Università degli studi di Cagliari

PH.D. DEGREE
Business and Economics – Quantitative field
Cycle XXXIII

General Sentiment Decomposition: opinion mining based on raw Natural Language text

Thesis in Statistics
(SECS-S/01)

Ph.D. Student: Maurizio Romano
Supervisor: Prof. Claudio Conversano

Final exam. Academic Year 2019 – 2020
Thesis defence: April 2021 Session

"Disce, sed a doctis, indoctos ipse doceto" (Cato)

To my mother, to my sister, to my brother, to my beloved relatives, to my very best friends and to my research group.

A mia madre, mia sorella, mio fratello, ad i familiari a me più a cuore, ai miei migliori amici ed al gruppo di ricerca tutto.
Con questa breve dedica voglio ringraziare tutti quelli che mi hanno supportato in questo lungo percorso, che hanno creduto in me dall'inizio alla fine e che mi hanno sopportato sia nei momenti più bui e difficili, sia quanto –immerso nel mio mondo– incominciavo a raccontare le cose più disparate su cui stavo volgendo la mia ricerca.

Grazie in particolar modo a chi mi ha fatto conoscere ed appassionare a questo mondo, un mondo a cui guardavo con retticenza giacchè mai avrei voluto continuare a studiare e benchèmeno statistica! Ho scoperto negli anni quanto invece fosse faticoso, ma assolutamente emozionante ed appagante poter fare nuove scoperte anche solo partendo da piccole ed effimere cose.

Grazie, grazie davvero a tutti quelli che hanno creduto in me investendo quanto di più prezioso ci sia a questo mondo: Il tempo!

Acknowledgements

Maurizio Romano gratefully acknowledges Sardinian Regional Government for the financial support of his Ph.D. scholarship (P.O.R. Sardegna F.S.E. – Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014-2020 – Axis III Education and training, Thematic goal 10, Investment Priority 10ii), Specific goal 10.5.

Maurizio Romano sincerely acknowledges his research group for being professional, kind, and always available to support him over the years. In particular, he wants to be express his gratitude to Prof. F. Mola, to Prof. C. Conversano, to Dr. L. Frigau, and Dr. G. Contu for introducing him to the research, showing a new side of the world and a shining reality in which he loves to immerse fully.

Abstract

The importance of person-to-person communication about a certain topic (Word of Mouth) is growing day by day, especially for decision-makers. These phenomena can be directly observed in online social networks. For example, the rise of influencers and social media managers. If more people talk about a specific product, then more people are encouraged to buy it and vice versa. Forby, those people usually leave a review for it. Such a review will directly impact the product, and this effect is amplified proportionally to how much the reviewer is considered to be trustworthy by the potential new customer. Furthermore, considering the negative reporting bias, it is easy to understand how customer satisfaction is of absolute interest for a company (as well as citizens' trust is for a politician).

Textual data have then proved extremely useful, but they are complex, as the language is. For that, many approaches focus more on producing well-performing classifiers and ignore the highly complex interpretability of their models. Instead, we propose a framework able to produce a good sentiment classifier with a particular focus on the model interpretability. After analyzing the impact of Word of Mouth on earnings and the related psychological aspects, we propose an algorithm to extract the sentiment from a Natural Language text corpus. The combined approach of Neural Networks, characterized by high predictive power but at the cost of complex interpretation (usually considered as black-boxes), with more straightforward and informative models, allows not only to predict how much a sentence is positive (negative) but also to quantify a sentiment with a numeric value. In fact, the General Sentiment Decomposition (GSD) framework that we propose is based on a combination of Threshold-based Naïve Bayes (an improved version of the original algorithm), SentiWordNet (an enriched Lexical Database for Sentiment Analysis tasks), and the Words Embeddings features (a high dimensional representation of words) that directly comes from the usage of Neural Networks.

Moreover, using the GSD framework, we assess an objective sentiment scoring that improves the results' interpretation in many fields. For example, it is possible to identify specific critical sectors that require intervention to improve the offered services, find the company's strengths (useful for advertising campaigns), and, if time information is present, analyze trends on macro/micro topics.

Besides, we have to consider that NL text data can be associated (or not) with a sentiment label, for example: "positive" or "negative". To support further decision-making, we apply the proposed method to labeled (Booking.com, TripAdvisor.com) and unlabelled (Twitter.com) data, analyzing the sentiment of people who discuss a particular issue. In this way, we identify the aspects perceived as critical by the people concerning the "feedback" they publish on the web and quantify how happy (or not) they are about a specific problem. In particular, for Booking.com and TripAdvisor.com, we focus on customer satisfaction, whilst for Twitter.com, the main topic is climate change.

Contents

List of Figures

Chapter 1

Introduction

"Scientia potestas est"" (Sir Francis Bacon (1597)), since long time people recognize that "knowledge is power". Forby, we constantly try to increase our knowledge. Furthermore, collecting an expert's opinion before taking an important decision might make a huge difference between failure and success on our tasks.

Thanks to the technology improvement, our experience-sharing (opinion-sharing) capabilities are far beyond our ancients' expectations. Considering this immense amount of valuable data produced day by day, this thesis will start with a detailed explanation of the words' power (Word-of-Mouth, Chapter 2). Moreover, after an overview of the main techniques that are available for analyzing those types of data (Machine Learning, Chapter 3), we will show how this opinion (general textual data) can be used to produce new and valuable information (Natural Language Processing, Chapter 4). Furthermore, considering the specific task of identifying if someone is talking positively (or not) of a particular topic (Sentiment Analysis), we propose our approach (General Sentiment Decomposition, Chapter 5) able to identify the "sentiment" event without providing examples of concepts that are "positive" ("negative") and, for that, "General". We will then show applications in different cases based on real data and exciting results (Chapter 6). In particular, with this approach, we propose a framework able to "decompose the sentiment": transforming a sentiment into a score and then specifically highlighting (and quantifying) which are the positive (negative) aspects of the considered topics. For that, we called it "General Sentiment Decomposition".

Despite the encouraging results, still many things can be done to improve the framework. In the last section (Conclusions, Chapter 7), we describe with more detail some possible strategies that might be adopted for increasing the performances of the proposed classifier and for adding further ex-post analysis.

Chapter 2

Word Of Mouth & Marketing

Considering that the central framework of this Ph.D. is Statistics, but it is enrolled in the Business and Economics field, with this chapter, we are offering a "contact point" between those two frameworks. Furthermore, those concepts represent the ground fields of the thesis and, for that, this chapter is the answer to the following questions:

- How can this thesis be placed in the context of a Ph.D. in Business and Economics? (Section 2.1)
- Which type of data are we going to use? (Section 2.2)
- Why is it so important to analyze it? (Section 2.3)

"People enjoy stories because stories satisfy people's needs" (McKee 2003; Yuan et al. (2020)). Following what has been said in the introduction, probably no other quote can describe so well the concept of how important it is to share the experience. For that, after an overview of the motivations that place this thesis in the context of a Ph.D. in Business and Economics and a description of the Word of Mouth, with this chapter, we investigate the important relationship that exists between marketing and the related psychological aspects.

2.1 Motivations

Word Of Mouth (WoM) is defined by Arndt (1967b) as "oral person-to-person communication concerning a brand, product, or service between a receiver and a communicator whom the receiver perceives as non-commercial". In more recent research, WOM definition was extended to include "the exchange of ephemeral oral or spoken messages between a contiguous source and a recipient who communicates directly in real-life" (Yuan et al., 2020). Customers, in fact, exchange their experience regarding a service quality or product performance (Buttle, 1998).

In other words, WOM is a product of perceived service quality and value (Hartline and

Jones (1996); Yuan et al. (2020)). Although WOM has many benefits, it also presents several problems. Harrison-Walker (2001) attempted to conceptualize WOM by using a set of measures, but the results were criticized for being overly superficial (Mazzarol Tim et al., 2007).

According to Sırma (2009), nowadays, the proliferation of online customers' reviews (so, an electronic-WOM) has been reported to being as one of the most important information sources in the industry. However, sources of eWOM are less credible and transparent than those of traditional WOM. For example, we could consider a hotel business in a tourism-related area. In fact, it impacts the consumers' decision-making process while influencing the hotel's performance; for that, it has gained considerable attention (Schuckert et al. (2015)). As reported by Nielsen (2007), instead of trust messages received through mass-market communications or general advertisements, 78% of consumers find that recommendations from other consumers are more reliable. In fact, while only 14% of people believe in what they perceive from advertisements, 90% find trustworthy a reported experience from their friends and acquaintances (Rusticus (2007)). Definitely, this happens because consumers find it challenging to evaluate a product or a service before actually using it or trying it by themselves (O'Connor (2010); Yang et al. (2016)). Compared to traditional marketing channels, this makes Word Of Mouth marketing more robust whilst trustworthy.

According to Phillips et al. (2015), User Generated Content (UGC), and in particular, online reviews, have seen rapid growth in recent years. Furthermore, online reviews have transformed consumer behavior in information searching and sharing, looking for other consumers' feedbacks (Jun et al. (2010)). Forby, user-generated reviews should be used to improve the quality of the products and the services (Phillips et al. (2015)), cause they have been found to help within the identification of customer needs, and the implementation of new marketing strategies (Yacouel and Fleischer (2012); Loureiro and Kastenholz (2011)). Reviews significantly influence customer behavior, particularly in a service context (Mudambi and Schuff, 2010), where the quality of the services cannot be assessed before consumption (Murray, 1991). In such cases, consumers rely even more on the experiences of other customers (Reimer and Benkenstein, 2016).

Moreover, as reported by Phillips et al. (2015) and O'Connor (2008), in February 2014 TripAdvisor.com issued a press release announcing a UGC milestone, making it the first travel site to offer consumers 150 millions of reviews and opinions. The increase in online reviews echoes was found to have a similar pattern to the growth of hotel room e-bookings (Schegg and Scaglione (2013); Toh et al. (2011)). Considering this tremendous amount of reviews that are still growing day by day and the many studies that have been made to

analyze them, it is still possible to provide broader insights of eWOM as a determinant for many purposes (Cantallops and Salvi (2014)).

Furthermore, marketers have increasingly turned towards Word Of Mouth marketing by enlisting consumers to talk about brands, products, and services within their social networks (Carl and Oles, 2007), demonstrating how analyzing the feedback from customers has become crucial to predict their prospect behavior. In particular, hospitality and tourism fields have been widely studied to perceive the most influential features w.t.r. to a particular service or hotel (Aakash and Gupta Aggarwal, 2020).

It is noticeable how, according to Yang et al. (2018), to reduce decision-making costs, people tend to rely more on review ratings than textual comments. This means that consumers who are reading a vast amount of reviews are likely to focus on the reviews' score or volume, which are considered to be proxies for underlying product quality (Chaiken and Maheswaran (1994)) and hotel reputation (Anderson and Lawrence (2014)). Other studies have, in fact, demonstrated that this information has an influence over the tourist's decision-making process and within the organizations' pricing strategies and performance (Liu and Park (2015); Xie et al. (2016)).

Furthermore, as showed in Jun et al. (2010), reviews are a resource that contribute to the effective management of the entire tourism sector and to the achievement of competitive businesses' advantage (H. et al. (2009); Leung et al. (2013); Lu and Stepchenkova (2012); Robson et al. (2013)).

Moreover, Thorne (2016) reported that Word Of Mouth marketing is not only about stimulating customers to talk about a product, but it also involves learning how to make Word Of Mouth fit a marketing objective. In fact, consumers share their experiences and their Word Of Mouth will drive more sales than advertisements. Sernovitz and Kawasaki (2006), and Sırma (2009) demonstrated that it is possible to use it against the negative Word Of Mouth, channeling it in the desired direction improving customers satisfaction.

Considering those marketing-related aspects that involve WoM, some companies try to use money rewards to incentivize customers to spread a positive WoM. Such interference in the customer-to-customer interaction process by the company causes a loss of credibility (Martin, (2014); Reimer and Benkenstein (2016)) on those rewarded online reviews. The suspect of the review author's potential personal interest skeptic the review reader, leading to a decrease in trust in both the source (Godes et al., 2005) and the company itself that provide the rewards (Campbell and Kirmani, 2008). This phenomenon well represents some of the psychological impacts that we have on the customers, and that we analyze more in detail in Section 2.3.

2.2 From Word-of-Mouth (WoM) to electronic-WoM (eWOM)

As suggested by Kirby (2006), "Word Of Mouth" is defined as "oral communication", "oral publicity" and "speaking" (Oxford English Dictionary, 1998).

Other studies (Halstead, 2002), have defined Word Of Mouth as "talking of a product" (Blankertz and Cox, 1969), "knowing a certain service or product from friends" (Sırma, 2009), "interpersonal information exchange about a product, service or retailer" (Higie et al., 1987), "interpersonal communications in which none of participants are marketing sources" (Bone, 1995),"the act of telling at least one friend, acquaintance, or family member about a satisfactory or unsatisfactory product experience" (Halstead, 2002) and "oral, person-to-person communication between a perceived non-commercial communicator and a receiver concerning a brand, product, or service offered for sale" (Arndt, 1967a). WOM, which is different from communication initiated by merchants and advertisers, is a form of "informal communication directed at other consumers about the ownership, usage or characteristics of particular goods and services and their sellers" (Westbrook, 1987). It is often considered to be one of the oldest and most powerful forms of marketing that has always existed. For example, we can even consider "a human pointing to a cave painting to share the location of a good hunting ground with his family" as the start of WoM (O'Leary and Sheehan, 2008). That explains why WOM has been widely examined in the marketing field, even though the WOM offered by the message sender, may not necessarily be positive (Arndt, 1967a). Even with that, it provides a special way for understanding consumers' attitudes toward a certain brand.

However, Moore and Lafreniere (2020), as shown in Fig. 2.1, highlights that WOM is more than a simple sender-to-receiver exchange as we were thinking until now. In fact, we can observe three distinct subjects that can influence both the sender and the receiver and, for that, define how the online WOM impacts the receivers:

- Platforms, that can manipulate the presentation of the reviews, showing some statistics that summarize the information (e.g., number of users who liked that review), and request specific information from senders that will be displayed to the customers;
- Sellers, that respond to messages or that try to interfere in the process, rewarding customers that do positive reviews;
- Other customers, that comment and rate other reviews according to their experience

With advances in information technology, WOM has taken on an electronic form (Electronic Word Of Mouth, eWOM) and has an enhanced effect on the business as eWOM can

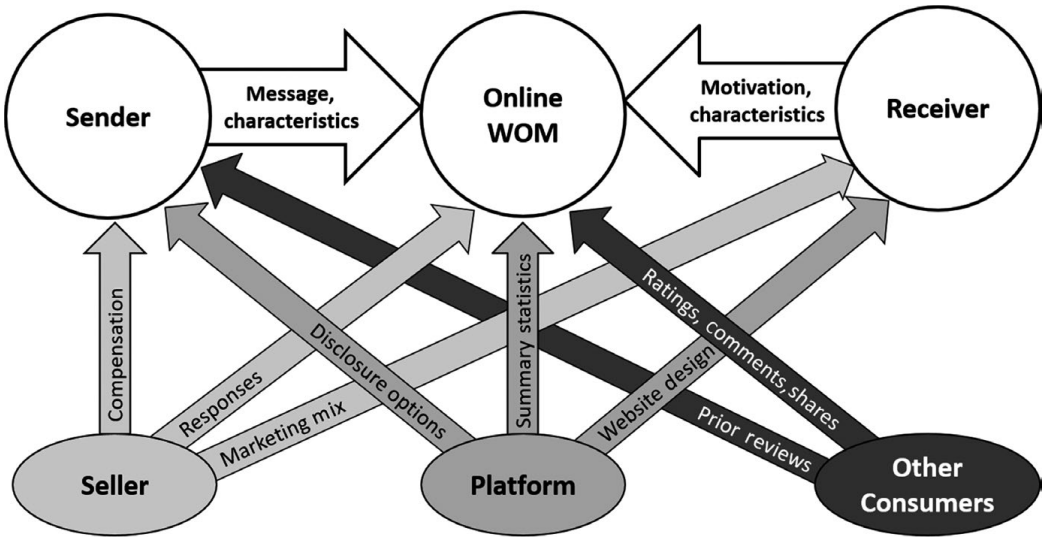


Figure 2.1: How online WOM impacts receivers (Moore and Lafreniere, 2020)

reach a broader audience with limited geographic and time barriers. In fact, it is accessible to everyone for more extended periods and is not limited to just personal contacts (Hennig-Thurau et al., 2004b).

Moreover, considering the evolution that the Internet has permitted, eWOM is defined as "any statement made by customers about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau et al., 2004a).

Forby, the spreading of the internet technologies have transformed WOM into eWOM (Yuan et al., 2020). The popularization of blogs, forums, and more in general social media, has given more influential power to eWOM. In fact, Hart and Blackshaw (2006) explains that such a transformation was led by the fact that propagation of traditional WOM was limited by the physical social networks (e.g., family, friends, and acquaintances), while instead eWOM transcends this limit spreading globally (it is possible to obtain information about companies, products, or brands from many users (King et al., 2014)). For that, it is evident how WOM and eWOM differ: eWOM is not limited to a physical area or by time and is free of the brand, product, and service constraints (Buttle, 1998). Apart from this, we have to consider that, because of the vast number of users, eWOM usually contains positive and negative commentary (Xie et al., 2011).

Considering those aspects, with the words of Hennig-Thurau et al. (2004a), Nam et al. (2020) define eWOM as "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet", highlighting how eWOM users can provide their opinions online, forgetting the problem of incrimination from identity as in face-to-face

WOM (Wang and Fesenmaier, 2004).

Differently from the past, as shown by King et al. (2014), electronic Word Of Mouth (eWOM) has become a major source of information thanks to:

1. enhanced volume,
2. dispersion,
3. persistence and observability,
4. anonymity and deception,
5. salience of valence,
6. community engagement.

There is evidence that consumers in the "social commerce era" are definitely more influenced by eWOM, especially over the social media platforms (Hajli et al., 2014). In particular, consumers born after 1995 (Generation Z), who are particularly active on social media, share opinions and experiences with videos and other internet sources. In fact, they actively search, collect and spread information over the web within the most popular social platforms (Sajjanit, 2018).

A recent report by Yang et al. (2018) revealed that consumer review websites had been identified as the second most frequently used information source apart from search engines (e.g., Google) when travelers are researching a trip. With the recent proliferation of social media websites that facilitate the sharing of travel experiences with others, the role of online consumer reviews has become increasingly pertinent for the tourism and hospitality industry.

This phenomenon underscores the perceived uncertainty founded in consumers while deciding to make a journey in a new place. This leads to a fundamental need to obtain reliable, trustworthy, and helpful information when considering travel options such as booking a room or renting a car (Yin et al., 2014).

Reimer and Benkenstein (2016) highlights how "eWOM represents a powerful and low-cost tool to acquire new customers" for companies. It is more credible and even more persuasive than traditional advertising (Trusov et al., 2009). Given the importance of eWOM, companies should proactively encourage consumers to write reviews through eWOM marketing. Therefore, eWOM is renowned as a powerful marketing instrument (Hussain et al. (2017), Saleem and Ellahi (2017)) that companies have been extensively used for their business (Khwaja and Zaman, 2020). As an example, Aakash and Gupta Aggarwal (2020) shown how, considering that the second most frequently used information source for the

travelers is the guests' review, service providers have started to encourage consumers to post online their experience, utilizing eWom as a marketing system (Sharma et al. (2019); Stringam and Gerdes (2010)).

At the same time, Khwaja and Zaman (2020) highlighted how social media is the leading eWOM platform that is widely used by customers. For that, marketers considered it as a valuable transmitter able to spreading information and gently approaching the targetted audience (Hussain et al. (2018); Bashir et al. (2019)). Users love social media because they can do something that is not possible elsewhere: interact with each other and knowing the person that is giving the information. For the above-mentioned concept, the anonymous information is, of course, less trustworthy.

Forby, it was identified that eWOM can increase transactions, sales, and profits (Cheung and Thadani (2012); Chevalier and Mayzlin (2006); Clemons et al. (2006); Duan et al. (2008); Zhu (2010); Nam et al. (2020)). The interaction between subjects of the eWOM framework, is an example of value co-creation between consumers, service providers, and review websites (Grönroos and Voima (2013); See-To and Ho (2014); Zwass (2010)). However, not all online interactions co-create value, instead another potential results might be achieved: co-destruction. In other words, the opposite of the co-creation process, "that results in a decline in the well-being of at least one member of the system" (Edvardsson et al. (2011); Plé Loïc and Chumpitaz Cáceres Rubén (2010)).

Considering all those aspects, it is possible to notice that eWOM provides to companies a new way to listen to the consumers' needs, reassessing promotion, products, and services (Cheung and Thadani 2012) whilst allowing to understand which are the crucial factors that motivate consumers to post online their opinion (Nam et al., 2020) or, more importantly, to explore the impact of such a comment on other potential consumers (Sparks and Browning, 2011).

2.3 Impact of WoM on consumers' buying decision

Over the years, many social media platforms are born, and some of them are made explicitly for collecting consumer' reviews. In that way, WOM conversations have migrated to the web (Brown et al., 2007), creating new information in the form of online reviews (Schindler and Bickart, 2004), and these reviews provide evaluations of the products from the customers' perspective. It is important to notice how such reviews have a strong influence on consumers' attitudes and purchase behaviour (Chevalier and Mayzlin (2006); Duan et al. (2008); Senecal and Nantel (2004)), definitely more than marketer-generated information (Chiou and Cheng (2003); Willemsen et al. (2011)).

This happens, because consumers are believed to have no particular interest in recommending a certain product or service, thus reviews are more credible and consequently more useful than marketer-generated information (Bickart and Schindler (2001); Ha (2002); Herr et al. (1991); Willemsen et al. (2011)).

On the other hand, as reviews gain in popularity and numbers, it becomes harder for the users to assess the usefulness of the information offered, resulting in an information overload (Duan et al., 2008). To deal with that problem, many review websites have created peer-rating systems that enable consumers to produce a score assessing how much a review was considered helpful in their purchase decision-making process. At the same time, these votes are an indicator of review diagnostics, helping the users to filter relevant opinions, thus achieving their objective more efficiently (Ghose and Ipeirotis (2011); Mudambi and Schuff (2010)).

Forby, analyze reviews is a tricky problem for companies. In fact, customers do not follow a structured format while posting their reviews (Park and Kim (2008); Pollach (2008)). Because of that, reviews' contents might differ a lot: e.g., a simple recommendation with extremely positive/negative words or an accurate product analysis and evaluation supported by considerable reasoning (Mudambi and Schuff, 2010). Especially considering that Willemsen et al. (2011) findings indicate that differences in the perceived usefulness of reviews are related to differences in the content of reviews and not in other characteristics such as star rating or price of the product.

However, Ahmad et al. (2014) shown that there is much evidence that WOM (eWOM) directly produces an impact on consumers while they decide what to do about a certain product or service. In particular, it is possible to find in the academic literature that:

- The increasing use of social media in the tourism industry has resulted in eWOM reviews having a strong influence in consumer decision-making (Blal and Sturman, 2014)

- both eWOM valence and volume will influence consumers' willingness to pay (Nieto-García et al., 2017)

- online reviews' perceived effectiveness and reliability, reduce search costs for consumers and enhance sellers' trustworthiness, which persuades people to pay more for products and ultimately increases sales (Pavlou and Dimoka, 2006).

- Nieto et al. (2014) said that numerous studies have empirically examined the relationship between eWOM and hotel performance and that it was discovered that customer ratings boost hotel performance and affect hotel room prices.

- Nieto et al. (2014) showed that several studies measured hotel performance by the proxy variable of the number of reviews for a property, and they found a 10% increase in review ratings posted on a major Chinese online travel agency (OTA) increased online hotel bookings (measured by the number of consumer reviews on hotels) by more than 5%.

- Yang et al. (2018) highlight how necessary it is for hotels to maintain a high eWOM valence level to attract customers. This result can be explained by the experiential nature of hotel products: no one knows the product's quality until consuming it. Customers, therefore, require supplemental independent reviews to make a decision and reduce the associated risks.

- positive online hotel reviews can enhance customers' trust in the hotel (Sparks and Browning, 2011) resulting in improved financial performance (Öğüt and Bedri Kamil Onur Taş, 2012).

- Recent TripAdvisor.com industry data indicates that about 53% of travelers would not make a reservation until they read hotel online reviews and 77% of prospective guests report reading reviews before they choose a hotel either "always" or "usually" (Yang et al., 2018).

- Vermeulen and Seegers (2009) showed that online hotel reviews increase customers' awareness of the hotels and enhance their consideration in the customers' mind.

- High review scores convey: both product quality and social validation (Cialdini, 2009).

- Yuan et al. (2020) said that many studies were validating the immense effects of WOM on consumer behaviors. In particular, Homer and Yoon (1992) highlight how consumers are more careful when viewing negative messages rather than positive ones.

- Moreover, Yuan et al. (2020) said that, although other studies have found that negative WOM heavily influenced consumer evaluations of brand value (Mizerski 1982; Richins 1983) and purchase intention (Park and Lee 2009), positive eWOM is the main factor that influenced consumers' product purchase intentions. This is in contrast to Allard et al. (2020), which said that "Both types of reviews, negative and positive, greatly influence consumer decision making (Basuroy, Chatterjee, and Ravid 2003; Godes and Mayzlin 2009; Trusov, Bucklin, and Pauwels 2009), with negative reviews often being the most impactful (Chen and Lurie 2013; Mizerski 1982)".

- Aakash and Gupta Aggarwal (2020) said that "satisfaction of guests affects the sales, revisit intension, and market reputation of the firm" (Radojevic et al., 2018).

- Khwaja and Zaman (2020) stated that "Wu and Lin (2017) and Matute Jorge et al. (2016) explained that the patterns of online buying are now dependent upon the comments/suggestions being provided by the previous customers. Considerable research studies have been conducted on examining eWOM on the social media platforms (Choi and Kim (2019); Fatma Mobin et al. (2020); Gurney et al. (2019)). Researchers argue that the transmission of eWOM plays a crucial role in the decision-making of customers (Khwaja et al. (2020b); Bashir et al. (2020))."

- Ahmad and Laroche (2015) and Li et al. (2011) demonstrate that "recent research has also shown that online reviews affect pricing strategy of a firm for repeat-purchase products and can therefore affect product demand".

- Roy et al. (2020) shown that previous studies have modeled the impact of the eWOM environment on financial outcomes like book sales (Godes and Mayzlin, 2004) or non-financial consequences like online purchase intention (Cheung & Thadani, 2012; Roy et al., 2019)

It might now be definitely evident how the literature has reached consensus on finding that higher review scores positively affect demand for hotel and, consequently, increase sales and revenues (e.g., Chevalier and Mayzlin (2006); Phillips et al. (2017); Sparks and Browning (2011)), while negative reviews impact customers' attitudes negatively (Lee et al., 2008). As customers become more discerning, they use online reviews to specify better their service requirements and uncover the best value propositions in the market. As a result, it is common for people to read comments about other's experiences to reduce uncertainty before making a purchase (Archak et al. (2011); Zhang et al. (2011))

In fact, according to Aakash and Gupta Aggarwal (2020), eWOM has a sentiment that refers to the reviewer's emotions, which are expressed through words. This sentiment is defined as the satisfaction level in terms of positive, negative, or neutral behavior that customers represent by spreading eWOM. For that, eWOM contains a substantial amount of information able to influence customers in their buying decision (Aakash & Aggarwal (2019); Aggarwal & Aakash (2017); Ma et al. (2018)). This is extremely useful for companies, especially while making strategic marketing decisions.

Considering the sentiment information, Reimer and Benkenstein (2016) notice that eWOM is able to influence customers' attitudes and behavior (Chevalier and Mayzlin (2006)). Forby, a company might find it interesting, especially while making its advertising plan; in fact, customer-to-customer communication is more persuasive and credible (Trusov et al. (2009)).

A vast amount of research studies have then explored how eWOM impacts online sales and purchase decisions (Cheung & Thadani, (2012); Roy et al. (2020)). However, it is mandatory to consider that the effect of valence and volume of eWOM is influenced by the type of product/service considered. In particular, valence becomes critical for evaluating commodity goods experience on review websites (You et al. (2015)). For example, while perceiving the quality of a movie, the volume of eWOM becomes vital for the first moments; afterward, consumers start to look for a qualitative review from many available sources, thus making valence more important (Bae & Kim, 2013). In the specific case of hotel bookings, tourists usually consider both valence and volume while deciding in which hotel they will make a reservation (Manes & Tchetchik (2018)). In fact, the volume is a proxy for the popularity of hotels among the past customers. After noticing the number of reviews of a hotel, a customer looks for its valence, reading them to understand if the reported experience is positive, negative, or mixed. So, an extensive reviews volume with an excellent eWOM valence score creates, in a potential new customer, a definitely positive expectation toward the hotel and its services, affecting the consumer perception of product/service quality and making a relevant impact on his buying decision (Blal & Sturman (2014); Shen et al. (2012)).

However, beyond review valence, argumentation is an essential proxy of the perceived usefulness of reviews (Willemsen et al., 2011). A high level of argument density and diversity influences the usefulness perception positively more than other characteristics that can be processed with minimal cognitive effort (i.e., the star rating) (Chevelier & Mayzlin (2006); Ghose & Ipeirotis (2008)).

Forby, online reviews have transformed consumer behavior in information searching and sharing. Their growing popularity has enabled new differentiation strategies for lodging operators. More subtly, online review systems have forced hotel managers to compete through the effective use of information systems that they have not created or purchased. Therefore, managers must adapt to the widespread use of external systems, incorporate them in their strategy and evaluate their effects (Pavlou and Dimoka, 2006).

Nevertheless, dealing with such information is tricky for his "hard to measure" nature of reviewers sentiment in an objective way. For instance, above all the other studies, Nam et al. (2020) highlight how negative WOM has a more decisive influence than positive WOM on consumers' brand evaluation (Arndt 1967). In particular, the impact of eWOM on purchase decisions is more significant for negative eWOM than for positive eWOM (Sparks and Browning 2011). There is no interest for a company to reward someone after writing a negative review. On the other hand, a company might try to reward those customers who do a positive review. For that, consumers tend to trust more on negative eWOM

and, in particular, this is confirmed for experience goods such as hotels and for utilitarian products. In general, the effect of sentiment depends on the type of product and on any prior expectations the reader has about a particular product or service (Sen and Lerman (2007)). For that, Ahmad and Laroche (2015) report that, since negative messages are likely to have more substantial effects than positive ones (Chakravarty et al., 2009), a review that has two-sided messages is viewed to be more credible (Jensen et al., 2013).

Summarizing the above-mentioned concepts, Ahmad and Laroche (2015) stated that:

- The opinions expressed in the review have the potential to influence consumers' purchase decisions (Li et al., 2013),
- Studying consumers' reviews is extremely important from a retailing point of view (Pan and Zhang, 2011),
- Both valence and volume of product reviews influence the sales of a product (Chevalier and Mayzlin (2006); Ghose and Ipeirotis (2006)),
- the helpful reviews tend to influence the potential customers' decision more than the others that are perceived as useless (Li et al., 2013).

In conclusion, we agree that "since the helpful product reviews strongly influence other potential customers and, in turn, sales, examining the content characteristics that make a product review helpful is managerially and theoretically important" (Ahmad and Laroche, 2015).

2.3.1 Psychology of Word of Mouth Marketing

Recalling that, joint to the WoM, there is a sentiment component, in this subsection we investigate how that affects the potential new customers while deciding to buy, or not, a specific product.

Ismagilova et al. (2020) have investigated the relationship between emotions and rationality. It has been found that emotional states influence people's reasoning processes and their logical rationality (Pham (2007)). For that, if consumers impress emotions while doing reviews, the eWOM receivers can interpret them as an indicator of rationality.

The emotional influence above the attitude of customers has been demonstrated to be of particular interest for companies. For example, when a consumer perceives a price as fair, it will engage in eWOM communications in order to advise others on the product/service. On the other hand, when it feels that such a price or attitude is unfair, it will engage in eWOM communications to deal with his negative feelings, punishing the company (Wetzer et al. (2007)).

To understand how to investigate the phenomena of sentiment insight eWOM properly, it is essential to highlight some psychological aspects that lead the customer to provide a review and contribute to the eWOM phenomena. As it is possible to notice in Fig. 2.2, Zhou et al. (2020) show the psychological three-step process that affects customers while participating to eWOM:

1. after dealing with a product or a service, a customer feels a certain emotion
2. following a motivation, he will produce an eWOM text corpora
3. this will lead to a change in the behavior in himself and in other users too.

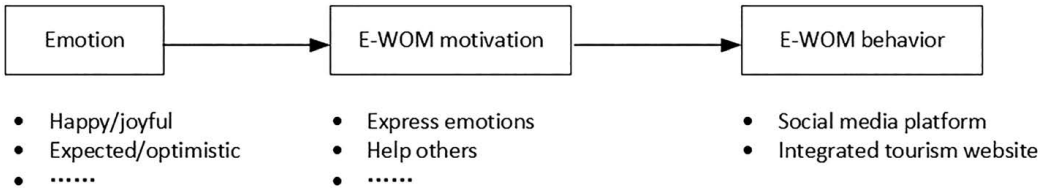


Figure 2.2: The research routine following customers eWOM psychological process (Zhou et al., 2020)

This model finds a confirmation also in Reimer and Benkenstein (2016), where the motivation is defined as a "key determinant of general behavior" (Deci and Ryan (1987)) and as "an internal state or condition that serves to arouse or energize behavior and give it a goal-direction" (Kleinginna and Kleinginna (1981)). In other words, "an inner desire to make an effort" (Dowling and Sayles (1978)).

An exciting aspect of the motivation was shown in some recent papers reported by Chen and Yuan (2020): "the language senders adopt when crafting their WOM is largely influenced by their motivation". For example, while attempting to persuade others, senders use more emotional appeal because of a learned association between emotionality and persuasion (Rocklage et al., 2018). For that, reviews that use simple and direct language (instead of a complex one) tend to elicit greater engagement (i.e., likes, comments, and shares) by facilitating processing fluency (Pancer et al., 2019). WOM that follows a storytelling path tends to be more persuasive and well-received since it "allows receivers to immerse themselves in the review experience" (van Laer et al., 2018).

The self-determination theory (SDT) makes a distinction between two types of motivation, according to the different sources of producing a certain action: extrinsic and intrinsic motivation (Ryan and Deci, 2000):

- Extrinsic motivation means that engagement in an activity is goal-driven, that is, done in order to attain a separable outcome.

- Intrinsically motivated people do an activity for their own sake rather than for external rewards, referring to the fact that the engagement itself provides pleasure and hedonistic satisfaction (Amabile et al. (1994); Huang (2003)).

Recently, academics have started to consider the role of intrinsic motivation in the field of eWOM (e.g., Georgi and Mink, 2013; Sun and Chen, 2014; Yoo et al., 2013), rejecting the idea that external rewards were essential to motivate desired behavior, addressing extrinsic motivation (Steers et al., 2004). This is supported from Reimer and Benkenstein (2016), assessing that when a recommendation is rewarded by a company, "the independence of the reviews is undermined, which leads to the perception that the review has been bought by the company". In other words, it is compared to a firm-generated communication, and for that is not considered to be authentic, relevant, and unbiased like WOM is (Friestad and Wright 1994; Godes and Mayzlin 2004; Allard et al. (2020)).

Forby, consumers who are intrinsically motivated enjoy writing a review and, in other words, contributing to sharing information about a product or a service. If they enjoy the process, then they will be more likely to write a new one. However, Reimer and Benkenstein (2016) highlight how, until now, "there has been no empirical evidence of whether altruistic motivation really is a substantial part of the motivation to write online reviews".

In such a context, altruistic motivation is activated when consumers want to help other users with their buying decisions for the two possible outcomes (Engel et al., 1993):

- positive WOM to enable the same positive experience,
- negative WOM to save them from miss-purchases.

Another possible motivation for creating online reviews is given through the consumer's satisfaction with a product or service. The customer wants that the company becomes or remains successful (Sundaram et al., 1998), "returning the favor" to the company for such a positive experience.

With a similar concept in Fig. 2.3, Van Kleef (2010) proposes a model that shows how social interactions are often ambiguous and that emotions help to disambiguate the situation by providing information about the expresser's feelings, desires, motives, and intentions.

Now, considering that, as a fundamental part of social communication, WoM is the way of sharing ideas, believes, and experiences among each other (Ahmad et al., 2014). While knowing that customers want to reduce their own risk of wasting money while trying to make a decision regarding their willingness to pay (Banerjee, 1992), a social contagion phenomenon occurs (Iyengar et al., 2011b). In fact, people are likely to mimic others' situational behavior in order to reduce their own risk (Iyengar et al., 2011a).

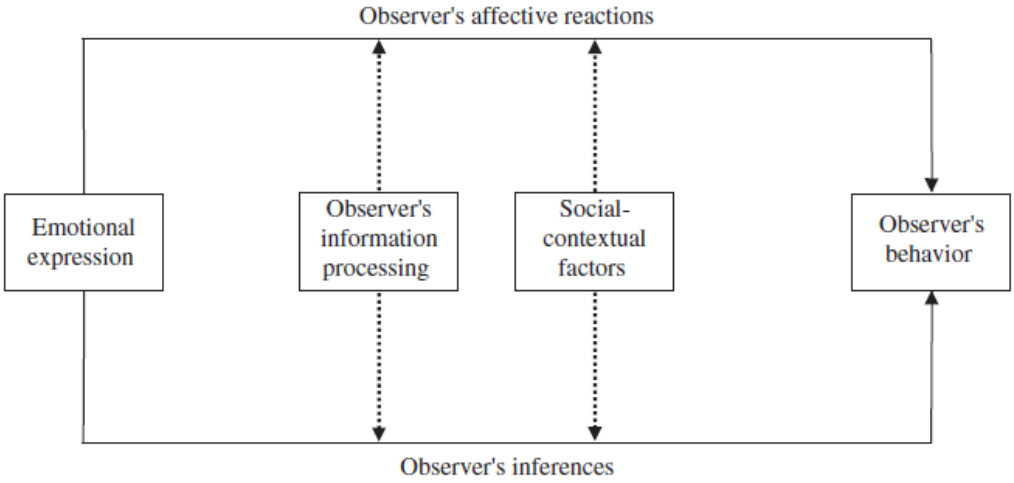


Figure 2.3: The emotions as social information (EASI) model (Van Kleef, 2010)

It is interesting to notice that, while people prefer to share positive self-enhancing content in front of a large audience, they prefer to share useful other-focused WOM when facing a small audience (Barasch and Berger (2014); Chen and Yuan (2020)). Forby, the amount of people in a particular area affects sharing: people perceive a loss of control in a crowded area. For that, they produce more WOM transmissions, trying to restore a sense of control in such a situation (Consiglio et al., 2018). Furthermore, considering that writing instead of talking have an effect on sharing an opinion, it is possible to notice how "people tend to mention more interesting products when writing (versus talking) since the former gives them more time to deliberate and pick WOM topics that reflect well on the self" (Berger and Iyengar, 2013).

As an example, Yang et al. (2018) shown that consumers are more likely to leave comments when they stay in higher-end hotels versus lower-end accommodations (Miguéns et al., 2008). Blal and Sturman (2014) found that eWOM volume appears to exert positive effects on economy, midscale, and upper-midscale hotels, whereas effects are negative for upscale and luxury hotels. Furthermore, the star ratings reflecting hotel classes serve as another signaling factor (Lu and Ye, 2014). A five-star hotel is thought to provide guests with high-quality facilities and services, which affects customers' expectations for service consumption. Moreover, while comparing a five-star hotel with negative reviews and a lower-star hotel with a higher consumer rating, consumers perceive a better quality for the five-star hotel.

Furthermore, Ladhari and Michaud (2015) findings suggest that when a hotel has more positive online comments, customers tend to:

- have a more positive attitude toward the property,

- develop a sense of trust,
- perceive a positive service quality,
- is more inclined to make a reservation with this property.

On the other hand, when the customer selects the hotel based on various eWOM stimuli, he develops an expectation about the hotel (Roy et al., 2020). After the check-in, he perceives the service outcome in the form of "process consumption" (Gronroos, 1988; Grönroos, 2001; Kettinger & Lee, 1994) according to the received service at the hotel. If the expectation matches such a perception, then he will be satisfied, recommending the hotel to others (Lin & Mattila, 2010) and manifesting the intention to repeat the purchase (Hernández-Lobato et al., 2006).

There is thus a personal gratification side aspect that has to be considered. In fact, Bronner and de Hoog (2011) shown that self-directed, social benefits like helping other tourists, helping companies, and consumer empowerment, are a relevant motivation factor that guides them to WoM (eWOM), and Aakash and Gupta Aggarwal (2020) support this, recalling that "good experience about product/service can increase the perception toward product/service quality, which is the antecedents for guest satisfaction, while the bad experience is an antecedent of guest dissatisfaction" (Dai et al., 2015).

There is ample evidence about people encountering emotions while in a consumption process (Ahmad and Laroche, 2015), and these emotions vary from extremely positive to stronger disappointment (Richins, 1997). Furthermore, they report a strong evidence that discrete emotions such as love, joy, and hopefulness (all positive emotions, but with a different scale) have distinct influence on decision making (Nabi (2003); Raghunathan et al. (2006)) or information processing (Tiedens and Linton, 2001).

Also, Nieto et al. (2014) analyzed the effect of marketing decisions by Spanish rural lodging establishments on eWOM and the effects of eWOM on business performance (measured as the owner's perceptions). They found that those customer ratings and the number of reviews positively influenced the perceived satisfaction, profitability and market perceptions.

Moreover, Katz and Lazarsfeld (1955) shows that a consumer perceives received information from someone in consumers' social network (a family member, friend, or a peer) as objective. In other words, away from manipulation and commercial purposes.

Furthermore, Khwaja and Zaman (2020) findings confirmed what Chu and Kim (2011) explained: the dimensions of eWOM include tie strength, homophile, trust, normative influence and informational influence. Tie strength can be referred to as the potency of bonds between members of a network (Ngarmwongnoi Chananchida et al., 2020). The social ties

can be of a weak or strong nature (Luo et al., 2013). The ties of family and friends are of strong nature and cannot be sidelined at all. The interpersonal networks of individuals with family and friends are quite strong, and they also provide emotional and substantive support (Khwaja et al. (2019);Yun et al. (2020)). Weak ties include taking information from colleagues, acquaintances, and facilitation of information on diverse topics from various people (Khwaja et al., 2020a). The role of tie strength in eWOM communications is critical, as individuals consider the information of strong ties to be of a more credible nature (Hwangbo and Kim, 2019). The inclination of individuals with people having strong tie characteristics is certainly strong, powerful, and impactful (Melancon and Dalakas, 2018).

That is why WOM (eWOM) is a much more persuasive and powerful tool for marketing, particularly in comparison with traditional marketing communication channels such as advertising.

On the other side, when the product or service does not match the consumer's expectation, the result of this experience will be dissatisfaction. Solomon et al. (2006) findings shown that, while feeling highly disappointed in a psychological context, the consumer will take actions to reduce his/her discomfort. There are several ways through which a consumer can express their dissatisfaction:

- They can directly complain to the company (voice response or a negative review),
- take legal action against the company,
- write a letter about their complaint to a newspaper (third-party response),
- express the dissatisfaction in an implicit way such as boycotting product/service by switching the brand (private response).

As reported from Solomon et al. (2006), according to a study by the White House Office of Consumer Affairs, 90% of dissatisfied consumers will not do business with a company again, and each of these unsatisfied consumers are very likely to share their negative experiences with at least nine other people, and 13% of those customers will go on, telling that to more than 30 people. Even though it has been argued that consumers with bad experiences shares that with people than those with good experiences (Hart et al., 1990), equally there are some arguments to suggest that positive events produce a more robust response (Holmes and Lett, 1977).

Chapter 3

Machine Learning

Considering that the author has a Master's Degree in computer science, this chapter well represents such a nature, mixing concepts from the author's background within the core focus of this Ph.D.: Statistics. Moreover, those concepts represent the description of the instruments that will be used for analyzing the data and, more specifically, the type of data that we have described in Chapter 2. In fact, this chapter is the answer to the following questions:

- How do we use past data to "predict the future"? (Section 3.1)
- Despite trying to make predictions, what else can we do if this is not possible? (Section 3.2)
- How can those concepts represent the base of an Artificial Intelligence? (Section 3.3)

In this chapter, we will analyze the main techniques that are used for transforming data into new information and often try to "predict the future". Before starting, it is essential to highlight that this chapter should be considered as an introduction to Machine Learning, based on the books of Hastie et al. (2009) and Alpaydın (2010). It is definitely not exhaustive, but it provides the basic knowledge and elements for an overview of it, thus being able to understand the other chapters that are the core of this thesis.

"Machine learning is programming computers to optimize a performance criterion using example data or past experience" (Alpaydın, 2010).

When people are facing a particular problem, which they try to deal with in the fact that human expertise does not exist or it is not sufficient to solve it, or such an experience exists, but they do not know how to explain it to a computer program, they need to use machine learning. Usually, such an approach consists of collecting a large amount of data that are analyzed to create a mathematical "model" able to learn from the given data and predict new results that will come in the future.

Like a newborn baby, a computer, despite the incredible calculus power or immense storage, is an empty board (what we might call a *tabula rasa*). It might appear stupid, but in fact, it is able to absorb all the information that we will "feed" to him, and with that to learn from them, using the past experience to produce his domain-knowledge that will be extremely helpful while dealing with in the new – expected or not – events that will come in the coming days.

This kind of metaphor well represents how we deal with a computer while we try to solve some particular problems. An example:

- converting an acoustic speech signal to ASCII text. Almost every person can do this without any difficulty, but we are unable to explain how they do it. Just think about how hard it might be for anyone to hear some completely different language and to understand it. Moreover, it is far more complicated if we consider that the same word can have a different sound depending on the characteristics of the subject that is pronouncing it (age, accent, gender, ...);

- In addition to retail, in finance banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market;

- In manufacturing, learning models are used for optimization, control, and troubleshooting;

- In medicine, learning programs are used for medical diagnosis;

- In telecommunications, call patterns are analyzed for network optimization and to maximize the quality of service;

- In science, large amounts of data in physics, astronomy, and biology can only be analyzed fast enough by computers;

- routing packets over a computer network. Here the problem to be solved changes in time cause it depends on the environment. Being able to predict when to use a certain path rather than another, it is mandatory to improve the efficiency of systems like that.

Considering that an algorithm is "a sequence of instructions that should be carried out to transform the input to output" (Alpaydın, 2010), to solve a problem with a computer we need an algorithm. Many can be created to solve a specific problem, thus providing a required input-output functionality as expected. For that, for choosing one, a comparison

between them for finding the most efficient one should be made. In fact, they might differ in terms of computational complexity, the precision of the results, or some other performance parameters that act as an indicator of efficiency.

However, for some tasks, an algorithm cannot be made without the usage of some example data. For example, we are categorizing emails into spam or non-spam. Input and output combinations are well known: a file/email document (input) and a spam/non-spam (output). Some email categorization examples are needed even for doing that by hand; so, for that, we need to provide such a piece of information to the computer to obtain such a program. Data can then take care of our lack of knowledge, and the computer learns from it which one constitutes spam or not. In that way, we are using a Machine to extract an algorithm Learning automatically from some past experience (data).

The more computer technology improves, the more increasing our capability to store large amounts of data and to process them from physically distant locations over the world. Moreover, as we have observed in Chap. 2, an immense amount of data is produced, stored, and processed every day. People today are considered to be a "living product" for the data that they can provide, and Google or Facebook are good examples of how companies are making billions with free-users, analyzing the data that they produce.

We do not know why some people love more a particular product rather than another but, by collecting some data and processing it, we believe that there is something that will explain such a phenomenon. This strategy can be applied in particular when we know that events are not entirely random (like consumer behavior). The target then is to find patterns in the data.

Even though finding the exact real pattern of a phenomenon is almost impossible, we can produce a good approximation, detecting some regularities. Such an approximation will not explain all the data that we have, but it will consistently reduce the uncertainness that we have, helping to understand it better and to make predictions. In fact, we can assume that at least the near future will not be substantially distant from the past data that we are observing.

However, Machine Learning is not just a matter of data; it is the base system of Artificial Intelligence (AI). To be considered valuable, an AI should be able to learn and adapt to changes that come from the environment. In particular, for those events that are not directly observed in the past data, but that can be inferred. This allows a machine to know what to do for all possible situations.

Considering that the main task is to make the inference from a sample, according to Alpaydın (2010), Machine Learning is based on:

- statistics: a key role for producing mathematical models;

- computer science: efficiency and optimization problems during the training of the model and in the algorithmic solution provided by the model. “Space and time complexity, maybe are important as its predictive accuracy” (Alpaydın, 2010)

To summarize, Machine Learning has a huge impact in many disciplines and for many problems:

- Predict whether a patient hospitalized due to a heart attack will have a second heart attack. The prediction is to be based on demographic, diet, and clinical measurements for that patient;
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data;
- Identify the numbers in a handwritten ZIP code from a digitized image;
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person’s blood;
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

While facing a problem, usually we have an outcome variable (quantitative or categorical) that we want to predict with some features. After collecting some data, we have a proper training set to observe how such features impact the outcome variable as a collection of examples. Using this data in general, we build a prediction model for use those features to predict the outcome for new and not observed objects. The higher the precision of the model in the prediction of those unobserved events, the higher it will be considered in understanding the phenomena.

That summary describes what is called the Supervised Learning problem. “It is called “supervised” because of the presence of the outcome variable to guide the learning process” (Hastie et al., 2009). On the other hand, if we just observe the features without having any measure of the outcome, we are facing an Unsupervised Learning problem. Such cases, that are less developed in the literature (Hastie et al., 2009), places the focus on describing how the data are clustered.

3.1 Supervised Learning

In this section, we will see a general overview of the Supervised Learning problems. Recalling what we have said in the introduction, here we have some data composed of many examples

of input-output combinations. In other words, we have predictors and outcome variables, features and responses, or independent and dependent variables. The goal is to use the inputs to predict the output.

We will observe now a general supervised machine learning algorithm as described by Alpaydın (2010).

We have a sample:

$$X = \{x^t, y\}$$

with $x_i^t = (x_{i1}, x_{ij}, x_{ip})$, $i = 1, \cdots, n$ and $j = 1, \cdots, p$ observed on n instances on which p measurements are made.

The sample is composed of independent and identically distributed (iid) random variables; the ordering is not important, and all instances are drawn from the same joint distribution $p(x, y)$.

The other parameters are as follows:

- i indexes one of the n instances,
- j indexes one of the p features
- x^t is the arbitrary dimensional input
- y is the associated desired output. That is a 0/1-form for two-class learning, a K-dimensional binary vector (where exactly one of the dimensions is 1 and all others 0) for $(K > 2)$-class classification, and is a real value in regression.

The aim is to build a good and useful approximation to y using the model $g(x|\theta)$. In doing this, there are three decisions we must make:

1. The model we use in learning, denoted as

   $$g(x|\theta)$$

   Where,

   - $g(\cdot)$ is the model,
   - x is the input,
   - θ are the parameters.

So, $g(\cdot)$ defines the hypothesis class H, and a particular value of θ instantiates one hypothesis $h \in H$. The model (inductive bias), or H, is fixed by the machine learning system designer based on his or her knowledge of the application and the hypothesis h is chosen (parameters are tuned) by a learning algorithm using the training set, sampled from $p(x, y)$.

2. The loss function, $L(\cdot)$, to compute the difference between the desired output, y , and our approximation to it, $g(x|\theta)$, given the current value of the parameters, θ. The *approximation error*, or loss, is the sum of losses over the individual instances:

$$E(\theta|x) = \sum_t L(y, g(x|\theta))$$

In class learning, where outputs are probabilities for retrieving a specific label, $L(\cdot)$ checks, in general, for the accuracy of the prediction, usually considering the label with the highest probability to be the predicted one and checking if that match, or not (Misclassification Error), with the label that was supposed to predict. In regression, because the output is a numeric value, we have to order information for distance, and one possibility is to use the square of the differences.

3. The optimization procedure to find θ^* that minimizes the total error

$$\theta^* = \underset{\theta}{\operatorname{argmin}} E(\theta|x)$$

Where argmin returns the argument that minimizes θ^*. In regression, we can solve analytically for the optimum. With more complex models and error functions, we may need to use more complex optimization methods, for example, gradient-based methods, simulated annealing, or genetic algorithms.

Moreover, following Hastie et al. (2009) we might state that the optimization method is usually characterized by some loss function $L(y, \hat{y})$, for example, $L(y, \hat{y}) = (y, \hat{y})^2$. If one supposes that (x, y) are random variables represented by some joint probability density $P(x, y)$, then supervised learning can be formally characterized as a density estimation problem where one is concerned with determining properties of the conditional density $P(y|x)$.

Usually the parameters of interest are the "location" parameter μ that minimizes the expected error at each x:

$$\mu(x) = \underset{\theta}{\operatorname{argmin}}\, E_{y|x} L(y|\theta)$$

Conditioning, one has:

$$P(x, y) = P(y|x) \cdot P(x)$$

Where $P(x)$ is the joint marginal density of the x values alone.

In supervised learning, $P(x)$ is typical of no direct concern. One is interested mainly in the properties of the conditional density $P(y|x)$. Since y is often of low dimension (usually one), and only its location $\mu(x)$ is of interest, the problem is greatly simplified.

Furthermore, Alpaydın (2010) defines three conditions that should be satisfied:

1. the hypothesis class of $g(\cdot)$ should have enough capacity to include the unknown function that generated the data that is represented in a noisy form in y;

2. the amount of training data should be of considerable amount, at least sufficient to pinpoint, from the hypothesis class, the closest approximation to the correct hypothesis (if not exactly the correct one);

3. considering the data used for training the model, we need a good optimization method able to find the correct hypothesis with that data. In fact, Machine Learning algorithms can differ not only for the model that they use or the loss measure that they calculate for the performance of the fit but also for the optimization procedure adopted.

3.1.1 Naïve Bayes Classifier

One of the most famous, simple, but powerful Machine Learning model is the Naïve Bayes Classifier (Lewis, 1998). Many classifiers have been developed over the years but, considering its popularity, and that in Section 5.4, we propose an adapted version of it for our framework, it will be relevant to discuss this model.

As the name suggests, this classifier relies on the Bayes' theorem, which is of fundamental importance for inferential statistics (Berrar, 2019). Let A and B be two events from a finite (or countably infinite) sample space Ω. Let $P : \Omega \to [0, 1]$ be a probability distribution on

Ω such that $0 < P(A) < 1$ and $0 < P(B) < 1$ and $P(\Omega) = 1$. Then, the Bayes' theorem states that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A|B)$ is the conditional probability of A given B, and $P(B)$ is the marginal probability.

According to Berrar (2019), for the Total probability theorem, Bayes' theorem can be used to derive the posterior probability of a hypothesis given observed data:

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

where $P(data|hypothesis)$ is the likelihood of the data given the hypothesis, $P(hypothesis)$ is the prior probability of the hypothesis, and $P(data)$ is the probability of observing the data, disregarding the specific hypothesis.

With the same considerations, we can move on the continuous case. Let x and y denote two continuous random variables with joint probability density function $f_{xy}(x, y)$. Then the Bayes' theorem states that:

$$f_{x|y}(x|y) = \frac{f_{y|x}(y|x)f_x(x)}{f_y(y)}$$

where $f_{x|y}(x|y)$ and $f_{y|x}(y|x)$ are the conditional probability density functions for x and y such that:

$$f_{x|y}(x|y) = \frac{f_{xy}(x, y)}{f_y(y)} \qquad \text{and} \qquad f_{y|x}(y|x) = \frac{f_{xy}(x, y)}{f_x(x)}$$

Considering an input set of n measurement of x, and a categorical output y with K classes, we can now define a Bayesian probabilistic model that assigns a posterior class probability to an instance: $P(Y = y_k|X = x_i)$. Where applying the Bayes' theorem, we obtain:

$$P(y_k|x_i) = \frac{P(x_i|y_k)P(y_k)}{P(x_i)}$$

Furthermore, to define a (simple) Naïve Bayes classifier, we make a strong assumption: the individual x_i are independent from each other. We call this classifier as *naïve* because this assumption is usually violated in real applications. This implies that $P(x_1|x_2, x_3, \ldots, x_n, y_K) = P(x_1|y_k)$. Thus, with all this considerations, Berrar (2019) shows that we can define a model as follows:

$$\hat{y} = \underset{y_k}{\text{argmax}} \prod_{i=1}^{n} P(x_i|y_k)P(y_k)$$

Which is a classifier that, for the observation x, predicts the *Maximum A Posteriori* (MAP) class.

Following those considerations, many versions of the Naïve Bayes have been defined according to the type of data for which they are applied. In Section 5.4 we propose a new version of this classifier.

3.2 Unsupervised Learning

The previous section was concerned with predicting the values of one output or response variable y for a given set of input or predictor variables $x^t = (x_1, \ldots, x_p)$. Denote by $x_i^t = (x_{i1}, \ldots, x_{ip})$ the inputs for the i-th training case, and let y_i be a response measurement.
The predictions are based on the training sample $(x_1, y_1), \ldots, (x_n, y_n)$ of previously solved cases, where the joint values of all of the variables are known. This is called supervised learning or "learning with a teacher". Under this metaphor, the "student" presents an answer $\hat{y}_i$ for each x_i in the training sample, and the supervisor or "teacher" provides either the correct answer and an error associated with the student's answer.

As discussed previously, there are many approaches for successfully addressing supervised learning in a variety of contexts. Here we address unsupervised learning or "learning without a teacher".

In this case one has a set of n observations $(x_1, x_2, \ldots, x_n)$ of a random p-vector x having joint density $P(x)$. The goal is to directly infer the properties of this probability density without the help of a supervisor or teacher providing correct answers or degree-of-error for each observation.

The dimension of x is sometimes much higher than in supervised learning, and the properties of interest are often more complicated than simple location estimates. These factors are somewhat mitigated by the fact that x represents all of the variables under consideration; one is not required to infer how the properties of $P(X)$ change, conditioned on the changing values of another set of variables.

While in low-dimensional problems ($p \leq 3$), it is possible to apply some non-parametric methods to directly estimate the density $P(x)$ and representing it graphically. Here we are dealing with a high-dimensional problem.

One must settle for estimating rather crude global models, such as Gaussian mixtures or various simple descriptive statistics that characterize $P(x)$. Generally, these descriptive statistics attempt to characterize x-values, or collections of such values, where $P(x)$ is relatively large. Principal components, multidimensional scaling, self-organizing maps, and principal curves, for example, attempt to identify low-dimensional manifolds within the x-space that represents high data density. This provides information about the associations among the variables and whether or not they can be considered as functions of a smaller set of "latent" variables.

Cluster analysis attempts to find multiple convex regions of the x-space that contain modes of $P(x)$. This can tell whether or not $P(x)$ can be represented by a mixture of simpler densities describing distinct types or classes of observations. Mixture modeling has a similar goal. Association rules attempt to construct simple descriptions (conjunctive rules) that describe regions of high density in the particular case of very high dimensional binary-valued data.

With supervised learning, there is a precise measure of success or lack thereof, that can be used to judge adequacy in particular situations and to compare the effectiveness of different methods over various situations. Lack of success is directly measured by expected loss over the joint distribution $P(x)$.

This can be estimated in a variety of ways, including cross-validation. In the context of unsupervised learning, there is no such direct measure of success. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms.

One must resort to heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. This uncomfortable situation has led to heavy the proliferation of proposed methods, since "effectiveness is a matter of opinion and cannot be verified directly" (Hastie et al. (2009)).

3.2.1 Cluster Analysis

The goal of cluster analysis is to partition the observations into groups ("clusters") so that the pairwise dissimilarities between those assigned to the same cluster tends to be smaller than those in different clusters.

Clustering algorithms fall into three distinct types:

- Combinatorial algorithms, that work directly on the observed data with no direct reference to an underlying probability model.

- Mixture modeling, that supposes that the data is an i.i.d. sample from some population described by a probability density function. This density function is characterized by a parametrized model taken to be a mixture of component density functions; each component density describes one of the clusters. This model is then fit to the data by maximum likelihood or corresponding Bayesian approaches.

- Mode seeking, that takes a non-parametric perspective, attempting to estimate distinct modes of the probability density function directly. Observations "closest" to each respective mode then define the individual clusters.

Here will follow the definition of the most popular clustering algorithm, the K-means proposed in Hastie et al. (2009).

This algorithm directly assigns each observation to a group or cluster without any reference to a probability model describing the data. Each observation is uniquely labelled by an integer $i \in 1, \ldots, n$. A pre-specified number of clusters $K < n$ is postulated, and each one is labelled by an integer $k \in 1, \ldots, K$.

Each observation is assigned to one and only one cluster. These assignments can be characterized by a many-to-one mapping, or encoder $k = C(i)$, that assigns the i-th observation to the k-th cluster.

One seeks the particular encoder $C^*(i)$ that achieves the required goal, based on the dissimilarities $d(x_i, x_{i'})$ between every pair of observations. These are specified by the user as described above.

Generally, the encoder $C(i)$ is explicitly delineated by giving its value (cluster assignment) for each observation i. Thus, the "parameters" of the procedure are the individual cluster assignments for each of the n observations. These are adjusted so as to minimize a "loss" function that characterizes the degree to which the clustering goal is not met.

The K-means algorithm is one of the most popular iterative descent clustering methods. It is intended for situations in which all variables are of the quantitative type, and squared Euclidean distance is chosen as the dissimilarity measure:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

where $d(\cdot)$ is the dissimilarity function between two objects or observations (x_i and $x_{i'}$).

The goal is to specify a mathematical loss function directly and attempt to minimize it through a combinatorial optimization algorithm. Since the objective is to assign close

points to the same cluster, a natural loss function would be the within-point scatter function $W(\cdot)$:

$$\begin{aligned} W(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}\|x_i - x_{i'}\|^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2 \end{aligned} \tag{3.1}$$

where $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$ is the mean vector associated with the k-th cluster, and $n_k = \sum_{i=1}^{n} I(C(i) = k)$. Thus, the criterion is minimized by assigning the n observations to the K clusters in such a way that within each cluster, the average dissimilarity of the observations from the cluster mean, as defined by the points in that cluster, is minimized. This criterion characterizes the extent to which observations assigned to the same cluster tend to be close to one another. For that, it is referred to as the "within cluster" point scatter.

In the same way, we define the "between cluster" point scatter function $B(\cdot)$, which is larger when observations assigned to different clusters are far apart:

$$\begin{aligned} B(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k}\|x_i - x_{i'}\|^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)\neq k}\|x_i - \bar{x}_k\|^2 \end{aligned} \tag{3.2}$$

As well, we define the "total" point scatter, a constant (T) given the data, independent of cluster assignment:

$$T = W(C) + B(C)$$

It is relevant to highlight that, by definition, minimizing $W(\cdot)$ is equal to maximizing $B(\cdot)$.

The K-means Clustering algorithm can then be summarised with those three steps:

---

1. For a given cluster assignment C, the total cluster variance is minimized with respect to each cluster mean $m_1, \ldots, m_K$.

2. Giving a current set of means $m_1, \ldots, m_K$, the total cluster variance is minimized by assigning each observation to the closest (current) cluster mean:

$$C(i) = \underset{i \leq k \leq K}{\operatorname{argmin}} \|x_i - m_k\|^2$$

3. Steps 1 and 2 are iterated until the assignments do not change any more.

Each of steps 1 and 2 reduces the value of the criterion so that convergence is assured. However, the result may represent a suboptimal local minimum. For that, one should start the algorithm with many different random choices for the starting means and choose the solution having the smallest value of the objective function.

3.3 Neural Networks

In this section, we now describe a class of learning methods that were developed separately in various fields based on two main frameworks: statistics and artificial intelligence. The central idea is to extract linear combinations of the inputs as derived features and then model the target as a non-linear function of these features. The result is a powerful learning method with various applications in many fields.

The term neural network has evolved to encompass a large class of models and learning methods. Here we describe the most widely used "vanilla" neural net, sometimes called the single hidden layer back-propagation network or single-layer perceptron.

There has been a great deal of hype surrounding neural networks, making them seem magical black-box. Instead, they are just non-linear statistical models: a neural network, in fact, is a two-stage regression or classification model, typically represented by a network diagram as in Fig. 3.1.

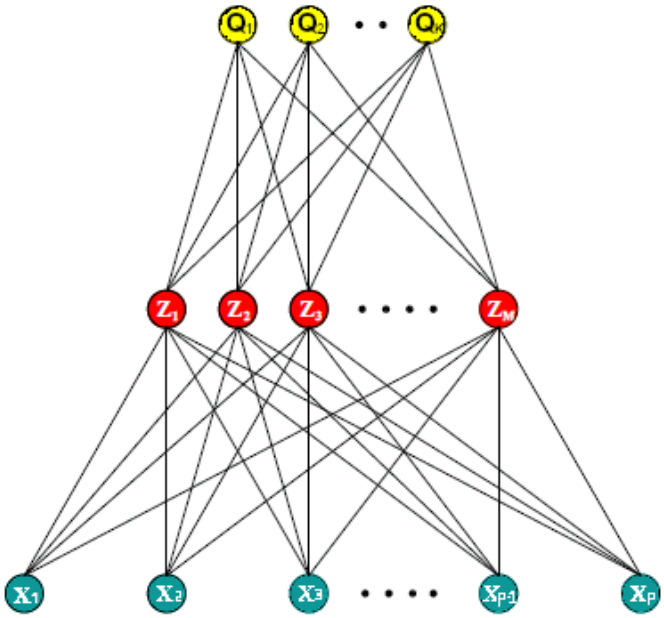


Figure 3.1: Adapted Hastie et al. (2009) Basic Feed-Forward Neural Network scheme

This network applies both to regression or classification. For regression, K = 1 and there is only one output unit Q_1 at the top. However, these networks can handle multiple quantitative responses in a seamless fashion so that we will deal with the general case.

For K-class classification, there are K units at the top, with the k-th unit modeling the probability of class k. There are K target measurements $Q_k, k = (1, \ldots, K)$, each being coded as a 0 – 1 variable for the k-th class.

Derived features Z_m are created from linear combinations of the inputs, and then the target Q_k is modeled as a function of linear combinations of the Z_m:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^Y X),\ m = 1, \ldots, M, \\ Y_k &= \beta_{0k} + \beta_k^Y Z,\ k = 1, \ldots, K, \\ f_k(X) &= g_k(Y),\ k = 1, \ldots, K \end{aligned} \tag{3.3}$$

The activation function $\sigma(v)$ is usually chosen to be the *sigmoid* $\sigma(v) = \dfrac{1}{(1 + e^{-v})}$.

Neural network diagrams like Fig. 3.1 are sometimes drawn with an additional bias unit feeding into every unit in the hidden and output layers. Thinking of the constant "1" as an additional input feature, this bias unit captures the intercepts α_{0m} and β_{0k} in the model.

The output function $g_k(Y)$ allows a final transformation of the vector of outputs Y.

For regression we typically choose the identity function $g_k(Y) = Y_k$. Early work in K-class classification also used the identity function, but this was later abandoned in favour of the *softmax* function, producing positive estimates that sum to one:

$$g_k = \frac{e^{Y_k}}{\sum_{l=1}^{K} e^{Y_l}}$$

The units in the middle of the network, computing the derived features Z_m, are called hidden units because the values Z_m are not directly observed. We can think of the Z_m as a basis expansion of the original inputs x; the neural network is then a standard linear model, or linear multilogit model, using these transformations as inputs.

Notice that if σ is the identity function, then the entire model collapses to a linear model in the inputs. Hence a neural network can be thought of as a non-linear generalization of the linear model, both for regression and classification.

The name "neural networks" derives from the fact that they were first developed as models for the human brain. Each unit represents a neuron, and the connections represent synapses. In early models, the neurons fired when the total signal passed to that unit exceeded a certain threshold, which here corresponds to the use of the sigmoid function.

While dealing then with the problem of fitting a neural network model, we should consider that he have unknown parameters, usually called weights, and we seek values for them that make the model fit the training data well. We denote the complete set of weights by θ, which consists of:

$$\begin{aligned} \{\alpha_{0m}, \alpha_m; m = 1, 2, \ldots, M\} \qquad & M(p+1)\ weights \\ \{\beta_{0k}, \beta_k; k = 1, 2, \ldots, K\} \qquad & K(M+1)\ weights \end{aligned}$$

For regression, we use sum-of-squared errors as our measure of fit (error function):

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

For classification we use either squared error or cross-entropy (deviance):

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log f_k(x_i)$$

where the corresponding classifier is $G(x) = \text{argmax}_k\, f_k(x)$.

With the softmax activation function and the cross-entropy error function, the neural network model is exactly a linear logistic regression model in the hidden units, and all the parameters are estimated by maximum likelihood.

It is essential to highlight how we do not search for the global minimizer of $R(\theta)$, as this is likely to be an overfit solution. Instead, some regularization is needed: this is achieved directly through a penalty term or indirectly by early stopping. The generic approach to minimizing $R(\theta)$ is by gradient descent, usually called back-propagation. Thanks to the compositional form of the model, the gradient can be easily derived using the chain rule for differentiation. This can be computed by a forward and backward sweep over the network, keeping track only of quantities local to each unit.

To summarize, Neural Networks are a "family of learning techniques that were historically inspired by the way computation works in the brain, and which can be characterized as

learning of parametrized differentiable mathematical functions" (Goldberg (2017)). They provide a powerful learning machinery that is very appealing for use in natural language problems, in fact Recursive models were shown to produce state-of-the-art or near state-of-the-art results for sentiment classification tasks (Hermann and Blunsom (2013), Socher et al. (2013a), Socher et al. (2013b)).

Chapter 4

Natural Language Processing

The main focus of this Ph.D. thesis is placed on analyzing textual data (Chapter 2) with machine Learning methods recalled in Chapter 3. For that, with this chapter, we present a literature review of the most common and specific techniques that support models for analyzing textual data. In this chapter, we propose the answer to the following questions:

- what is Natural Language Processing (NLP)?
- what kind of knowledge databases exists for supporting the models in the NLP field? (Section 4.1)
- how can we deal with the complex rules that we find in the language field? How should we represent this type of textual data to make it suitable for the analysis but avoiding losing too much information during the transformation process? (Section 4.2)

"Natural Language Processing (NLP) is the field of designing methods and algorithms that take as input or produce as output unstructured, natural language data" (Goldberg, 2017).

This thesis has a central focus on the Natural Language Processing techniques application (Chapters 5 and 6), so, for that, with this chapter, we are giving an overview of what NLP is and how to use it.

"While we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language" (Brownlee, 2017).

Such kind of processing, used for understanding and producing language with a computer, is considered to be highly challenging (Goldberg, 2017). In fact, we have to consider that the written language's essential elements are combinations of characters. However, pragmatically, if we consider just the symbols of the individual characters themself, there is no relation between "potato" and "chips". Moreover, there is also a compositional component because we can make words with characters, phrases, and sentences. Furthermore,

despite the words that compose a specific sentence, the meaning could be metaphorically different or understandable only if considered within another group of sentences. Those are just examples of the intricate set of rules that we need to "teach" to a computer before it will be able to understand a particular language.

However, rules are not the only problem here: the type of data that we are considering, if treated in the usual way, will never be sufficient. In fact, while combining characters, we can produce an immense amount of words, and the combination of words can lead to infinity. This "data sparseness" phenomenon makes it hard to work with the usual way of "learning from examples" (Supervised Learning, Chap. 3) with just raw data.

Some preprocessing is then mandatory for reducing the complexity of the data and the intrinsic ambiguity of the human (natural) language. Those preprocessing steps, which will be analyzed more in detail in Chapter 5, can be summarized as follows:

- Filtration: where all the non-necessary pieces of information (like stop-words) are removed;
- Lexical analysis: decomposition of language expressions into tokens;
- Grammatical analysis: association to every considered word with his "role" in the considered text;
- Syntactic analysis: where tokens are inserted in a syntactic structure for considering their relationships;
- Semantic analysis: new semantic information is achieved from the analysis of the syntactic structure that was created at the previous step.

Despite that machine learning methods "excel at problem domains where a good set of rules is tough to define but annotating the expected output for a given input is relatively simple" (Goldberg, 2017), here we are dealing with ambiguous and variable inputs of a non-specified set of rules.

The mandatory preprocessing phase helps while dealing with such type of data, but it is not sufficient for producing a suitable analysis. For that, over the last 20 years, but more particularly during the last decade, some strong supporting materials and technologies were developed.

We are now going to focus on two principal components of this "supporting material", that will also make the ground knowledge for this thesis' core (Chapters 5 and 6): Lexical Databases and Words Embeddings.

The firsts are databases usually produced by language experts with admirable hand-work. The seconds are instead representations of words in a high-dimensional space. Usually created by Neural Networks, despite the fact that humans cannot understand what any dimension means, a computer instead is able to use it to significantly improving his knowledge of the language and the words that compose it.

4.1 Lexical Databases

Lexical Databases are a set of information that represent psycholinguistic theories of human lexical memory in a suitable way for a Natural Language Processing task. With some hand-work, "English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets" (Miller et al., 1990).

Thus, an "useful additional source of information about word forms are lexical resources"(Goldberg, 2017). In fact, they are dictionaries that were created for being accessed by a program, a computer, instead of normal people. In such thesaurus, the most important thing is the fact that words are linked with other words, providing more information of a standard dictionary.

A very well-known lexical resource in English is WordNet. WordNet is "a very large manually curated dataset attempting to capture conceptual semantic knowledge about words" (Goldberg, 2017).

4.1.1 WordNet

WordNet began more than a decade ago as an experiment of a psycholinguist: (Miller, 1986).

"Miller and his co-workers in the Cognitive Science Laboratory at Princeton University wanted to find out whether a semantic network could be constructed not just for a handful of words but for the better part of the vocabulary of a natural language. Over the years, WordNet has grown into a large lexical database that has become the tool of choice for researchers in many areas of computational linguistics in dozen of countries" (Fellbaum, 1998).

Nowadays, people have gone much further: instead of just using it, they are creating new versions for other languages. WordNet, in fact, has become a *standard de facto* of the majority of lexical databases, extremely useful for Natural Language Processing networks.

WordNet is a "large electronic lexical database for English" (Miller (1995)). The author was inspired by experiments in Artificial Intelligence for understanding the human semantic memory (Fellbaum, 2010). Those experiments were proposing a hierarchical structure of concepts where more general concepts are superordinate to more specific ones. The intuition of Miller and his team was to transport those concepts in a network structure. The result of this intuition is WordNet:"a large, manually constructed semantic network where words that are similar in meaning are interrelated" (Fellbaum, 2010).

Formally speaking, WordNet is an acyclic graph representing a semantic network (Fellbaum, 2012). Although it might look like an ordinary dictionary for certain things, it is not. WordNet, in fact, links words not only for their form (the letters that compose the words) but instead following the specific sense that they have. So, "words that are found in close proximity to one another in the network are semantically disambiguated" (Fellbaum, 2012). In particular, the link between words represents the semantic relationship, and something like that is hard to find in a standard dictionary.

"The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms are grouped into unordered sets, dubbed *synsets*" (Fellbaum, 2012).

The most crucial relation among synsets is the super-subordinate relation (also called hyperonymy).

As shown in Fig. 4.1, all noun hierarchies ultimately go up to the entity root node. "The hyperonymy relation is transitive: If an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture" Fellbaum (2012).

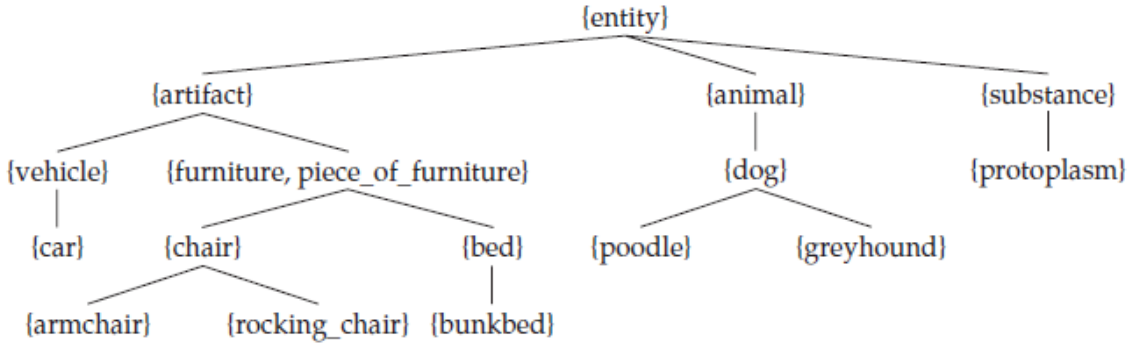


Figure 4.1: WordNet tree-structure example (Fellbaum, 2012)

Now, many versions of WordNet have been developed over the word for many languages, and even exists an organization that tracks all of them: The Global WordNet Organization

(www.globalwordnet.org).

In particular, a relevant effort was made in the production of MultiWordNet (Pianta et al., 2002), a version that not only connects the Italian language with the original English version of WordNet but also allows to compare (for similarities and divergences) corresponding entries of many languages, including Latin, Portuguese, and Spanish.

All of this explains why WordNet is so popular and why it is defined as a thesaurus, where "the arcs among words and synsets express a finite number of well-defined and labeled relations" (Fellbaum, 2010).

Moreover, the electronic format design allows automatic systems to detect and measure the semantic relatedness of words co-occurring in a context, facilitating alternative or complementary symbolic approaches to word sense discrimination (Fellbaum, 2010).

To summarize, in the WordNet structure:

- nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms: synsets;
- each synset expresses a distinct concept, and is interlinked by means of conceptual-semantic and lexical relations;
- the resulting network of related concepts and words can be navigated with the browser.

Moreover, despite not all the versions of the other languages are openly available, WordNet is also freely and publicly available for download.

Nevertheless, today there is a far better and automatic way (like Words Embeddings) to model the human semantic organization, but it is still an essential and valuable tool for research focused on Natural Language Processing. For example, it is essential to highlight that it has been developed as a public library (NLTK library, Loper and Bird (2002), Bird et al. (2009)) that uses the synset as an interface for accessing words in WordNet.

4.1.2 SentiWordNet

Publicly available on www.github.com/aesuli/SentiWordNet, SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. SentiWordNet is described in details in the papers: Esuli and Sebastiani (2006b), Esuli and Sebastiani (2006a), and Baccianella et al. (2010)

Developed by Esuli and Sebastiani (2006b), SentiWordNet is a lexical resource in which each WordNet synset is associated with three numerical scores: Obj(s), Pos(s), and Neg(s).

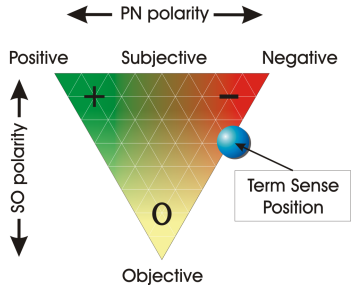
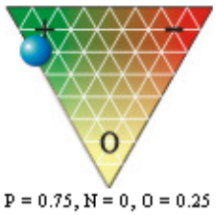
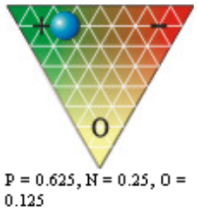


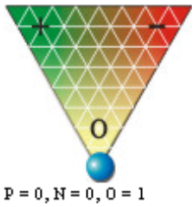
Figure 4.2: SentiWordNet graphical representation (Baccianella et al., 2010)



estimable(1)
deserving of respect of high regard



honourable(5) good(4) respectable(2) estimable(2)
deserving of esteem and respect;
"all respectable companies give guarantees";
"ruined the family's good name"



computable(1) estimable(3)
may be computed or estimated; "a calculable risk";
"computable odds";
"estimable assets"

Figure 4.3: SentiWordNet content example (Baccianella et al., 2010)

In Fig. 4.2 and Fig. 4.3 we can observe that those scores describe how objective, positive, and negative the synset terms are.

According to `ontotext.fbk.eu` (that partially supported the creation of Baccianella et al. (2010) project), typical use of SentiWordNet is to enrich the text representation in opinion mining (OM) applications, adding information on the sentiment-related properties of the terms in the text. OM is a recent subdiscipline at the crossroads of information retrieval, and computational linguistics, which is concerned not with the topic a document is about but with the opinion it expresses. OM has a rich set of applications, ranging from tracking users' opinions about products or about political candidates as expressed in online forums to customer relationship management. In order to aid the extraction of opinions from the text, recent research has tried to automatically determine the "PN-polarity" of subjective terms, i.e., identify whether a term that is a marker of opinionated content has

a positive or a negative connotation. Research on determining whether a term is indeed a marker of opinionated content (a subjective term) or not (an objective term) has been, instead much scarce. SentiWordNet is the first lexical resource that provides such a specific level of detail (the word sense represented by a synset) and such a broad coverage (all the 115,000+ WordNet synsets).

Before describing how this lexical database was generated, it is important to highlight that many upgrades have been made over the years on WordNet. For that, the authors have systematically improved SentiWordNet. The current version of SentiWordNet is 3.0, which is based on WordNet 3.0.

Since the last version is the most updated one, and it is the version that is used in this thesis (Chapters 5 and 6), that version will be discussed more in detail, highlighting the differences between the first version and the last one. According to the authors (Baccianella et al., 2010), the main differences are the following:

---

1. Version 1.0 consists of an annotation of the older WordNet 2.0, while version 3.0 is an annotation of the newer WordNet 3.0;

2. For SentiWordNet 1.0, automatic annotation was carried out via a weak-supervision, semi-supervised learning algorithm. Conversely, for SentiWordNet 3.0, the results of this semi-supervised learning algorithm are only an intermediate step of the annotation process, since they are fed to an iterative random-walk process that is run to convergence. SentiWordNet 3.0 is the output of the random-walk process after convergence has been reached;

3. Version 1.0 uses the glosses of WordNet synsets as semantic representations of the synsets themselves when a semi-supervised text classification process is invoked. That classifies the (glosses of the) synsets into categories Pos, Neg and Obj. In SentiWordNet 3.0 both the semi-supervised learning process (first step) and the random-walk process (second step), instead of the raw glosses, uses the manually disambiguated glosses from the Princeton WordNet Gloss Corpus2.

So, the generation of SentiWordNet 3.0 consists of 2 phases: a weak-supervision, a semi-supervised learning step, and a random-walk step.

The semi-supervised learning step "consists in turn of four sub-steps: (1) seed set expansion, (2) classifier training, (3) synset classification, and (4) classifier combination" (Baccianella et al., 2010).

During those four sub-steps, two small groups of sets of synsets are taken from WordNet: one group with all the synsets that contains seven "paradigmatically positive" terms and the other with seven "paradigmatically negative" terms. Then those groups are expanded following the relationships of WordNet, preserving or inverting the Pos/Neg proprieties. Then, using what can be defined as a "Bag of synsets" (similar to a Bag-of-Words, but with synsets instead of words), a ternary classifier is trained with the expanded groups to classify a synset in Pos, Neg, or Obj. Once the classifier is trained, all WordNet synsets are classified. Furthermore, to improve the classification performance, eight ternary classifiers are generated and trained. The final classification is the average of the Pos, Neg, or Obj values.

The random-walk step "consists of viewing WordNet 3.0 as a graph, and running an iterative, "random-walk" process in which the Pos(s) and Neg(s) (and, consequently, Obj(s)) values, starting from those determined in the previous step, possibly change at each iteration. The random-walk step terminates when the iterative process has converged" (Baccianella et al., 2010).

4.2 Words Embeddings

Nothing helps more to face the problem of the representation of the words in a formal way that is suitable from a computer than this sentence: "you shall know a word by the company it keeps" (Firth, 1957). In fact, like a child facing a new word, we are able to infer the meaning of a new word (or a symbol) considering his context. That is something that we start to do from when we are newborns. Babies learn their mother language in that way, and sometimes they surprise their parents while using unexpected words that probably they have to listen to just one single time. Despite ignoring his proper meaning, they use it in the correct context. This philosophy is the leading guide of this section and keeping that in mind definitely helps to understand the following concepts.

While trying to deal with the words-component of the language, scientists have tried many ways. We will now see an excursus that starts with the most simple representation, that involves the usage of basic statistics over a set of words (Bag-of-Words, BoW) and shows the most harder ones in a high-dimensional space. Neural Networks really improved this field, significantly improving the computer knowledge of the language and the words that compose it.

The main idea for such a representation is based on a older – but definitely still valid – linguistic theory: the "distributional hypothesis" of Firth (1957) and Harris (1954). Such a theory stated that if some words have a similar context, then they have a similar meaning.

While in simple representations such as BoW or a document matrix, we are able to interpret each dimension (e.g., each dimension corresponds to a particular language/letter), the neural network approach for producing a word embedding (so mathematical vectors from text source), increases the generalization power but prevents any interpretation of the dimensions. On the other hand, the distance between words (that now is possible to calculate) allows to easily generalize which word is similar (or not) to another one. The usage of such a representation has been proved to lead to superior classification accuracy (Kim, 2014).

4.2.1 Distributional Information

Discussing the common and simple choice of representing textual data, we have:

- bag-of-items: a collection of unique items, composed by all the items in the considered data;
- n-grams: a sequence of N consecutive items (a single letter, a pair, a word).

It should be intuitively clear why word-bigrams is more informative than individual words: it captures structures such as New York, not good, and Paris Hilton.

Indeed, a bag-of-bigrams representation is much more powerful than bag-of-words, and in many cases, proves very hard to beat. Of course, not all bigrams are equally informative; bigrams such as "of the", "on a", "the boy", are very common and, for most tasks, not more informative than their individual components. However, "it is very hard to know a-priori which n-grams will be useful for a given task" (Goldberg, 2017). The common solution is to include all n-grams up to a given length and let the model regularization discard the less interesting ones.

Up until now, our treatment of words was as discrete and unrelated symbols: the words "pizza", "burger", and "chair" are all equally similar (and equally dissimilar) to each other as far as the algorithm is concerned.

We achieve some form of generalization across word types by:

- mapping them to coarser-grained categories such as parts-of-speech or syntactic roles ("the", "a", "an", "some" are all determiners);
- generalizing from inflected words forms to their lemmas ("book", "booking", "booked" all share the lemma book);
- looking at membership in lists or dictionaries ("John", "Jack", and "Ralph" appear in a list of common U.S. first names);

- looking at their relation to other words using lexical resources such as WordNet.

However, this kind of approach lacks generalization power, and usually, they rely on ad hoc manually compiled dictionaries. In fact, if we do not have a detailed list of foods, we are not able to understand that a "burger" is more similar to a "pizza" rather than to a "chair".

Considering the aforementioned distributional hypothesis of Firth (1957) and Harris (1954), we should infer the meaning of a word taking into account the context in which it is used. For example, many clustering algorithms have used bigger co-occurrence matrices to discover that the context of "pizza" is more similar to those in which there is a "burger" rather than those of the "chair". Intuitively, "when people encounter a sentence with an unknown word such as the word "wampimuk" in *Marco saw a hairy little wampinuk crouching behind a tree*, they infer the meaning of the word based on the context in which it occurs" (Goldberg, 2017).

In contrast to the so-called count-based methods described above, the neural networks community advocates the use of distributed representations of word meanings. In a distributed representation, each word is associated with a vector in $\mathbb{R}^d$, where the "meaning" of the word with respect to some task is captured in the different dimensions of the vector as well as in the dimensions of other words. Unlike the explicit distributional representations in which each dimension corresponds to a specific context the word occurs in, the dimensions in the distributed representation are not interpretable, and specific dimensions do not necessarily correspond to specific concepts. The distributed nature of the representation means that a given aspect of meaning may be captured by (distributed over) a combination of many dimensions and that a given dimension may contribute to capturing several aspects of meaning. (Goldberg, 2017)

We are now going to formalize the creation of a words embedding representation with the Neural Networks, following the definition of Collobert and Weston (2008) as suggested by Goldberg (2017).

Let w be a target word, $c_{1:k}$ be an ordered list of context items (the context of a word is the k-gram of words preceding it, each word is associated with a vector, and their concatenation is encoded into a d_{hid} dimensional vector $\boldsymbol{h}$ using a non-linear transformation), and $v_w(w)$ and $v_c(c)$ embedding functions mapping word and context indices to d_{emb} dimensional vectors (from now on we assume the word and context vectors have the same number of dimensions).

The model computes a score $s(w, c_{1:k})$ of a word-context pair by concatenating the word and the context embeddings into a vector $\boldsymbol{x}$, which is fed into a Neural Network (a Multi-Layer Perceptron) with one hidden layer whose single output is the score assigned to the word-context combination:

$$s(w, c_{1:k}) = g(\boldsymbol{x}\boldsymbol{U}) \cdot \boldsymbol{v}$$

where

$$\boldsymbol{x} = [v_c(c_1); \ldots; v_c(c_k); v_w(w)]$$

$$\boldsymbol{U} \in \mathbb{R}^{(k+1)d_{emb} \times d_{hid}}$$

$$\boldsymbol{v} \in \mathbb{R}^{d_{hid}}$$

The network is trained with the following loss function to score correct word-context pairs $(w, c_{1:k})$ above incorrect word-context pairs $(w', c_{1:k})$ with a margin of at least 1. Such a function, defined as a margin-based ranking loss function $L(w, c_{1:k})$ for a given word-context pair, is given by:

$$L(w, c, w') = max(0, 1 - (s(w, c_{1:k}) - s(w', c_{1:k})))$$

Where w' is a random word from the vocabulary. The training procedure repeatedly goes over word-context pairs from the corpus, and for each one samples a random word w', compute the loss $L(w, c, w')$ using w', and updates the $\boldsymbol{U}, \boldsymbol{v}$ parameters, the word, and context embeddings to minimize the loss.

Now we will describe the most famous and popular algorithms that implement such a process for creating a words embedding representation suitable for a Natural Language Processing task: Word2Vec, Continuous Bag-of-Words Model (CBOW), Skip-gram Model, GloVe.

4.2.2 Google: Word2Vec

The Word2Vec algorithms are very effective in practice and are highly scalable, allowing to train word representations with huge vocabularies over billions of words of text in a matter of hours, with very modest memory requirements. The connection between the SGNS variant (Skip-Gram Negative-Sampling) of Word2Vec and word-context matrix-factorization approaches ties the neural methods and the traditional "count-based" ones, suggesting that

lessons learned in the study of "distributional" representation can transfer to the "distributed" algorithms and vice versa, and that in a deep sense, the two algorithmic families are equivalent. (Harris, 1954)

Inside the textscWord2Vec software package, the Word2Vec algorithm was developed by Mikolov et al. (2013b) and Mikolov et al. (2013a) at Google. Like the previous state-of-the-art algorithm of Collobert and Weston (2008), Word2Vec also starts with a neural language model and modifies it to produce faster results. Moreover, textscWord2Vec is not a single algorithm: it is a software package implementing two different context representations (Continuous Bag-of-Words and Skip-Gram) and two different optimization objectives (Negative-Sampling and Hierarchical Softmax). Here, for comparison purposes, we focus on the Negative-Sampling objective (NS).

Like the Collobert and Weston (2008) algorithm, the NS variant of Word2Vec works by training the network to distinguish "good" word-context pairs from "bad" ones. However, Word2Vec replaces the margin-based ranking objective with a probabilistic one. Consider a set D of correct word-context pairs, and a set $\bar{D}$ of incorrect word-context pairs. The goal of the algorithm is to estimate the probability $P(D = 1|w, c)$ that the word-context pair (w, c) came from the correct set D. This should be high (P=1) for pairs from D and low (P=0) for pairs from $\bar{D}$. The probability constraint dictates that $P(D = 1|w, c) = 1 - P(D = 0|w, c)$. The probability function is modelled as a sigmoid over the score $s(w, c)$:

$$P(D = 1|w, c) = \frac{1}{1 + e^{-s(w,c)}}$$

The corpus-wide objective of the algorithm is to set the parameters Θ s.t. maximize the log-likelihood of the data $D \cup \bar{D}$:

$$L(\Theta; D, \bar{D}) = \sum_{(w,c)\in D} \log P(D = 1|w, c) + \sum_{(w,c)\in \bar{D}} \log P(D = 0|w, c)$$

The positive examples D are generated from a corpus, while instead, the negative examples $\bar{D}$ can be generated in many ways. In Word2Vec, they are generated in this way: for each good pair $(w, c) \in D$, sample k words $w_{1:k}$ and add each of (w_k, c) as a negative example to $\bar{D}$.

This results in the negative samples data $\bar{D}$ being k times larger than D. The number of negative samples k is a parameter of the algorithm. Moreover the negative words w can be sampled according to their corpus-based frequency

$$\frac{\#(w)}{\sum_{w'} \#(w')}$$

or, as done in the Word2Vec implementation, with a smoothed version:

$$\frac{\#(w)^{0.75}}{\sum_{w'} \#(w')^{0.75}}$$

According to the authors' empirical results, the usage of the constant 0.75 produces a smoothed version that gives more relative weight to less frequent words and results in better word similarities.

4.2.3 Continuous Bag-of-Words Model (CBOW)

Other than changing the objective from margin-based to a probabilistic one, Word2Vec also considerably simplifies the definition of the word-context scoring function, $s(w, c)$. For a multi-word context $c_{1:k}$, the CBOW variant of Word2Vec defines the context vector $\boldsymbol{c}$ to be a sum of the embedding vector of the context components: $\boldsymbol{c} = \sum_{i=1}^{k} \boldsymbol{c_i}$. It then defines the score to be simply $s(w, c) = \boldsymbol{w} \cdot \boldsymbol{c}$, resulting in:

$$P(D = 1|w, c_{1:k}) = \frac{1}{1 + e^{-(\boldsymbol{w}\cdot\boldsymbol{c_1}+\boldsymbol{w}\cdot\boldsymbol{c_2}+\cdots+\boldsymbol{w}\cdot\boldsymbol{c_k})}}$$

Recently, Choudhari and Veenadhari (2020) well summarized it saying that: "Word2Vec is a shallow two-layer neural network where there is an input layer, hidden layer, and output layer... The network is trained based on the maximum likelihood principle... the CBOW model finds the probability of the target word given the neighboring words and then tries to maximize the probability".

4.2.4 Skip-Gram Model

The Skip-Gram variant of Word2Vec scoring decouples the dependence between the context elements even further. For a k-elements context $c_{1:k}$, the skip-gram variant assumes that the elements c_i in the context are independent from each other, essentially treating them as k different contexts, i.e., a word-context pair $(w, c_{i:k})$ will be represented in D as k different contexts: $(w, c_1), \ldots, (w, c_k)$. The scoring function $s(w, c)$ is defined as in the CBOW version, but now each context is a single embedding vector:

$$\begin{aligned}
P(D = 1|w, c_i) &= \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{c_i}}} \\
P(D = 1|w, c_{1:k}) &= \prod_{i=1}^{k} P(D = 1|w, c_i) = \prod_{i=1}^{k} \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{c_i}}} \\
\log P(D = 1|w, c_{1:k}) &= \sum_{i=1}^{k} \log \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{c_i}}}
\end{aligned}$$

While introducing strong independence assumptions between the elements of the context, the skip-gram variant is very effective in practice and very commonly used.

4.2.5 GloVe

"Many variants on the Word2Vec algorithms exist, none of which convincingly produce qualitatively or quantitatively superior word representations" Goldberg (2017).

The GloVe algorithm (Pennington et al., 2014) constructs an explicit word-context matrix, and trains the word and context vectors $\boldsymbol{w}$ and $\boldsymbol{c}$ attempting to satisfy:

$$\boldsymbol{w} \cdot \boldsymbol{c} + \boldsymbol{b}_{[w]} + \boldsymbol{b}_{[c]} = \log \#(w, c) \qquad \forall (w, c) \in D$$

where $\boldsymbol{b}_{[w]}$ and $\boldsymbol{b}_{[c]}$ are word-specific and context-specific trained biases. The optimization procedure is based on a weighted least-squares loss and, while focussing on observed word context pairs and skipping zero count events, it assigns more weight to the correct reconstruction of frequent items.

Chapter 5

General Sentiment Decomposition

Being the core component of this thesis, this chapter highlights the methodology used for analyzing the data described in Chapter 2 together with the techniques that are possible to find in Chapters 3 and 4. Here, we provide answers to the following questions:

- which type of framework is used for analyzing textual data? (Section 5.1)
- which one are we going to use? (Section 5.2)
- is it possible to use the raw data directly, or some necessary preprocessing steps are needed? What should we do first before analyzing such a type of data? (Section 5.3)
- which is the innovative component concerning the other approaches? (Section 5.4)
- what kind of model are the authors using? And why? (Section 5.5)
- how do we interpret the results that comes from the usage of the selected model? (Section 5.6)

Starting from Natural Language text corpora, considering data related to the same context, we define a process to extract the sentiment component with a numeric transformation. Considering that the Naïve Bayes model, despite its simplicity, is handy in related tasks such as spam/ham identification, we have created an improved and adapted version to solve a NLP task: Threshold-based Naïve Bayes (Romano et al. (2018) and Conversano et al. (2019)).

The new version of the Naïve Bayes classifier has proven to be really good with respect to the standard version and the other most commons classifiers (Fig. 5.7). However, according to our previous studies, we have a strong limit:

- A response variable is needed: we need to know a priori the "Positive" ("Negative") label of a consistent amount of comments in the data;
- There is some heavy handwork: consistently reducing the problem's dimensionality is

a keystone for a sentiment classification task. That means to "merge words by their meanings", and usually, it is done by hand. This leads to major problems in terms of subjectivity while those words are merged.

From the literature review, we discover that, except for really new and few studies, they solve those problems (that we can face, i.e., in a tweet sentiment classification task) with handwork. For example, they usually manually classify a certain proportion of tweets, and then they train their model with this hand-constructed response variable. Nevertheless, this leads the authors to put some subjectivity inside the data. Let us think about a practical example: "They have put some pineapple in my pizza". That sentence might appear as positive if the authors are not Italians, or more appropriately should sound like a nightmare: definitely negative. "*De gustibus non disputandum est*".

This toy example highlights that doing handwork produces subjectivity problems. For that computer scientists, usually prefer to adopt other types of models based on Neural Networks or using Words Embeddings representations of the data for their model. However, in this strategy, they use a hand-constructed response variable, and the produced output is less interpretable (if not impossible).

We propose then a solution to deal with this subjectivity problem. While generalizing the concept that we have developed in our previous papers (Romano et al. (2018); Conversano et al. (2019)), we use SentiWordNet to automatically (and objectively) produce a temporary [-1; 1] sentiment score. This not only produces a not any more hand-constructed response variable, like what we have found in the literature review but allows us to identify (and remove) noise (for example, neutral sentiment comments) from the training set for our Threshold-based Naïve Bayes model. Furthermore, to reduce the problem's dimensionality, we use a K-means cluster analysis over the Words embedding representation of the data both before the training phase and in the interpretation phase. That allows us to "merge words by their meaning", similarly to our previous related works, but automatically and objectively.

The versatility of the Threshold-based Naïve Bayes model output permits us to rearrange it in categories of words of interest (with a substantially lossless transformation). Moreover, while plotting those categories data in a time-series, it is possible to observe how trends evolve through time. With that, the interpretation of the results has been simplified and made more objective and valuable, achieving the possibility of exploring interdisciplinary fields and obtaining results that can improve environmental, political, business, and social area studies.

5.1 Literature Review

Following the preface, there are not many methods or algorithms related to this kind of new approach that works not only for labeled datasets but also for unlabelled ones. Those that are *close* to the GSD framework adopt a clustering technique over a Words Embedding representation of a labeled dataset only. Next, they feed that clustering-output data as an input to a Machine Learning algorithm for classify the unlabeled data.

Example of interesting related works are: Choudhari and Veenadhari (2020), Mudinas et al. (2018), Acosta et al. (2017), Alshari et al. (2017), and Zhang et al. (2015). To have a better idea of what the authors of those works do and summarize all those approaches, we will now compare them. First, we will analyze their frameworks, and then we will rearrange them in a table to highlights the main similarities and differences.

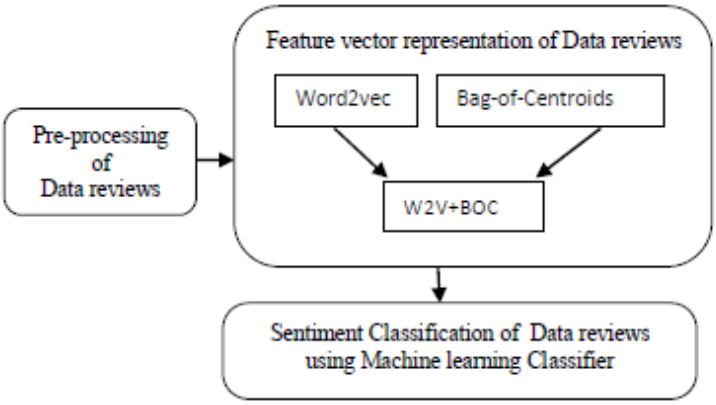


Figure 5.1: The Choudhari and Veenadhari (2020) framework

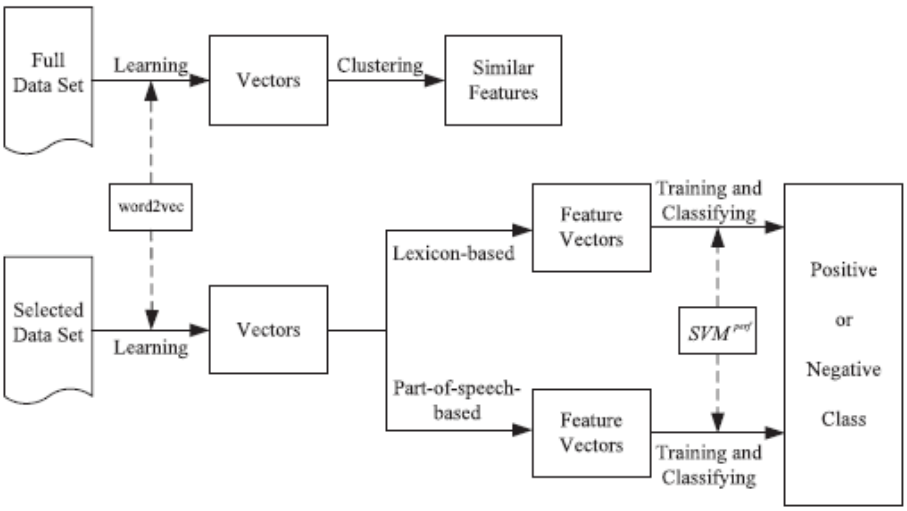


Figure 5.2: The Zhang et al. (2015) framework

As it is possible to notice in Fig. 5.1 and in Fig. 5.2, the works of Choudhari and Veenadhari (2020) and Zhang et al. (2015) both use Word2Vec to create a Words Embedding of their data. Moreover, they use Cluster Analysis over the vectorial representation of the words in their dataset to improve the performance of the chosen Machine Learning classifier. Despite Acosta et al. (2017) authors did not provide a referenced framework, they adopt a similar strategy for unlabelled data. However, they first classify the data with some handwork, and they do not use Cluster Analysis.

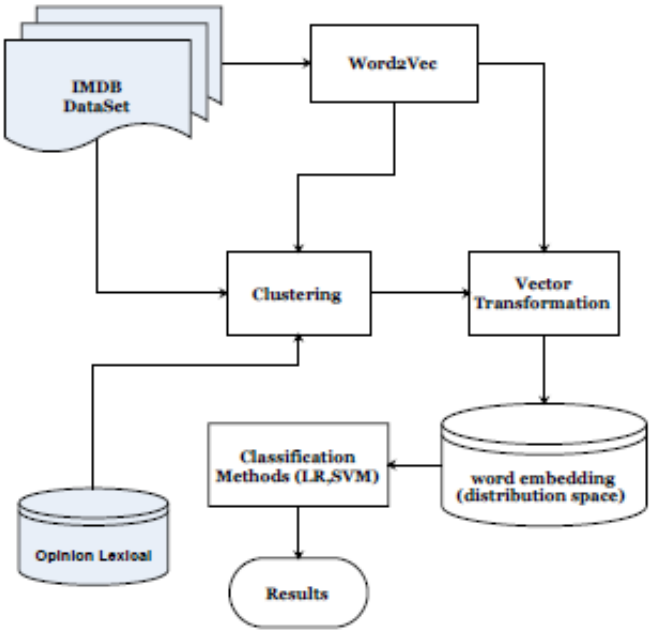


Figure 5.3: The Alshari et al. (2017) framework

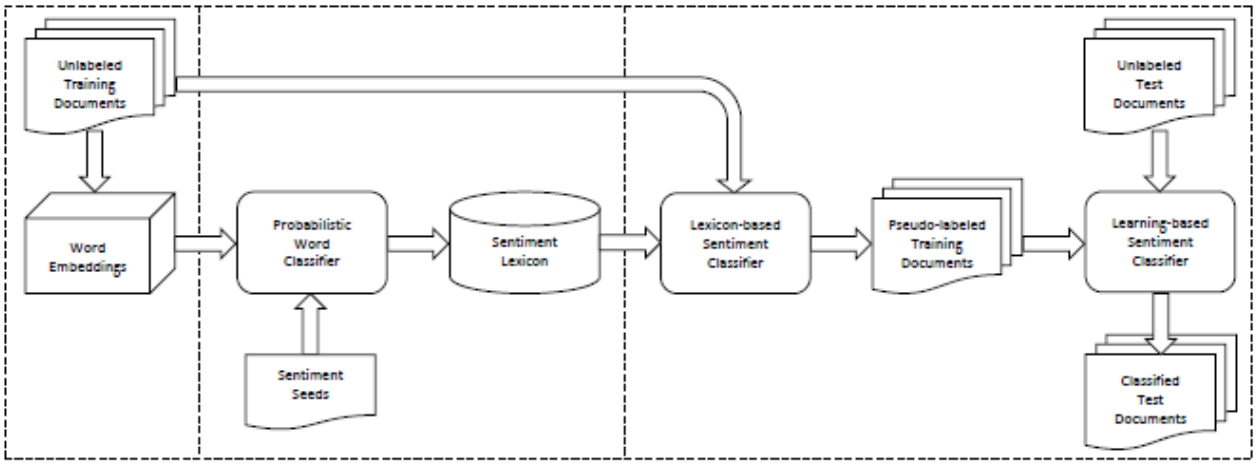


Figure 5.4: The Mudinas et al. (2018) framework

The works of Alshari et al. (2017) (labelled data) and Mudinas et al. (2018) (unlabelled data), that we see represented in Fig. 5.3 and Fig. 5.4, are more sophisticated. In fact, not only they use Word2Vec to produce the vectorial representation of their words and apply Cluster Analysis to improve their results, but they place the focus on the usage of a Lexical Database to increase either the clustering effectiveness and for pre-classifying the data, abolishing handwork.

Article	Type of Data	Hand-made Response Variable	Lexical Database	Words Embeddings	Clustering	ML Models
Choudhari and Veenadhari (2020)	Labelled	No	No	Word2Vec (SG & CBOW)	K-Means	**LRCV**, MLP, RF, DT, GNB
Zhang et al. (2015)	Labelled	No	No	Word2Vec (SG)	Cosine similarity, calculated by the Word2Vec tool	**SVM**
Acosta et al. (2017)	Unlabelled	Yes	No	Word2Vec (SG)	No	**LR**, **SVM**
Alshari et al. (2017)	Labelled	No	Yes	Word2Vec (SG)	Cosine similarity, calculated by the Word2Vec tool	**LR**, SVM
Mudinas et al. (2018)	Unlabelled	No	Yes	Word2Vec (SG)	Used but not specified	**SVM**, pSenti, ProbLex-DCM, CNN, LSTM

Table 5.1: Related works comparison

With Table 5.1 we are able to compare all of those strategies for many perspectives. More in detail:

- Type of data: Labelled or Unlabelled that allows us to know if the original textual data was already provided with a Positive/Negative sentiment label, suitable for the analysis;
- Hand-made response variable: to know if the authors define, with some handwork, a response variable if the data did not have any label (i.e., manually classifying some tweets with Positive/Negative label);
- Lexical Database: if the authors use or not some lexical databases for preprocessing the data and improving the performance of the classifier;
- Words Embeddings: representing if the authors are using at least one Words Embedding technique. All of them are using the Word2Vec algorithm (SkipGram or Continuous-Bag-Of-Words version);
- Clustering: allowing to know if a clustering technique is used in the related work and which type of Cluster analysis (if specified by the authors);

- ML Models: to know which Machine Learning model has been used in the studies. In bold, the model that the authors reported to have better performances among the others treated in their analysis.
  The used acronyms represent the following models:
  - LR: Logistic Regression,
  - LRCV: Logistic Regression CV,
  - RF: Random Forest,
  - DT: Decision Trees,
  - GNB: Gaussian Naïve Bayes,
  - pSenti: a concept-level lexicon-based sentiment, classifier (Mudinas et al. (2012)),
  - ProbLex-DCM: probabilistic lexicon-based, classification using the Dirichlet Compound Multinomial (DCM) likelihood to reduce effectively counts for repeated words (Eisenstein (2017)),
  - MLP: Multi Layer Perceptron,
  - CNN: Convolutional Neural Network,
  - LSTM: Long Short-Term Memory, a Recurrent Neural Network (RNN) that can remember values over arbitrary time intervals (Hochreiter and Schmidhuber (1997)).

5.2 GSD main features

While going more deeply into the philosophy of our proposed General Sentiment Decomposition method, which will be defined more in detail within the following sections, we can observe the Fig. 5.5. There are two inputs: "natural language text" corpora and "fields of interest" from the user, but the mandatory one is just the first one. Starting from the so defined text, we check if we face a supervised learning problem (so we have a response variable: labeled text) or not (we do not have a response variable, a.k.a. unlabeled text). If we have a labeled dataset, we go directly to the next step; otherwise, the process follows a pre-classification of the data for automatically (and objectively) produce a temporary label. Such a pre-classification phase use SentiWordNet over the Words Embedding representation of the words in the input dataset. The next step consists of using the Threshold-based Naïve Bayes classifier (Chapter 5.4) with the labeled (or temporary labeled) data, obtaining a quantitative sentiment value for each of the words that are presents in the input dataset. The final step is the interpretation phase, in which the numeric information is rearranged to produce time-series and other types of valuable and interesting plots for the final users.

As we can notice, while comparing the General Sentiment Decomposition logic (Fig. 5.5) with the Table 5.1, the most related paper is Mudinas et al. (2018), that also works for unlabelled data and have the most similar framework to the one of this thesis.

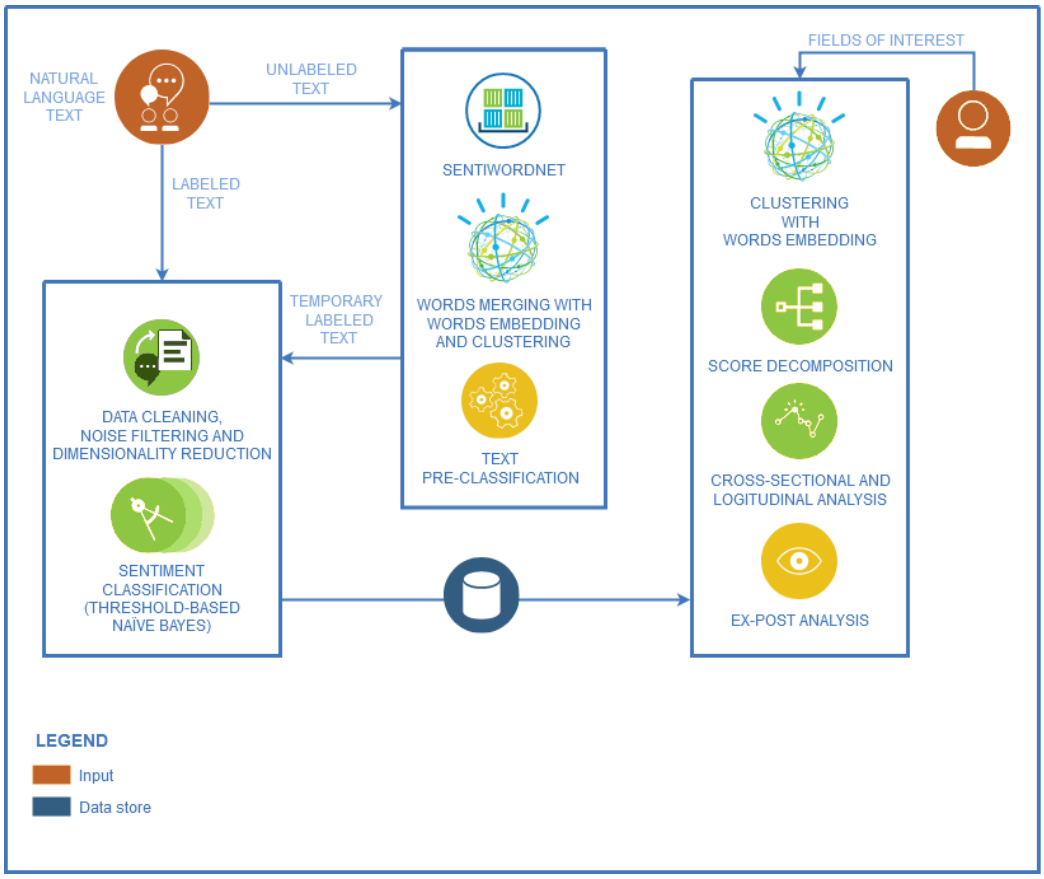


Figure 5.5: General Sentiment Decomposition Logic

It is also mandatory to mention that a really recent work Ahmed et al. (2020) could provide an improvement on the performance of the actual process; in particular, here, the authors create an alternative sentiment lexical database to SentiWordNet. In fact, they have compared their performance with the new sentiment lexical database that they have created to SentiWordNet, and results suggest that their proposed version is providing much better performance. We are planning, as future work, to include this as an improvement of the process described in this thesis.

5.3 Data Cleaning

The raw data that is downloaded is usually not suitable for the analysis. It has many unnecessary words like stop words that do not explain the meaning of the sentence and acronyms whose meaning is difficult to decipher and hence tend to confuse the algorithm.

Moreover, it contains emojis which have helpful information, so they have to be converted into meaningful text. This phase is the first step for preprocessing the data and make it worthwhile for the analysis. Below mentioned are the details of all the processes used for the data cleaning phase in a subsequent way for every single observation in the dataset.

1. preprocessing: some basic – but necessary – filtration is done before going to the next step. It is important to remove links (especially if just a partial one), acronyms (as their meaning is difficult to decipher), or recurrent and meaningless keywords (like RT (re-tweets), **@**username, uninterested #hashtags);

2. emoticons conversion: such valuable information, especially regarding the sentence sentiment polarity, is contained inside emoticons (like :-) or :-( ) and in emojis (like ☺ or ☹). In order to consider them in the same way, conversion of emoticons to emoji is needed.

3. emoji replacement: Once all emoticons have become emojis, the next step is to replace the emojis with their corresponding meaning so that they can be further treated and analyzed together within the normal text. In that way, all the meaningful symbols are now converted into text.

4. stop words & alphanumeric characters: the incoming text is first tokenized into separate words, and any punctuation adjacent to the words is also separated. Thereafter, these punctuation symbols, along with some alphanumeric characters that might be present, are then detected and removed. Cases of all alphabets are normalized to lowercase. Next, stop words like “a”, “the”, “do”, “to” are removed from the data as they do not provide any valuable information about the text’s sentiment. However, negative words like “not” are kept as they completely alter a sentence’s meaning and profoundly impact its sentiment.

5. stemming: as the last data cleaning phase, the tokens are stemmed; in other words, they are reduced to their root or base form. For example, “fishing,” “fished,” “fisher” are all reduced to the stem “fish”. In that way, we merge words that are related to the same topic by their root or base form.

5.4 SentiWordNet & Words Embedding combination

The main goal of Words Embedding is to reduce the dimensionality of text data. In order to achieve this goal, we go through the hypernyms and lemmas phases. We will now recall a few fundamental concepts and terminologies to understand the next steps better:

- synsets: As defined earlier, these are a collection of words that have a similar meaning. These inbuilt vectors of words are used to find out to which synset a certain word belongs.

- hypernyms: These are more abstract terms concerning the name of particular synsets. While organizing synsets in a tree-like structure based on their similarity to each other, the hypernyms allow to categorize and group words. In fact, such a structure can be traced all the way up to a root hypernym.

- lemmas: A lemma is a WordNet's version of an entry in a dictionary: A word in canonical form, with a single meaning. E.g., if someone wanted to look up "mouses" in the dictionary, the canonical form would be "mouse", and there would be separate lemmas for the nouns meaning "animal" and "pc component";

- merging words by their meaning: we iterate through every word of the received text, and, for everyone, we fetch the synset to which it belongs to. Using the synset name, we fetch the hypernym related to that word. Finally, the hypernym name is used to find the most similar word, replacing the actual word in the text.

So, while using the hypernyms proprieties, we adopt a newspaper pre-trained Words Embedding produced by Google with Word2Vec SG for obtaining the vectorial representation of all the words in the dataset (after the data cleaning process). Moreover, to finalize the "merging words by their meaning" step, we use the K-Means clustering technique.

In fact, it produces a number of clusters λ and computes the centroid-word as the word that replaces all the other words present in that cluster. In this way, the model is trained with a Bag-of-Centroids (of the clusters produced over the Words Embedding representation of the dataset) instead of a general Bag-of-Words.

The λ-value can be estimated by a cross-validation process, calculating the best accuracy (or other performance metrics) within a labeled dataset (E.g., Booking.com or TripAdvisor data). More details can be found in Chapter 6.

Once the data is correctly cleaned and all the words with the same meaning are merged in a single one, it is finally possible to calculate each observation's total sentiment score.

For this purpose, the Lexical Database SentiWordNet permits to obtain the positive (pos_score) as well as the negative score (neg_score) of a particular word. The sentiment score (neg_score – pos_score) allows us to determine the polarity of each word. So, the overall score of a particular observation (i.e., a comment, a review, a tweet) is defined as the average of all the scores of all the words present in the parsed observation.

In that way, with this strategy, we have created a temporary sentiment label while using a simple threshold over the so produced overall score. Such a temporary label is the valuable base for training the Threshold-based Naïve Bayes Classifier.

5.5 Threshold-based Naïve Bayes Classifier

This classifier is defined in a labeled context in which each text corpora is composed of two components: one positive and one negative (like in the case of Booking.com data, Section 6.1). Forby, we formalize it as follows.

Considering a Natural Language text corpora as a set of reviews $\boldsymbol{r}$ s.t.:

$$r_i = comment_{pos_i} \cup comment_{neg_i}$$

where $comment_{pos}$ ($comment_{neg}$) are set of words (comments) composed by only positive (negative) sentences, and one of them can be equals to an empty set $\emptyset$. An ad-hoc classifier has been implemented able to predict, as accurately as possible, a comment as negative or positive based on the words included in its content. This Threshold-based Naïve Bayes classifier derives from a modification of the original Naïve Bayes classifier having the same name and resulted as the best performing one in terms of generalizability among several of the most commonly used classifiers. The essential features of Threshold-based Naïve Bayes applied to reviews' content are as follows. For a specific review r and for each word w ($w \in$ *Bag-of-Words*), we consider the log-odds ratio of w,

$$\begin{aligned} LOR(w) &= \log\left[\frac{P(c_{neg}|w)}{P(c_{pos}|w)}\right] \approx \\ &\approx \log\left[\frac{P(w|c_{neg})}{P(\bar{w}|c_{neg})} \cdot \frac{P(w|c_{pos})}{P(\bar{w}|c_{pos})} \cdot \frac{P(c_{neg})}{P(c_{pos})}\right] = \ldots = \\ &\approx pres_w + abs_w \end{aligned}$$

where $c_{pos}(c_{neg})$ are the proportions of observed positive (negative) comments whilst $pres_w$ and abs_w are the log-likelihood ratios of the events ($w \in r$) and ($w \notin r$), respectively.

Likewise, for the set of J words included in a comment c, the log-odds ratio of c is defined as:

$$\begin{aligned} LOR(c) &= \sum_{w_i \in J} pres_{w_i} + \sum_{w_{i'} \notin J} abs_{w_{i'}} = \\ &= \sum_{w_i \in \mathcal{J}} \log P(w_i|c_{neg}) - \log P(w_i|c_{pos}) + \sum_{w_{i'} \notin \mathcal{J}} \log P(\bar{w_{i'}}|c_{neg}) - \log P(\bar{w_{i'}}|c_{pos}) + \\ &\quad + \log P(c_{neg}) - \log P(c_{pos}) \end{aligned}$$

While calculating those values for all the w ($w \in$ *Bag-of-Words*) words, it is possible to obtain an output such that reported in Table 5.2, where we have c_{pos}, c_{neg}, $pres_w$ and abs_w for each word included in a *Bag-of-Words*.

	w_1	w_2	w_3	w_4	w_5	...
$P(w_i\|c_{neg})$	0.011	0.026	0.002	0.003	0.003	...
$P(w_i\|c_{pos})$	0.007	0.075	0.005	0.012	0.001	...
$pres_{w_i}$	0.411	-1.077	-1.006	-1.272	1.423	...
abs_{w_i}	-0.004	0.052	0.003	0.008	-0.002	...

Table 5.2: Example of a Threshold-based Naïve Bayes output

We then use cross-validation to estimate a parameter τ such that: c is classified as "negative" if $LOR(c) > \tau$ or as "positive" if $LOR(c) \leq \tau$.
According to the type of data and the goal of the analysis, the value of τ can be chosen in at least three ways: minimizing only the Type I error, minimizing only the Type II error, minimizing them both at the same time. For example, while considering Booking.com data (Section 6.1) in Fig. 5.6, the selected value of τ ($\hat{\tau} = 1.138$) is that minimizing simultaneously both the Type I and the Type II errors.

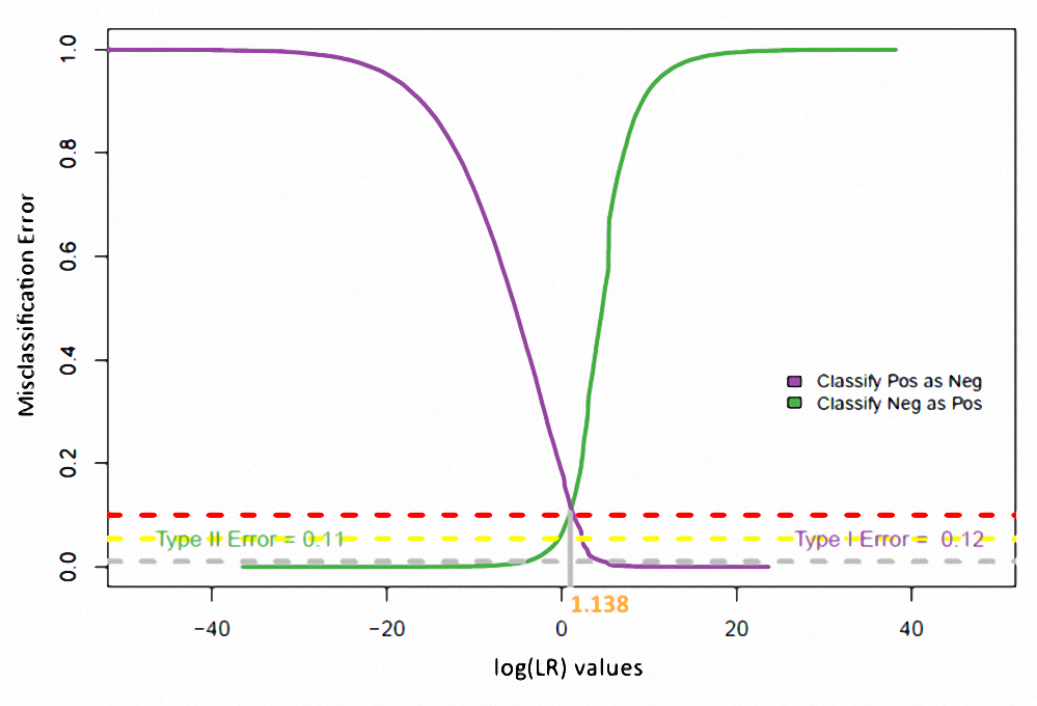


Figure 5.6: *LOR* values vs Misclassification Error
for minimizing Type I and Type II Errors. Example of data retrieved from Booking.com

The same approach can be used for the set of K words composing a review r, thus

computing $pres_{w_i}$ and $abs_{w_{i'}}$ for all the words appearing and not appearing in the review, and comparing $LOR(r)$ with the value of τ obtained from the classification of the comments into "positive" and "negative".

To benchmark Threshold-based Naïve Bayes, we compared its prediction accuracy (estimated with Cross Validation) when classifying comments as positive or negative with that of alternative classifiers, in particular: Logistic Regression, Random Forest, standard Naïve Bayes, Decision Trees and Linear Discriminant Analysis. While considering both the use cases of Chapter 6, and using the text data only, The Threshold-based Naïve Bayes classifier performed considerably better than competitors as it provided a Matthews correlation coefficient (Accuracy) of 0.813 (0.9111) versus an average value of 0.327 (0.8081) obtained from the alternatives as shown in Table 5.3 and Fig. 5.7.

Model	*Accuracy*	*Sensitivity*	*Fall-out*	*F1 score*	*Matthews Correlation Coefficient*
Tb-Naïve Bayes	**0.911**	**0.929**	**0.117**	**0.926**	**0.813**
Logistic	0.850	0.884	0.532	0.877	0.361
Random Forest	0.811	0.873	0.591	0.849	0.303
Naïve Bayes (e107)	0.806	0.804	0.389	0.834	0.390
Naïve Bayes (klaR)	0.806	0.804	0.389	0.834	0.390
CART	0.768	0.842	0.587	0.815	0.272
LDA	0.764	0.860	0.641	0.816	0.246

Table 5.3: benchmarking of the Threshold-based Naïve Bayes Classifier

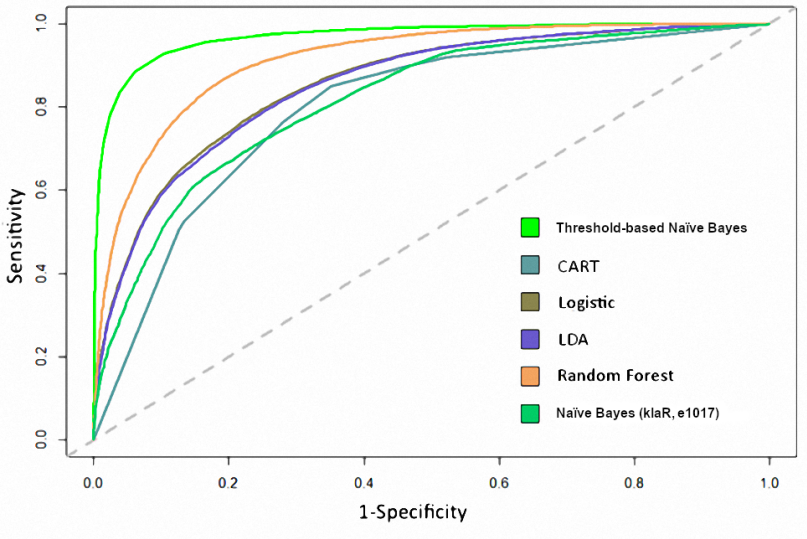


Figure 5.7: ROC curve for benchmarking of the Threshold-based Naïve Bayes Classifier

5.5.1 Iterative Threshold-based Naïve Bayes – A simple, but robust improvement

We knew now that the Threshold-based Naïve Bayes Classifier has engaging performances with respect to other algorithms. However, still, there are some problems during the classification in Pos/Neg of text data that have a LOR-value close to the selected value of τ. In particular, we can observe this in Fig. 5.8: more the value is far from the estimated threshold, then lower are the chances to make a wrong classification.

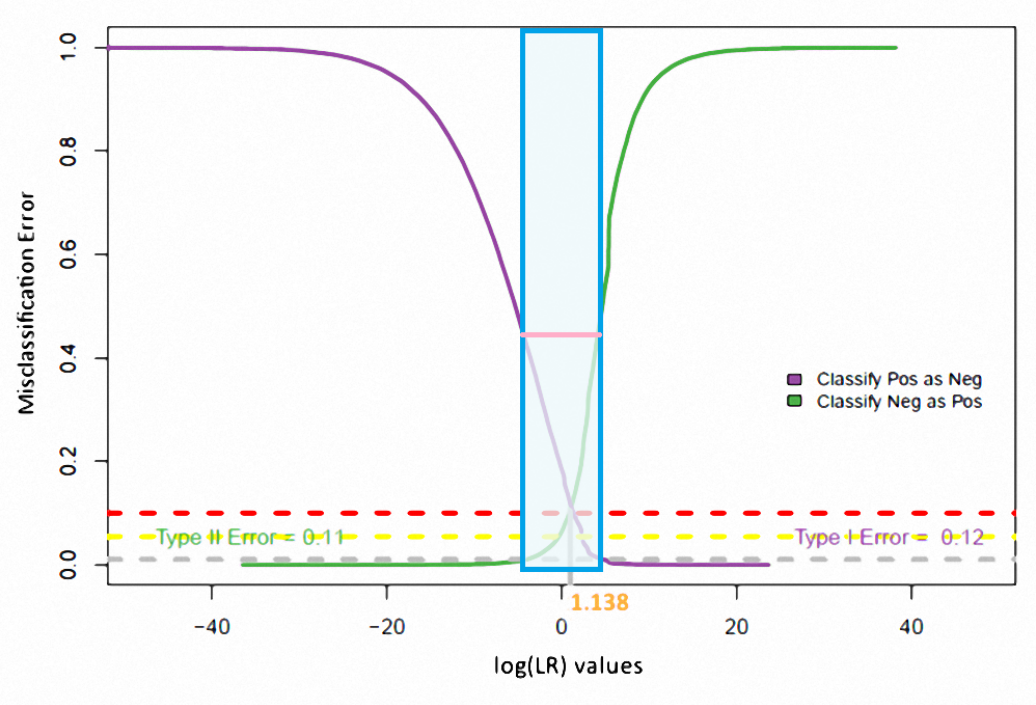


Figure 5.8: Threshold-based Naïve Bayes uncertainness area

Considering that the LOR-value is the numeric sentiment value that we have calculated, then it is pretty obvious to understand that if this value is close to the threshold, then even a word with an irrelevant sentiment value could reverse the classification. Furthermore, if we have stronger sentiment numeric information, then it is easy to classify the comments.

To improve the performance of the Threshold-based Naïve Bayes Classifier, we need to add a new step for refining the threshold while classifying only those comments that are close to the selected value of τ. Then, we classify again those comments with a new estimated value of τ.

The "augmented" Iterative Threshold-based Naïve Bayes works as follows:

1. a proportion of observations close to τ, and located in the uncertainness area, are marked for being reclassified

2. A new decision rule is created for those observations only, estimating a new value of τ

3. Observations in the uncertainness area are re-classified and Steps 1-3 are repeated until the proportion of cases in the uncertainness area reduces consistently

We can then transform that logic in a algorithm that better describes such a process:

---

1. Map the LOR-value probabilities distributions into functions:
   $f_{pos}(x), f_{neg}(x) \qquad \forall x \in [0; 1]$

2. Compute the intersection point i s.t. $f_{pos}(x) = f_{neg}(x') \qquad \forall x \in (0 + \epsilon \approx 0; 1]$

3. Define max_{pos}, max_{neg} s.t. $f_{pos}(max_{pos})$ and $f_{neg}(max_{neg})$ return the absolute max of f_{pos}, f_{neg}

4. if $max_{pos} < max_{neg}$ then update the decision rule:
   $LOR(\pi_c) > i \rightarrow$ c is classified as "Negative"
   $LOR(\pi_c) \leq i \rightarrow$ c is classified as "Positive"

5. otherwise, if $max_{pos} > max_{neg}$ then update the decision rule:
   $LOR(\pi_c) > i \rightarrow$ c is classified as "Positive"
   $LOR(\pi_c) \leq i \rightarrow$ c is classified as "Negative"

6. Repeat all the steps with the a proportion p of observations that are close to i until a stopping criteria is satisfied.

The proportion of observations marked for being re-classified, is estimated while considering the uncertainness area around τ. Using a too much larger p produces similar results of the Threshold-based Naïve Bayes classifier, whilst using a too much lower p might considering only a few observations, thus producing an over-fitting decision rule. Empirically, a good trade-off is reached while using $p = 0.20$.

Furthermore, we need to understand when we should stop the iterative process. For that, there are some considerations:

1. we need an acceptable amount of data (positive and negative comments) for being able to produce a reliable distribution (and the associated function that map it);

2. if there is no intersection point between $f_{pos}(x),\ and\ f_{neg}(x)$ then the distributions are well separated, and we can just use the τ estimated by the standard Threshold-based Naïve Bayes Classifier at the previous step to classify those comments;

3. if $max_{pos} = max_{neg}$ then we cannot distinct the positive distribution from the negative one. Forby, there is not a geometric solution to this problem and the classification of those comments has to be done with the τ estimated by the Threshold-based Naïve Bayes Classifier.

All of those situations have to be considered as stop criteria for the algorithm. Moreover, it is interesting to highlight how empirical results suggest that the stop criteria usually is reached after no more than two iterations.

5.5.2 Performance metrics for different scenarios

We have to consider that a classifier's performances might vary while changing the datasets' size, thus the amount of training and test data. Forby, while considering Booking.com data (Section 6.1), we have estimated the performance indicators for the considered classifiers in some relevant scenarios. In fact, we consider all the combinations of the following specifics:

- sample size (N): 20.000, 50.000, 100.000;
- training-test set proportions: 50–50, 67–33, 80–20;
- removing (or not) short comments (with less than three words).

Hence, we estimate the performance indicators 100 times while resampling every time from the original dataset of 106.800 observations that was used for assessing the performance while developing the Threshold-based Naïve Bayes classifier (Booking.com data, section 6.1). Forby, in Table 5.4 and in Table 5.5 we can observe that Tb-NB and ITb-NB have good performances in most of the cases. Moreover, their results are more relevant when short comments are removed. Therefore, despite having good performance with N = 20.000, the proposed classifier outperforms the others also when N = 100.000. Furthermore, many other tables like the two that follow have been produced while taking into account the other scenarios. Nevertheless, considering that results do not vary enough respect to those reported in Table 5.4 and Table 5.5, we have not inserted them in this thesis.

5.5.3 Robustness

In the context of sentiment analysis, we classify NL text comments into *positive* or *negative*. Forby, while training our model, we consider a binary response variable. Moreover, we define a numeric sentiment value (LOR-value), and we estimate two probabilities distributions: one for the *positive* labeled comments, and one for the *negative* labeled ones. Furthermore, if those distributions are well separated, then even a simple classifier will show good performances. Otherwise, even a complex classifier has some performance problems

N.		20.000					50.000					100.000				
SAMPLE	MODEL	PROP	ACC	TPR	TNR	F1	PROP	ACC	TPR	TNR	F1	PROP	ACC	TPR	TNR	F1
LEARNING	TB NB	80	0.878	0.910	0.839	0.894	80	0.879	0.915	0.833	0.894	80	0.878	0.917	0.830	0.893
	ITB NB		**0.910**	0.900	**0.928**	**0.925**		**0.919**	0.917	**0.924**	**0.932**		**0.922**	0.922	**0.922**	**0.934**
	NB(KLAR)		0.866	**0.938**	0.790	0.878		0.867	**0.937**	0.793	0.879		0.866	**0.936**	0.793	0.878
	NB(E1017)		0.866	**0.938**	0.790	0.878		0.867	**0.937**	0.793	0.879		0.867	**0.936**	0.793	0.879
	RF		**0.939**	**0.945**	**0.930**	**0.948**		**0.940**	**0.948**	**0.928**	**0.948**		**0.939**	**0.949**	**0.925**	**0.947**
	SVM		0.858	0.929	0.784	0.871		0.874	0.923	0.816	0.888		0.885	0.920	0.841	0.899
	CART		0.780	0.890	0.684	0.790		0.777	0.891	0.680	0.786		0.776	0.893	0.678	0.785
	LDA		0.903	0.920	0.881	0.917		0.902	0.919	0.878	0.915		0.901	0.919	0.877	0.915
	LOG		0.583	0.583	NaN	0.737		0.584	0.584	NaN	0.737		0.583	0.583	NaN	0.737
TEST	TB NB	20	0.867	**0.926**	0.804	0.880	20	0.869	**0.929**	0.804	0.882	20	0.867	**0.931**	0.800	0.880
	ITB NB		0.875	0.915	0.827	0.890		0.877	0.918	0.828	0.892		0.877	0.920	0.826	0.891
	NB(KLAR)		0.862	**0.934**	0.787	0.874		0.865	**0.936**	0.790	0.878		0.867	**0.936**	0.793	0.879
	NB(E1017)		0.862	**0.934**	0.787	0.874		0.865	**0.936**	0.790	0.878		0.866	**0.936**	0.793	0.878
	RF		0.878	0.876	0.882	**0.898**		0.891	0.895	**0.886**	0.908		0.896	0.902	**0.887**	**0.912**
	SVM		0.855	**0.926**	0.780	0.868		0.872	0.922	0.814	0.887		0.884	0.919	0.839	0.898
	CART		0.776	0.887	0.680	0.786		0.776	0.890	0.679	0.786		0.776	0.894	0.678	0.785
	LDA		**0.890**	0.908	**0.865**	**0.905**		**0.897**	0.914	0.872	**0.911**		**0.898**	0.917	**0.874**	**0.912**
	LOG		**0.891**	0.908	**0.869**	**0.907**		**0.901**	0.916	**0.879**	**0.915**		**0.903**	0.919	0.881	**0.916**

Table 5.4: Classifier's performance. Shorts comments (less than three words) NOT removed from the dataset

N.		20.000					50.000					100.000				
SAMPLE	MODEL	PROP	ACC	TPR	TNR	F1	PROP	ACC	TPR	TNR	F1	PROP	ACC	TPR	TNR	F1
LEARNING	TB NB	80	0.910	0.929	0.881	0.925	80	0.911	0.930	0.882	0.926	80	0.911	0.930	0.882	**0.926**
	ITB NB		0.912	0.901	0.932	0.930		**0.922**	0.918	**0.929**	**0.937**		**0.925**	0.923	**0.928**	**0.939**
	NB(KLAR)		0.901	0.946	0.844	0.915										
	NB(E1017)		0.901	0.946	0.844	0.915										
	RF		**0.958**	0.963	**0.950**	**0.965**										
	SVM		0.878	**0.940**	0.803	0.893		0.891	**0.940**	0.829	0.906		0.899	**0.939**	0.846	0.913
	CART		0.808	0.869	0.734	0.834		0.809	0.869	0.734	0.835		0.810	0.875	0.731	0.835
	LDA		**0.916**	**0.939**	**0.883**	**0.929**		**0.914**	**0.939**	**0.879**	**0.928**		**0.914**	**0.939**	0.878	**0.927**
	LOG		0.600	0.600	NaN	0.750		0.601	0.601	NaN	0.751		0.601	0.601	NaN	0.751
TEST	TB NB	20	0.859	0.941	0.780	0.868	20	0.840	**0.945**	0.751	0.845	20	0.755	**0.948**	0.667	0.709
	ITB NB		0.873	0.931	0.810	0.886		**0.873**	**0.935**	0.808	**0.886**		**0.874**	**0.934**	**0.809**	**0.886**
	NB(KLAR)		0.853	**0.943**	0.767	0.863										
	NB(E1017)		0.853	**0.943**	0.767	0.863										
	RF		0.878	0.891	**0.861**	**0.895**										
	SVM		0.842	**0.940**	0.752	0.850		0.845	0.940	0.759	0.852		0.793	0.930	0.710	0.771
	CART		0.786	0.873	0.703	0.800		0.779	0.876	0.694	0.788		0.736	0.903	0.654	0.694
	LDA		**0.881**	0.926	0.828	**0.893**		**0.875**	0.930	**0.816**	**0.886**		0.824	0.919	**0.754**	0.815
	LOG		**0.881**	0.921	**0.833**	**0.894**		**0.878**	0.926	**0.825**	**0.890**		**0.827**	0.916	0.760	**0.820**

Table 5.5: Classifier's performance. Shorts comments (less than three words) removed from the dataset

while training with those comments. However, within this context, the distributions are not perfectly separated, and they are even less separated if we consider a response variable made of temporary labels (like in the proposed GSD framework).

Forby, we present a robustness analysis. We estimate (and compare) the performance of different classifiers while reducing the difference between the distributions. In other words, considering a positive (negative) comment with a negative (positive) label (or temporary label) as an outlier, we compare different classifiers while increasing the percentage of outliers from 1% to 50% of the sample size. We have computed the most used performance indicators: Misclassification Error (ME), Accuracy (ACC), True Positive Rate (TPR), True Negative Rate (TNR), F1-score (F1), Matthews Correlation Coefficient (MCC), BookMaker informedness (BM), MarKedness (MK) (Chicco and Jurman (2020); Tharwat (2020); Powers (2011); Ting (2010); Fawcett (2006); Stehman (1997)). However, considering that recently MCC has been proven to be a more appropriate indicator than the others in the context of a binary classification (Chicco and Jurman (2020); Chicco et al. (2021)), we compare the classifiers with respect to their MCC-value. The MCC performance measure is an index that vary between -1 (worst classifier) and 1 (best classifier), where MCC= 0 indicates that the classifier is performing like a "toss-a-coin" model (that randomly assigning an equiprobable class). Forby, if the associated MCC-value is between 0 and 1, then a classifier is considered to be "usable", otherwise it is "useless".

In Figure 5.9, we can observe how our proposed classifiers (Threshold-based Naïve Bayes (Tb-NB) and Iterative Threshold-based Naïve Bayes (ITb-NB)) are more robust than the others. In fact, until 33% of outliers, they lose few points within their performance indicators, while the other classifiers lose more. In particular, while being the principal competitor at 0% outliers, the random forest is more sensitive to outliers. Forby, at 25% outliers, it is 27.5% worst than before (in terms of MCC), and 41.4% worst at 33% outliers. Further details are available in Table 5.6 for comparison purposes.

Furthermore, in Figure 5.10, we notice that despite having similar performance, the improved version of the Naïve Bayes classifier (ITb-NB) is more robust than Tb-NB. In fact, it is more consistent in terms of MCC than the other one when the percentage of outliers is greater than 33%.

Forby, considering that MCC is a relatively new score, in Figure 5.11 we have estimated other typical performance indicators that will be useful for comparison purposes. Moreover, we notice that all of them follow the same trend. All the considerations that we have made in terms of MCC are reasonably extendable to all the other common performance indicators.

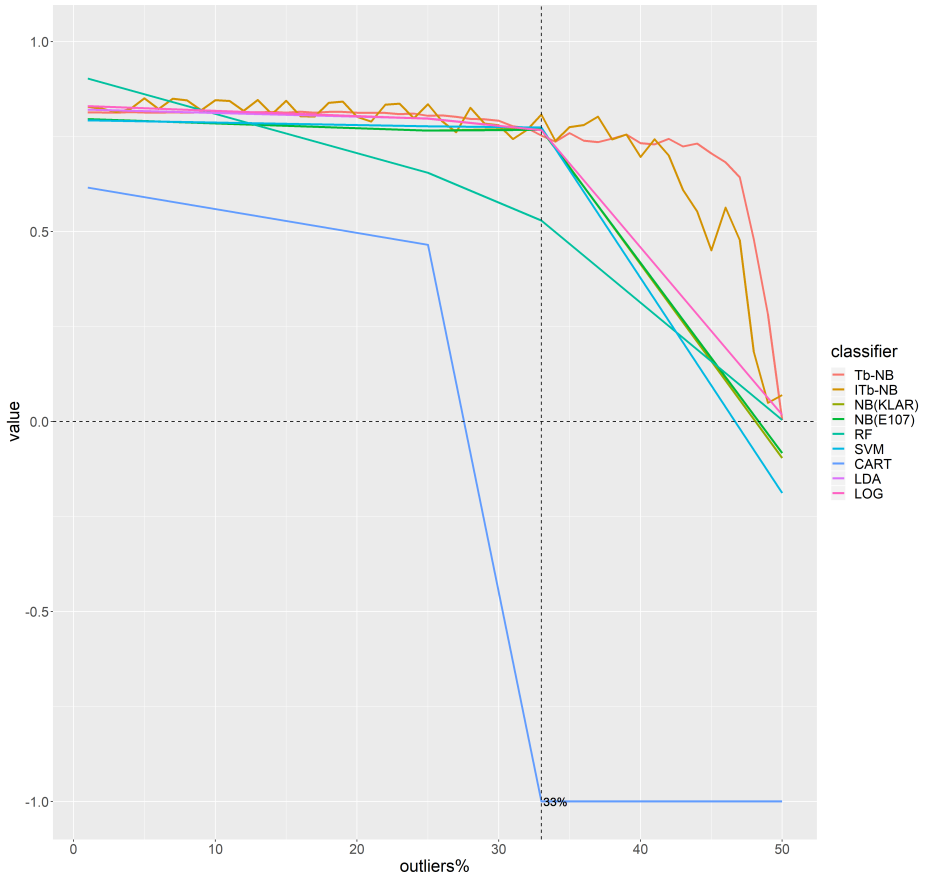


Figure 5.9: Classifiers' robustness – MCC value over outliers percentage variation

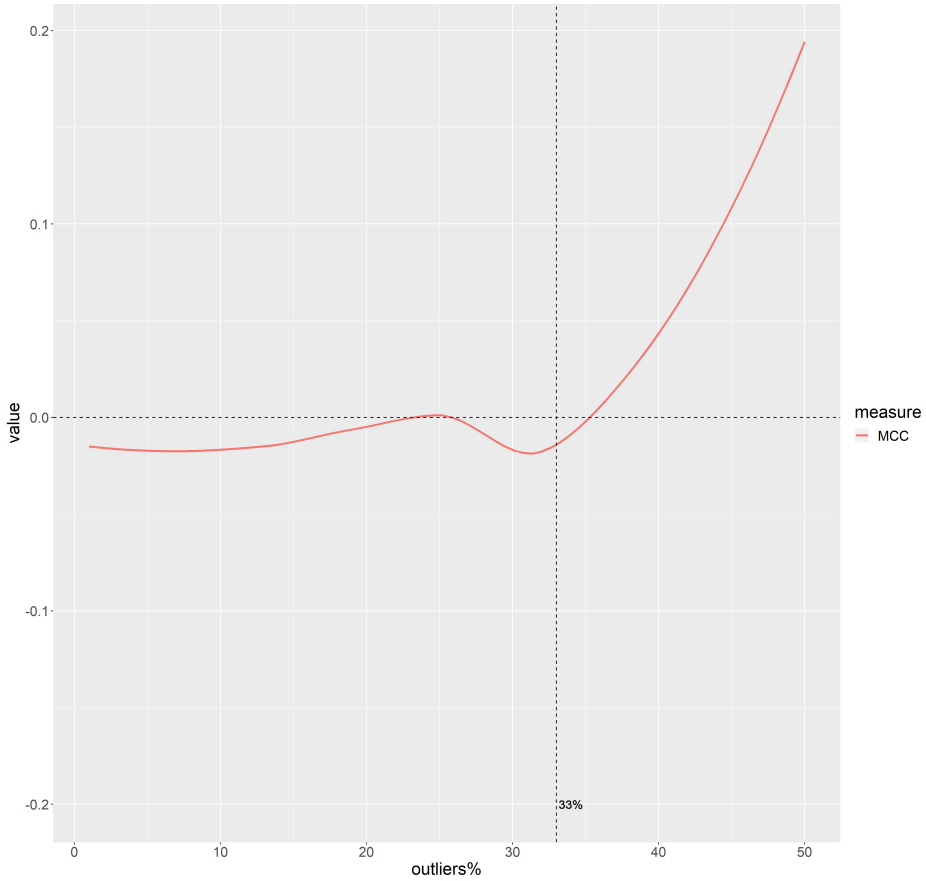


Figure 5.10: Δ MCC Iterative Threshold-based NB − Threshold-based NB

N	SAMPLE	MODEL	PROP OUTLIERS	ACC	TPR	TNR	F1	MCC	Δ ACC	Δ TPR	Δ TNR	Δ F1	Δ MCC
80.000	LEARNING	TB-NB	0	0,911	0,930	0,882	0,926	0,814	-	-	-	-	-
		ITB-NB		**0,925**	0,923	**0,928**	**0,939**	**0,829**	-	-	-	-	-
		NB(KLAR)		0,899	**0,947**	0,838	0,913	0,796	-	-	-	-	-
		NB(E1017)		0,899	**0,947**	0,838	0,913	0,796	-	-	-	-	-
		RF		**0,953**	**0,963**	**0,939**	**0,961**	**0,903**	-	-	-	-	-
		SVM		0,899	0,939	0,846	0,913	0,793	-	-	-	-	-
		CART		0,810	0,875	0,731	0,835	0,616	-	-	-	-	-
		LDA		0,914	0,939	0,878	0,927	0,821	-	-	-	-	-
		LOG		0,918	0,937	0,891	0,932	**0,830**	-	-	-	-	-
	LEARNING	TB-NB	25	**0,909**	0,912	**0,903**	**0,926**	**0,808**	-0,212%	-1,975%	2,422%	0,000%	-0,737%
		ITB-NB		**0,921**	0,946	**0,884**	**0,934**	**0,835**	-0,423%	2,461%	-4,779%	-0,556%	0,724%
		NB(KLAR)		0,878	**0,964**	0,786	0,891	0,766	-2,336%	1,795%	-6,205%	-2,410%	-3,769%
		NB(E1017)		0,878	**0,964**	0,786	0,891	0,766	-2,336%	1,795%	-6,205%	-2,410%	-3,769%
		RF		0,832	0,880	0,769	0,857	0,655	-12,697%	-8,619%	-18,104%	-10,822%	-27,464%
		SVM		0,886	**0,957**	0,806	0,900	0,777	-1,413%	1,959%	-4,723%	-1,469%	-2,018%
		CART		0,685	0,915	0,565	0,665	0,465	-15,444%	4,547%	-22,751%	-20,322%	-24,513%
		LDA		0,902	0,928	0,865	0,918	0,798	-1,272%	-1,170%	-1,453%	-0,992%	-2,801%
		LOG		0,902	0,927	0,866	0,918	0,797	-1,743%	-1,067%	-2,806%	-1,502%	-3,976%
	LEARNING	TB-NB	33	0,883	0,858	**0,937**	**0,909**	0,757	-3,066%	-7,779%	6,279%	-1,860%	-7,002%
		ITB-NB		**0,900**	0,923	**0,870**	**0,917**	**0,793**	-2,694%	-0,030%	-6,287%	-2,366%	-4,343%
		NB(KLAR)		0,881	**0,964**	0,791	0,893	**0,771**	-2,002%	1,795%	-5,609%	-2,191%	-3,141%
		NB(E1017)		0,879	**0,965**	0,787	0,892	0,768	-2,225%	1,901%	-6,086%	-2,300%	-3,518%
		RF		0,768	0,836	0,685	0,798	0,529	-19,412%	-13,188%	-27,050%	-16,961%	-41,417%
		SVM		0,885	0,955	0,805	0,899	0,774	-1,525%	1,746%	-4,842%	-1,578%	-2,396%
		CART		0,601	0,601	NaN	0,751	-1,000	-25,813%	-31,330%	-	-10,018%	-262,338%
		LDA		**0,887**	0,915	0,848	0,905	0,767	-2,914%	-2,554%	-3,389%	-2,394%	-6,577%
		LOG		**0,887**	0,915	0,848	0,905	0,767	-3,377%	-2,348%	-4,826%	-2,897%	-7,590%
	LEARNING	TB-NB	50	**0,606**	**0,606**	0,313	**0,755**	0,005	-33,475%	-34,865%	-64,498%	-18,487%	-99,386%
		ITB-NB		**0,564**	**0,624**	**0,447**	0,654	**0,076**	-39,021%	-32,414%	-51,851%	-30,368%	-90,832%
		NB(KLAR)		0,417	0,562	0,343	0,555	-0,096	-53,615%	-40,655%	-59,069%	-39,211%	-112,060%
		NB(E1017)		0,477	0,567	0,350	0,560	-0,083	-46,941%	-40,127%	-58,234%	-38,664%	-110,427%
		RF		0,502	0,603	0,401	0,501	0,004	-47,324%	-37,383%	-57,295%	-47,867%	-99,557%
		SVM		0,526	0,517	0,122	**0,684**	-0,188	-41,471%	-44,919%	-85,578%	-25,116%	-123,707%
		CART		0,502	NaN	NaN	0,386	-1,000	-38,033%	-	-	-53,751%	-262,338%
		LDA		0,497	0,610	**0,408**	0,518	0,018	-45,601%	-35,036%	-53,518%	-44,132%	-97,808%
		LOG		0,497	0,610	**0,408**	0,518	0,018	-45,861%	-34,899%	-54,209%	-44,421%	-97,831%

Table 5.6: Classifiers' robustness – common indicators value over outliers percentage variation. Estimated by Cross Validation.

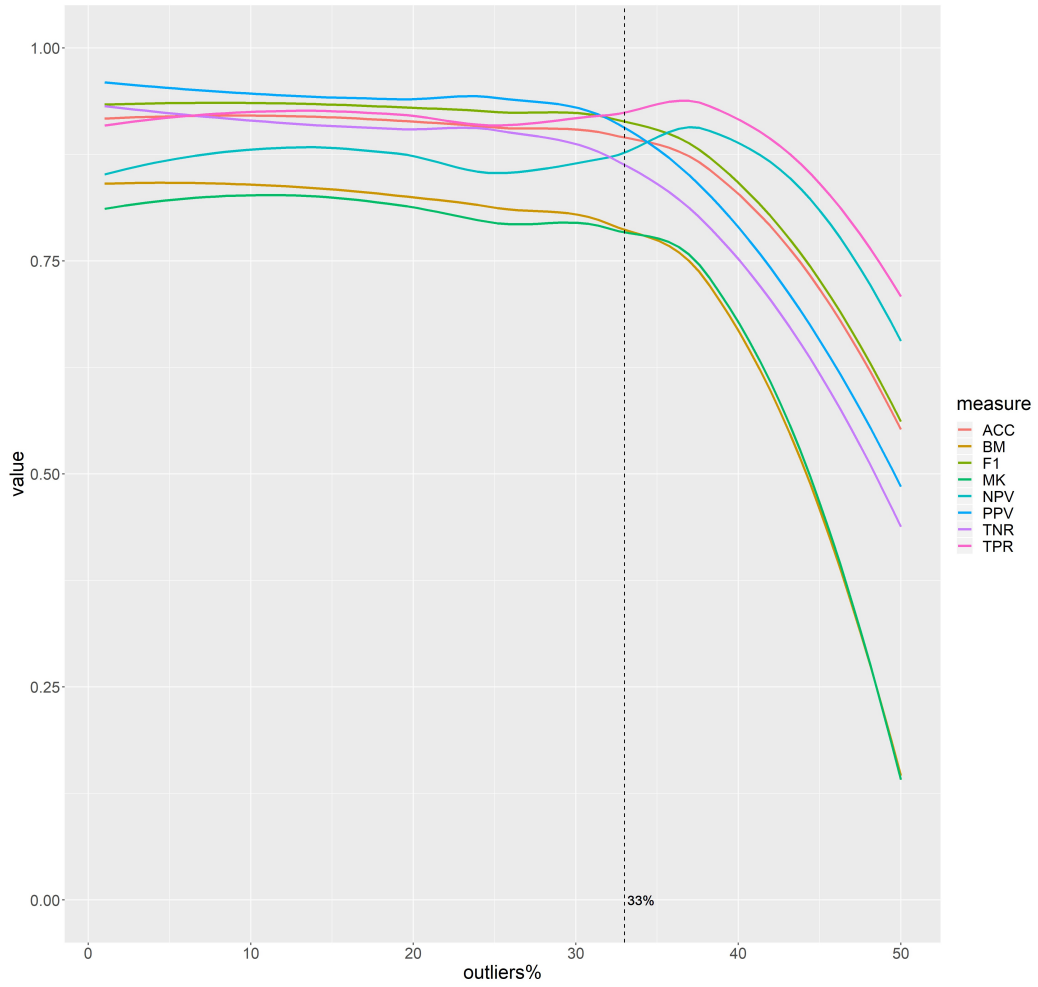


Figure 5.11: Iterative Threshold-based NB other common performance indicators

5.5.4 Computational complexity

Finally, after assessing how reliable and robust the proposed classifier is, we consider another interesting aspect: the computational complexity. Forby, for assessing how much time is needed to train a classifier and produce the relative predictions, we consider a dataset of 100 random observations. Moreover, we randomly divide the sample into two parts: 50% observations for training the model and 50% for testing. Table 5.7 shows that the proposed classifier is considerably quicker both in training and predicting time compared to the others. In particular, it is interesting to notice how:

- the standard implementation of Naïve Bayes is ∼288 times slower in predicting, and ∼5 times slower in training than the proposed one;
- Tb-NB (ITb-NB) is ∼19 times quicker in training and ∼82 times quicker in predicting than logistic regression (LOG);

- Tb-NB (ITb-NB) is ∼12 times quicker in training and ∼26 times quicker in predicting than random forest (RF);
- both in training and predicting time, CART is almost quickly as Tb-NB (ITb-NB). However, we recall that it is really sensible to outliers, and it has considerably lower performance than the proposed classifier.

MODEL	TRAINING TIME	PREDICTING TIME	TRAINING TIME / Tb-NB	PREDICTING TIME / Tb-NB
Tb-NB	5.273	0.371	1.000	1.000
ITb-NB	5.434	0.371	1.030	1.000
NB(KLAR)	26.999	106.608	5.120	287.654
NB(E1017)	26.999	106.608	5.120	287.654
RF	64.876	9.718	12.303	26.222
SVM	35.019	24.329	6.641	65.645
CART	11.177	4.162	2.120	11.229
LDA	-	-	-	-
LOG	102.026	30.260	19.348	81.650

Table 5.7: Computational performance. Training and predicting time (in seconds) of the considered classifiers for a training set of 50 observations and a test set of the same size; executed 100 times.

5.6 Interpretation

As it is possible to notice, the Threshold-based Naïve Bayes classifier presents a good performance while classifying a comment into positive or negative with a numerical value. What actually is even more interesting, compared to other kinds of approaches, are the LOR values that the model must estimate in order to be able to produce the classification. According to what we have shown before, those values have a "versatile nature"; in fact, they can be summed together matching specific criteria. The criteria that the model uses to merge words' values is, for a given set c ($c \in BoW$), to check which word belongs (or not) to c. While using the same LOR values with the same approach but different criteria, it is possible to apply a dimensionality reduction technique to produce some valuable plots.

Usually, natural language text data that are taken from the web comes with temporal information (data only or time too), so a time-series of the LOR words values will take full advantage of that.

As better shown in Chapter 6, considering that plotting all the massive number of words inside BoW is meaningless, we produce a small set of interesting categories. Each word inside BoW will be then mapped to one and only one category. It is then possible to calculate how the LOR for a specific category changes in time. In fact, a single category

can be considered to be a comment composed of all the words that are mapped to such a category.

There are many ways to identify which are the best categories:

- Arbitrary specific words/categories, identified by some context-domain knowledge;
- Relevant words/categories that come from the literature;
- Using a semi-supervised clustering method over a Word Embedding representation of the words contained in the BoW as shown in Chapter 6

Once the categories are identified, each word in the BoW has to be mapped in one and only one category. While we could do some manual work for achieving such a result, we suggest preferring a better objective and automatic way. For instance, a semi-supervised clustering over the Words Embedding representation of the words considered in our BoW. In fact, we consider every category to be a centroid of a fixed number of clusters representation in that Words Embedding space. Because we already know the number of categories, we already know how many clusters we need: one and only one cluster for a category. After the clustering process, to conclude, all the words inside a certain cluster-category are mapped with the corresponding category.

Chapter 6

Use Cases

In the previous chapters, we have illustrated the General Sentiment Decomposition method from a theoretical perspective. Now we apply the proposed method to analyze a labeled dataset (Booking.com data). Moreover, we use the same process for an unlabeled dataset (Twitter.com data) concerning a completely different topic. Furthermore, we use a labeled dataset (TripAdvisor.com) to assess the General Sentiment Decomposition method's performance in the unlabeled field. In fact, we use it like an unlabelled dataset for training the model, and then we compare the estimated label of the model with the real labels that comes within the dataset (but not used by the model). In this way, we demonstrate the effectiveness of the General Sentiment Decomposition method and his generalization capability, which allows us to use it in many fields.

6.1 Labeled Data: Booking

Big data has been touted as a new research paradigm that utilizes diverse sources of data and analytical tools to make inferences and predictions about reality (Boyd and Crawford, 2012). Mainly, with increasingly powerful natural language processing and machine learning capabilities, textual contents from the Web provide a huge shared cognitive and cultural context and, thus, have been analyzed in many application domains (Halevy et al., 2009). This phenomenon has also characterized the tourism sector, where product review forums about tourism topics have become commonplace. An increasing number of websites provide platforms for tourists to publicize their personal evaluations and opinions of products and services. This information is of great interest to both companies and consumers. Companies spend a considerable amount of money to find customers' opinions and sentiments since this information is helpful to exploit their marketing mix in order to affect consumer satisfaction. Individuals are interested in others' opinions when purchasing a product or hiring a service.

From the business viewpoint, online reviews, including their peripheral cues, such as user-supplied photos and the reviewer's personal information, are intended as means of persuasive communication in order to build credibility and influence user behavior (Sparks et al., 2013). From an operational point of view, this situation has raised many NLP challenges, commonly referred to as Sentiment Analysis, such as subjectivity detection (Wiebe et al., 1999), polarity recognition (Schmunk et al., 2013) and rating inference (Esuli and Sebastiani, 2006a). Focusing on product review classification, various approaches have been proposed during the last decade. Most of them only consider the polarity of the opinions (i.e., negative vs. positive) and rely on machine learning techniques trained over vectors of linguistic feature frequencies.

Within the above-described framework, we have retrieved data about clients' reviews hosted in accommodations, hereafter hotels, located in Sardinia whose information is available on Booking.com. A routine to scrape all the essential information concerning a hotel listed on Booking.com has been implemented in Python. Next, to define a benchmarking tool for hotels, we have processed reviews' content with natural language to comprehend the items leading towards increasing or decreasing customer satisfaction. The results of this analysis were the basis of the design of a search engine based on clients' reviews (Conversano et al., 2019), that allows its users to select the most suitable hotel on the basis of an, even negligible, set of keywords, i.e., a sentence. We have chosen Booking.com as a reference platform as the reviews there available come from customers who effectively stayed in a hotel. Booking.com utilizes a score defined in the [2.5, 10.0] interval, and each review is split into two parts: a positive comment and a negative one.

Our analysis of Booking.com data has several purposes. The first goal is defining an ad-hoc classifier able to classify a comment as positive or negative from its content. The same classifier allows us to quantify the (positive or negative) impact of a specific word within a review. The second goal is developing a prediction model for the score obtained by a hotel based on the reviews reported on Booking.com. Predicted scores constitute a benchmarking tool for a hotel to be evaluated. Last but not least, an additional goal is defining a search engine based on clients' reviews that would allow users to select the most suitable hotel concerning their preferences. For all these goals to be accomplished, we followed three basic steps:

1. data collection: data were scraped from Booking.com and cleaned in preparation of the statistical analysis;

2. reviews' classification: the reviews' content was processed through an ad-hoc defined Naïve Bayes classifier, hereafter Threshold-based Naïve Bayes, in order to obtain a

predicted score and the polarity for each review based on its specific content;

3. search engine design and implementation: it derives from the results obtained in step 2.

The three above-mentioned steps are described in detail in what follows.

A Python extractor has been implemented that, through web-scraping, has retrieved all the valuable information publicly available on Booking.com. Retrieved data have next been organized into flat tables. They concern 619 hotels operating in Sardinia. For them, it was possible to scrape 66,237 reviews consisting of 106,800 positive and negative comments. Data about each hotel concerns the hotel information usually available on Booking.com (e.g., type of accommodation, postcode, city, scores about cleaning, comfort, room, restaurant). Data about each review concerns its content together with the information about the reviewer and the type of customer (e.g., business trip, pleasure trip).

More in detail, we observe:

- 619 hotels located in Sardinia (Tab. 6.1)
- 66,237 reviews (Tab. 6.2)
  - 106,800 comments (in Italian or English) (Tab. 6.2)
    - 44,509 negative + 62,291 positive
- Observation period: Jan 3, 2015 - May 27, 2018 (1,240 days)

Data is organized in two datasets with a total of 127 features:

- Hotels dataset (86 features):
  - Hotel (3),
  - Review (8), Reviewer (2),
  - Booking's score (11), Score components (12),
  - Guest (8), Accommodation (32), Length of stay (6),
  - Other info (4).
- Comments dataset (41 features):
  - Hotel (2),
  - Comment (6), Reviewer (2),
  - Score components (6),

– Guest (4), Accommodation (16), Length of stay (3),

– Other info (2).

Name	Type	Postal Code	City	Total Review	Total C_{Pos}	Totale C_{Neg}	...
Hotel 1	Other Facilities	09044	Sant'Isidoro	35	35	19	...
Hotel 2	3–Star	09049	Villasimius	289	286	104	...
Hotel 3	3–Star	07013	Mores	125	123	42	...
Hotel 4	4–Star	09123	Cagliari	725	678	492	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
Hotel 619	4–Star	07026	Olbia	2147	1975	1545	...

Table 6.1: Hotel data: $n = 619$

Name	ID_C	ID_R	Comment	Neg–Pos	Score	Business	Stay	...
Hotel 1	1	1	christina was the best...	Pos	10.0	Yes	1–3 Days	...
Hotel 1	2	2	we travelled into cagliari...	Pos	9.2	No	4–7 Days	...
Hotel 1	3	3	it was fantastic...	Pos	10.0	No	$>$ 7 Days	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
Hotel 619	106,800	66,237	il wifi e le zanzariere non erano presenti...	Neg	5.0	Yes	4–7 Days	...

Table 6.2: Comments data: $n = 106,800$ comments from $66,237$ reviews

Collected data have been cleaned by removing conjunctions, punctuation, numbers, and all the stopwords, according to the specified approach described in Chapter 5.2. Next, words with similar meanings have been merged together in macro-words, and all the macro-words composing each comment have been joined in the Bag of Words (BoW). Each element of BoW had its own frequency, and these macro-words have been subsequently assigned individually to the following reference categories: "bar", "cleaning", "comfort", "food", "hotel", "position", "price-quality-rate", "room", "services", "sleep-quality", "staff", "wifi" and "other". As an example, the words "breakfast", "restaurant", "lunch", etc. all belong to the "food" category. The definition of these reference categories allowed us to process the data in an aggregate manner, as well as to quantify pros and cons concerning the hotel services associated with each category.

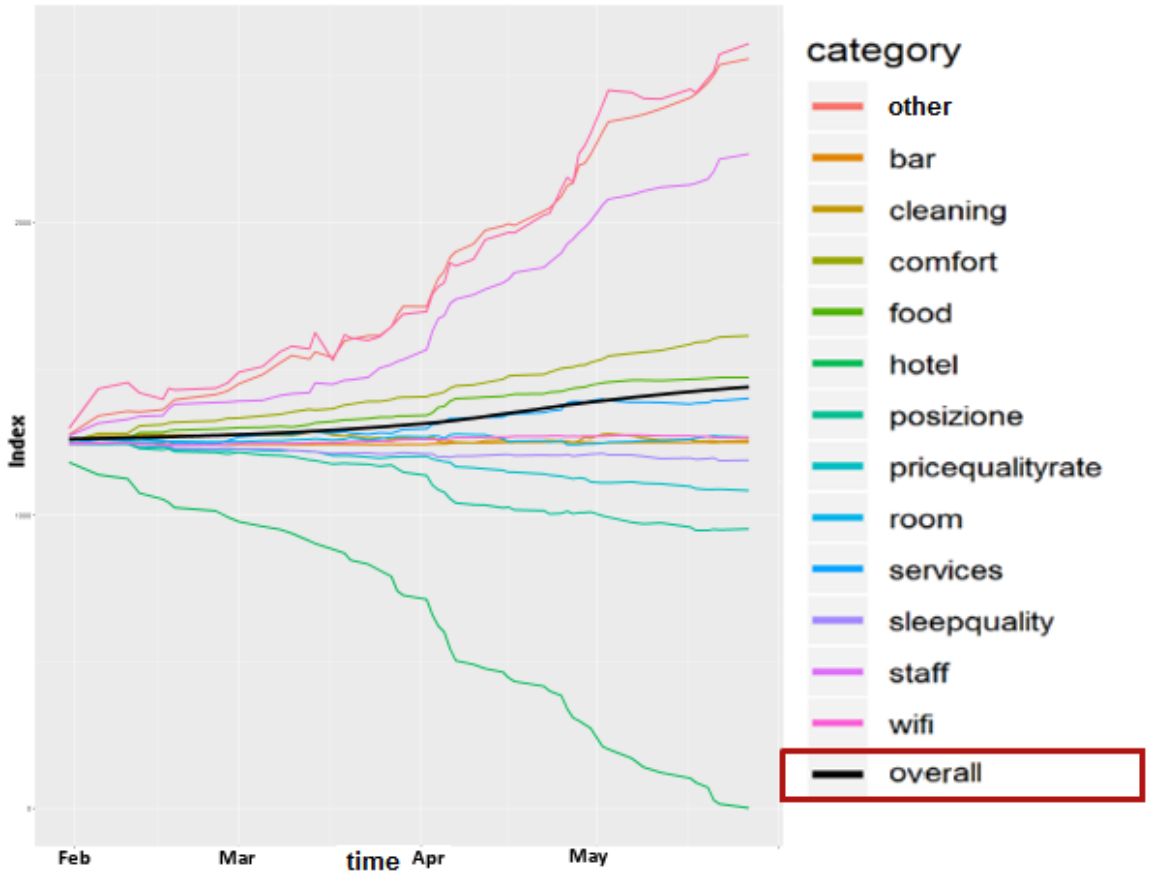


Figure 6.1: Score of categories of words observed in time;
the higher is a score, the positive is the sentiment and vice versa.
Data from a Sardinian hotel with a 10/10 score evaluation on Booking.com.
In black, the overall sentiment score.

Following the procedure defined in Chapter 5.6, which defines how to proceed within the interpretation phase, we have produced many types of plots like those in Fig. 6.1 where we can easily understand the phenomena for a single hotel. Moreover, thanks to the versatility of the output, we can aggregate it to have a macro-overview. For example we could consider the Sardinian provinces (Fig. 6.2 and Fig. 6.3), and notice similarities and differences.

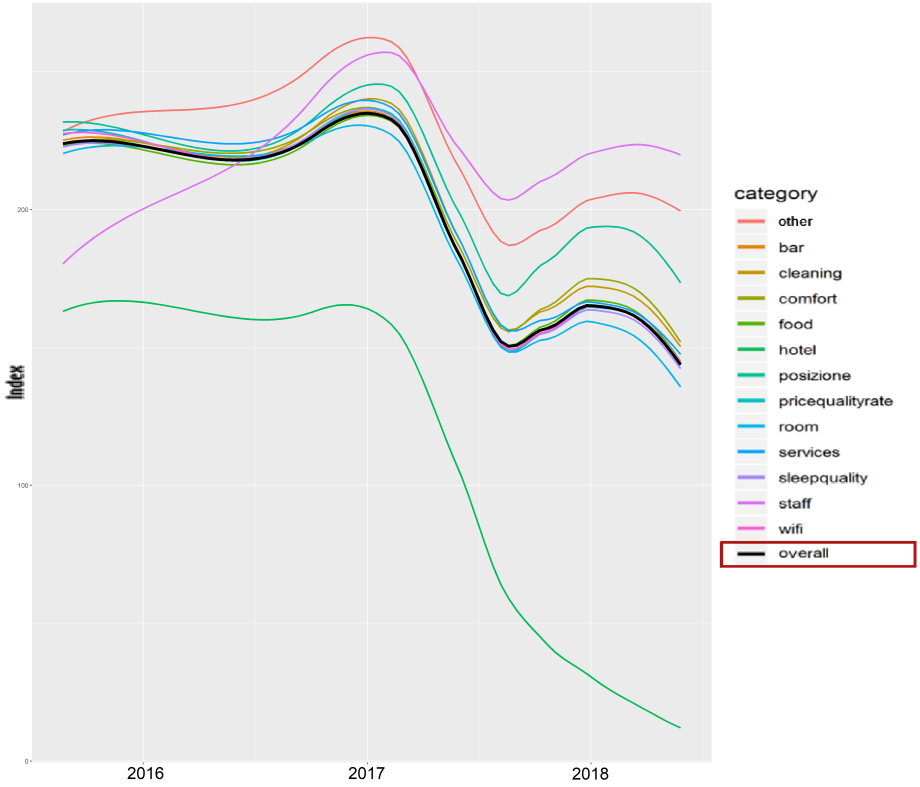


Figure 6.2: Score of categories of words observed in time;
the higher is a score, the positive is the sentiment and vice versa.
Data from the Sardinian hotels located in province of Cagliari.
In black, the overall sentiment score.

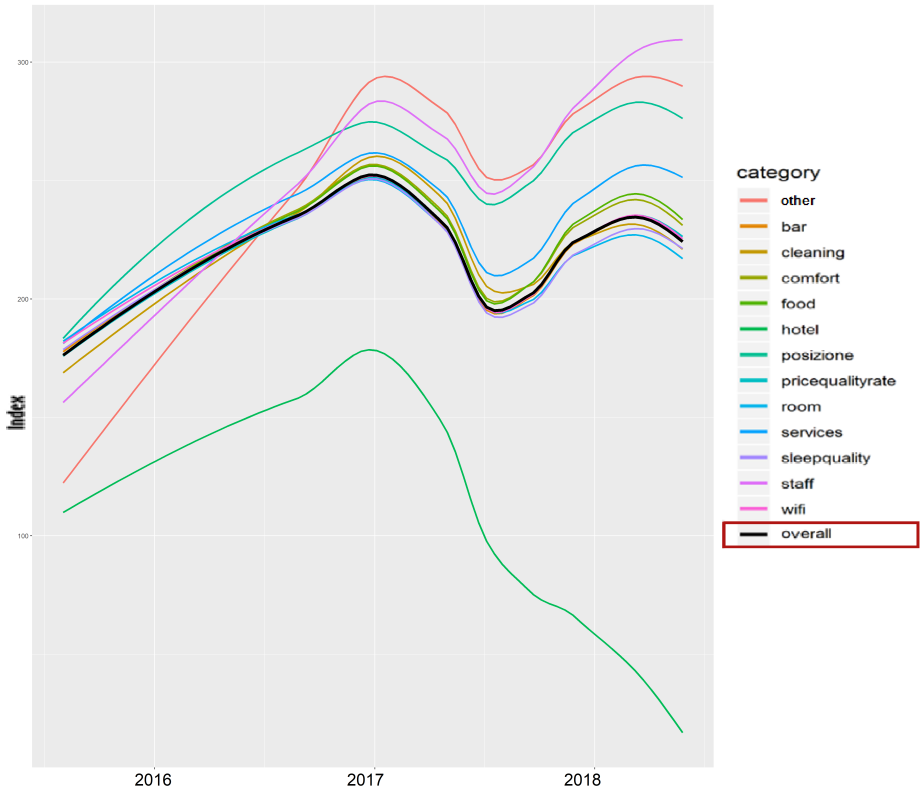


Figure 6.3: Score of categories of words observed in time;
the higher is a score, the positive is the sentiment and vice versa.
Data from the Sardinian hotels located in province of Sassari.
In black, the overall sentiment score.

In an additional experiment, we have used the quantities $\sum_{w_i \in K} pres_{w_i} + \sum_{w_i' \notin K} abs_{w_{i'}}$ computed for a review r to predict the score the reviewer who generated r assigned to a hotel on Booking.com. Again, we learned the same set of alternative classifiers, and the best performing one was Random Forest, with MSE = 0.6704. We have noticed that considering the log-likelihood ratios of categories originated by merging similar words included in the BoW as inputs for the classifier, instead of the single words ($w \in BoW$), considerably improved the accuracy of the classifier. As an additional predictor, the polarity of a review (positive or negative) estimated in the previous step with the Threshold-based Naïve Bayes classifier has been considered. Since Booking.com provides scores for each service offered by a hotel, the same predictive approach has been applied for each service, thus obtaining predicted scores arising from reviews' content for each hotel service individually.

6.2 Unlabeled Data: Tripadvisor and Twitter

Sentiment Analysis, also known as Opinion Mining or Emotion Artificial Intelligence, is the use of natural language processing techniques such as text analysis to extract and study the sentiment information (Pang and Lee, 2008). In other words, it is a method that allows categorizing peoples' reactions starting from a natural language text (a review, a Facebook or Twitter post, et similia) into positive, neutral, or negative responses.

As shown in Gibert et al. (2018a) and in Gibert et al. (2018b), Data Science can add value to Environmental Sciences in many different ways, and Data Mining (DM) is a fundamental component of the Data Science process. Using DM (as well as opinion mining) techniques to identify the Twitter users' sentiment about a specific ecological problem improves the decision-making process of, for example, politicians or academics (Hauthal et al. (2020); Agarwal et al. (2011)). When applying Data Science methods and, in particular, Sentiment Analysis to climate change-related problems, it is of primary importance to design a support system able to understand what people think about a specific issue, identify if a certain task is appreciated, or find the critical aspects identified by a consistent group of people.
As an example, Sentiment Analysis:

- is widely used for social media monitoring as it allows us to gain significant insights into a public reaction towards social topics;
- is used in companies to understand customers' reactions towards their advertisement campaign and their products to take the necessary steps to improve them;
- was used by the Obama administration to gauge public opinion to policy announcements and campaign messages ahead of the 2012 Presidential Election (Rajput et al., 2019);

- has been used to demonstrate that peoples' changing reaction towards a brand/company is directly correlated with shifts in the stock market.

Nowadays, climate change is a significant issue, and many people use Twitter or other social media platforms to discuss various aspects of it. Here, climate change has been chosen to observe how people worldwide react to discussions about it on social media. In such a framework, specific hashtags are used to discuss various perspectives on climate change. To demonstrate ho GSD works on unlabeled data, we consider two topics and the related hashtags, as shown in Table 6.3.

Topic	Data
Eco Policies (political perspective)	#climateemergency #parisagreement #zeroplastic
Environment (citizen perspective)	#climatechange #savetheplanet #fridaysforfuture

Table 6.3: Data topic groups

Hashtags in Table 6.3 have been chosen to portray a different reaction to the same issue of Climate Change, i.e.:

- #climatechange: People tend to focus on the consequences of climate change and which human activities lead to this problem. So, the reaction here is mostly negative.
- #savetheplanet: In this case, people seem to have a positive outlook, wherein they try to suggest various ways to reduce pollution, deforestation, et similia, which are the prime causes of climate change.
- #fridaysforfuture: This is an international movement started by students that go off from class on Fridays to participate in demonstrations, demanding political leaders to focus on preventing climate change. Specifically, they demand to make a mandatory transition from fossil fuel to renewable energy.

Selected hashtags were chosen out, taking into account the frequency of posts and the area of focus of these hashtags, according to the `www.best-hashtags.com` website. i.e.:

- #climatechange (209 posts per hour)
- #savetheplanet (188 posts per hour)
- #fridaysforfuture (78 posts per hour)

Our final goal is classifying tweets into positive/negative based on their specific content. For this purpose, we use the Threshold-based Naïve Bayes classifier that has been used to classify hotel guests' reviews that appeared on Booking.com (Romano et al., 2018). As a matter of fact, tweets are not classified a-priori into positive/negative. To further enforce results obtained for booking.com data, we use TripAdvisor data to compare this approach within Naïve Bayes. In fact, TripAdvisor data have a 1 to 5 Stars variable, which is a proxy of the positive/negative comment of Booking.com data and, even though the model does not use this variable, it allows us to validate the predicted classification.
The data used in this study have been extracted using a Twitter API with a Python interface and with an ad-hoc web scraping python program for TripAdvisor. We collected more than 400,000 tweets (divided between all the hashtags) and related them to March and April 2020 for the Eco Policies and Environment topics. Furthermore, we collected 39,000 reviews of Sardinian hotels from TripAdvisor.com to validate the usage of the Threshold-based Naïve Bayes classifier.

Applying the General Sentiment Decomposition method, we obtain some preliminary results. Figure 6.4 depicts the first few tweets from the file #savetheplanet. We can see in the processed text that the stopwords, acronyms, words like 'RT (re-tweet)', and all punctuations are removed. The username, other hashtags, and the links mentioned, which can be confusing for the analysis, have also been removed. Moreover, the emojis are converted into meaningful texts. Finally, next to each tweet, we see a score. If the magnitude of the score is very close to zero, it is considered to be neutral. Otherwise, a negative score depicts a positive sentiment, whereas a positive score means negative sentiment. Then, the score is multiplied by -1 to avoid his tricky nature. So, the higher the positive and vice versa. Hence, from the above output, we can observe that:

- The first tweet (score +0.025) is just a Good night having globe and whale emojis. Although it does not say much, it cannot be considered as negative. It shows an overall positive sentiment with a shallow magnitude.

- The second one (score +0.071) says "Great Interview", followed by a negative comment. Thus, it is a mixture of positive and negative interviews, and the score suggests the same thing.

- The third one says that a lot of time and money is wasted for commuting long distances, which is an evident negative score(-0.028).

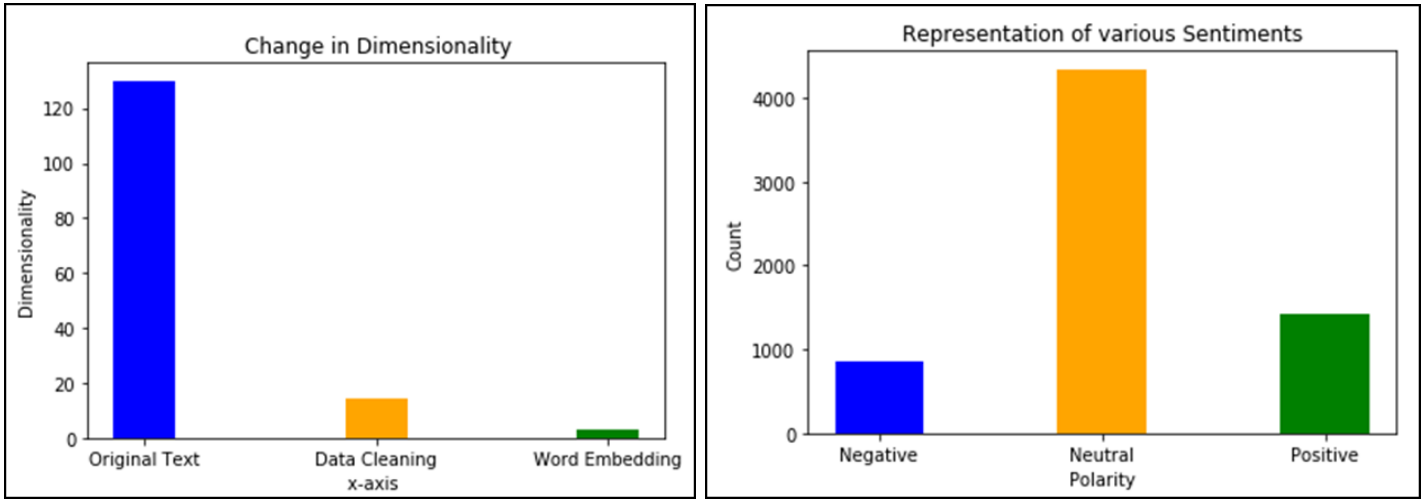
Furthermore, while considering both the plots in Figure 6.5, we can observe the effects of the application of the two first step of General Sentiment Decomposition: "Data Cleaning" (Chapter 5.3) and "SentiWordNet & Words Embedding combination" (Chapter 5.4). More in detail:

- Figure 6.5a represents the reduction in dimensionality achieved within the Data Cleaning step. As we can see, the raw text data contains much-unwanted information that does not provide any significant insight and tends to confuse the algorithm. In fact, according to the procedure described more in detail in Chapter 5.3, almost 80% of the words are removed from the corpus. Moreover, while merging words by their meaning (Chapter 5.4) with the Words Embedding help, we further reduce the corpus size by replacing all similar words with one single word. Dimensionality is reduced, but information is preserved;

- Figure 6.5b shows the count of positive, negative, and neutral tweets included in the corpus of #savetheplanet. According to the procedure described in Chapter 5.4, those are the temporary sentiment labels that we produce while defining a threshold for the overall score.

- text: “RT @FreddyBeltranP: Good Night ☽\n Twitter Planet 🌍\n\n#Uyuni #Potosi #OurPlanet #Photography\n#EnssaysDiseings #whale #DiscoverAdventure \n#...”
  - after clean text: “whale spout globe meridian planet twitter crescent moon night”
  - hypernyms: {large_person, opening, advantage, globe, degree, celestial_body, sound, curve, moon, time_period}
  - score: 0.025
- text: “@priyamenon96 Great interview, though I do disagree with the usa of palm oul as a primary feedstock for biodiesel...”
  - after clean text: “feedstock oil palm though primary biodiesel use disagree interview great”
  - hypernyms: {raw_material, lipid, area, though, primary, biodiesel, activity, disagree, interrogation, achiever}
  - score: 0.0714
- text: “RT @QUBEcc: Do you spend half of your day travelling for work?\nHave you worked out how much time and money are lost on your commute?\nThere...”
  - after clean text: “lost traverl half spend time money much work day”
  - hypernyms: {people, motion, common_fraction, spend, case, medium_of_exchange, large_indefinite_quantity, activity, time_unit}
  - score: -0.0277

Figure 6.4: An excerpt of the output from #savetheplanet



(a) Dimensionality reduction results

(b) Polarities counts

Figure 6.5: Twitter Data output dimensionality reduction effectiveness (a) and temporary labels frequencies (b)

Following the General Sentiment Decomposition procedure, we should now use the Threshold-based Naïve Bayes Classifier. Considering that, until now, this classifier was applied in a labeled context (Booking.com data), it is first essential to assess the performance of the model when used in this original without-labels context. We can achieve that with the TripAdvisor hotel data with the following considerations:

- we have data for Sardinian hotels reviews', and each review is associated to a number of 1 to 5 Stars;
- similar to what we have done until now, following the General Sentiment Decomposition procedure, we produce a temporary sentiment label;
- we compute the most used performance indicators: Misclassification Error, Accuracy, True Positive Rate, True Negative Rate, F1-score, Matthews Correlation Coefficient, Bookmaker Informedness, MarKedness (Chicco and Jurman (2020); Tharwat (2020); Powers (2011); Ting (2010); Fawcett (2006); Stehman (1997));
- despite we did not use it to produce the temporary sentiment label, we transform the reviews' associated number of 1-5 Stars, into a "real label": negative label [1-2-3 stars], and positive label [4-5 stars]. We consider 3 stars to be a negative value considering the usual behaviour of the platform' users and the frequencies of the stars levels. This "real label" permits to compute the performance indicators for the "real" label that, in an unlabelled context (such as the Twitter.com data), is not possible to know a priori;
- we could stop here and directly use the temporary sentiment label for classification. For that, in Table 6.4 we have computed the performance indicators while matching the temporary sentiment label with the "real" label;
- but, if instead of stopping, we continue with the General Sentiment Decomposition method, we obtain more interesting results;
- so, while feeding the text data (and the temporary sentiment label), we train the Threshold-based Naïve Bayes Classifier with those reviews that have a positive (or negative) temporary sentiment label;
- computing the same performance indicators as before, we can observe in Table 6.5 that sometimes the Threshold-based Naïve Bayes classifications are not equal to the temporary label;

- finally, computing one last time the performance indicators, we compare the Threshold-based Naïve Bayes classifications with respect to the "real" label. Considering that, despite the unlabeled context such as Twitter.com, we do not have a "real" label, we would like to classify the tweets correctly; this is the most interesting one (Table 6.6). In fact, it permits to observe which are the performance of the model on producing as accurate as possible, a response variable that he did not have observed during the training phase.

ME	ACC	TPR	TNR	F1	MCC	BM	MK
0.092	0.908	0.936	0.398	0.951	0.268	0.334	0.215

Table 6.4: Performance on using the temporary sentiment label to predict the "real" label. Notice that to calculate the temporary sentiment label we use text data only, and the "real" label is not provided for the training phase. In fact, it is determined within the GSD approach by a threshold applied on the overall sentiment score (Section 5.4

ME	ACC	TPR	TNR	F1	MCC	BM	MK
0.070	0.930	0.941	0.645	0.963	0.402	0.586	0.275

Table 6.5: Threshold-based Naïve Bayes performance on predicting the temporary sentiment label – trained with the temporary sentiment label, performance estimated with 10-fold CV

ME	ACC	TPR	TNR	F1	MCC	BM	MK
0.055	**0.945**	**0.973**	**0.503**	**0.973**	**0.475**	**0.476**	**0.474**

Table 6.6: Threshold-based Naïve Bayes performance on predicting the real label – trained with the temporary sentiment label, performance estimated with 10-fold CV

Considering the excellent performance reported in Table 6.5 and Table 6.6 (MCC in particular), we can now trust the results that we see with the Twitter data using the same approach, the same algorithm, and the same model. As an additional confirmation of the methodology's appropriateness, we can also notice in Figure 6.6 that the τ-value well separates the log-odds distributions of positive and negative Tweets. In particular, comparing Table 6.4 and Table 6.6, it is evident that using the Threshold-based Naïve Bayes classifier in that way allows us to trust a valuable interpretation phase (Figure 6.7) whilst improving the performance on the prediction of the "real" label.

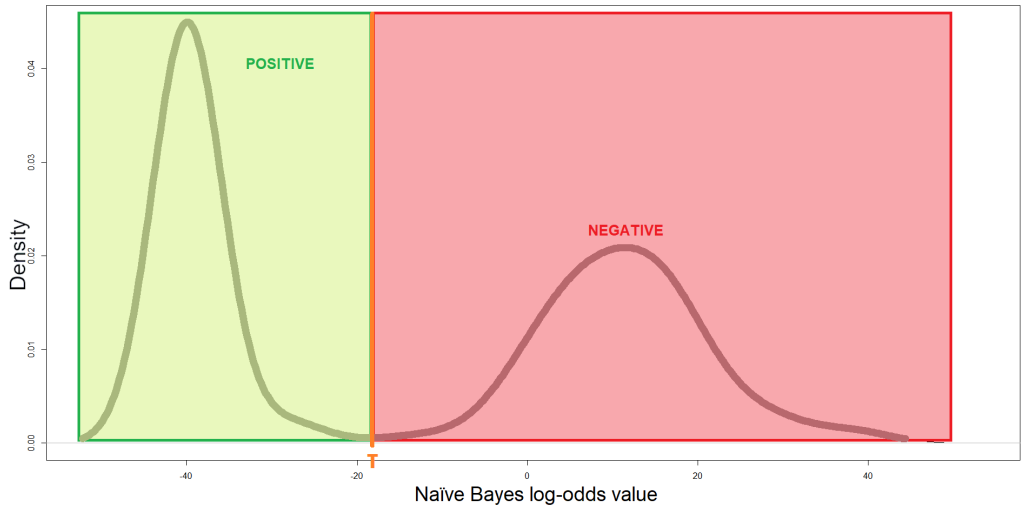


Figure 6.6: Threshold-based Naïve Bayes log-odds distribution for the twitter data

Moreover, for the versatile nature of the output, it is possible to produce a vast number of interesting plots. In the following, we focus on some main results.

To make the plot in Fig. 6.7, we have focused on the essential category of words in terms of positive/negative contribution, applying the same dimensionality reduction process showed before to obtain a better view of the results. Considering the above mentioned versatile nature of the outputs, it is possible to make similar plots with:

- Arbitrary identified words/categories;
- Relevant words/categories that we find in literature;
- Using the same Words Embedding + Clustering approach with a fixed small λ, to identify the most critical group of words (=clusters) and the categories (=centroids of those clusters); or start from a fixed set of categories (word-topics of interest) and then generate agglomerative clusters using them as a starting point.

More in detail, in Fig. 6.7, the whole set of words are segmented into ten categories where each word belongs to one of the categories by applying the K-means clustering method. The following are the categories explanations:

- Action: it explains the policy action or the policy implementations done by the government or even what kind of opinion do people have in general about the government actions.
- Earth: it expresses the thoughts on how the earth is getting affected by the change in the environment and its future causes and opinions on those factors.

- Environment: Currently, we face many environmental challenges like global warming, pollution, climate change, acid rain, ozone depletion. In this particular category, we can see the people's view on these Environmental problems.

- Initiatives: it expresses decisions of the government bodies like what are the initiatives they will take to protect the environmental issues in the future and what is the public opinion on those initiatives, whether it is better to take those initiatives or has some other good initiatives.

- People: the word people is very general in nature, and this helps to know that the environmentalists, climate experts, or government bodies want to convey something to the public and even vice-versa. From this, we can able to analyze the opinions, and government unions can make better decisions.

- Plastic: Nowadays, Plastic plays an important role in everyday life, and using the excess of plastic materials, bags, ..., is affecting every organization in the world directly or indirectly. This leads to an imbalance in the eco-system, and there are different types of plastics that cannot be recycled further.

- Policy: The Eco-Policies generally alert people and spread awareness among the public. So this category is chosen to see the response of the people who have tweeted positive or negative opinions about the Eco-policies.

- Pollution: As the population is increasing day by day, cities are growing, and there are different types of pollutions which affect our environment every day. Thus, from this category we can analyze the opinions about pollution.

- Recycle: Recycle of waste is an important thing because wastes impact negatively on the environment. By recycling products, it is possible to reduce pollution. This category, expresses how people are reacting to recycling.

- Weather: the first thing which is affected directly or indirectly by any environmental issues is the climatic conditions. From this category, it is possible to understand people's opinions on the weather.

Thus, In Figure 6.7, it is possible to notice that our approach can correctly identify major events: March 2^{nd}, Australian firefighters announced the fire extinguish; April 17^{th}, the first Friday after the Notre Dame fire anniversary (April 15^{th}), and this hashtag is related to Fridays for future. The effect, in particular, with respect to the "pollution" category, on people talking negatively in that category, is noticeable after a few days of

those major events. It is essential to see that, although the problem itself does not change (air pollution due to the extensive fire), while reading the tweets we notice that people have a positive attitude while talking about the end of the Australian fire. In fact, they think that the air pollution will reduce now that the fire is over. Instead, while talking about the Notre Dame fire anniversary, they are definitely more negative, hoping that something like this will never happen again.

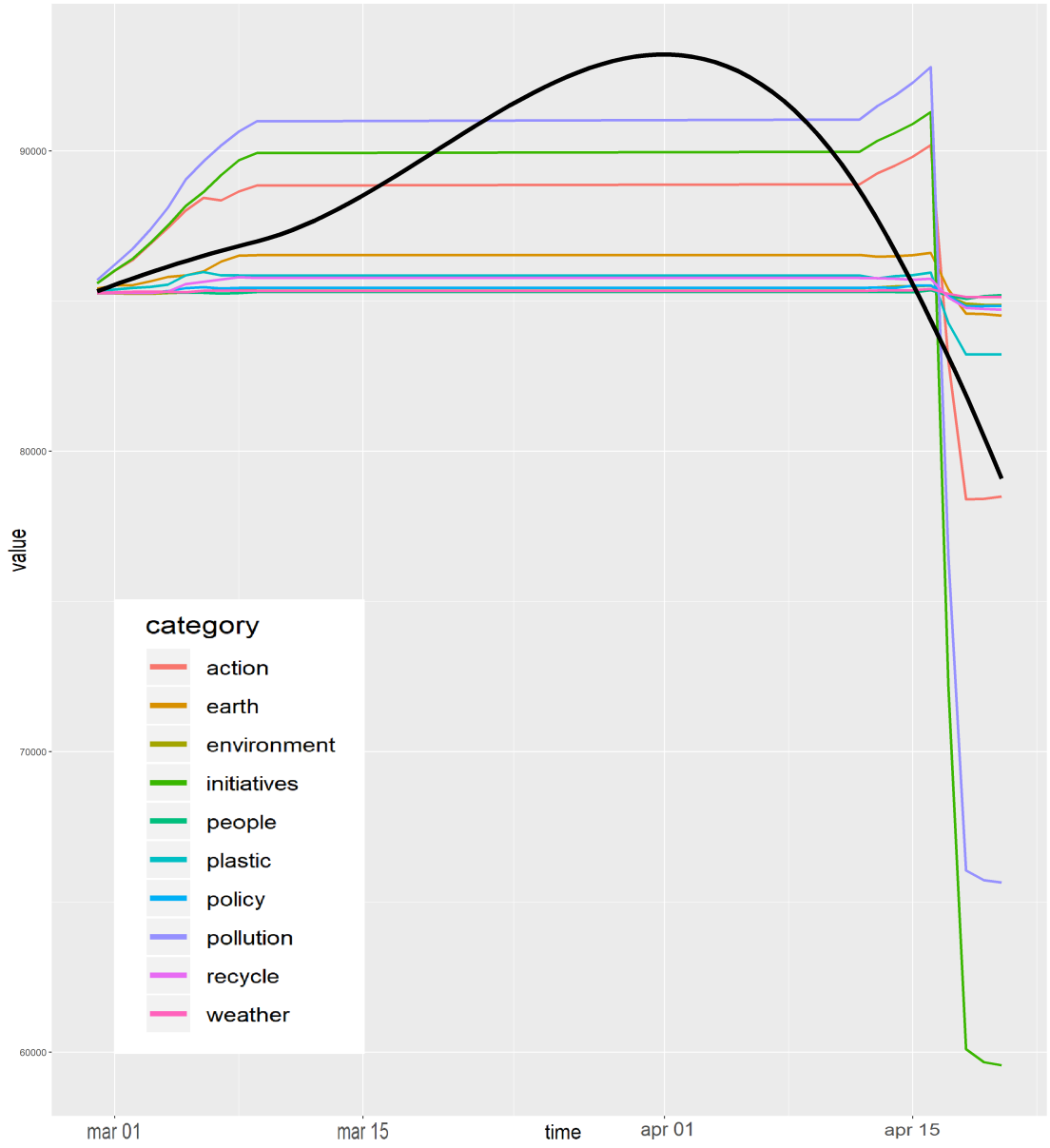


Figure 6.7: Scores of categories of words evolution in time;
the higher is a score, more positive is the sentiment and vice versa.
#savetheplanet Twitter data, Environment topic (Citizen perspective).
In black, the general sentiment score.

Chapter 7

Conclusions

With this thesis, we have shown how the Word of Mouth's importance is crucial to understand if what people think is right (or not) and quantifying the feeling that they post over the web about a specific topic. Mixing the high prediction power of the Neural Networks to produce the Words Embedding within the high interpretation capabilities of a model like Threshold-based Naïve Bayes was fundamental for that goal. With this approach, it is possible to notice many interesting, relevant conclusions in many fields cause just a single plot can contain multiple topics that can be compared simultaneously with an objective quantity. We will then continue to investigate that field further to improve the Threshold-based Naïve Bayes classifier performances and find more satisfying results. In particular, we have already planned to investigate some paths that can improve the General Sentiment Decomposition methodology. More in detail, we plan to:

- replicate all the "Use Cases" analysis (Chapter 6) but with the Iterative Threshold-based Naïve Bayes classifier instead of the Threshold-based Naïve Bayes classifier;
- produce our constructed Words Embedding despite not having much data, and then merging it within the pre-trained News-Paper Google version. This can improve the precision of the sentiment quantification;
- substitute, while automatically choosing the categories of words, the standard K-Means algorithm with a Semi-supervised clustering/community detection approach over the Words Embeddings representation;
- explore the analysis of trends that originates within the interpretation phase;
- add weights to the Threshold-based Naïve Bayes model to leverage the sentiment effect that words have. Recently we have observed how the COVID-19 had a significant impact on redefining the perception of certain words. For example, people were buying fewer "Corona" beers cause of the associated name. A model who bases his "sentiment

identification power" too much on past experience might be slow to identify a flex (point of major interest for decision-makers). A well-placed weight can probably help in such a flex-point identification;

- change the τ selection criteria. Recently, it was demonstrated that the MCC performance indicator is more reliable than ME within this context. Thus, we should update the procedure for selecting the best threshold (τ). In other words, we plan to substitute the minimization of the ME-value with the maximization of the MCC-value. Forby, the resulting classifier will make more balanced predictions.

Bibliography

Aakash, Aakash and Anu Gupta Aggarwal (2020), "Assessment of Hotel Performance and Guest Satisfaction through eWOM: Big Data for Better Insights." *International Journal of Hospitality & Tourism Administration*, 1–30.

Acosta, Joshua, Norissa Lamaute, Mingxiao Luo, Ezra Finkelstein, and Andreea Cotoranu (2017), "Sentiment Analysis of Twitter Messages Using Word2Vec." *Proceedings of Student-Faculty Research Day CSIS Pace University*, 7.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau (2011), "Sentiment Analysis of Twitter Data." *Association for Computational Linguistics*, 30–38.

Ahmad, Nawaz, Jolita Vveinhardt, and Rizwan Raheem Ahmed (2014), "Impact of Word of Mouth on Consumer Buying Decision." *European Journal of Business and Management*, 12.

Ahmad, Shimi Naurin and Michel Laroche (2015), "How Do Expressed Emotions Affect the Helpfulness of a Product Review? Evidence from Reviews Using Latent Semantic Analysis." *International Journal of Electronic Commerce*, 20, 76–111.

Ahmed, Murtadha, Qun Chen, and Zhanhuai Li (2020), "Constructing domain-dependent sentiment dictionary for sentiment analysis." *Neural Computing and Applications*, 32, 14719–14732.

Allard, Thomas, Lea H. Dunn, and Katherine White (2020), "Negative Reviews, Positive Impact: Consumer Empathetic Responding to Unfair Word of Mouth." *Journal of Marketing*, 84, 86–108.

Alpaydın, Ethem (2010), *Introduction to Machine Learning*, second edition. Adaptive Computation and Machine Learning, MIT Press, Cambridge, Mass.

Alshari, Eissa M., Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mustafa Alkeshr (2017), "Improvement of Sentiment Analysis Based on Clustering of Word2Vec

Features." In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, 123–126, IEEE, Lyon, France.

Anderson, Chris K. and Benjamin Lawrence (2014), "The Influence of Online Reputation and Product Heterogeneity on Service Firm Financial Performance." *Service Science*, 6, 217–228.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2011), "Deriving the pricing power of product features by mining consumer reviews." *Management Science*, 57, 1485–1509.

Arndt, J. (1967a), "Role of product-related conversations in the diffusion of a new product." *Journal of Marketing Research*, 4, 291–295.

Arndt, Johan (1967b), "Role of Product-Related Conversations in the Diffusion of a New Product." *Journal of Marketing Research*, 4, 291–295.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani (2010), "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta.

Banerjee, A. V. (1992), "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, 107, 797–817.

Barasch, Alixandra and Jonah Berger (2014), "Broadcasting and Narrowcasting: How Audience Size Affects What People Share." *Journal of Marketing Research*, 51, 286–299.

Bashir, Shahid, Muddasar Ghani Khwaja, Yasir Rashid, Jamshid Ali Turi, and Tariq Waheed (2020), "Green Brand Benefits and Brand Outcomes: The Mediating Role of Green Brand Image." *SAGE Open*, 10, 215824402095315.

Bashir, Shahid, Muddasar Ghani Khwaja, Jamshid Ali Turi, and Hira Toheed (2019), "Extension of planned behavioral theory to consumer behaviors in green hotel." *Heliyon*, 5, e02974.

Berger, Jonah and Raghuram Iyengar (2013), "Communication Channels and Word of Mouth: How the Medium Shapes the Message." *Journal of Consumer Research*, 40, 567–579.

Berrar, Daniel (2019), "Bayes' Theorem and Naive Bayes Classifier." In *Encyclopedia of Bioinformatics and Computational Biology*, 403–412, Elsevier.

Bickart, Barbara and Robert M Schindler (2001), "Internet forums as influential sources of consumer information." *Journal of Interactive Marketing*, 15, 10.

Bird, Steven, Ewan Klein, and Edward Loper (2009), *Natural Language Processing with Python*, first edition. O'Reilly Media Inc.

Blal, Inès and Michael C. Sturman (2014), "The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales." *Cornell Hospitality Quarterly*, 55, 365–375.

Blankertz, D. F. and D. Cox (1969), "Risk taking and information handling in consumer behavior." *Journal of Marketing Research*, 6, 110–111.

Bone, Paula Fitzgerald (1995), "Word-of-mouth effects on short-term and long-term product judgments." *Journal of Business Research*, 32, 213–223.

Boyd, Danah and Kate Crawford (2012), "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society*, 15, 662–679.

Bronner, Fred and Robert de Hoog (2011), "Vacationers and eWOM: Who Posts, and Why, Where, and What?" *Journal of Travel Research*, 50, 15–26.

Brown, Jo, Amanda J. Broderick, and Nick Lee (2007), "Word of mouth communication within online communities: Conceptualizing the online social network." *Journal of Interactive Marketing*, 21, 2–20.

Brownlee, Jason (2017), *Deep Learning for Natural Language Processing: Develop Deep Learning Models for Your Natural Language Problems*, 1.7 edition. Machine Learning Mastery.

Buttle, Francis A. (1998), "Word of mouth: Understanding and managing referral marketing." *Journal of Strategic Marketing*, 6, 241–254.

Campbell, Margaret and Amna Kirmani (2008), "I know what you're doing and why you're doing it: The use of Persuasion Knowledge Model in consumer research." In *Handbook of Consumer Psychology*, 549–573, Taylor & Francis Group/Lawrence Eribaum Associates.

Cantallops, A. S. and F. Salvi (2014), "New consumer behavior: A review of research on eWOM and hotels." *International Journal of Hospitality Management*, 36, 41–51.

Carl, M. McGlinn, W. J. and J. Oles (2007), "Measuring the ripple: Creating the G2XRelay rate and an industry-standard methodology to measure the spread of word-of-mouth conversations and marketing-relevant outcomes." *Measuring Word of Mouth*, 3, 36–46.

Chaiken, S. and D. Maheswaran (1994), "Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment." *Journal of Personality and Social Psychology*, 66, 460–473.

Chakravarty, Anindita, Yong Liu, and Tridib Mazumdar (2009), "Persuasive Influences of Online Word of Mouth and Professional Reviews." *Advances in Consumer Research*, 8, 124–125.

Chen, Zoey and May Yuan (2020), "Psychology of word of mouth marketing." *Current Opinion in Psychology*, 31, 7–10.

Cheung, Christy M.K. and Dimple R. Thadani (2012), "The impact of electronic word-of-mouth communication: A literature analysis and integrative model." *Decision Support Systems*, 54, 461–470.

Chevalier, Judith A. and Dina Mayzlin (2006), "The effect of word of mouth on sales: Online book reviews." *Journal of Marketing Research*, 43, 345–354.

Chicco, Davide and Giuseppe Jurman (2020), "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC Genomics*, 21, 6.

Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman (2021), "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation." *BioData Mining*, 14, 13.

Chiou, Jyh-Shen and Cathy Cheng (2003), "Should a company have message boards on its Web sites?" *Journal of Interactive Marketing*, 17, 50–61.

Choi, Youngtae and Junga Kim (2019), "Influence of Cultural Orientations on Electronic Word-of-Mouth (eWOM) in Social Media." *Journal of Intercultural Communication Research*, 48, 292–313.

Choudhari, Poonam and S. Veenadhari (2020), "Sentiment classification of online mobile reviews using combination of word2vec and bag-of-centroids." In *Machine Learning and Information Processing* (Debabala Swain, Prasant Kumar Pattnaik, and Pradeep K. Gupta, eds.), 69–80, Springer, Singapore.

Chu, Shu-Chuan and Yoojung Kim (2011), "Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites." *International Journal of Advertising*, 30, 47–75.

Cialdini, Robert B (2009), *Influence: Science and Practice*, volume 4. Pearson education Boston, MA.

Clemons, Eric K., Guodong Gordon Gao, and Lorin M. Hitt (2006), "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry." *Journal of Management Information Systems*, 23, 149–171.

Collobert, Ronan and Jason Weston (2008), "A unified architecture for natural language processing: Deep neural networks with multitask learning." In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 160–167, Association for Computing Machinery, New York, NY, USA.

Consiglio, Irene, Matteo de Angelis, and Michele Costabile (2018), "The Effect of Social Density on Word of Mouth." *Journal of Consumer Research*.

Conversano, Claudio, Maurizio Romano, and Francesco Mola (2019), "Hotel search engine architecture based on online reviews' content." In *Smart Statistics for Smart Applications*, 213–218, Pearson, Milano.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008), "Do online reviews matter? — An empirical investigation of panel data." *Decision Support Systems*, 45, 1007–1016.

Edvardsson, Bo, Bård Tronvoll, and Thorsten Gruber (2011), "Expanding understanding of service exchange and value co-creation: A social construction approach." *Journal of the Academy of Marketing Science*, 39, 327–339.

Eisenstein, Jacob (2017), "Unsupervised learning for lexicon-based classification." In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 3188–3194, San Francisco, CA, USA.

Esuli, Andrea and Fabrizio Sebastiani (2006a), "Determining term subjectivity and term orientation for opinion mining." In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy.

Esuli, Andrea and Fabrizio Sebastiani (2006b), "SENTIWORDNET: A publicly available lexical resource for opinion mining." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), Genoa, Italy.

Fatma Mobin, Ruiz Andrea Perez, Khan Imran, and Rahman Zillur (2020), "The effect of CSR engagement on eWOM on social media." *International Journal of Organizational Analysis*, 28, 941–956.

Fawcett, Tom (2006), "An introduction to ROC analysis." *Pattern Recognition Letters*, 27, 861–874.

Fellbaum, Christiane (1998), "A Semantic Network of English: The Mother of All WordNets." In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* (Piek Vossen, ed.), 137–148, Springer Netherlands, Dordrecht.

Fellbaum, Christiane (2010), "WordNet." In *Theory and Applications of Ontology: Computer Applications* (Roberto Poli, Michael Healy, and Achilles Kameas, eds.), 231–243, Springer Netherlands, Dordrecht.

Fellbaum, Christiane (2012), "WordNet." In *The Encyclopedia of Applied Linguistics* (Carol Chapelle, ed.), 8, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Firth, J. R. (1957), "A synopsis of linguistic theory 1930-1955." *Studies in linguistic analysis (special volume of the philological society)*, 1952-59, 1–32.

Ghose, A. and P. G. Ipeirotis (2011), "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics." *IEEE Transactions on Knowledge and Data Engineering*, 23, 1498–1512.

Ghose, Anindya and Panagiotis G Ipeirotis (2006), "Designing Ranking Systems for Consumer Reviews:." *Proceedings of the Sixteenth Annual Workshop of Information Technology and Systems*, 6.

Gibert, Karina, Jeffery S. Horsburgh, Ioannis N. Athanasiadis, and Geoff Holmes (2018a), "Environmental Data Science." *Environmental Modelling & Software*, 106, 4–12.

Gibert, Karina, Joaquín Izquierdo, Miquel Sànchez-Marrè, Serena H. Hamilton, Ignasi Rodríguez-Roda, and Geoff Holmes (2018b), "Which method to use? An assessment of data mining methods in Environmental Data Science." *Environmental Modelling & Software*, 110, 3–27.

Godes, David, Dina Mayzlin, and Yubo Chen (2005), "The Firm's Management of Social Interactions." *Marketing Letters: A Journal of Research in Marketing*, 16, 415–428.

Goldberg, Y. (2017), *Neural Network Methods in Natural Language Processing*. Synthesis Lectures on Human Language Technologies.

Grönroos, Christian and Päivi Voima (2013), "Critical service logic: Making sense of value creation and co-creation." *Journal of the Academy of Marketing Science*, 41, 133–150.

Gurney, Laura, John JD Eveland, and Indira R. Guzman (2019), ""What you say, I buy!": Information Diagnosticity and the Impact of Electronic Word-of-Mouth (eWOM) Consumer Reviews on Purchase Intention." In *Proceedings of the 2019 on Computers and People Research Conference*, 183–189, ACM, Nashville TN USA.

H., Kim L., Qu H., and Kim D. (2009), "A study of perceived risk and risk reduction of purchasing air-tickets online." *Journal of Travel and Tourism Marketing*, 26, 203.

Ha, Hong-Youl (2002), "The Effects of Consumer Risk Perception on Pre-purchase Information in Online Auctions: Brand, Word-of-Mouth, and Customized Information." *Journal of Computer-Mediated Communication*, 8.

Hajli, Nick, Xiaolin Lin, Mauricio Featherman, and Yichuan Wang (2014), "Social Word of Mouth: How Trust Develops in the Market." *International Journal of Market Research*, 56, 673–689.

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009), "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems*, 24, 8–12.

Halstead, D. (2002), "Negative word of mouth: Substitute for or supplement to consumer complaints?" *The Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior*, 15, 1.

Harris, Zellig S. (1954), "Distributional Structure." WORD, 10, 146–162.

Harrison-Walker, L. Jean (2001), "The Measurement of Word-of-Mouth Communication and an Investigation of Service Quality and Customer Commitment As Potential Antecedents." *Journal of Service Research*, 4, 60–75.

Hart, C., J. L. Heskett, and W. Sasser (1990), "The profitable art of service recovery." *Harvard business review*, 68, 148–156.

Hart, Christopher and P. Blackshaw (2006), "Internet Inferno-One customer can take down your company, but you can turn the potential nightmare into a boon." *Marketing Management*, 15.

Hartline, Michael D. and Keith C. Jones (1996), "Employee performance cues in a hotel service environment: Influence on perceived service quality, value, and word-of-mouth intentions." *Journal of Business Research*, 35, 207–215.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics)*, second edition. Springer-Verlag, New York.

Hauthal, Eva, Dirk Burghardt, Carolyn Fish, and Amy Griffin (2020), "Sentiment analysis." In *International Encyclopedia of Human Geography*, second edition, 169–177, Elsevier, Amsterdam, Netherlands.

Hennig-Thurau, T., K. Gwinner, G. Walsh, and Dwayne D. Gremler (2004a), "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?" *Journal of Interactive Marketing*, 18, 38–52.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler (2004b), "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?" *Journal of Interactive Marketing*, 18, 38–52.

Hermann, Karl Moritz and Phil Blunsom (2013), "The role of syntax in vector space models of compositional semantics." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 894–904, Association for Computational Linguistics, Sofia, Bulgaria.

Herr, Paul M., Frank R. Kardes, and John Kim (1991), "Effects of Word-of-Mouth and Product-Attribute Information on Persuasion: An Accessibility-Diagnosticity Perspective." *Journal of Consumer Research*, 17, 454.

Higie, R. A., L. F. Feick, and L. L. Price (1987), "Types and amount of word-of-mouth communications about retailers." *Journal of Retailing*, 63, 260–278.

Hochreiter, Sepp and Jürgen Schmidhuber (1997), "Long short-term memory." *Neural Computation*, 9, 1735–1780.

Holmes, J. H. and John D. Lett (1977), "Product sampling and word of mouth." *Journal of Advertising Research*, 17, 35–40.

Hussain, Safdar, Wasim Ahmed, Rana Muhammad Sohail Jafar, Ambar Rabnawaz, and Yang Jianzhou (2017), "eWOM source credibility, perceived risk and food product customer's information adoption." *Computers in Human Behavior*, 66, 96–102.

Hussain, Safdar, Wang Guangju, Rana Muhammad Sohail Jafar, Zahida Ilyas, Ghulam Mustafa, and Yang Jianzhou (2018), "Consumers' online information adoption behavior: Motives and antecedents of electronic word of mouth communications." *Computers in Human Behavior*, 80, 22–32.

Hwangbo, Hyunwoo and Jonghyuk Kim (2019), "A Text Mining Approach for Sustainable Performance in the Film Industry." *Sustainability*, 11, 3207.

Ismagilova, Elvira, Yogesh K. Dwivedi, and Emma Slade (2020), "Perceived helpfulness of eWOM: Emotions, fairness and rationality." *Journal of Retailing and Consumer Services*, 53, 13.

Iyengar, Raghuram, Christophe Van den Bulte, and Jeonghye Choi (2011a), "Distinguishing between drivers of social contagion: Insights from combining social network and co-location data." *Dartmouth College, Hanover, NH.*

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W. Valente (2011b), "Opinion Leadership and Social Contagion in New Product Diffusion." *Marketing Science*, 30, 195–212.

Jensen, Matthew L., Joshua M. Averbeck, Zhu Zhang, and Kevin B. Wright (2013), "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective." *Journal of Management Information Systems*, 30, 293–324.

Jun, Soo Hyun, Christine A. Vogt, and Kelly J. MacKay (2010), "Online information search strategies: A focus on flights and accommodations." *Journal of Travel & Tourism Marketing*, 27, 579–595.

Katz, E. and F. P. Lazarsfeld (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communications.* Taylor & Francis Ltd.

Khwaja, Muddasar Ghani, Ahmad Jusoh, and Khalil Md Nor (2019), "Does Electronic word-of-mouth (eWOM) on Social Media leads to Information Adoption? Empirical Evidence from the Emerging Markets!" *International Journal of Recent Technology and Engineering*, 8, 3281–3288.

Khwaja, Muddasar Ghani, Saqib Mahmood, and Ahmad Jusoh (2020a), "Online information bombardment! How does eWOM on social media lead to consumer purchase intentions?" *International Journal of Grid and Utility Computing*, 11, 857–867.

Khwaja, Muddasar Ghani, Saqib Mahmood, and Umer Zaman (2020b), "Examining the Effects of eWOM, Trust Inclination, and Information Adoption on Purchase Intentions in an Accelerated Digital Marketing Context." *Information*, 11, 478.

Khwaja, Muddasar Ghani and Umer Zaman (2020), "Configuring the Evolving Role of eWOM on the Consumers Information Adoption." *Journal of Open Innovation: Technology, Market, and Complexity*, 6, 125.

Kim, Yoon (2014), "Convolutional Neural Networks for Sentence Classification." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751, Association for Computational Linguistics, Doha, Qatar.

King, R., Pradeep Racherla, and Victoria D. Bush (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, 167–183.

Kirby, Justin, ed. (2006), *Connected Marketing: The Viral, Buzz and Word of Mouth Revolution*, transferred to digital print edition. Butterworth-Heinemann, Amsterdam.

Ladhari, Riadh and Mélissa Michaud (2015), "eWOM effects on hotel booking intentions, attitudes, trust, and website perceptions." *International Journal of Hospitality Management*, 46, 36–45.

Lee, Jumin, Do-Hyung Park, and Ingoo Han (2008), "The effect of negative online consumer reviews on product attitude: An information processing view." *Electronic Commerce Research and Applications*, 7, 341–352.

Leung, Daniel, Rob Law, Hubert van Hoof, and Dimitrios Buhalis (2013), "Social media in tourism and hospitality: A literature review." *Journal of Travel & Tourism Marketing*, 30, 3–22.

Lewis, David D. (1998), "Naive (Bayes) at forty: The independence assumption in information retrieval." In *Machine Learning: ECML-98* (Jaime G. Carbonell, Jörg Siekmann, G. Goos, J. Hartmanis, J. van Leeuwen, Claire Nédellec, and Céline Rouveirol, eds.), volume 1398, 4–15, Springer Berlin Heidelberg, Berlin, Heidelberg.

Li, Mengxiang, Liqiang Huang, Chuan-Hoo Tan, and Kwok-Kee Wei (2013), "Helpfulness of Online Product Reviews as Seen by Consumers: Source and Content Features." *International Journal of Electronic Commerce*, 17, 101–136.

Li, Xinxin, Lorin M. Hitt, and Z. John Zhang (2011), "Product Reviews and Competition in Markets for Repeat Purchase Products." *Journal of Management Information Systems*, 27, 9–42.

Liu, Z. and S. Park (2015), "What makes a useful online review? Implication for travel product websites." *Tourism Management*, 47, 140–151.

Loper, Edward and Steven Bird (2002), "NLTK: The Natural Language Toolkit." In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* -, volume 1, 63–70, Association for Computational Linguistics, Philadelphia, Pennsylvania.

Loureiro, S. and E. Kastenholz (2011), "Corporate reputation, satisfaction, delight, and loyalty towards rural lodging units in Portugal." *International Journal of Hospitality Management*, 30, 575–583.

Lu, Qi and Qiang Ye (2014), "Moderating effects of product heterogeneity between online word-of-mouth and hotel sales." *Journal of Electronic Commerce Research*, 15, 12.

Lu, W. and S. Stepchenkova (2012), "Ecotourism experiences reported online: Classification of satisfaction attributes." *Tourism Management*, 33, 702–712.

Luo, Chuan, Xin (Robert) Luo, Laurie Schatzberg, and Choon Ling Sia (2013), "Impact of informational factors on online recommendation credibility: The moderating role of source credibility." *Decision Support Systems*, 56, 92–102.

Matute Jorge, Polo-Redondo Yolanda, and Utrillas Ana (2016), "The influence of EWOM characteristics on online repurchase intention: Mediating roles of trust and perceived usefulness." *Online Information Review*, 40, 1090–1110.

Mazzarol Tim, Sweeney Jillian C., and Soutar Geoffrey N. (2007), "Conceptualizing word-of-mouth activity, triggers and conditions: An exploratory study." *European Journal of Marketing*, 41, 1475–1494.

Melancon, Joanna Phillips and Vassilis Dalakas (2018), "Consumer social voice in the age of social media: Segmentation profiles and relationship marketing strategies." *Business Horizons*, 61, 157–167.

Miguéns, J, R Baggio, and C Costa (2008), "Social media and Tourism Destinations: TripAdvisor Case Study." In *Advances in Tourism Research*, 1–6, Aveiro.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a), "Efficient Estimation of Word Representations in Vector Space." *arXiv:1301.3781 [cs]*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013b), "Distributed Representations of Words and Phrases and their Compositionality." *Advances in Neural Information Processing Systems*, 26, 3111–3119.

Miller, George A. (1986), "Dictionaries in the mind." *Language and Cognitive Processes*, 1, 171–185.

Miller, George A. (1995), "WordNet: A lexical database for English." *Communications of the ACM*, 38, 39–41.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990), "Introduction to WordNet: An On-line Lexical Database." *International Journal of Lexicography*, 3, 235–244.

Moore, Sarah G. and Katherine C. Lafreniere (2020), "How online word-of-mouth impacts receivers." *Consumer Psychology Review*, 3, 34–59.

Mudambi and Schuff (2010), "Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com." *MIS Quarterly*, 34, 185.

Mudinas, Andrius, Dell Zhang, and Mark Levene (2012), "Combining lexicon and learning based approaches for concept-level sentiment analysis." In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12*, 1–8, ACM Press, Beijing, China.

Mudinas, Andrius, Dell Zhang, and Mark Levene (2018), "Bootstrap domain-specific sentiment classifiers from unlabeled corpora." *Transactions of the Association for Computational Linguistics*, 6, 269–285.

Murray, Keith B. (1991), "A test of services marketing theory: Consumer information acquisition activities." *Journal of Marketing*, 55, 10–25.

Nabi, Robin L. (2003), "Exploring the Framing Effects of Emotion: Do Discrete Emotions Differentially Influence Information Accessibility, Information Seeking, and Policy Preference?" *Communication Research*, 30, 224–247.

Nam, Kichan, Jeff Baker, Norita Ahmad, and Jahyun Goo (2020), "Dissatisfaction, Disconfirmation, and Distrust: An Empirical Examination of Value Co-Destruction through Negative Electronic Word-of-Mouth (eWOM)." *Information Systems Frontiers*, 22, 113–130.

Ngarmwongnoi Chananchida, Oliveira João S., AbedRabbo Majd, and Mousavi Sahar (2020), "The implications of eWOM adoption on the customer journey." *Journal of Consumer Marketing*, 37, 749–759.

Nielsen (2007), "Trust in Advertising. A global Nielsen consumer report."

Nieto, Jannine, Rosa M. Hernández-Maestro, and Pablo Muñoz-Gallego (2014), "Marketing decisions, customer reviews, and business performance: The use of the Toprural website by Spanish rural lodging establishments." *Tourism Management*, 45, 115–123.

Nieto-García, Marta, Pablo A. Muñoz-Gallego, and Óscar González-Benito (2017), "Tourists' willingness to pay for an accommodation: The effect of eWOM and internal reference price." *International Journal of Hospitality Management*, 62, 67–77.

O'Connor, P. (2008), "User-generated content and travel: A case study on Tripadvisor.Com." In *Information and Communication Technologies in Tourism 2008*.

O'Connor, Peter (2010), "Managing a Hotel's Image on TripAdvisor." *Journal of Hospitality Marketing & Management*, 19, 754–772.

Öğüt, Hulisi and Bedri Kamil Onur Taş (2012), "The influence of internet customer reviews on the online sales and prices in hotel industry." *The Service Industries Journal*, 32, 197–214.

O'Leary, Steve and Kim Sheehan (2008), *Building Buzz to Beat the Big Boys: Word of Mouth Marketing for Small Businesses*. Westport, Conn.: Praeger Publishers.

Pan, Yue and Jason Q. Zhang (2011), "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews." *Journal of Retailing*, 87, 598–612.

Pancer, Ethan, Vincent Chandler, Maxwell Poole, and Theodore J. Noseworthy (2019), "How Readability Shapes Social Media Engagement." *Journal of Consumer Psychology*, 29, 262–270.

Pang, Bo and Lillian Lee (2008), "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval*, 2, 1–135.

Park, Do-Hyung and Sara Kim (2008), "The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews." *Electronic Commerce Research and Applications*, 7, 399–410.

Pavlou, Paul A. and Angelika Dimoka (2006), "The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation." *Inf. Syst. Res.*, 17, 392–414.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014), "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Association for Computational Linguistics, Doha, Qatar.

Phillips, Paul, Stuart Barnes, Krystin Zigan, and Roland Schegg (2017), "Understanding the impact of online reviews on hotel performance: An empirical analysis." *Journal of Travel Research*, 56, 235–249.

Phillips, Paul, Krystin Zigan, Maria Manuela Santos Silva, and Roland Schegg (2015), "The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis." *Tourism Management*, 50, 130–141.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002), "MultiWordNet: Developing an aligned multilingual database." In *Proceedings of the First International Conference on Global WordNet*.

Plé Loïc and Chumpitaz Cáceres Rubén (2010), "Not always co-creation: Introducing interactional co-destruction of value in service-dominant logic." *Journal of Services Marketing*, 24, 430–437.

Pollach, Irene (2008), "Electronic word-of-mouth: A genre approach to consumer communities." *International Journal of Web Based Communities*, 4, 284.

Powers, David (2011), "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies*, 24.

Raghunathan, Rajagopal, Michel T. Pham, and Kim P. Corfman (2006), "Informational Properties of Anxiety and Sadness, and Displaced Coping." *Journal of Consumer Research*, 32, 596–601.

Rajput, Dharmendra Singh, Ramjeevan Singh Thakur, S. Muzamil Basha, and Madjid Tavana, eds. (2019), *Sentiment Analysis and Knowledge Discovery in Contemporary Business:*. Advances in Business Information Systems and Analytics, IGI Global.

Reimer, Thomas and Martin Benkenstein (2016), "Altruistic eWOM marketing: More than an alternative to monetary incentives." *Journal of Retailing and Consumer Services*, 31, 323–333.

Richins, Marsha L. (1997), "Measuring Emotions in the Consumption Experience." *Journal of Consumer Research*, 24, 127–146.

Robson, Karen, Mana Farshid, John Bredican, and Stephen Humphrey (2013), "Making Sense of Online Consumer Reviews: A Methodology." *International Journal of Market Research*, 55, 521–537.

Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2018), "Persuasion, Emotion, and Language: The Intent to Persuade Transforms Language via Emotionality." *Psychological Science*, 29, 749–760.

Romano, Maurizio, Luca Frigau, Giulia Contu, Francesco Mola, and Claudio Conversano (2018), "Customer Satisfaction from Booking." In *Selected paper GARR_18 Data (R)evolution*, 111–118, Consortium GAAR, Cagliari.

Roy, Gobinda, Biplab Datta, Srabanti Mukherjee, and Rituparna Basu (2020), "Effect of eWOM stimuli and eWOM response on perceived service quality and online recommendation." *Tourism Recreation Research*, 1–16.

Rusticus, Sven (2007), *Creating Brand Advocates*. Justin Kirby and Paul Marsden.

Sajjanit, Chonlada (2018), "The influence of perceived service performance on generation Z's eWOM intentions in the food service sector." *Journal of Technology*, 4, 231–256.

Saleem, Anum and Abida Ellahi (2017), "Influence of Electronic Word of Mouth on Purchase Intention of Fashion Products on Social Networking Websites." *Pakistan Journal of Commerce and Social Science*, 11, 597–622.

Schegg, Roland and Miriam Scaglione (2013), "Substitution effects across hotel distribution channels." In *Information and Communication Technologies in Tourism 2014*.

Schindler, Robert M and Barbara Bickart (2004), "Published Word of Mouth: Referable, Consumer-Generated Information on the Internet." In *Online Consumer Psychology: Understanding and Influencing Consumer Behavior in the Virtual World*, 35–61, Lawrence Erlbaum Associates.

Schmunk, Sergej, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen (2013), "Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC." In *Information and Communication Technologies in Tourism 2014* (Zheng Xiang and Iis Tussyadiah, eds.), 253–265, Springer International Publishing, Cham.

Schuckert, Markus, Xianwei Liu, and Rob Law (2015), "A segmentation of online reviews by language groups: How English and Non-English speakers rate hotels differently." *International Journal of Hospitality Management*.

See-To, Eric W.K. and Kevin K.W. Ho (2014), "Value co-creation and purchase intention in social network sites: The role of electronic Word-of-Mouth and trust – A theoretical analysis." *Computers in Human Behavior*, 31, 182–189.

Senecal, Sylvain and Jacques Nantel (2004), "The influence of online product recommendations on consumers' online choices." *Journal of Retailing*, 80, 159–169.

Sernovitz, A. and G. Kawasaki (2006), *Word of Mouth Marketing: How Smart Companies Get People Talking*. Kaplan Publishing.

Sharma, Himanshu, Aakash Aakash, and Anu G. Aggarwal (2019), "The Role of Website Quality and Social Ties EWOM in E-Services Adoption." In *Structural Equation Modeling Approaches to E-Service Adoption*, 268–298, Yakup Akgul.

Sırma, Eda (2009), *Word-of-Mouth Marketing from a Global Perspective*. Ph.D. thesis, Instituto Universitàrio de Lisboa.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts (2013a), "Recursive deep models for semantic compositionality over a sentiment treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642, Association for Computational Linguistics, Seattle, Washington, USA.

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts (2013b), "Supplementary Material: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Solomon, M, G J Bamossy, S Askegaard, and M K Hogg (2006), *Consumer Behaviour: A European Perspective (3rd Edition)*. Prentice Hall.

Sparks, Beverley A. and V. Browning (2011), "The impact of online reviews on hotel booking intentions and perception of trust." *Tourism Management*, 32, 1310–1323.

Sparks, Beverley A., Helen E. Perkins, and Ralf Buckley (2013), "Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior." *Tourism Management*, 39, 1–9.

Stehman, Stephen V. (1997), "Selecting and interpreting measures of thematic classification accuracy." *Remote Sensing of Environment*, 62, 77–89.

Stringam, Betsy Bender and John Gerdes (2010), "An Analysis of Word-of-Mouse Ratings and Guest Comments of Online Hotel Distribution Sites." *Journal of Hospitality Marketing & Management*, 19, 773–796.

Tharwat, Alaa (2020), "Classification assessment methods." *Applied Computing and Informatics*, ahead-of-print.

Thorne, Sally (2016), *Interpretive Description: Qualitative Research for Applied Practice*, second edition. Routledge, New York.

Tiedens, Larissa Z and Susan Linton (2001), "Judgment Under Emotional Certainty and Uncertainty: The Effects of Specific Emotions on Information Processing." *Journal of Personality and Social Psychology*, 81, 973–988.

Ting, Kai Ming (2010), "Precision and Recall." In *Encyclopedia of Machine Learning* (Claude Sammut and Geoffrey I. Webb, eds.), 781–781, Springer US, Boston, MA.

Toh, Rex S., Peter Raven, and Frederick DeKay (2011), "Selling Rooms: Hotels vs. Third-Party Websites." *Cornell Hospitality Quarterly*, 52, 181–189.

Trusov, Michael, Randolph E Bucklin, and Koen Pauwels (2009), "Effects of Word-of-Mouth versus Traditional Marketing: Findings from an Internet Social Networking Site." *Journal of Marketing*, 73, 90–102.

Van Kleef, Gerben A. (2010), "The Emerging View of Emotion as Social Information: Emotion as Social Information." *Social and Personality Psychology Compass*, 4, 331–343.

van Laer, Tom, Jennifer Edson Escalas, Stephan Ludwig, and Ellis A van den Hende (2018), "What Happens in Vegas Stays on TripAdvisor? A Theory and Technique to Understand Narrativity in Consumer Reviews." *Journal of Consumer Research*, 267–285.

Vermeulen, I. and Daphne Seegers (2009), "Tried and tested: The impact of online hotel reviews on consumer consideration." *Tourism Management*, 30, 123–127.

Wang, Youcheng and Daniel R. Fesenmaier (2004), "Modeling participation in an online travel community." *Journal of Travel Research*, 42, 261–270.

Westbrook, R. A. (1987), "Product/Consumption-Based affective responses and postpurchase processes." *Journal of Marketing Research*, 24, 258–270.

Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara (1999), "Development and use of a gold-standard data set for subjectivity classifications." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 246–253, Association for Computational Linguistics, College Park, Maryland, USA.

Willemsen, Lotte M., Peter C. Neijens, Fred Bronner, and Jan A. de Ridder (2011), ""Highly Recommended!" The Content Characteristics and Perceived Usefulness of Online Consumer Reviews." *Journal of Computer-Mediated Communication*, 17, 19–38.

Wu, Tai-Yee and Carolyn A. Lin (2017), "Predicting the effects of eWOM and online brand messaging: Source trust, bandwagon effect and innovation adoption factors." *Telematics and Informatics*, 34, 470–480.

Xie, Hui, Li Miao, Pei-Jou Kuo, and Bo-Youn Lee (2011), "Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition." *International Journal of Hospitality Management*, 30, 178–183.

Xie, K., Chihchien Chen, and Shin-yi Wu (2016), "Online consumer review factors affecting offline hotel popularity: Evidence from tripadvisor." *Journal of Travel & Tourism Marketing*, 33, 211–223.

Yacouel, Nira and A. Fleischer (2012), "The role of cybermediaries in reputation building and price premiums in the online hotel market." *Journal of Travel Research*, 51, 219–226.

Yang, Yang, Noah J. Mueller, and Robertico R. Croes (2016), "Market accessibility and hotel prices in the Caribbean: The moderating effect of quality-signaling factors." *Tourism Management*, 56, 40–51.

Yang, Yang, Sangwon Park, and Xingbao Hu (2018), "Electronic word of mouth and hotel performance: A meta-analysis." *Tourism Management*, 67, 248–260.

Yin, Dezhi, Samuel Bond, and Han Zhang (2014), "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews." *MIS Quarterly*, 38, 539–560.

Yuan, Yu-Hsi, Sheng-Hao Tsao, Jiin-Tian Chyou, and Sang-Bing Tsai (2020), "An empirical study on effects of electronic word-of-mouth and Internet risk avoidance on purchase intention: From the perspective of big data." *Soft Computing*, 24, 5713–5728.

Yun, JinHyo Joseph, DaeCheol Kim, and Min-Ren Yan (2020), "Open Innovation Engineering—Preliminary Study on New Entrance of Technology to Market." *Electronics*, 9, 791.

Zhang, Dongwen, Hua Xu, Zengcai Su, and Yunfeng Xu (2015), "Chinese comments sentiment classification based on word2vec and SVMperf." *Expert Systems with Applications*, 42, 1857–1863.

Zhang, T., R. Agarwal, and H. Lucas (2011), "The value of IT-Enabled retailer learning: Personalized product recommendations and customer store loyalty in electronic markets." *MIS Q.*, 35, 859–881.

Zhou, Simin, Qiang Yan, Mengling Yan, and Chuwen Shen (2020), "Tourists' emotional changes and eWOM behavior on social media and integrated tourism websites." *International Journal of Tourism Research*, 22, 336–350.

Zhu, Feng (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics." *Journal of Marketing*, 133–148.

Zwass, Vladimir (2010), "Co-Creation: Toward a Taxonomy and an Integrated Research Perspective." *International Journal of Electronic Commerce*, 15, 11–48.