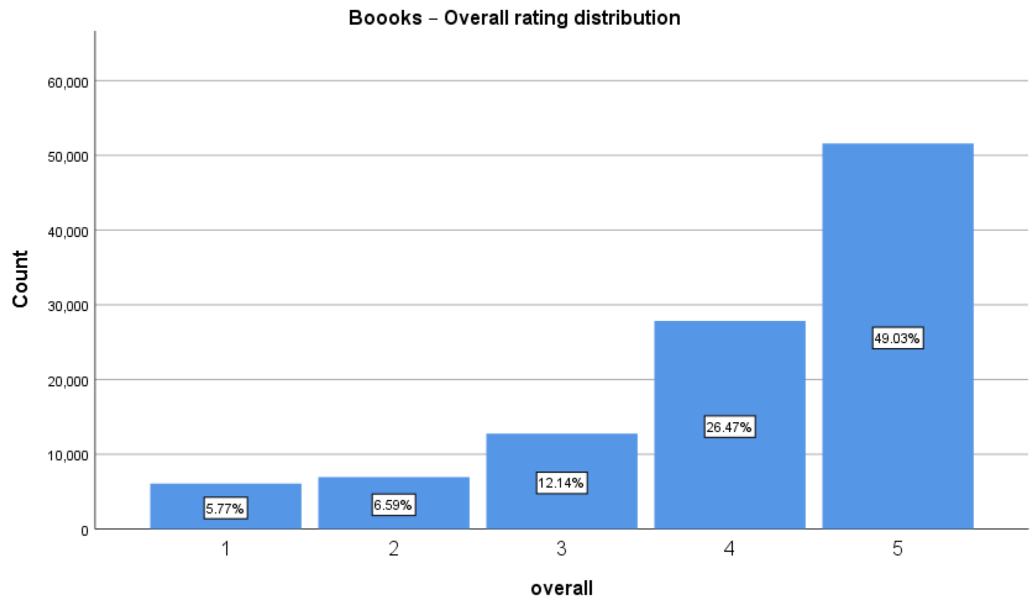


Figure 4 - Overall rating - Books

By way of example, Table 9 displays the values of the mean, standard deviation, minimum and maximum, for comparison purposes between incentivized and non-incentivized reviews.

Table 9 - Descriptive statistics overall - Electronics

Electronics – Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
non-incentivized	2797	1	5	4.16	1.219
incentivized	2797	1	5	4.26	1.058

Even though it is not as accentuated as in the ReviewMeta report, a difference in the mean of 0.10 stars is still a significant deviation, considering the presented distribution of this variable.

4.4. Data Preparation

Data preparation is a key step toward obtaining accurate models. Its main purpose is to structure the data so it can be manipulated by the algorithms of pattern extraction (Liu, 2008). In this stage, it is expected to evolve from the *corpus* to a structured database

(Delen and Crossland, 2008). The documents that constitute the *corpus* are usually poor in content quality, which is why it is crucial to perform all the selecting, cleaning and formatting tasks to prepare the data (Feinerer et al., 2008). The original data used in this analysis was structured in a JSON format. JSON is a way to store information in an organized, easy-to-access manner. It provides a human-readable collection of data that can be accessed in a logical manner, which simplified the process of creating the dataset.

For the final dataset, the text of the review is not included, since it is not necessary for the modeling phase and it would only make the dataset files larger and harder to process. The dataset variables will be explained in greater detail towards the end of this section.

Some of the tasks commonly executed in this phase are tokenization, normalization, part of speech, negation handling and dimensionality reduction. Following, there is a brief explanation of each of these tasks.

Tokenization

The first task in data preparation for a text mining process is usually tokenization (Aranha and Passos, 2008; Webster and Kit, 1992). Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens (Hassler and Fliedl, 2006). A token is an entity that cannot be further decomposed (Miner et al., 2012, p. 13; Webster and Kit, 1992), whether it is a word or a group of words – expression. (Hassler and Fliedl, 2006; Kaplan, 2005).

After this task, it is necessary to standardize the data and apply dimensionality reduction techniques, so it reduces the noise in the data and the processing time (Blake, 2011).

Normalization

It is common practice to map variants of a term to a single, normalized form in order to reduce the dimensionality of the corpus, creating only a version of a certain word or expression (Ananiadou, 2013). This solves the problem of upper and lower case and acronyms and it is an important part of the process of automatic term recognition. However, there are authors who say that it can be worthy to let upper case tokens be treated as such, since it can indicate a more intense feeling (Brooke et al., 2009).

Part of Speech

In the part of speech tagging, tokens are categorized by syntax (Jackson and Moulinier, 2007). A part-of speech tagger is a system that uses context to assign parts of speech to words (Cutting et al., 1992). There are several categories like nouns, verbs, adjectives, etc. Depending on the investigation question, different analysis may be performed over different types of tags (Gimpel et al., 2011; Chopra and Bangalore, 2012). Previous studies have shown that adjectives are good indicators of the presence of opinion and help determine the sentiment attached (Hatzivassiloglou and Wiebe, 2000; Wiebe et al., 2001). Other studies revealed the not so obvious relevance of other tags that are not adjectives, such as Hu and Liu (2004), who used nouns to detect the different subjects addressed in a review.

Negation handling

Sentence analysis becomes even more complex with the negation factor. To analyze negation words, it is necessary to identify its scope, since it can be can be local (e.g., not good), involve longer-distance dependencies (e.g., does not look very good), negation of the subject (e.g., no one thinks that it’s good) or even a change in its role – depending on the context of the sentence, negation can be used to intensify (e.g., not only good but amazing) (Wilson et al., 2005). In order to find out the scope of the negation, the sequence of words in the sentence should be identified. Overall, it is not simply the negation of a word but negation of the sentence. In addition to this, there are also contrast words, like “but” or “however”. These words usually change the polarity of the sentiment, giving more emphasis to the idea after the contrast word.

Dimensionality reduction

This operation has the same purpose of normalization, which is to reduce the noise of the data. Some of the tasks performed include removing hyperlinks, sites and notation (Das and Chen, 2007; Guerreiro et al., 2015; Lin et al., 2012). Another task is removing stopwords. Stopwords are extremely common words that have little value in the analysis, or words that do not carry information, like articles, connectors or prepositions (Feinerer et al., 2008). Besides these specific types of words, there are other words that simply do not contribute to the analysis and thus should be eliminated (Delen and Crossland, 2008).

Another task that plays an essential part is stemming. Stemming is the process of transforming a word in its normalized form. This includes different verb forms or words of the same family of derivation with similar meaning, like the words "fishing", "fished", and "fisher" that would be reduced to the root word "fish". One of the most popular stemming algorithms is the Porter stemmer. It was proposed by Martin Porter in 1980 and is still the default go-to stemmer. This algorithm performs about sixty rules in six steps and provides a good trade-off between speed, readability and accuracy (Porter, 1980). However, it is controversial whether stemming improves quality of text mining result. It significantly reduces the bag-of-words together with stop word removal, but due to the lack of consideration of syntax, the context information attached to the original word is eliminated by stemming.

For this work, the sentiment analysis was performed using the VADER algorithm (Hutto and Gilbert, 2014) on two different levels for each review in the datasets. First it was determined the sentiment scores of the review. These scores describe the sentiment intensity of the review as a whole. The compound score of a review is the easiest measure to understand, because it gives direct information of how positive or negative a given review is. Secondly, a sentiment analysis was carried out on a sentence level. For each sentence of a review, the analysis would compute the sentiment scores and depending on the highest score, it would classify a sentence as being positive, negative or neutral. After every sentence was classified, the final scores were calculated as being the average of the scores of each polarity. For instance, if a review had 10 sentences classified as being 3 positive, 1 negative and 6 neutral, the average scores calculated for this review would be:

avg_positive_sentences = average of the positive scores of the 3 positive sentences.

avg_negative_sentences = average of the negative scores of the 1 negative sentence.

avg_neutral_sentences = average of the neutral scores of the 6 neutral sentences.

The purpose of this estimates is to have some exponentiated scores of the polarity of each review, since most reviews are mainly neutral, as it will be seen further ahead.

After performing sentiment analysis in the datasets created in the previous phase, a few more variables were added to the final data sets. Besides the sentiment scores, calculated through VADER as explained before, 3 length-related variables were also added: number of characters, number of words and number of sentences in a review and also a flag variable that indicated whether a review was incentivized or not.

Table 10 shows a more detailed explanation of the variables that constitute the final dataset:

Table 10 - Variables description

Variable	Description
reviewid	ID of a review. Assigned number during the extraction process to easily access the text of a review.
reviewerID	Amazon reviewer ID, provided in the original dataset.
asin	Amazon product ID, provided in the original dataset.
unixReviewTime	Time when the review was written in unix format, provided in the original dataset. It places a review in a specific moment in time and for that reason it will be excluded from the model, since it cannot be used as a predictive variable to classify new reviews.
reviewTime	Date when the review was written, in raw format, provided in the original dataset.
number_chars_review	Number of characters in the review text. Calculated in the extraction process.
number_words_review	Number of words in the review text. Calculated in the extraction process.
number_sentences_review	Number of sentences in the review text. Calculated in the extraction process.
overall	Star rating given by the reviewer to the product, scored from 1 to 5, provided in the original dataset.
helpful	Helpfulness rate of a review. This score is determined by other customers, who vote if the review was helpful in their decision process. In the original dataset, 2 values are provided: number of helpful votes and number of total votes (e.g. 2/3, which means that 3 people voted that review: 2 found it helpful and the other one did not). In order to have a normalized score, this parameter was converted in a scale from 0 to 1.
helpful_votes	Number of helpful votes of a review, provided in the original dataset.
helpful_total_votes	Number of total votes of a review (helpful and not helpful), provided in the original dataset.
overall_positive_value	Positive score of the review as a whole, calculated through the VADER algorithm.
overall_negative_value	Negative score of the review as a whole, calculated through the VADER algorithm.
overall_neutral_value	Neutral score of the review as a whole, calculated through the VADER algorithm.
overall_compound_value	Compound score of the review as a whole, calculated through the VADER algorithm. This is a single measure of polarity of the review, ranging in a scale from -1 (extremely negative) to 1 (extremely positive).
number_positive_sentences	Number of positive sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the positive score is the highest out of the three, the sentence is classified as positive.
number_negative_sentences	Number of negative sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the negative score is the highest out of the three, the sentence is classified as negative.

number_neutral_sentences	Number of neutral sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the neutral score is the highest out of the three, the sentence is classified as neutral.
avg_positive_sentences	Average of the positive scores of the positive sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
avg_negative_sentences	Average of the negative scores of the negative sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
avg_neutral_sentences	Average of the neutral scores of the neutral sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
compound_sentences_average	Average of the compound scores of all the sentences on a review, calculated through VADER algorithm.
summary_positive	Positive score of the review summary, calculated through the VADER algorithm.
summary_negative	Negative score of the review summary, calculated through the VADER algorithm
summary_neutral	Neutral score of the review summary, calculated through the VADER algorithm.
incentivized	Categorical variable that labels the review as being incentivized (1) or non-incentivized (0).

Hypothesis Testing

***H1a**: Incentivized reviewers tend to be more extensive than regular reviewers.*

To analyze the relation between the length of a review and the fact that it is incentivized, a histogram of frequencies was built in SPSS Statistics. By way of example, the charts displayed in Figures 5 and 6 show the distribution of one of the length variables – the number of characters of a review.

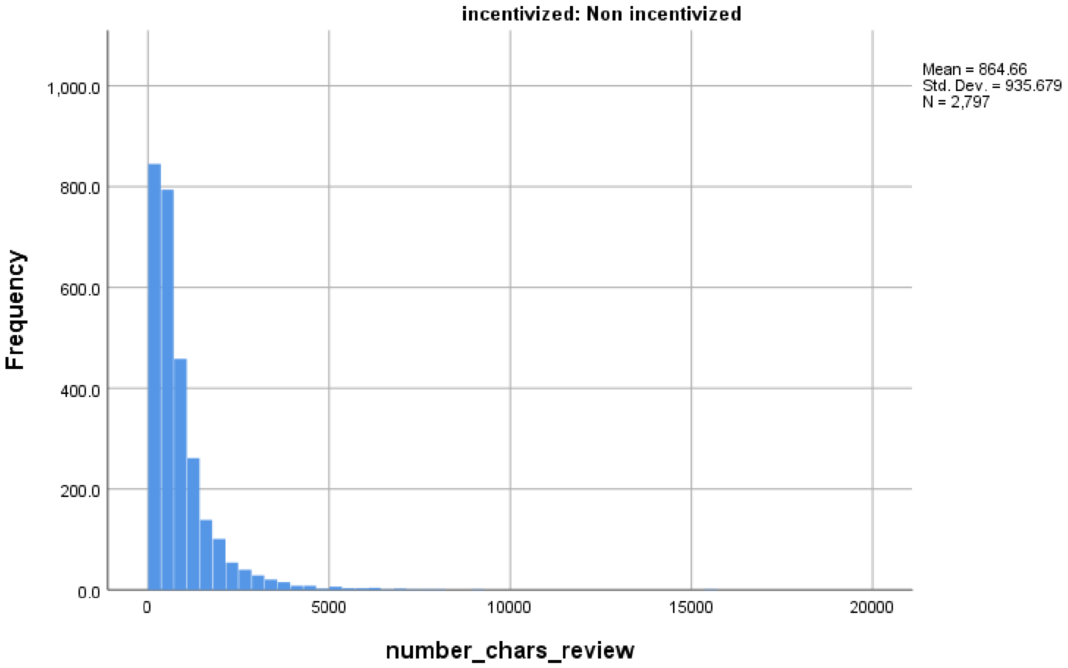


Figure 5 - Number of characters in non-incentivized reviews – Electronics

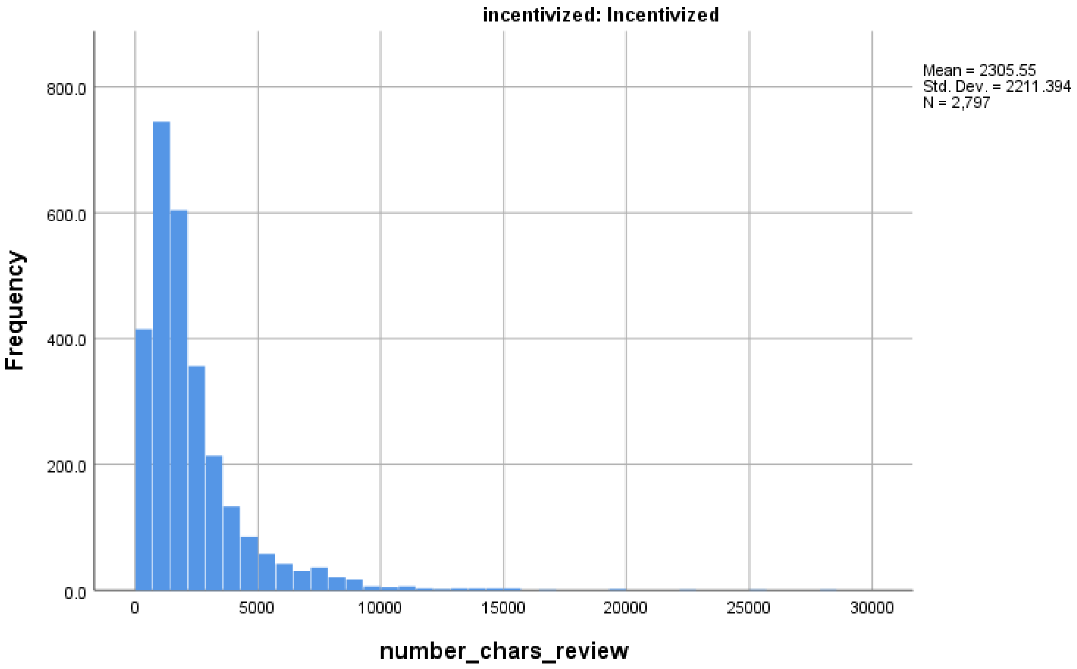


Figure 6 - Number of characters in incentivized reviews – Electronics

In average, a non-incentivized review has 865 characters while an incentivized has 2306, which is almost three times more. When it comes to word count, the results have a similar ratio: the average for a non-incentivized review is 159 words, whereas an incentivized has an average of 412 words. Finally, as for the number of sentences in a review, the average in an incentivized review is 18, which is more than the double of the non-incentivized – 8 sentences.

A Krustal-Wallis test was also used to verify the relationship between these two variables. The lower the asymptotic significance is, the more significant the test is. For both datasets, the test carried is significant and it is possible to say that incentivized reviewers do write longer reviews.

Table 11 - Kruskal-Wallis test - Electronics

Ranks – Electronics

	incentivized	N	Mean Rank
number_chars_review	Non-incentivized	2797	1937.10
	Incentivized	2797	3657.90
	Total	5594	

Table 12 - Kruskal-Wallis test statistics - Electronics

Test Statistics – Electronics

	number_chars_review
Kruskal-Wallis H	1587.737
df	1
Asymp. Sig.	.000

Table 13 - Kruskal-Wallis test - Books

Ranks – Books

	incentivized	N	Mean Rank
number_chars_review	Non-incentivized	52601	39265.11
	Incentivized	52601	65937.89
	Total	105202	

Table 14- Kruskal-Wallis test statistics - Books

Test Statistics – Books

	number_chars_review
Kruskal-Wallis H	20287.559
df	1
Asymp. Sig.	.000

In light of this analysis, it can be said the hypothesis H1a is supported.

***H1b**: Incentivized reviewers tend to be more emotional than regular reviewers.*

To verify the second part of the hypothesis, concerning the emotional charge in a review, the variable used was the overall compound score, since it is the most understandable measure of the sentiment-related ones. Since the variable takes values between -1 (extremely negative) and +1 (extremely positive), the strength of the sentiment will be considered as the absolute value of this variable, meaning that neutral reviews are the ones with the lower sentiment charge.

In order to have a better understanding of the differences of sentiment scores between the incentivized and the non-incentivized reviews, a box plot graph was created for the overall compound scores. This type of plot is very suggestive to visual interpretation, since it shows how the scores are distributed. The median is a measure of central tendency, but unlike the mean, it is not so influenced by cases with extreme values. The black line in the middle of the blue box represents the median of the scores. 50% of the cases lie within the box. The range between the whiskers is expected to comprehend approximately 95% of the cases, if the data is normally distributed. Observing the graphs in Figures 7 and 8, it is possible to see that almost the whole incentivized subset, both in Books and Electronics, scored just as high as only the top half of the non-incentivized reviews.

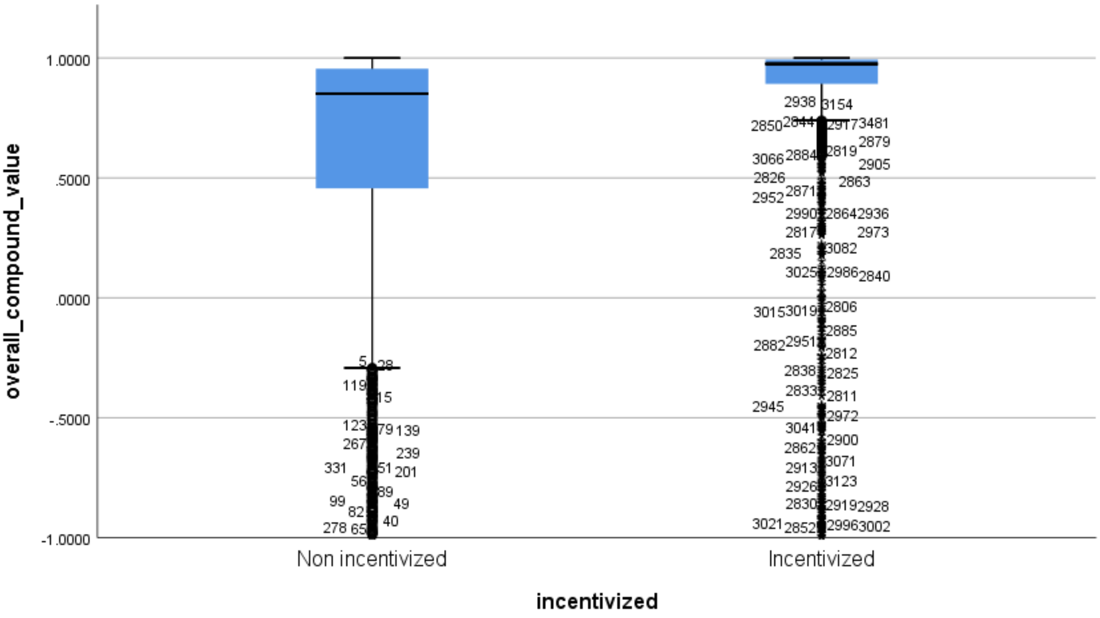


Figure 7 - Box plot overall compound score - Electronics

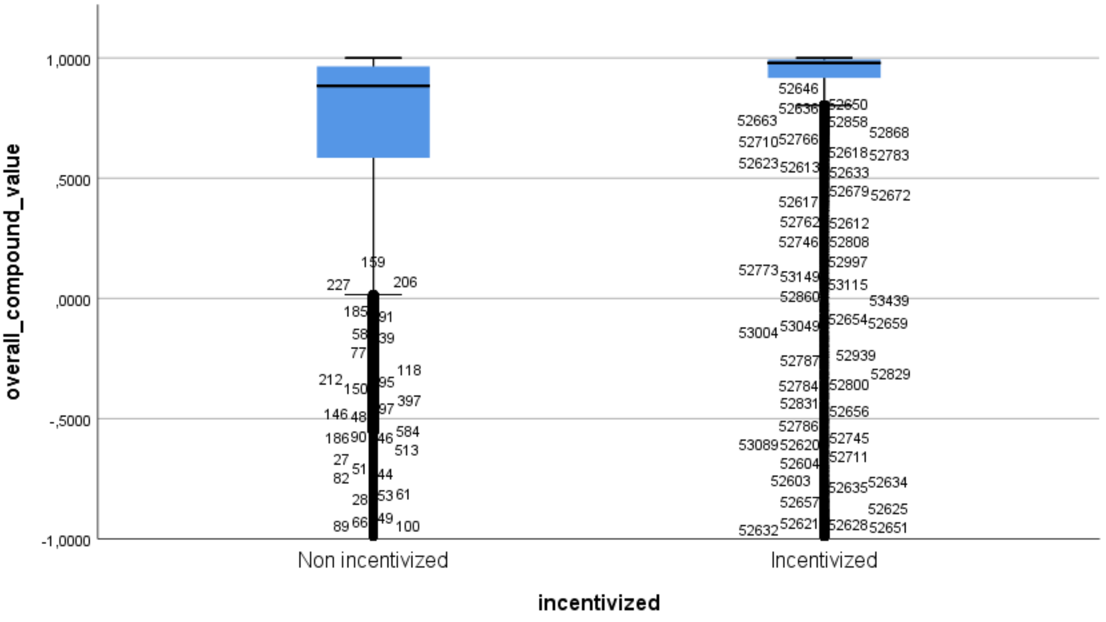


Figure 8 - Box plot overall compound score - Books

The descriptive statistics of this variable in both datasets, split by incentivized and non-incentivized cases, are displayed in the following tables.

Table 15 - Descriptive statistics overall compound score - Electronics

Electronics – Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
non-incentivized	2797	-0.9906	0.9999	0.5955	0.5258
incentivized	2797	-0.9977	0.9999	0.8167	0.4182

Table 16 - Descriptive statistics overall compound score - Books

Books – Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
non-incentivized	52601	-0.9997	0.9999	0.6310	0.5400
incentivized	52601	-0.9997	0.9999	0.8047	0.4657

Analyzing this metrics in a Kruskal-Wallis test, it is possible to verify that in both cases, the incentivized reviews have a higher score in the overall compound value. Since the asymptotic significance is null, this test does demonstrate that the overall compound score is higher if the review is incentivized. These results are presented in the next tables.

Table 17 - Kruskal-Wallis test - Electronics

Ranks – Electronics

	incentivized	N	Mean Rank
overall_compound_value	Non-incentivized	2797	2157.00
	Incentivized	2797	3438.00
	Total	5594	

Table 18 - Kruskal-Wallis test statistics - Electronics

Test Statistics – Electronics

	overall_compound_value
Kruskal-Wallis H	879.879
df	1
Asymp. Sig.	.000

Table 19 - Kruskal-Wallis test - Books

Ranks – Books

	Incentivized	N	Mean Rank
overall_compound_value	Non-incentivized	52601	41119.69
	Incentivized	52601	64083.31
	Total	105202	

Table 20 - Kruskal-Wallis test statistics - Books

Test Statistics – Books

	overall_compound_value
Kruskal-Wallis H	15037.449
df	1
Asymp. Sig.	.000

***H2**: There is a positive correlation between sentiment score and helpfulness of reviews.*

To test this hypothesis, the same assumption from the previous case was made: reviews with elevated sentiment are the ones with a high absolute value for the overall compound score. The graph used to analyze this case was a scatter plot, presented in Figure 9, where it was possible to see how the observations were distributed in this relation. In this situation, no split between incentivized and non-incentivized data was required.

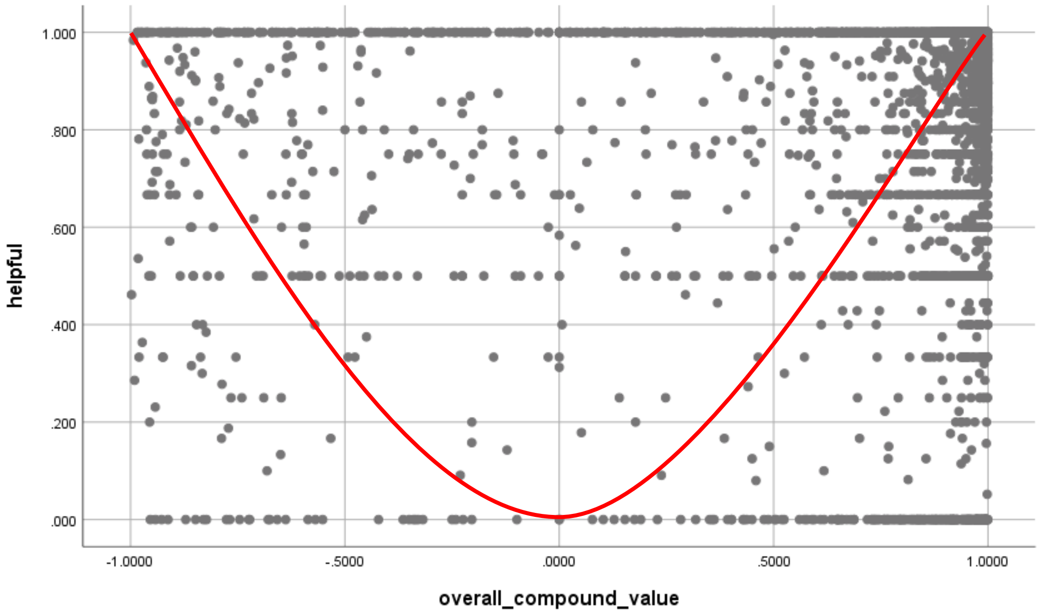


Figure 9 - Relation between overall compound and helpfulness rate – Electronics

For the hypothesis H2 to be proven, the scatter distribution should go along with the red line. This illustrative line has the purpose of comparing the real dispersion of the data with the desired one. On the top right corner of the graph of the Electronics dataset there is a concentration of data, where the sentimental charge is elevated and the helpfulness rate is high.

Although it would be reasonable to think this was just the crossing of two independent variables' distributions, a Pearson correlation test shows that there is in fact a significant positive correlation between the two variables. The Pearson Correlation indicates whether a statistically significant linear relationship between a pair of continuous variables exists and measures its strength and direction. The null hypothesis (H0) and alternative hypothesis (H1) of the significance test for correlation can be expressed in the following way:

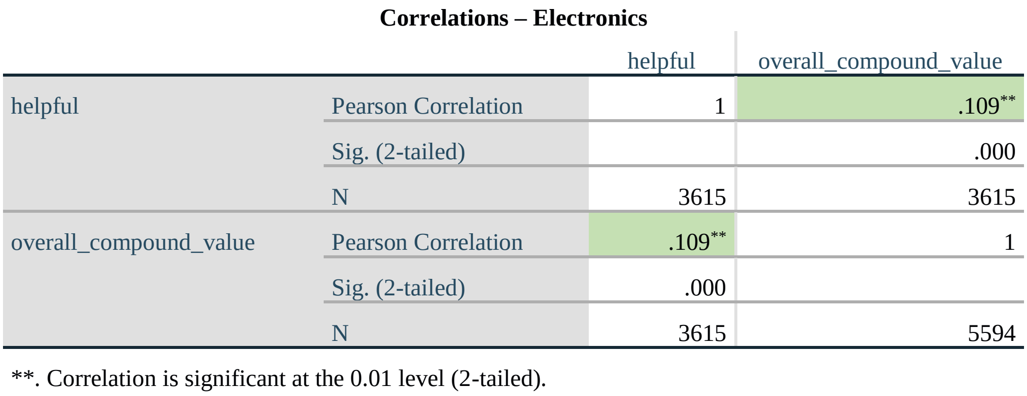
- H0: ρ = 0 ("the population correlation coefficient is 0; there is no association")
- H1: ρ ≠ 0 ("the population correlation coefficient is not 0; a nonzero correlation could exist")

where ρ is the population correlation coefficient. Correlation can take a value in the range [-1, 1]. The sign of the correlation coefficient indicates the direction of the relationship, while the magnitude of the correlation (how close it is to -1 or +1) indicates the strength of the relationship. The strength can be assessed by the general guidelines provided by Cohen (1988):

- $0.1 < |r| < 0.3$ … small / weak correlation
- $0.3 < |r| < 0.5$ … medium / moderate correlation
- $0.5 < |r|$ … large / strong correlation.

The Pearson Correlation value for these two variables is the displayed in Table 21.

Table 21 - Pearson correlation - Electronics



Correlations – Electronics

		helpful	overall_compound_value
helpful	Pearson Correlation	1	.109**
	Sig. (2-tailed)		.000
	N	3615	3615
overall_compound_value	Pearson Correlation	.109**	1
	Sig. (2-tailed)	.000	
	N	3615	5594

**. Correlation is significant at the 0.01 level (2-tailed).

Even though the correlation is weak (green cells), it does exist and it is significant at a level of 0.01: $r = 0.109$, $p < 0.01$. The correlation is positive, which means that these variables tend to increase together.

For the Books dataset, the graph is less clear than for the Electronics, but the Pearson Correlation coefficient, presented in Table 22, is higher.

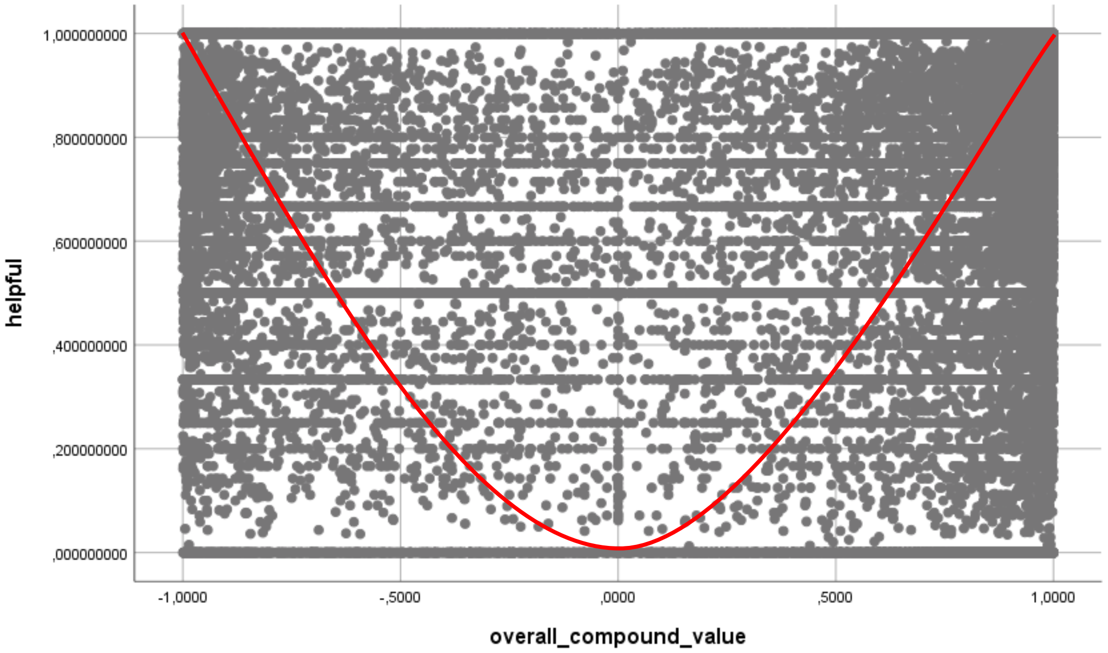


Figure 10 - Relation between overall compound and helpfulness rate – Books

Table 22 - Pearson correlation - Books

Correlations – Books

		helpful	overall_compound_value
helpful	Pearson Correlation	1	,150**
	Sig. (2-tailed)		,000
	N	60132	60132
overall_compound_value	Pearson Correlation	,150**	1
	Sig. (2-tailed)	,000	
	N	60132	105202

**. Correlation is significant at the 0.01 level (2-tailed).

In this case, the correlation is r = 0.150, p < 0.01, which means that there is also a positive correlation between these two variables.

In light of these results, it is possible to conclude that the hypothesis H2 is supported, although the correlation is weak, for both datasets.

4.5. Modeling

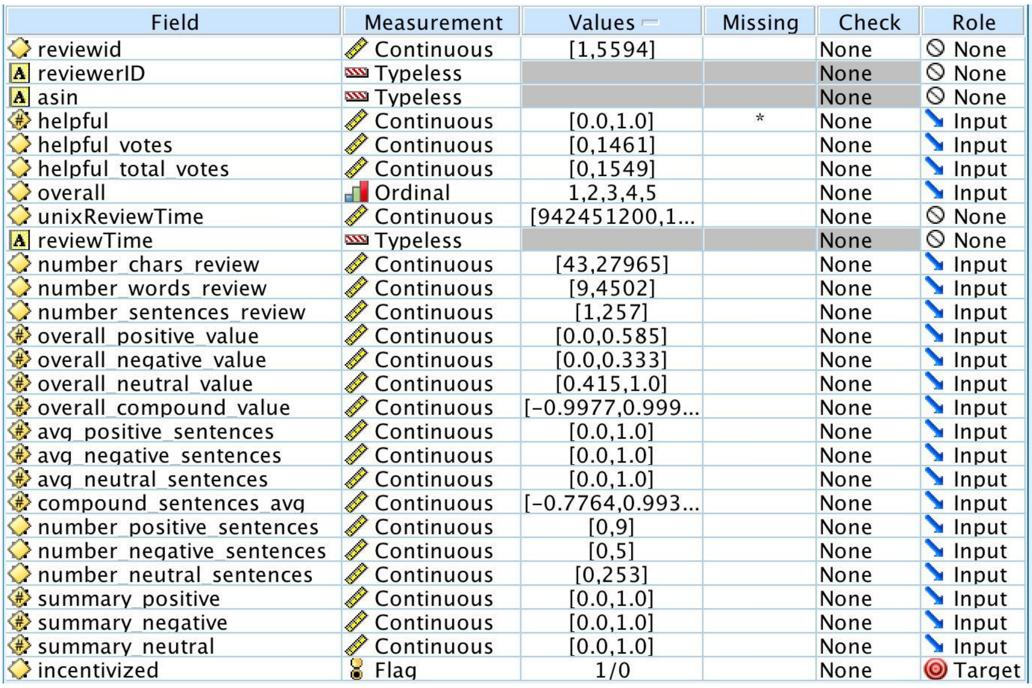
In this step, the datasets built and prepared in the previous phases were used to serve a modeling tool, with the purpose of finding the model that would best explain the existing data and predict the target variable for new observations.

Since the purpose of the analysis in this work was to predict whether a non-disclosed review was incentivized or not, the required model is a classification model. For this type of problems, the most common models are decision trees, regression, neural networks, support vector machines and Bayesian networks.

Configurations

Since both streams (Electronics and Books) were set up the same way, the demonstration and explanations were only made for Electronics, to avoid redundancy of information.

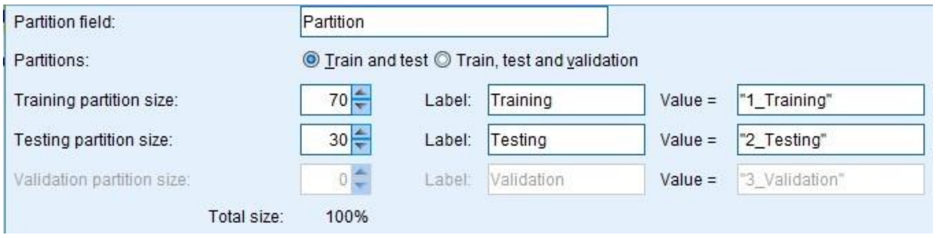
Concerning the variables' types and roles, Figure 11 shows how the configuration was made. The target variable for this model was "incentivized", since it was the variable to predict. The rest of the variables were labeled as input data, except for "reviewid", "reviewerID", "reviewTime" and "asin" that, due to their type, cannot be accounted.



Field	Measurement	Values	Missing	Check	Role
reviewid	Continuous	[1,5594]		None	None
reviewerID	Typeless			None	None
asin	Typeless			None	None
helpful	Continuous	[0.0,1.0]	*	None	Input
helpful_votes	Continuous	[0,1461]		None	Input
helpful_total_votes	Continuous	[0,1549]		None	Input
overall	Ordinal	1,2,3,4,5		None	Input
unixReviewTime	Continuous	[942451200,1...		None	None
reviewTime	Typeless			None	None
number_chars_review	Continuous	[43,27965]		None	Input
number_words_review	Continuous	[9,4502]		None	Input
number_sentences_review	Continuous	[1,257]		None	Input
overall_positive_value	Continuous	[0.0,0.585]		None	Input
overall_negative_value	Continuous	[0.0,0.333]		None	Input
overall_neutral_value	Continuous	[0.415,1.0]		None	Input
overall_compound_value	Continuous	[-0.9977,0.999...		None	Input
avg_positive_sentences	Continuous	[0.0,1.0]		None	Input
avg_negative_sentences	Continuous	[0.0,1.0]		None	Input
avg_neutral_sentences	Continuous	[0.0,1.0]		None	Input
compound_sentences_avg	Continuous	[-0.7764,0.993...		None	Input
number_positive_sentences	Continuous	[0,9]		None	Input
number_negative_sentences	Continuous	[0,5]		None	Input
number_neutral_sentences	Continuous	[0,253]		None	Input
summary_positive	Continuous	[0.0,1.0]		None	Input
summary_negative	Continuous	[0.0,1.0]		None	Input
summary_neutral	Continuous	[0.0,1.0]		None	Input
incentivized	Flag	1/0		None	Target

Figure 11 - "Type" node configuration

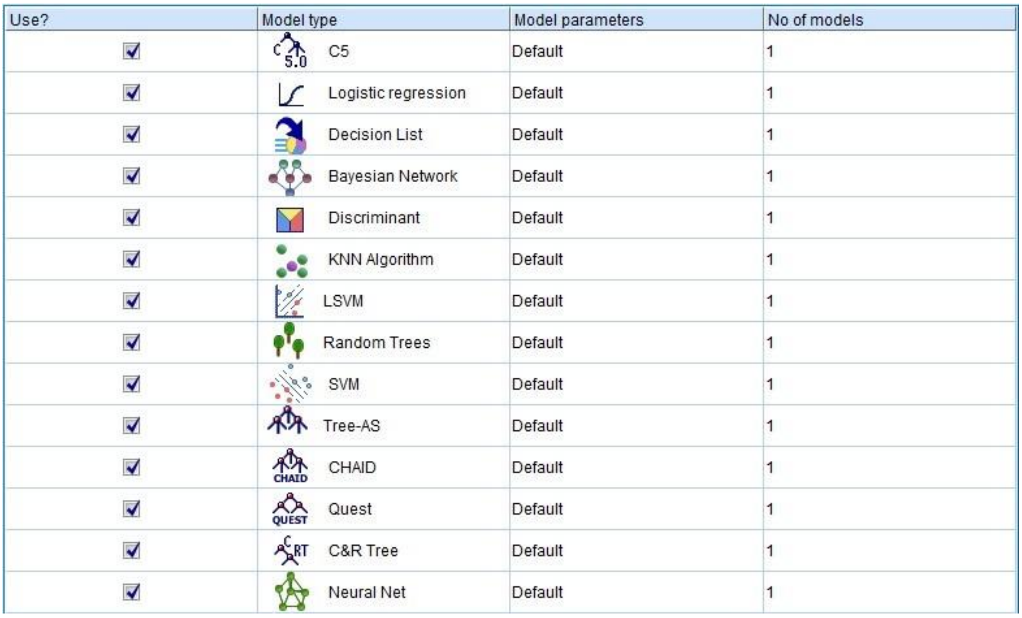
In the partition node, the training dataset was set to 70% and the testing to 30%, as shown in Figure 12.



Partition field:	Partition				
Partitions:	● Train and test ○ Train, test and validation				
Training partition size:	70	Label:	Training	Value =	"1_Training"
Testing partition size:	30	Label:	Testing	Value =	"2_Testing"
Validation partition size:	0	Label:	Validation	Value =	"3_Validation"
Total size:	100%				

Figure 12 - "Partition" node configuration

In order to have a general overview of the accuracy of the several models, the Auto Classifier node available in SPSS Modeler was set to test the 14 models, as shown in Figure 13 and, based on the overall accuracy, rank the top five.



Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Default	1
☑	LSVM	Default	1
☑	Random Trees	Default	1
☑	SVM	Default	1
☑	Tree-AS	Default	1
☑	CHAID	Default	1
☑	Quest	Default	1
☑	C&R Tree	Default	1
☑	Neural Net	Default	1

Figure 13 - "Auto Classifier" node configuration: models to test

4.6. Evaluation

After modeling using the datasets, in this stage models created are evaluated. The first step is to assess the degree to which the model meets the business objectives. Some of the tasks to perform at this stage include understanding mining results, interpreting the results in terms of the application, evaluating if the discovered information is new and useful and validating if the model achieved the original business objectives (Chawla, 2009). After these tasks, it is time to summarize the process review and identify failures, deviations and possible alternative actions.

Examining the result of the "Auto classifier", it is possible to see that four of the five most accurate models were the same for both datasets – four decision tree models, although there was a slight difference in accuracy values from one dataset to the other. However, the fifth model generated for Books dataset was a Neural Network, while for Electronics was another decision tree. After running the "Auto classifier", and based on its findings, the five models of each stream were individually generated, in order to evaluate and compare their performances, since the only comparison measure this node provides is accuracy. The decision tree models were generated setting the maximum tree depth to 5 and the stopping rules to the default values – 1% of minimum records in a child branch and 2% in a parent branch. For Electronics dataset, C5.0 model achieved 77.19% of accuracy for the training data and 76.61% for testing data, which denotes a rather accurate result, while for Books dataset the Neural Network achieved 75.85% for training data and 75.58% for testing data. In fact, all five models for both datasets performed almost as good. But to assess the quality of a model, there are more measures than just accuracy. Based on the information on the confusion matrix, it is possible to calculate several performance measures. Tables 23 and 24 comprise the values for the confusion matrixes of all generated models.

Table 23 - Confusion matrix values - Electronics models

Electronics	Training				Testing			
	TP	TN	FP	FN	TP	TN	FP	FN
C5.0	1465	1546	447	443	682	615	189	207
Tree AS	1440	1520	473	468	669	615	189	220
C&R	1492	1465	528	416	687	596	208	202
QUEST	1341	1591	402	567	628	653	151	261
CHAID	1440	1553	440	468	656	617	187	233

Table 24 - Confusion matrix values - Books models

Books	Training				Testing			
	TP	TN	FP	FN	TP	TN	FP	FN
Neural net	28134	27807	9166	8662	12128	11630	3998	3677
Random Trees	28314	28601	8372	8482	11888	11747	3881	3917
C5.0	28764	28267	8706	8032	12017	11570	4058	3788
Tree AS	28127	27632	9341	8669	11997	11501	4127	3808
C&R	27428	27429	9544	9368	11783	11536	4092	4022

With this in mind, these measurements were calculated for all 5 models in each dataset, so that it is possible to compare their performances. The results are shown in the Tables 25 and 26:

Table 25 - Performance metrics - Electronics

Electronics	Accuracy		Precision		Recall/ sensitivity		F-measure		Specificity	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
C5.0	77.9%	76.6%	76.6%	78.3%	76.8%	76.7%	76.7%	77.5%	77.6%	76.5%
Tree AS	75.9%	75.8%	75.3%	78.0%	75.5%	75.3%	75.4%	76.6%	76.3%	76.5%
C&R	75.8%	75.8%	73.9%	76.8%	78.2%	77.3%	76.0%	77.0%	73.5%	74.1%
QUEST	75.2%	75.8%	76.9%	80.6%	70.3%	70.6%	73.5%	75.3%	79.8%	81.2%
CHAID	76.7%	75.2%	76.6%	77.8%	75.5%	73.8%	76.0%	75.8%	77.9%	76.7%

Table 26 - Performance metrics - Books

Books	Accuracy		Precision		Recall/ sensitivity		F-measure		Specificity	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Neural net	75.8%	75.6%	75.4%	75.2%	76.5%	76.7%	75.9%	76.0%	75.2%	74.4%
Random Trees	77.2%	75.2%	77.2%	75.4%	76.9%	75.2%	77.1%	75.3%	77.4%	75.2%
C5.0	77.3%	75.0%	76.8%	74.8%	78.2%	76.0%	77.5%	75.4%	76.5%	74.0%
Tree AS	75.6%	74.8%	75.1%	74.4%	76.4%	75.9%	75.7%	75.1%	74.7%	73.6%
C&R	74.4%	74.2%	74.2%	74.2%	74.5%	74.6%	74.4%	74.4%	74.2%	73.8%

In terms of accuracy, it is possible to see that all models achieved very close values. Apart from C&R for Books dataset, they all correctly classified over 75% of the data. One good indicator that is clear just by looking at these values is the fact that, for every model generated, the training and testing metrics were very similar, which means there was no overfitting.

For the Electronics dataset, the model that performed the best was C5.0. This model properly categorized 77.9% of the training data and 76.6% of the testing data. In this case, since the dataset was balanced, this was already a good and reliable metric on its own. However, in addition to this, the other reference measurements all reached values over 75%, both in training and testing subsets. The most important variables in this model were related to the length of the review, followed by the number of helpful votes and the average compound score of the sentences. The two length related variables were clearly more significant in determining the classification than all the others, as can be seen in Figure 14, summing up to 0.70 in a scale from 0 to 1, where 1 is the sum of all variables' importance.

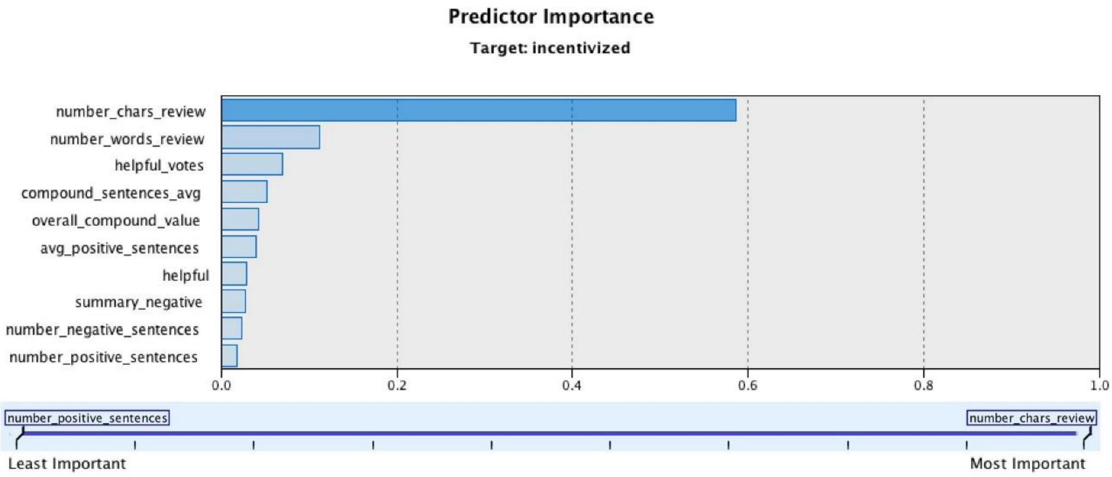


Figure 14 - Predictor variables - C5.0 model - Electronics

As for the Books dataset, the most accurate model for the testing data was a Neural Network. It had an 75.8% accuracy rate for training data and 75.6% for testing data. All the other reference measurements attained very similar values. The most significant variables for processing this model were the number of neutral sentences and the number of characters of the review, although in this model the case was not so disparate as in C5.0. In Figure 15 are presented the 10 most relevant variables for the classification.

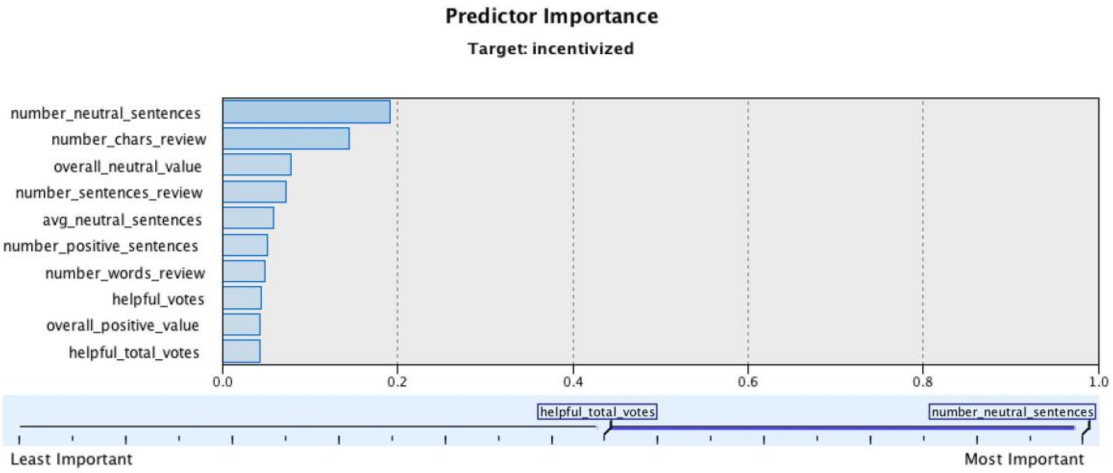


Figure 15 - Predictor variables - Neural network - Books

Another rather illustrative measurement to assess the performance of a model is the ROC curve (Receiver Operating Characteristic). This is a graphical demonstration of how well the model separates the positive cases from the negative ones and it is used to identify the best threshold for the parting. The metric used to evaluate this curve is AUC – area under the curve. An AUC value of 0.5 indicates no discriminative value (i.e. 50% sensitive and 50% specific) and is represented by a straight diagonal which is called the baseline. On the other hand, an AUC value of 1 indicates a perfect model. The further the ROC curve is from the baseline the better the model is. The C5.0 model for the Electronics dataset had an AUC value of 0.82 for both training and testing data, which represents a fair model. The graphs presented in Figure 16 show the ROC curves.

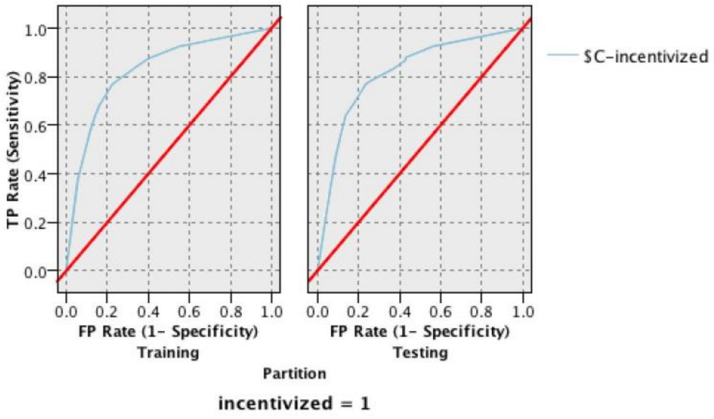


Figure 16 - C5.0 ROC curves

As for the Neural network model, the values of AUC were 0.83 for both training and testing data. In Figure 17, the ROC curves for this model are displayed.

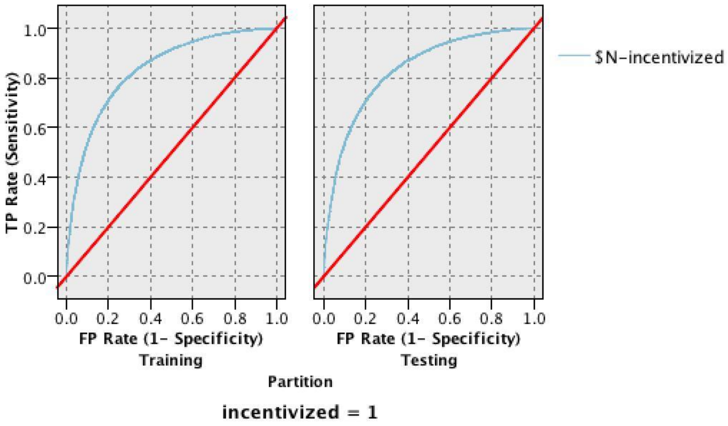


Figure 17 - Neural networks ROC curves

After analyzing the results of the models and considering every metric calculated and evaluated, it is fair to say that the models generated describe the existing data well enough to be considered reliable in the task of predicting future cases.

4.7. Discussion (Deployment)

At this final stage, the evaluation results and the knowledge acquired are put to use and applied to the business, whether it refers to a more practical approach like implementing a new model or a more theoretical one, like presenting a report with the findings and conclusions of the work. This report can be a summary of the project and its experiences or it may be a comprehensive presentation of the data mining results. In the present case, the purpose of this section is to take the conclusions of the modeling phase and understand the results obtained.

The reason why this work was applied to two different datasets was to see if there were significant differences in the final results. Although the considerable difference on the dataset size could be enough reason for the dissimilarity on the models' choice and accuracy, there are other possible causes. One of them may be related to the difference of the nature of the datasets itself. One dataset contains reviews about books and the other about electronic products. It is not hard to acknowledge that the target customers on these datasets is probably very different. That sole fact may very well be part of the reason why some variables are more important than others in the building process of a model. The category of a product may (and probably does) influence the attitude of a customer towards a product and its evaluation.

Taking this into account and considering the obtained results, the model that proved to have the best results for both these datasets was the C5.0. Although this model was not the best fit for the Books dataset, its results were very similar to the Neural Network and so it is a rather accurate model to describe the data in general, not considering the product category. As it can be seen in Figure 18, the three most important predictor variables in this model for the Books dataset are the number of characters, the overall compound value and the total number of helpful votes, which summed up to 0.80 (i.e., the three variables combined contribute to around 80% of the model).

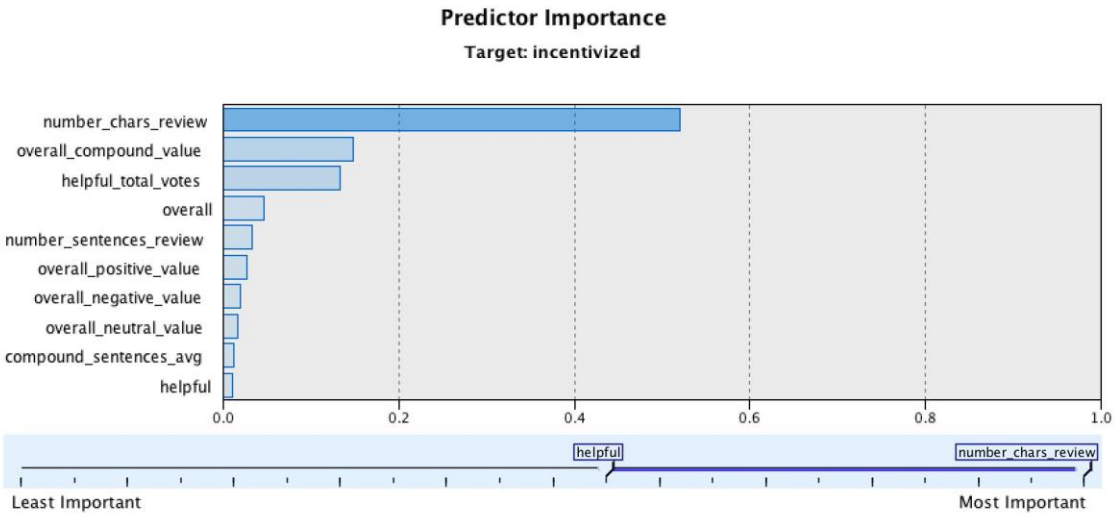


Figure 18 - Predictor variables importance - Books

In order to determine whether a review is incentivized, the first criterion analyzed was the length of the review. After that, based on the number of characters, the model would either check the amount of helpful votes and the helpfulness rate or the overall compound value and the overall score of the review. An example of a rule to classify a review as incentivized is shown in Figure 19: if the review has more than 778 characters, two or less helpful votes, and an overall score of 3.0 or higher, then there is a 79% probability of it being incentivized.

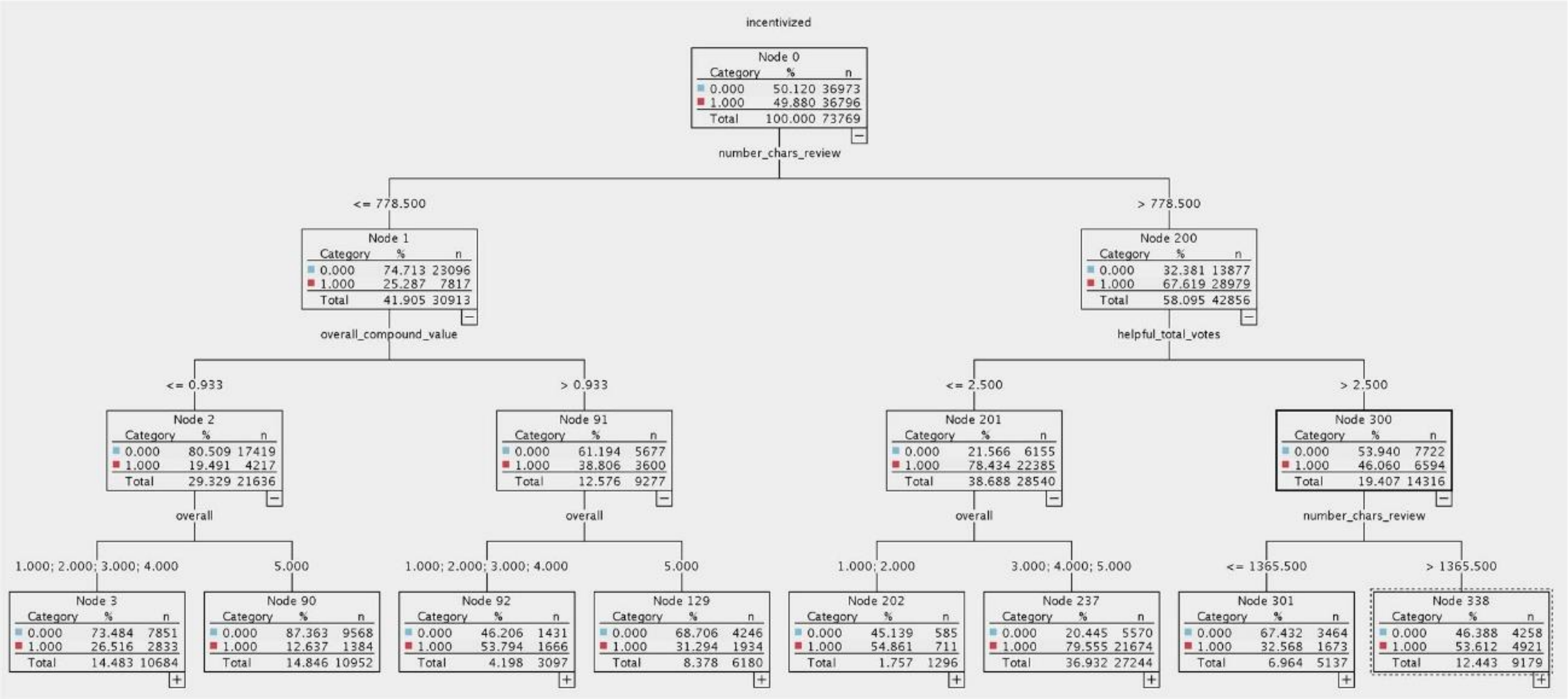


Figure 19 - C5.0 decision tree - Books

On the other hand, if a review has 778 characters or less and the overall compound value, calculated through VADER, is lower than 0.93 it is very likely that the review is not incentivized, regardless of the overall rating. However, if the overall compound value is higher than 0.93, then the model will check the overall rating of the review, to determine whether it is incentivized or not.

According to the Electronics decision tree, in order to determine whether a review is incentivized, the first criterion analyzed was also the length of the review. Then, depending on the number of characters, the model would check attributes like the amount of helpful votes, the helpfulness rate or the overall compound value. An example of a rule to classify an Electronics review as incentivized is: if the review has more than 1197 characters and less than 4 helpful votes, but the overall compound value is higher than 0.921, there is a probability of 86% of the review being incentivized. In this dataset, there is also a very determinant rule for non-incentivized reviews: if the review has less than 517 characters then there is an 86% probability of that review being categorized as non-incentivized regardless of any other variable. The C5.0 decision tree for this dataset is presented in Figure 20.

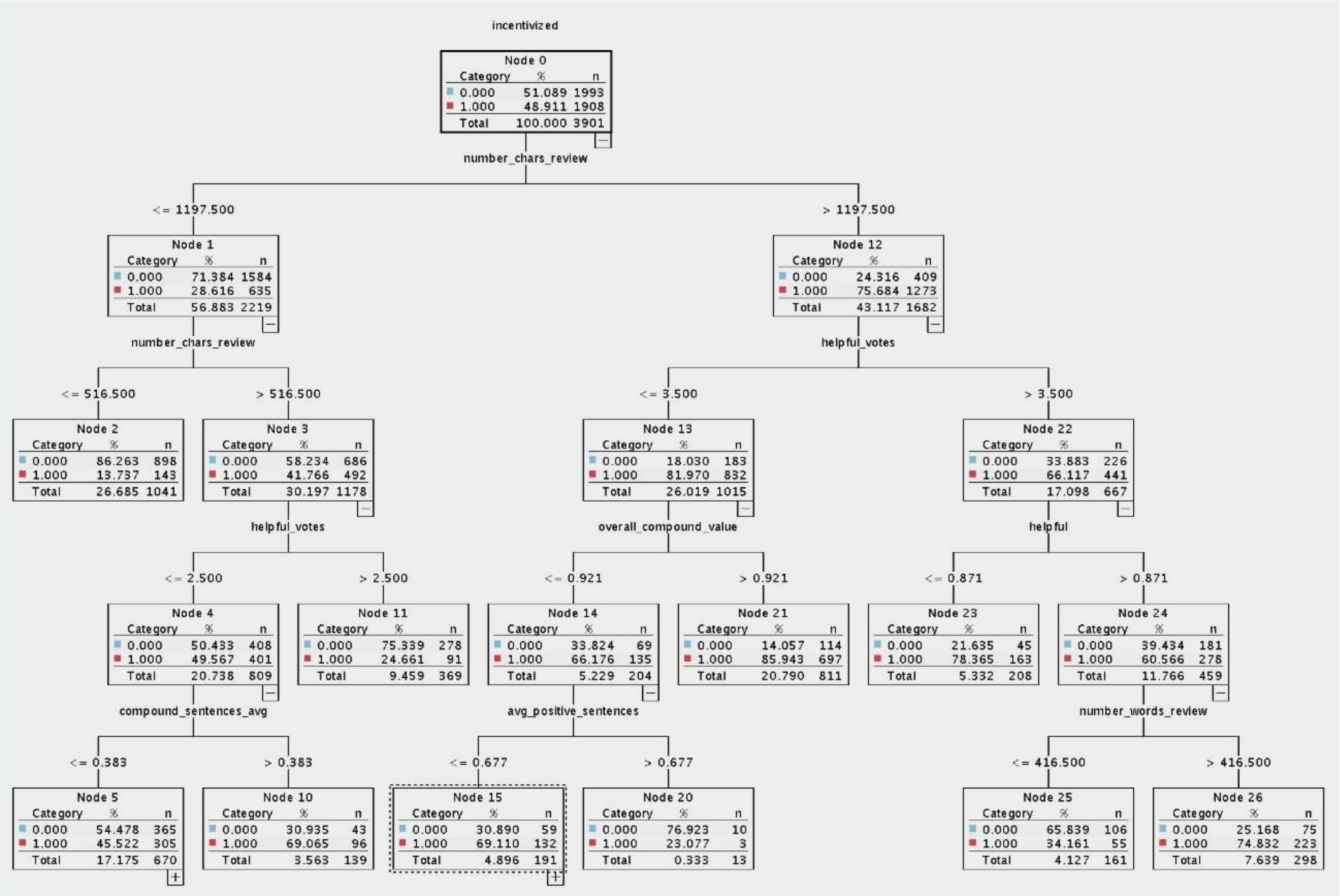


Figure 20 - C5.0 decision tree - Electronics

These models also support the hypothesis (1a), stating incentivized reviews are lengthier. This is evident just by looking at the decision trees, considering the number of characters is the most important variable in both. Although the boundary numbers are different for the two datasets, it is clear to see that the higher the number, the higher the probability of the review being categorized as incentivized. For books' reviews, this number is a little lower than for electronics', but still complies with what Moore (2012) stated about incentivized reviewers tending to be more explanatory and therefore more extensive in their reviews.

For Electronics reviews with a number of characters in a range from 516 to 1197, a higher number of helpful votes is a good predictor of non-incentivized reviews. But if there are less than 3 helpful votes, then the variable that determines whether or not the review is incentivized is the sentences compound average score, where a value higher than 3.8 means that the review is most likely incentivized.

In the Books dataset decision tree, there is a rule showing that a review with less than 778 characters, but with a high compound score (above 0.93) and a rating of 5.0 stars is probably non-incentivized. This is not an evident fact, since according to the literature, a high sentiment charge and rating almost certainly would reveal an incentivized review. So, in light of these results, this is probably where the real honest good reviews show in the decision tree.

Considering these results, it is possible to conclude that the most important variables in predicting bias in a review are mainly structured ones, like the length of the review or the rating. However, it was also shown that sentiment analysis can have an important part in adjusting the model's accuracy.

For future work, it would be interesting to see the model behavior with only sentiment metrics to test its accuracy and cross validate the most determinant variables at sentiment level.

5. Conclusions

When shopping online, product reviews are known to be an important decision factor to any potential customer. Since there is no interaction with the product itself, these reviews are strong influencers because they reflect true experiences reported by customers who acquired the item in question. Therefore, it is one of the key factors that sellers absolutely need to take into consideration. By processing this information on a regular basis, companies will be able to start acting more efficiently, which not only makes the customers happier, but it also prevents a waste of money in unfocused campaigns or improvements. As previously mentioned, some sellers on Amazon took advantage of this reviewing tool in order to favor their low rated, or not rated at all, products with the purpose of increasing its sales, by actually controlling the reviews (Butterworth, 2016).

The uniqueness of this study relies on the use of VADER algorithm to sentimentally classify a review and finding patterns in the behavior of incentivized reviewers, with the purpose of predicting bias in new-coming reviews, even if there is no disclaimer.

One of the limitations of this study is the assumption that the disclaimers used to identify incentivized and non-incentivized reviews are enough to perform a reliable extraction. It is possible that some of the observations on the datasets were misclassified on the extraction process, but considering the size of the dataset and the consequent impossibility of manually checking every review, this was considered an acceptable limitation. The other limitation is related to the fact that only reviews from two categories of products were used. The extent to which this may affect the data is unknown since it was not studied.

The main contributions of this work were the insights taken from the analysis of the hypothesis and the models generated. According to the analysis performed, it was possible to conclude that incentivized reviewers do write lengthier and more sentiment charged reviews. The helpfulness rate was also proved to be positively correlated with the overall polarity score of a review. Although it is not a strong correlation, it was shown to be a significant one.

The type of model that described the datasets better was decision trees, since in the top accurate models for both datasets, four out of five were decision trees. These are very simple and easy to read models and the information about the importance of each variable is very straightforward. Unlike neural networks, this process is easy to understand and interpret.

The models generated were able to correctly predict bias in a review over 75% of the times, based on some characteristics like the length of a review, the helpfulness rate and the polarity scores calculated through VADER. The most important variable, in both cases, was the number of characters of the review, which was related to one of the hypothesis intended to test. The most important sentiment-related variable was the overall compound score, which was higher on the incentivized reviews. Looking at the decision trees, it is simple to infer several decision rules to classify new incoming reviews.

With the conclusion of this work, a model was created to identify a possibly incentivized review even if it does not have a disclaimer, since that practice is now banned by Amazon.

6. Bibliography

Aggarwal, C. C., and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Aghaei, S., Nematbakhsh, M. A., and Farsani, H. K. (2012). *Evolution of the World Wide Web: From WEB 1.0 TO WEB 4.0*. International Journal of Web & Semantic Technology, 3(1), 1.

Akhlaq, A., and Ahmed, E. (2014). *Online shopping A Global Perspective*. Journal of Basic and Applied Scientific.

Akkaya, C., Wiebe, J., and Mihalcea, R. (2009). *Subjectivity word sense disambiguation*. In Proc. EMNLP-09.

Amazon.com (2016a), Update on Customer Reviews, https://www.amazon.com/p/feature/abpto3jt7fhb5oc

Amazon.com Community Guidelines (2016b) retrieved from https://www.amazon.com/gp/help/customer/display.html?nodeId=14279631

Amazon.com (2016c), Vine Program, retrieved from https://www.amazon.com/gp/vine/help

Ananiadou, S. (2013). *Term Normalization, Text Mining*. In Encyclopedia of Systems Biology (pp. 2155-2155). Springer New York.

Aranha, C., and Passos, E. (2008). *Automatic NLP for Competitive Intelligence*. Emerging Technologies of Text Mining: Techniques and Applications (pp. 54–76).

Araujo, M., Reis, J., Pereira, A., and Benevenuto, F. (2016, April). *An evaluation of machine translation for multilingual sentence-level sentiment analysis*. In Proceedings of the 31st Annual ACM Symposium on Applied Computing (pp. 1140-1145). ACM.

Berthon, P. R., Pitt, L. F., Plangger, K., and Shapiro, D. (2012). *Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy*. Business horizons, 55(3), 261-271.

Blake, C. (2011). *Text Mining*. Annual review of information science and technology, 45(1), 121–155.

Blazevic, V., Hammedi, W., Garnefeld, I., Rust, R. T., Keiningham, T., Andreassen, T., Carl, W. (2013). *Beyond traditional word-of-mouth: an expanded model of customer-driven influence*. Journal of Service Management, 24(3), 294-313.

Blei, D. M. (2012). *Probabilistic topic models*. Communications of the ACM, 55(4), 77-84.

Blei, D., Carin, L., and Dunson, D. (2010). *Probabilistic topic models*. IEEE signal processing magazine, 27(6), 55-65.

Bollen, J., Pepe, A., and Mao, H. (2009). *Modeling public mood and emotion: Twitter sentiment and socio economic phenomena*. arXiv:0911.1583 [cs].

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). *Cross-linguistic sentiment analysis: From english to spanish*. International Conference RANLP.

Bulmer, D., and DiMauro, V. (2009). *The new symbiosis of professional networks: Social media's impact on business and decision-making*. Society for New Communications Research.

Cao, Q., Duan, W., and Gan, Q. (2011). *Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach*. Decision Support Systems, 50(2), 511-521.

Calheiros, A. C., Moro, S., and Rita, P. (2017). *Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling*. Journal of Hospitality Marketing & Management, 26(7), 675-693.

Cambria, E., Fu, J., Bisio, F., and Poria, S. (2015, January). *AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis*. In AAAI (pp. 508-514).

Campbell, N. (2012), retrieved from Quora (2012), https://www.quora.com/What-percentage-of-buyers-write-reviews-on-Amazon

Chaney, A., and Blei, D. (2012). *Visualizing Topic Models*. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (pp. 419–422).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chawla, N. V. (2009). *Data Mining for Imbalanced Datasets: An Overview*. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 875-886). Springer US.

Cheng, T., Chakrabarti, K., Chaudhuri, S., Narasayya, V., and Syamala, M. (2013, April). *Data services for E-tailers leveraging web search engine assets*. In Data Engineering (ICDE), 2013 IEEE 29th International Conference on (pp. 1153-1164). IEEE.

Church, K. W., and Rau, L. F. (1995). *Commercial applications of natural language processing*. Communications of the ACM, 38(11), 71-79.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Constantinides, E., and Fountain, S. J. (2008). *Web 2.0: Conceptual foundations and marketing issues*. Journal of direct, data and digital marketing practice, 9(3), 231-244.

Cox, DR (1958). *The regression analysis of binary sequences (with discussion).* J Roy Stat Soc B. 20: 215–242. JSTOR 2983890.

Cristea, D. (2016). *Natural Language Processing versus Logic. Pros and cons on the dispute whether logic is useful in the computational interpretation of language*. Computer Science Journal of Moldova, 24(3).

Das, S. R., and Chen, M. Y. (2007). *Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web*. Management Science, 53(9), 1375–1388.

Davis, J., and Goadrich, M. (2006, June). *The relationship between Precision-Recall and ROC curves*. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

Delen, D., and Crossland, M. D. (2008). *Seeding the survey and analysis of research literature with text mining*. Expert Systems with Applications, 34(3), 1707–1720.

Dellarocas, C. (2003). *The digitization of word of mouth: Promise and challenges of online feedback mechanisms*. Management science, 49(10), 1407-1424.

Dennison, G., Bourdage-Braun, S., and Chetuparambil, M. (2009). *Social commerce defined*. White paper, 23747.

Duan, W., Gu, B., and Whinston, A. B. (2008). *Do online reviews matter? —An empirical investigation of panel data*. Decision support systems, 45(4), 1007-1016.

Ehsani, F., and Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm.

Feinerer, I., Hornik, K., and Meyer, D. (2008). *Text Mining Infrastructure* in R. Journal of Statistical Software, 25(5), 1–54.

Feldman, R., and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*.

Feldman, S. (1999). *NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval*. ONLINE-WESTON THEN WILTON-, 23, 62-73.

Fournier, S., and Avery, J. (2011). *The uninvited brand. Business horizons*, 54(3), 193-207.

Frawley, W.J. and Piatetsky-Shapiro, G. eds. (1991). *Knowledge discovery in databases* (Vol. 37). Menlo Park, CA: AAAI Press.

Freeman, L. (2010). Social shopping study reveals changes in consumers’ online shopping habits and usage of customer reviews.

Garland, E. (2009). *Social Media and Authority-The Intelligence Collaborative*. Washington DC, October, 22.

Gefen, D. (2002). *Reflections on the dimensions of trust and trustworthiness among online consumers*. ACM Sigmis Database, 33(3), 38-53.

Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yoogatama, D., Fanigan, J., Smith, N. A. (2011). *Part-of-speech tagging for twitter: Annotation, features, and experiments.* Em Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 42–47). Association for Computational Linguistics.

Grimes, S. (2013). *Unstructured Data and the 80 Percent Rule*. Breakthrough Analysis. Retrieved 2015-02-23.

Guerreiro, J., and Moro, S. (2017). A*re Yelp's tips helpful in building influential consumers?*. Tourism Management Perspectives, 24, 151-154.

Guerreiro, J., Rita, P., and Trigueiros, D. (2016). *A text mining-based review of cause-related marketing literature*. Journal of Business Ethics, 139(1), 111-128.

Hassler, M., and Fliedl, G. (2006). *Text preparation through extended tokenization*. Data Mining VII: Data, Text and Web Mining and their Business Applications, 37, 13–21.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Overview of supervised learning*. In The elements of statistical learning (pp. 9-41). Springer New York.

Hatzivassiloglou, V., and Wiebe, J. (2000). *Effects of adjective orientation and gradability on sentence subjectivity*. Proceedings of the 18th conference on Computational Linguistics-Volume 1 (pp. 299–305). Association for Computational Linguistics.

He, R., and McAuley, J. (2016, April). *Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering*. In Proceedings of the 25th International Conference on World Wide Web (pp. 507-517). International World Wide Web Conferences Steering Committee.

Hinton, G., Vinyals, O., and Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv preprint arXiv:1503.02531.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). *A Brief Survey of Text Mining.* LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20, 19–62.

Holleschovsky, N. I., and Constantinides, E. (2016). Impact of online product reviews on purchasing decisions.

Hu, N., Liu, L., and Sambamurthy, V. (2011). *Fraud detection in online consumer reviews*. Decision Support Systems, 50(3), 614-626.

Hu, N., Liu, L., and Zhang, J. J. (2008). *Do online reviews affect product sales? The role of reviewer characteristics and temporal effects*. Information Technology and Management, 9(3), 201-214

Hu, X., and Liu, H. (2012). *Text analytics in social media*. In mining text data (pp. 385-414). Springer US.

Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A. (2014). *Interactive topic modeling*. Machine learning, 95(3), 423-469.

Huang, Z., and Benyoucef, M. (2013). *From e-commerce to social commerce: A close look at design features*. Electronic Commerce Research and Applications, 12(4), 246-259.

Hutto, C.J. and Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Jackson, P., and Moulinier, I. (2007). *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization*. (J. B. Publishing, Ed.) (2nd ed.).

Kaplan, R. M. (2005). *A Method for Tokenizing Text*. Complexity and Education: Inquiries into Words, Constraints and Contexts (pp. 55–64).

Kass, Gordon V.; *An Exploratory Technique for Investigating Large Quantities of Categorical Data*, Applied Statistics, Vol. 29, No. 2 (1980), pp. 119–127

Katariya, M. N. P., Chaudhari, M. S., Subhani, B., Laxminarayana, G., Matey, K., Nikose, M. A., ... and Deshpande, S. P. (2015). *Text preprocessing for text mining using side information*. International Journal of Computer Science and Mobile Applications, 3(1), 01-05.

Kim, D. J., Ferrin, D. L., and Rao, H. R. (2008). *A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents*. Decision support systems, 44(2), 544-564.

Kim, J., and Gupta, P. (2012). *Emotional expressions in online user reviews: How they influence consumers' product evaluations*. Journal of Business Research, 65(7), 985-992.

Kim, J., Naylor, G., Sivadas, E., and Sugumaran, V. (2015). *The unrealized value of incentivized eWOM recommendations*. Marketing Letters, 1-11.

Kim, Y., and Srivastava, J. (2007, August). *Impact of social influence in e-commerce decision making*. In Proceedings of the ninth international conference on Electronic commerce (pp. 293-302). ACM.

King, R. A., Racherla, P., and Bush, V. D. (2014). *What we know and don't know about online word-of-mouth: A review and synthesis of the literature*. Journal of Interactive Marketing, 28(3), 167-183.

Lallana, E., Quimbo, R., and Andam, Z. R. (2000). *An Introduction to eCommerce: definition adapted and expanded*. Philippines: DAI-AGILE, 17.

Larose, D.T. (2005). *Discovering knowledge in data: an introduction to data mining*. John Wiley, NY.

Lee, J., Park, D. H., and Han, I. (2008). *The effect of negative online consumer reviews on product attitude: An information processing view*. Electronic commerce research and applications, 7(3), 341-352.

Liang, T. P., Ho, Y. T., Li, Y. W., and Turban, E. (2011). *What drives social commerce: The role of social support and relationship quality*. International Journal of Electronic Commerce, 16(2), 69-90.

Lin, C., He, Y., Everson, R., and Rüger, S. (2012). *Weakly supervised joint sentiment-topic detection from tex*t. IEEE Transactions on Knowledge and Data Engineering, 24(6), 1134–1145.

Liu, B. (2008). *Web Data mining: Exploring Hyperlinks, Contents, and Usage Data*. (2nd ed.). Springer Berlin Heidelberg New York.

Liu, B. (2010). *Sentiment Analysis and Subjectivity*. In N. Indurkhya and F. Damerau (Eds.), Handbook of Natural Language Processing (2nd ed.). Boca Raton, FL: Chapman & Hall.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, B., and Zhang, L. (2012). *A survey of opinion mining and sentiment analysis*. In Mining text data (pp. 415-463). Springer US.

Loh, W.-Y. and Shih, Y.-S. (1997). *Split selection methods for classification trees*. Statistica Sinica, 7:815–840.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). *Promotional reviews: An empirical investigation of online review manipulation*. The American Economic Review, 104(8), 2421-2455.

McAuley, J., Pandey, R., and Leskovec, J. (2015, August). *Inferring networks of substitutable and complementary products*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

McCulloch, Warren; Walter Pitts (1943). *A Logical Calculus of Ideas Immanent in Nervous Activity*. Bulletin of Mathematical Biophysics. 5 (4): 115–133. doi:10.1007/BF02478259.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (1st ed.). Academic Press.

Moore, S. G. (2012). *Some things are better left unsaid: how word of mouth influences the storyteller*. Journal of Consumer Research, 38(6), 1140-1154.

Morente-Molinera, J. A., Pérez, I. J., Chiclana, F., and Herrera-Viedma, E. (2015, October). *A novel group decision making method to overcome the Web 2.0 challenges*. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on (pp. 2233-2238). IEEE.

Moro, S., Rita, P., and Coelho, J. (2017). *Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip*. Tourism Management Perspectives, 23, 41-52.

Mostafa, M. M. (2013). *More than words: Social networks' text mining for consumer brand sentiments*. Expert Systems with Applications, 40(10), 4241-4251.

Mudambi, S. M., and Schuff, D. (2010). *What makes a helpful review? A study of customer reviews on Amazon.com*. MIS quarterly, 34(1), 185-200.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). *Text mining for market prediction: A systematic review*. Expert Systems with Applications, 41(16), 7653-7670.

Nielsen, J. (2010). *Global trends in online shopping*. A Nielsen Global Consumer Report, 1-10.

O'Reilly, T., and Battelle, J. (2009). *Web squared: Web 2.0 five years on*. O'Reilly Media, Inc.

Paltoglou, G., and Thelwall, M. (2012). *Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media*. ACM Trans. Intell. Syst. Technol., 3(4), 66:1-66:19.

Pang, B., and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Found. Trends Inf. Retr., 2(1-2), 1–135. doi:10.1561/1500000011.

Parise, S., and Guinan, P. J. (2008, January). *Marketing using web 2.0*. In Hawaii, International Conference on System Sciences, Proceedings of the 41st Annual (pp. 281-281). IEEE.

Porter, M. (1980). *An algorithm for suffix stripping*. Program: electronic library and information systems, 14(3), 130–137.

Power, D. J., and Phillips-Wren, G. (2011). *Impact of social media and Web 2.0 on decision-making*. Journal of decision systems, 20(3), 249-261.

Pradhan, B. (2013). *A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS*. Computers & Geosciences, 51, 350-365.

Rehling, J. (2011). How Natural Language Processing Helps Uncover Social Media Sentiment.

ReviewMeta (2016), retrieved from https://reviewmeta.com/blog/analysis-of-7-million-amazon-reviews-customers-who-receive-free-or-discounted-item-much-more-likely-to-write-positive-review/

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). *Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods*. EPJ Data Science, 5(1), 1-29.

Rose, K. (1998). *Deterministic annealing for clustering, compression, classification, regression, and related optimization problems*. Proceedings of the IEEE, 86(11), 2210-2239.

Sánchez, D., Martín-Bautista, M. J., Blanco, I., and Torre, C. J. D. La. (2008). *Text Knowledge Mining: An Alternative to Text Data Mining*. 2008 IEEE International Conference on Data Mining Workshops (pp. 664–672). Ieee.

Shafique, U., and Qaiser, H. (2014). *A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)*. Int. J. Innov. Sci. Res, 12(1), 217-222.

Sharda, R., Delen, D., and Turban, E. (2014). *Business Intelligence: A Managerial Perspective on Analytics*. Prentice Hall Press.

Tan, A. H. (1999, April). *Text mining: The state of the art and the challenges*. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70).

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) *Hierarchical Dirichlet processes*. Journal of the American Statistical Association 101(476):1566–1581.

Utz, S., Kerkhof, P., and van den Bos, J. (2012). *Consumers rule: How consumer reviews influence perceived trustworthiness of online stores*. Electronic Commerce Research and Applications, 11(1), 49-58.

Uysal, A. K., and Gunal, S. (2014). *The impact of preprocessing on text classification*. Information Processing and Management, 50, 104–112.

Webster, J., and Kit, C. (1992). *Tokenization as the initial phase in NLP*. Proceedings of the 14th conference on Computational linguistics-Volume 4 (pp. 6–10).

Wiebe, J., Bruce, R., Bell, M., Martin, M., and Wilson, T. (2001). *A corpus study of evaluative and speculative language*. Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16 (Vol. Proceeding, pp. 1–10). Association for Computational Linguistics.

Wigand, R. T., Benjamin, R. I., and Birkland, J. L. (2008, August). *Web 2.0 and beyond: implications for electronic commerce*. In Proceedings of the 10th international conference on Electronic commerce (p. 7). ACM.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005) *Recognizing contextual polarity in phrase-level sentiment analysis*. In: The conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada. Association for Computational Linguistics, 347-354.

Wilson, T., Wiebe, J., and Hwa, R. (2004). *Just how mad are you? finding strong and weak opinion clauses*. In Proc. NCAI-04s.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L. P. (2013). *Youtube movie reviews: Sentiment analysis in an audio-visual context*. IEEE Intelligent Systems, 28(3), 46-53.

Yap, B.W., Rani, K. A., Rahman, H. A. A., Fong, S. Khairudin, Z., and Abdullah, N. N. (2014). *Na Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets*. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013) (pp. 13-22). Springer.

7. Appendixes

7.1. Appendix A

List of diclaimers

```
disclaimer = ["disclaimer",
              "discount for review",
              "discount to review",
              "for the purpose of a review",
              "free for my review",
              "free for review",
              "free reviewer's sample",
              "free sample",
              "free to review",
              "freebie",
              "in exchange for a review",
              "in exchange for my honest",
              "in exchange of a review",
              "in return for a review",
              "in return of a review",
              "product for review",
              "product for test",
              "review for product",
              "review sample",
              "review unit",
              "reviewing purposes",
              "sample for an honest review",
              "sample for review",
              "sent this for review",
              "testing and review purposes",
              "product sent for review"]
```