Università degli studi di Cagliari

PH.D. DEGREE
Business and Economics - Quantitative track

Cycle XXXIV

Eye Tracking and Sentiment Analysis to evaluate user behavior and opinion

Thesis in Statistics
(SECS-S/01)

PhD Student: Gianpaolo Zammarchi
Supervisor: Prof. Claudio Conversano

Final exam. Academic Year 2020 - 2021
Thesis defence: April 2022 Session

This thesis is dedicated to Claudia

Acknowledgements

There are many people I would like to thank for their contribution to this thesis and to my course of study. First of all, I would like to thank Prof. Francesco Mola for having convinced me to pursue this path, even when I didn't know that the path existed. I would also like to thank Prof. Claudio Conversano for having pushed me to always give my best and for his guidance during the whole doctorate.

I wish to thank Dr. Luca Frigau for his useful and sincere advice and Prof. Jaromír Antoch for his time and help during my internship. Finally, I would like to thank the whole research group for always making me feel supported and welcomed.

Last but not least, I would like to thank my wife Claudia. Without her my life would not be the same.

Abstract

This dissertation concerns the use and integration of two different techniques, eye tracking and sentiment analysis, to improve our ability to extract information about human behavior and opinion. These techniques can be applied in several different fields in which it is desirable to be able to model behavior patterns or to conduct opinion mining using automated methods. While for some applications it is possible to ask information directly (for instance through a survey or an interview), in several situations this could either be not feasible or involve a high risk that questions could be misinterpreted, the answers may be deceptive, or the subject might not even know the answer. Eye tracking and sentiment analysis allow to obtain knowledge from different types of raw data, i.e. gaze position coordinates during visualization of a stimulus (eye tracking) and texts (sentiment analysis). However, there are several challenges related to the way in which data are collected, processed and analyzed. The main problem this thesis aims to address is how we can improve our ability to obtain knowledge on human behavior and opinion using eye tracking and sentiment analysis, and how these two methods can be integrated to address this task. Besides illustrating different studies in which we applied these two techniques to study the behavior of different types of users, we describe a new method to improve performance of sentiment analysis by leveraging eye tracking data. First, we focus on eye tracking and show how this technique can be used to identify aspects of web pages or digital flyers that might benefit of improvement, in order to provide a better user experience. We also show how eye tracking data can be useful to accomplish image classification tasks. Next, we apply sentiment analysis to understand how sentiment towards Italy shifted during the first phases of the COVID-19 outbreak by analyzing a large data set of tweets. We compare different sentiment analysis tools, identify a common breakpoint corresponding to the shift of sentiment scores and show that this change can serve as an early predictor of the evolution of stock exchange values. Finally, based on the hypothesis that the eye tracking technology can provide a substantial contribution to identify words that are able to attract more attention, and are thus potentially more relevant, we present a new dictionary

that allows to perform sentiment analysis leveraging eye tracking data. We apply the Eye dictionary to the classification of different types of texts, showing that this tool is able to achieve a good performance, even when compared with dictionaries implementing a much higher number of words.

Contents

Introduction

Sentiment analysis can be seen as a Natural Language Processing (NLP) and information extraction operation, where the aim is to automatically analyze a text. While being able to identify subjective opinions in textual data is highly desirable for a wide range of applications, manual processing of these data is often time-consuming, labor-intensive and at risk of subjective bias. Therefore, a growing number of automated methods have been developed to perform this task. While different methods are available, usually the main purpose is to assess if the overall sentiment or opinion expressed by the author is positive, neutral or negative. If this task can be accomplished with a relatively high degree of confidence in the reliability of the final result, then it can be repeated for all the texts that we need to evaluate. Sentiment analysis can be performed at different levels, depending on the definition of "text". For example, if we consider a book chapter as a text, at the end of the analysis we'll have a positive, neutral or negative value representing the whole chapter. Conversely, if we want to analyze that same text in a sentence-by-sentence fashion, at the end of the analysis we will obtain several values (one for each sentence). The input for our analysis is the same, but the focus has shifted from the whole document to each single sentence. We can also focus on a specific entity if we're only interested into a single aspect. For example, if a review of various phones or other electronic devices is negative for all brands except ours, we may consider that review as positive, even if most of the text contains negative elements.

The main approaches in sentiment analysis are lexicon-based approaches, machine learning approaches or, more recently, deep learning approaches. In the lexicon-based approach there is a dictionary in which each word is assigned a polarity (e.g. negative or positive) and, possibly, a weight that represents its importance (i.e. the intensity of the polarity). Following this approach, each word of the text is matched with the dictionary and a summarizing function is applied to obtain a value representative of the whole text. In the machine learning approach, we need to fit a model to perform a supervised classification task. There are many models proposed in the literature to carry out this type of task [e.g. Naïve Bayes (NB), Support Vector Machine (SVM)], but the phases and often the performances are similar (each model requires a transformation of the textual data into numerical data, a training phase and then the performance evaluation). In the deep learning approach the goal is to try to accomplish the same task of a supervised model but exploiting an artificial neural network that can learn directly from the data. Even if these methods often allow to obtain a good performance, one limitation is that the process is not fully interpretable, since it can be difficult to assess why a text has been classified in a certain way. While machine and deep learning methods have been increasingly applied to sentiment analysis, the lexicon-based methods are still widely used. However, the performance of available dictionaries often varies based on the type of text under study. One of the most challenging aspects in the development of a dictionary is the definition of weights to apply to each word. These weights reflect the relevance of a word in the overall text and therefore the importance that this word will play in the computation of the general sentiment.

In this dissertation we propose a new dictionary to perform lexicon-based sentiment analysis by leveraging information that can be retrieved using the eye tracking technique: the Eye dictionary. The main problem that we aimed to address through this new dictionary is: When we determine the overall polarity of a text using sentiment analysis, is it possible to take into account the attention that words might receive from a reader?

This is where the eye tracking technique comes at hand. In fact, eye tracking allows to detect the precise position of the eyes during the observation of any type of stimulus (e.g. an image, a video or a text) and visually represent these points as a series of coordinates on a plane. If an object is able to attract a person's attention, a higher number of points will be found on the portion of the screen that contains that object (e.g. a word). In this way it is possible to obtain information related to the preferences or tastes of the person, or to infer the level of attention gained by each object without asking direct questions. This work is based on the assumption that if people, on average, spend more time on a word, that word should be assigned a higher relevance in the evaluation of the sentiment of the sentence.

Indeed, eye tracking is used in several studies focusing on emotion recognition. Human emotions are usually evaluated through the interpretation of voice, body and visual features. The ability to recognize emotions helps us to better understand each other (Lopes et al., 2003). Eye tracking can provide a useful contribution in emotion recognition (for a recent review see (Lim et al., 2020) and has been suggested to allow to detect basic emotions (such as disgust, fun and interest) with a high level of accuracy (up to 90%). Eye tracking is also widely used in the evaluation of individuals with disorders characterized by differences in the ability of recognizing emotions such as autism spectrum disorders (ASD) [for a scientometric analysis see (Zammarchi and Conversano, 2021)]. As the eye tracking has proven to be useful to recognize emotions, it can also be useful if we analyze how humans read texts. For example, in this area, some researchers have focused on the automatic detection of the type of document based on eye tracking behavior (Kunze et al., 2013b) while other used eye tracking to detect the entry point and the reading strategy of people reading newspapers (Holsanova et al., 2006). While these data point to a potential utility of eye tracking to improve our ability to evaluate different aspects of texts, to date no study has integrated eye tracking data in a dictionary for lexicon-based sentiment analysis.

This thesis is composed of four chapters. The first chapter introduces eye tracking through the description of its history, the main types of eye trackers, as well as the main fields in which this technique can be useful. The main contributions of this chapter are three practical applications in which we used the eye tracking technique to different fields: web usability of a university portal, digital media marketing and classification of images depicting city or natural landscapes. The second chapter introduces sentiment analysis, the main methods that can be used to perform it, as well as the main challenges as regards to data processing and interpretation. The practical application illustrated in this chapter describes an analysis of the evolution of sentiment towards Italy before and after the COVID-19 outbreak based on a large data set of tweets that we collected and processed using different lexicon-based and machine learning methods. We also showed that a shift in sentiment could serve as an early predictor of the stock market index values. The third chapter proposes a new method to combine eye tracking and sentiment analysis through the development of the Eye dictionary. We leverage eye tracking data from two large data sets to develop weights based on how much time each word was observed, considering the observation time as a proxy of its importance. We compare the performance of this method with existing dictionaries, showing that it outperforms dictionaries with a similar number of included words. The last chapter is a side-project focused on gamification where a pharmacology Android app has been developed and launched on the Google Play store.

Chapter 1

Eye tracking

1.1 Eye functioning and eye movements

1.1.1 Visual attention

When a human being interacts with the environment, one of the fundamental objectives is to try to obtain as much information as possible to complete the various tasks he/she is called upon to carry out. Perception occurs through the senses and sight, together with hearing, is one of the main means of capturing the stimuli that derive from interpersonal communication (Levy et al., 2017).
Reception of signals from the surrounding environment is an active rather than a passive process. The term active derives from action and the two processes - perception and action - present a bidirectional link that causes one to influence the other. Given the presence of this link, the question is: how can we measure visual perception? One obvious answer is through sight, which gives us a reasonable approximation of our visual perception. To get a precise measurement of what the eye is observing we can make use of the eye tracker, i.e. a device for measuring eye positions and eye movements with high precision.

The reason why a human being moves the eyes is to focus on a particular area of the field of view in such a way that it can be observed in detail. This observation is useful to focus our concentration precisely on the element we are looking at. Although understanding human behavior is certainly a complex topic, the ability to accurately assess where a person's gaze focuses the most is a big help in understanding what attracts the observer's attention and which cognitive processes underly human actions (Duchowski, 2017). Visual attention is a topic that has been studied for many years now. In the last decades the technical improvements allowed to reach very high

precision. However, it is necessary to define what attention is. In this sense, a good definition was given by the psychologist W. James: *"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others. When the things are apprehended by the senses, the number of them that can be attended to at once is small, Pluribus intentus, minor est ad singula sensus"* (James, 1981). Basically, it is about focusing on a few elements (or, even better, a single element) and trying to exclude the others, in order for all our cognitive abilities to focus. This is necessary due to the limits of the human being, who cannot pay attention to everything at the same way, and also to the functioning of the visual apparatus that does not "see" the scene as a whole, but only small areas that are put together by our brain to create a representation of the whole (Duchowski, 2017).
Leaving aside for now the technical functioning of the visual apparatus (which will be briefly discussed in Section 1.1.3), it must be emphasized that sight, and in particular eye movements, largely depend on the purpose for which one is looking. A classic experiment in this sense was conducted by one of the most famous scholars and pioneers in this field: Alfred Lukyanovich Yarbus. In his experiment, Yarbus asked the same person to observe a painting, trying each time to determine different information on the basis of what was represented in the scene (Yarbus, 1967). Depending on what was asked to the person, his eye movements scanned the image following completely different paths (Figure 1.1). Therefore, in addition to the mere visual capacity of a subject, also his/her intelligence, culture and a very wide range of elements that can greatly influence the final result come into play.

1.1.2 History of the study of eye movements

Since the early 1900s, various scholars have explored techniques capable of detecting eye movements. One of the first was Von Helmholtz who wrote in 1925: *"We let our eyes roam continually over the visual field, because that is the only way we can see as distinctly as possible all the individual parts of the field in turn"* (Von Helmholtz, 1925). Von Helmholtz was convinced that eye movements were an indicator of the subject's willingness to inspect a particular object more carefully. William James, on the other hand, in 1981 wrote that attention is very much linked to will, although he wanted to put emphasis on the non-voluntary component (James, 1981). Actually, the two hypotheses were not so in contrast with each other. In fact, if we consider

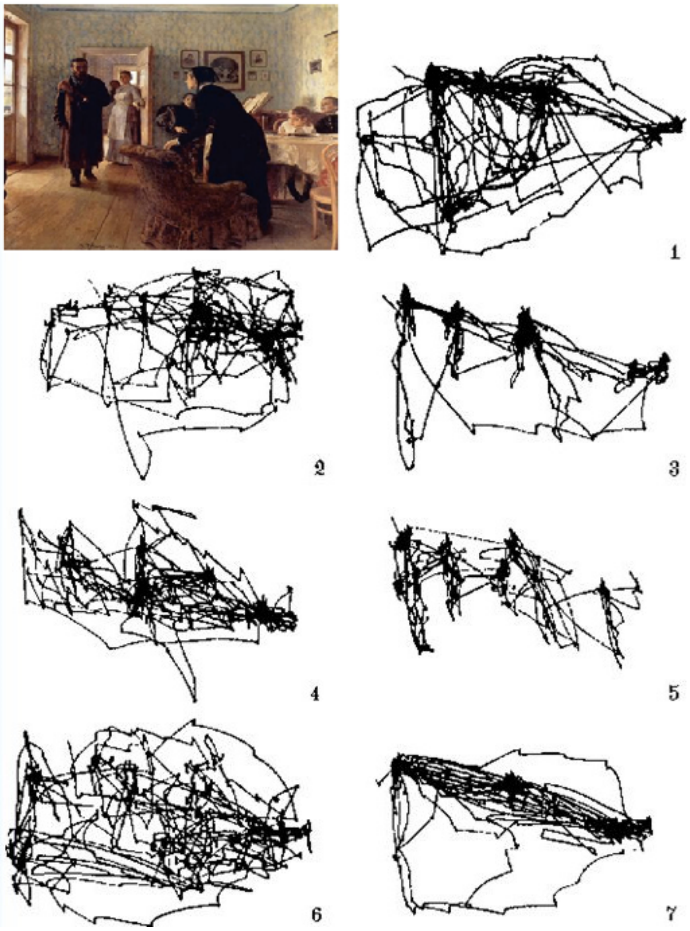


Figure 1.1: Eye movements from Yarbus experiment. The color image represents the painting by Il'ja Efimovic Repin called *They did not expect him* (1884-1888). Each number represents a path followed by the eyes of the participant in the Yarbus experiment who was required to look for specific information. Number 1: freely examine the painting; Number 2: estimate the economic condition of the family; Number 3: estimate the age of the people in the painting; Number 4: try to guess what people were doing before the guest arrived; Number 5: memorize people's clothing; Number 6: memorize the position of people and objects in the room; Number 7: estimate the time elapsed since the guest's last visit.

a stimulus coming from an image, this is first captured by peripheral vision (more closely linked to Von Helmholtz's concept of *where*) and then inspected with greater care (phase mostly linked to James's concept of *what*). This union of *where* and *what* forms the basis of the study of visual attention, which, however, on its own is not sufficient to fully explain the concept.

In 1941, Gibson introduced the concept of *how* into this discussion (Gibson, 1941). Our mind, in fact, partly influences the vision, as can be demonstrated with a simple experiment showing that an English term written in a deliberately incorrect way such as "sael" can be interpreted by the reader as "seal" if the text is related to the animal world or as "sail" if the context is seafaring. Several other scholars have contributed to refine the theoretical concepts behind the visual attention, including for example Posner (Posner et al., 1980), Treisman (Treisman and Gelade 1980; Treisman 1986) and (Kosslyn, 1994), providing their own point of view on this topic. Given the many contributions, it should therefore be expected that the topic does not require further study, but this is not the case. The model that describes the functioning of visual attention is still incomplete, as it explains some elements but not all of them. When an image or any other stimulus enters the eye's visual field, it is initially perceived by peripheral vision which, however, offers a low-resolution representation of the scene. This perception is enough to trigger a mechanism that interrupts the observation to move to the new object of interest. Finally, the elements that attract attention the most begin to be plumbed by the fovea to build a more defined image.

However, some factors still need to be clarified, such as: what types of features are needed to attract the attention of the human eye? If visual stimuli determine the movements of the human eye, why would we need voluntary movements? What exactly is the connection between attention and eye movements? Is attention always associated with the areas scanned by the fovea? In order to try to answer these and other questions, it is necessary to study not only the eye, but also the functioning of the areas of our brain that are activated during vision. These areas are responsible for the interpretation of visual stimuli captured by the eyes and not all stimuli cause the same reaction in the human visual system (Duchowski, 2017).

1.1.3 How the image reaches the brain

Vision is not only a technical process, but it involves a very strong component of interpretation that takes place in our brain. Since ancient times, philosophers considered the organ of sight to be the mind, not the eye (Pliny, Natural History, Latin: *Naturalis Historia*). Over time, many theories have been developed on how the eye and the brain work together and which are the respective tasks. To date, vision is considered a process where the reality that surrounds us is interpreted and transformed. The procedure begins in the retina, when the light hits millions of receptors and makes a complex, albeit very rapid, journey towards our brain. In order to reach the visual cortex, however, the signal needs to be coded to be able to travel as an electrical stimulus along the nerves. Along this journey, the visual information is modified, i.e. eliminating or emphasizing parts. In any case, each part of the visual system contributes to the composition of this information, defining aspects such as shapes, colors, movement of objects, position in space and much more. It is a complex system, with interacting components that participate in the process following a sort of hierarchy, using information coming from the lower-level areas, proceeding in an elaborate sequence of changes aimed at breaking down and recomposing the information received (Maffei, 2007). The visual pathways formed in the retina follow a route towards the brain, passing through the optic nerve, the optic tract and the optic chiasm. The complex system of lenses inside the eye presents an image that is both reversed and partial in the two eyes. Therefore, a phase of elaboration by the brain is needed. Specifically, the right hemisphere is in charge of processing the left visual field, while the left hemisphere processes the visual field of the right eye (Figure 1.2).

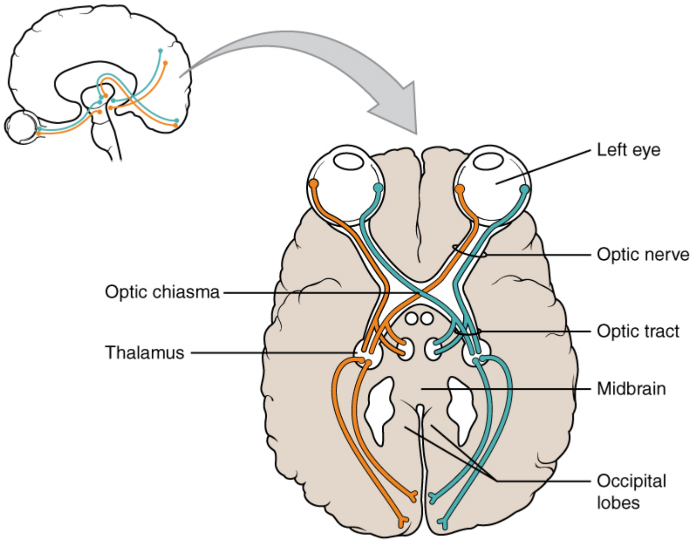


Figure 1.2: Visual pathway from the eye to the brain

Some of the nerves that cross the optic chiasm are also connected to their own hemisphere, thus allowing the autonomy of both eyes (Levy et al., 2017). Once the visual cortex is reached, the signals are processed to reconstruct the image. In this area there are links that ensure that the areas related to memory are connected, thus also activating the visual memory center. Different areas of the cortex are reached by the signal. In particular, the main area is the primary visual cortex (V1) which is located in the occipital lobe of the brain, within the so-called striated area or area 17. Any injury to the V1 area can cause blindness.

1.1.4 Main eye movements: fixations and saccades

The width of the visual field for a healthy person with steady head and eyes is about 200° horizontally and 135° vertically (Dagnelie, 2011). If we allow head and eyes to move, the visual field can be extended, but it must be considered that only a small portion of this field is used to capture details from elements around us, because the rest is mainly used to perceive movements in the surrounding environment. For this reason, head and eyes movements are crucial to get an idea, as broad as possible, of what surrounds us (Rayner, 1998). Specifically, the visual field is the area of the space that surrounds us where an individual is able to perceive visual stimuli.

In eye tracking studies, two main types of eye movements are generally considered: fixations (i.e. when the eye is staring at some element to explore it) and saccades (i.e. when the eye moves from one element to another). Fixations and saccades are much more frequent than we might expect since, to fully explore an element, the eye automatically makes various adjustments called micro-saccades. A series of fixations and saccades is called a *scanpath*. The oculomotor apparatus is the neural system that deals with eye movements through three pairs of muscle bundles: medial and lateral rectus, upper and lower rectus, upper and lower oblique. In general, the main (voluntary) movements that we can make to capture the stimuli coming from the visual field are:

1. fixations;
2. saccades (saccadic eye movements);
3. smooth pursuit.

Fixations are eye movements that stabilize the retina on a static object of interest. They are characterized by a combination of tiny eye movements: tremors, drifts and micro-saccades. Among the three, the micro-saccades represent the largest and fastest type of movement. They are also involuntary, with a movement dynamic similar to saccades. The drifts are movements that follow a curved trajectory and are between the micro-saccades. Finally, tremors (or ocular microtremors) are the smallest movements among the three and represent oscillations that occur simultaneously with the drifts (Martinez-Conde and Macknik, 2015).

Saccades or saccadic movements are rapid eye movements used to reposition the fovea on a new element of the environment. The word probably comes from a French term related to the rapid movement of a sail (Gregory, 1990). These are movements that can be both voluntary and reflex (i.e. activated by corrective optokinetic mechanisms). Saccades have durations ranging from 10 to 100 milliseconds (ms) and are defined as ballistic and stereotyped movements (Shebilske, 1983). The term ballistic means that in some cases the final destination is set before the movement is made, while the term stereotyped means that some specific movement patterns are repeated systematically (Duchowski, 2017).

Smooth pursuit are the typical movement made by the eye following a moving object. This action may seem trivial or obvious, but it is actually a very complex operation. One would think that the slow pursuit movements are the result of a series of fixations over time, or a more complex version of a single fixation, but this is not the case. First of all, the eye adjusts its movement based on the speed of the object observed, and then it reconstructs the movement of the object in a three-dimensional space by establishing its trajectory and trying to make predictions based on the information collected (Thier and Ilg, 2005).

1.2 The eye tracking technique

The eye tracking technique allows to measure and analyze eye movements with a high degree of precision, which would not be possible to reach with other methods. A relevant example of a field in which this can be highly useful could be measuring the attention gained by some products based on how they are arranged along the aisle of a supermarket. Without the eye tracker, it is almost impossible to accurately determine where the subject is actually looking or for how long he/she has observed a product. Even if we can ask the person directly, he/she may not remember or not report the observation in the correct way. Eye tracking allows researchers to study the eye movements of the subjects while they are engaged in different activities. This allows to draw conclusions about the cognitive process that takes place in our mind when we are observing something and can be useful to increase our knowledge on e.g. the learning or interaction process. It is also particularly useful in case the subject is unable to autonomously express concepts, as in the case of infants or people with specific conditions.

The devices used to track the movement of the eye are known as *eye trackers*. Eye trackers can measure eye movement in different ways:

- measurement of the movement of an object (usually, a particular type of contact lens) that moves along with the eye;
- measurement of the electrical potential of the areas around the eye using electrodes;
- optical tracking without the need for contact (less invasive).

Techniques that measure eye movements can assess the position of the eye with respect to the head or the orientation of the eye in space (POR, Point Of Regard) (Young and Sheena, 1975). The measurement of the latter is generally used when the aim is the identification of elements in a visual scene. The most used device for measuring POR is the eye tracker that uses corneal reflection. In addition, eye movement measurement methodologies can involve Electro-OculoGraphy (EOG), Scleral Contact Lens, Photo- or Video-oculography (POG or VOG), and video recording of pupillary / corneal reflection (Duchowski, 2017).
Electro-oculography is based on recordings of the electrical potential differences of the skin surrounding the eye socket. This technique measures eye movements as a function of the position of the head and is therefore not very precise in determining where the eye is actually looking. During the mid-1970s, this technique was the most widely applied method (Young and Sheena, 1975). Today the most applied eye movement technique, mainly used for POR measurements, is the method based on corneal reflection. The first method for objective eye measurements using corneal reflection was reported in 1901 (Robinson, 1968). To improve accuracy, contact lens techniques were implemented in the 1950s. They relied on physical contact with the eyeball, which provides very accurate measurements. The problem with these devices is related to the invasiveness of the technique for the subject involved in the experiment. The POG includes several variants that perform the measurement through: (i) the shape of the pupil, (ii) the position of the limbus (the border between iris and sclera) or (iii) other aspects. All these variants are not able to measure the POR and need, in many cases, to properly fix the head of the participant to avoid movement (Duchowski, 2017). To be aware of the exact point that the subject is observing, an alternative to fix the head of the participant (so that the allowed movements are only those of the eyes) is to use additional information related to the corneal reflex and the center of the pupil. Eye trackers based on video detection use cameras mounted on a stand in front of the subject or directly around the head. The corneal reflex in response to an infrared light beam is measured as a function of the position

of the center of the pupil. With adequate calibration, these instruments provide very precise measurements of the POR. Even using these devices, the problem of invasiveness for the subjects participating in the experiment still exists. Technical evolution of eye trackers has significantly improved this aspect. In fact, many of the more modern devices are very small and light, similar to normal eyeglasses, or do not require any physical contact at all (Duchowski, 2017). The modern eye tracker exploits technologies based on infrared light (*invisible near-infrared light*) and high-definition video cameras to project into the human eye a light beam that reaches the cornea. To track eye movement, the eye tracker uses infrared diodes that generate a reflection on the subject's cornea. These reflections, together with other data, are collected by the sensors of the instrument. Next, sophisticated algorithms involving complex mathematical calculations are used to determine the exact eye position and, therefore, the exact point where the participant is looking at. In this way it is possible to measure and study the gaze and all the movements that the eye makes when looking at an object. Since the position can be mapped several times per second, the eye tracker can produce a visual map of how the person looked at those objects. The sampling frequency of an eye tracker is the number of times per second in which the position of the eyes is captured by the device. Higher frequency means a higher number of points collected to describe the true path followed by the eye, but also means more data to analyze and noise to filter out (Duchowski, 2017). There are different types of eye trackers and each has its own peculiarities, although they function in a similar way. Each type of eye tracker is suitable for specific experiments, depending on the conditions in which they are carried out. Eye trackers can either be used as stand-alone devices or need to be mounted on other devices (e.g. a computer). The main types of eye trackers are:

Eye trackers for screen: these devices can be mounted on the screen of a laptop or a desktop PC monitor (Figure 1.3). This type of eye tracker is mainly used when the subjects participating in the study look at elements on a screen. Depending on the sampling frequency, the instrument can collect a large amount of data related to each participant's gaze shift. Some models are able to collect reliable data even in the case of large head movements and are useful in the case of studies including babies/children or patients with specific pathologies that prevent them from controlling their movements.

Wearable eye trackers: these devices can be worn as glasses (Figure 1.3) or integrated into virtual reality (VR) devices. The type mounted on glasses is ideal to study the behavior of the subject in "real" situations, where one is immersed in a physical environment. Typical examples are the observation of a supermarket shelf,

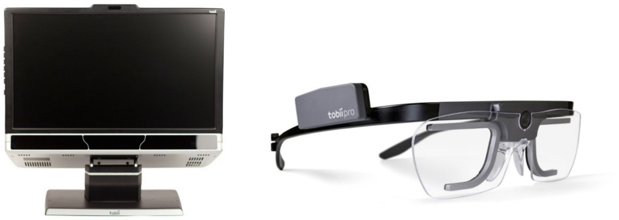


Figure 1.3: Eye tracker for screen (on the left) and wearable eye tracker (on the right)

a sporting event, the workplace and so on. The glasses do not prevent any movement and encourage a natural behavior of the subject. Virtual reality (VR) devices are similar to glasses, but they allow to conduct experiments without having to move from the test site. Reality is artificially created around the subject who can thus, for example, live situations of great danger in total safety. In addition to the safety aspect, a VR device allows, for example, to wander around a shopping center when it has not been built yet, to verify the correct location of the elements.

Webcam eye tracker: in this case the device is equipped with a webcam or is integrated into a computer and uses the PC webcam for detection. This type of eye tracker does not use infrared lights or special cameras but is based on the images captured by the webcam. Through an algorithm it is possible to determine the head and eyes position and, based on these, to establish which area of the screen is observed. The accuracy of the measurements obtained using this method is often lower, but it can be useful in pilot studies or in the design phase of a study.

The eye tracker produces a series of objective and quantifiable data, which are not influenced by the opinions of the subjects carrying out the study or by those who analyze the results. The technical evolution of these tools also allows to reproduce conditions similar to natural ones, that is, to ensure that the subject behaves naturally. It is also possible to conceive the experiment in a wide variety of different environments, making it a very versatile tool, being able to also collect real-time information that allow the researcher to correct errors even during the experiment, if necessary. In medicine, the eye tracker is also widely used in combination with other tools, such as electroencephalogram (EEG) or electrocardiogram (ECG), to conduct studies in which different types of data will be collected and analyzed jointly. The next section provides an overview of the main fields in which eye tracking can be applied.

1.3 Fields of eye tracking research

The eye tracking technique is used in several fields. Among the main applications there are (i) market research, (ii) user experience evaluation, and (iii) applications in medicine.

1.3.1 Market research

This type of research is mainly focused on how consumers act and which elements drive their decision-making processes. Studies can be carried out to investigate what happens below the most apparent level (i.e. the physical action), when a consumer sees or buys a product. It is possible to understand which are the elements that attract the attention the most and which are ignored. Unlike surveys or questionnaires, the data collected are not influenced by the opinion or will of the subject, and they reproduce in a more realistic way the purchasing process. The eye tracking is increasingly used to assess the impact of different aspects of packaging (color, size, brand positioning) on consumption habits for many products. For example, a recent review of the literature assessed whether a standardized packaging among the various cigarette manufacturers (making each manufacturer adopt the following standards: uniform color, absence of logos, presence of health hazard warnings and brand reported with predetermined font, color and size) could reduce tobacco consumption (McNeill et al., 2017). The data reported in the review refer to Australia (the first country to introduce the measure in 2012). The studies used eye tracking to compare classic packaging with standardized ones, showing that the brand is an element capable of attracting a large number of fixations and divert attention from health hazard messages (McNeill et al., 2017). Eye tracking was also used to assess the impact of nutrition labels' information and positioning on purchase decisions for different products (Popova et al. 2019; Bix et al. 2015; Antunez et al. 2013; Graham 2012; Miller and Cassady 2012; Graham 2011).

1.3.2 User experience

The best way to evaluate the user experience is to see what the user sees. Using the eye tracker, it is possible to understand how a typical user browses a website and how much a page is able to meet the user's expectations. This allows to identify design flaws and discover new uses that were not initially taken into consideration. Web usability can be defined as the approach aimed at designing websites that an end user can browse in a way as simple as possible, without the need of specific training (ec.europa.eu). In other words, the user should be able to intuitively understand which actions it is necessary to perform on a web page (for example, press a button, click on a link) to achieve a certain goal. The main objectives of web usability can be summarized in:

- present the information to the user in a clear and concise way;
- make the correct choices appear obvious;
- remove any ambiguity regarding the consequences of an action (for example, clicking on a link will complete the purchase);
- position the most important elements within a page or web application correctly.

It can therefore be said that the evaluation of web usability allows to obtain an estimate of the efficiency of a web page or site. Evaluation of the usability of a site is among the most interesting applications of eye tracking. Modern websites use high-level graphics, because part of the users' judgment also passes through the visual appeal. Graphics have become a means of communication and for many people, good graphics is an indicator of the efficiency of a website. Even when the efficiency is not excellent, websites with appealing graphics receive higher ratings compared with sites with better performance but less appealing aesthetics (Brady and Phillips, 2003). A cognitive bias causes people to rate the website with the best graphics more positively, regardless of whether this is well designed or not (Brady and Phillips, 2003). For this reason, the study of eye movements is useful to understand what people think or what mainly drives their attention.

Eye tracking is becoming a fundamental technique to evaluate web usability. In fact, the analysis of eye movements and users' interaction with the elements of a webpage is a profitable complement and expansion to the information collected through traditional metrics (for example the time required to complete a task) (Wang et al., 2018). Most studies have focused on evaluating websites or individual pages, although there are examples of further applications, including evaluating the ease of use of geographic maps on the web (Liao et al. 2019; Liao et al. 2017; Manson et al. 2012) or which representation techniques are more effective on small devices, such as smartphones (Kuhnel et al. 2017; Duh et al. 2006) and smartwatches (Park et al., 2020).

As regards to the evaluation of web pages' usability, several studies have examined the efficiency related to the execution of various tasks, including navigation, searching information within individual pages, and the use of search engines. The eye tracking provides various metrics that can be used to evaluate the efficiency of web pages. Many previous studies have measured web usability using the number of fixations or their duration (Poole et al. 2004; Jacob and Karn 2003; Goldberg 1999). Other metrics used include saccades (Vuori et al., 2004) and scanpath (Goldberg et al., 2002). A useful contribution to the use of eye tracking in evaluating web usability was offered by Ehmke (2007), who evaluated the efficiency of the BBC television broadcaster's website and of a travel booking website. The completion of the proposed tasks required navigation through links and menus, viewing images and searching for information within individual pages (Ehmke, 2007). The study included 19 participants and was carried out using a Tobii X50 eye tracker. The authors identified usability problems for many participants and correlated these problems with patterns recorded by eye tracking. For example, the authors correlated the "lack of information expected on a page" problem with a pattern that can be described as a large number of short fixations within that page (Ehmke, 2007). An important field that has seen a growing application of eye tracking to study web usability is tourism (Scott et al., 2019). In particular, it was studied how people perceive images related to tourist activities and landscapes (Wang and Sparks 2016; Li and Chen 2014) and how they interact with websites that advertise such activities (Muñoz-Leiva et al. 2019; Djamasbi et al. 2010) or accommodation facilities (Pan et al., 2013). Numerous studies have successfully used eye tracking to evaluate the efficiency of e-commerce portals (Oyekunle et al. 2020; Bach 2018; Hwang 2017; Roth et al. 2013). In fact, since in a virtual environment it's not possible to evaluate the physical characteristics of a product, consumers can only rely on the information and images shown on web pages. Available studies have tried to identify which visual elements influence consumers' purchasing behaviors the most and also reflect their cognitive processes (Hwang, 2017). Finally, a small number of authors used eye tracking to evaluate the web usability of websites of university libraries or other educational institutions (Zardari et al. 2020; Ritthiron and Jiamsanguanwong 2017; Lamberz et al. 2018). In particular, Lamberz and colleagues (2018) evaluated the web usability of the German educational institute "Bildungswerk Grafschafter Wirtschaft". The evaluation of the site was carried out through three tasks aimed at searching for specific courses. Each task was carried out by 8-12 students. The use of eye tracking made it possible to reconstruct and analyze the behavior of the participants during the navigation and information search phases (Lamberz et al., 2018). By evaluating the heat maps and metrics related to fixations, the authors highlighted which elements within the pages were confusing or slowed down the completion of tasks (for example the positioning of the navigation bar) (Lamberz et al., 2018).

1.3.3 Other applications

In the last few years, an increasing number of studies investigated gaze behavior in individuals with different neurological or psychiatric disorders during the visual perception of emotional stimuli (Zammarchi and Conversano, 2021). In the medical field, eye tracking has been applied to study eating disorders (Schag et al. 2021; Stott et al. 2021; Stojek et al. 2018; Scherr et al. 2017), autism (Zhang et al. 2021; Black et al. 2017; Frazier et al. 2017) and other psychiatric or neurological pathologies (Hunter and Chin 2021; Beltràn et al. 2018; Chen and Clarke 2017). As a specific example, patients with schizophrenia or bipolar disorder show distinct characteristics of eye movements during smooth pursuit and visual search (Morita et al., 2020). While studies on eye movement characteristics in these disorders have been conducted since the early 1900s, the advent of the eye tracking technology has allowed to more precisely evaluate these impairments. Finally, eye tracking has also recently been used in the field of learning, alongside more traditional educational techniques, for example in the study of scientific disciplines (Yang et al. 2018; Andrà et al. 2015), languages (Augereau et al. 2016; Kunze et al. 2013a) or in improving the ability to drive a vehicle (Kapitaniak et al. 2015; McDonald et al. 2015).

Eye tracking allows us to have a unique point of view of what, in many cases, occurs only at the subconscious level. Thanks to the observation of videos or graphical outputs such as gaze plots and heat maps, it is possible to pause or analyze something in detail to assess aspects that would be otherwise too complex to evaluate or, sometimes, just to perceive.

1.4 Application I: UniCa website

While different studies applied the eye tracking methodology to evaluate web usability, as reviewed in Section 1.3.2, in this study we applied for the first time Markov chain analysis to eye tracking data to thoroughly assess web usability of a university website through a comprehensive set of tasks. Specifically, we evaluated the web usability of the website of the University of Cagliari (www.unica.it) aiming to identify strengths as well as potential problems and suggest improvements, with a specific focus on users with limited knowledge of the website (Zammarchi et al., 2021).

Web usability of a website, and this is particularly true for websites of institutions, needs to be guaranteed to different kind of users. A university website can be used by (i) students needing to search information regarding courses, university buildings, offices, professors contacts, timetables and so on; (ii) prospective students looking for information regarding enrolling and different courses they might be interested into; (iii) university employees such as professors, researchers, technicians and PhD students; (iv) other users such as companies looking forward to establishing partnerships with the university for joint projects and contribute to the education of future employees.

In a website that offers a good web usability, users (even first-time users) should be able to find the information they need as quickly as possible. In order for this to be feasible, a website should therefore present a clear organization of contents and allow to quickly move between sections based on the needed information. This also implicates that single page need to be easily readable and accessible, in order for a user to be able to realize in a simple way whether the correct page has been reached. Our evaluation of the web usability of the portal of the University of Cagliari involved two groups of students representing two of the main types of users: currently enrolled students and prospective students (high school students). The two groups of students performed a comprehensive set of tasks aimed at evaluating different parts of the website. We combined qualitative and quantitative analysis of eye tracking data using different metrics (number of fixations, task duration and a difficulty ratio). Next, we used Markov chain to analyze the participants' behavior in respect to the main areas of interest (AOI) of the home page. In addition, we compared performances of university and high school students to assess whether users with low degrees of knowledge of the website are able to use it in an efficient way. Aims of the study were the following:

- evaluate differences in web usability and page efficiency based on the type of users and the relative level of knowledge of the website;
- understand which pages are characterized by high efficiency and which ones might be improved;
- understand viewing behaviors of users (e.g. check whether they scroll the page or limit their observation to the visible portion of the page), in order to improve positioning of relevant information;
- propose solutions to improve less efficient pages.

1.4.1 Design of the tasks

The website of the University of Cagliari (UniCa, Italy) *www.unica.it* was launched in 2017 and was the first Italian university portal to meet criteria established by

Agid (Agenza Italiana digitale del Consiglio dei Ministri). According to these criteria, a website should present a user-friendly layout, good readability and contents optimized for people with disability. The UniCa's website is responsive (pages can be accessed by any device) and offers a site section specifically designed for prospective students.

We designed ten tasks that required an active search within different sections of the web portal. A description of the ten tasks is reported in Table 1.1. These tasks, which all started from the home page and required a similar number of pages to be opened, either asked to reach a specific page of the website or to retrieve a specific information within a page.

These tasks cover a comprehensive set of pages that are considered to be relevant for students or prospective students.

Table 1.1: Description of the tasks

Task	**Description**	**# pages**[1]
Task 1	Find the main page of the Engineering course	3
Task 2	Find the page with instructions on how to use the University Wi-fi service	3
Task 3	Find the page with indication on how to participate to admission tests	3
Task 4	Find the deadline to register to admissions tests	2
Task 5	Find admission tests given in previous academic years in the Pharmacy course	4
Task 6	Find the email of the Secretary Office of the University	3
Task 7	Find the tuition fee regulation	3
Task 8	Find the page of the Human Sciences library	4
Task 9	Find the page of the University Sport Center (CUS)	3
Task 10	Find the instructions to enroll online	2

[1] Ideal (minimum) number of pages to open to complete the task

1.4.2 Participants

Two groups of students were included. The first group (prospective students) included randomly selected students enrolled in different types of high schools (i.e., in Italy high schools can be art-oriented, humanistic-oriented, tech-oriented, and so on, or a mix of these tracks) in the island of Sardinia, Italy, who took part to a university fair and open day to be informed about faculties. The second group included university students, mainly from economics and law departments, randomly selected in group study rooms (in different days of the week and different times of the day). Each student was randomly assigned one of the ten tasks. For each participant, information regarding age, gender, high school institute and, for university students, university course, were collected.

1.4.3 Data collection

For each participant, eye movements during the task were gathered with a screen-based Tobii X2-60 Compact eye tracker (Figure 1.4), which captures gaze data at 60 Hz, applied to a 25-inch monitor. This device is compact and is therefore useful for studies that need to be conducted outside a lab.

Figure 1.4: Eye tracker Tobii X2-60 Compact

This sensible and precise device is able to continue detection in case of interruptions. For example, in case a subject blinks, the eye tracker loses the ability to record eye movements since the eye and, in particular the pupil, is hidden by the eyelid. If the pupil is hidden only for a short time, then the software is able to continue the detection without interruption as soon as the pupil becomes visible again. However, the subject must have kept his head substantially motionless during the blinking. The device can be mounted on any type of screen, both on desktop and laptop PCs or even tablets, with a simple adhesive strip. The technical data sheet of the device used in the study is reported in Table 1.2 (Tobii Studio User Manual V. 3.3.1.).

The eye tracking can be monocular or binocular, depending on whether the data refer only to the dominant eye of the participant or the metrics are calculated as the average of viewing behavior of both eyes. Gaze detection precision refers to the angular spatial variation between a set of consecutive points. Accuracy, on the other hand, concerns the average angular distance between the real observed point and the one measured by the eye tracker. An example of measurements of accuracy and

Table 1.2: Technical sheet of Tobii X2-60 Compact eye tracker

Feature	**Description**
Sampling rate	60 Hz
Std. dev. of sampling rate	0.1 Hz (approx..)
Accuracy	Binocular: from 0.4° to 1.2° Monocular: from 0.4° to 1.9°
Precision	Monocular: 0.34 Binocular: 0.45
Freedom of head movement	Width x height: 50 x 36 cm (20 x 14”) @70 cm Operating distance: 40–90 cm (15.7–33.5”)
Latency	Total system latency: < 35 ms
Time to tracking recovery	For blinks: immediate
Recommended screen size	Up to 25” (16:9)
Data sample output	Timestamp Eye position Gaze point Pupil diameter Validity code

precision is shown in Figure 1.5, where the dashed red line represents the subject’s gaze direction, while the solid line the point measured by the eye tracker.

Another important element is the recovery time of the detection. In case the connection between the device and the eye is lost, it must be recovered as soon as possible. When the eye tracker is unable to detect the eyes in the area where they were last detected, a search will begin in the surrounding area after a few milliseconds. If the eye tracker is unable to detect the participant’s eyes after a minute, the system will enter a “slow search” mode, with longer recovery times. Finally, an important information to take into account during the experiments is the operating distance. This represents the range of distances [minimum: 40 cm (15.7 inches); and maximum: 90 cm (33.5 inches)] where the eyes of the participant can be detected by the eye tracker sensor (Tobii Studio User Manual V. 3.3.1.).

In order to analyze and collect the data, the device is equipped with specific software. For this study, Tobii Studio (ver. 3.3.1) was used, which allows to make recordings of what is happening on the screen (e.g. display of video, images) as well as to log time, position and other useful information for each observed point. Through the combination of all these data it is possible to produce different types of graphical outputs. Among the most used, heat maps are graphical representa-

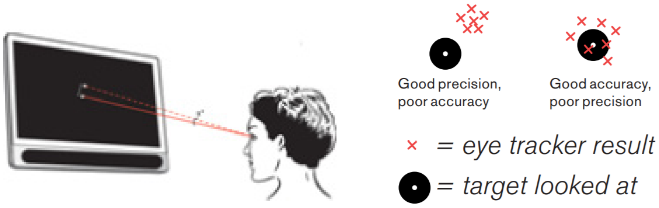


Figure 1.5: Eye tracker precision and accuracy. From. Tobii X2-60 Eye tracker technical specification

tion of data with colors ranging from red (areas of the page with more fixations) to green (areas of the page with less fixations), while gaze plots show fixations in the exact order in which they occurred (Dong et al., 2014). During the practical test, participants perform tasks without realizing what is happening in the software. At the end of the test it is possible to show some graphical output to the participant to further analyze the visual behavior. This part could be useful, for example, to a researcher who wants to add a qualitative part to the data collection or to increase the participant's level of involvement in the later stages of the study.

For the current study, participants received instructions on the study procedures and were told not to use search engines (internal or external to the site) to complete the task. Before each task, the instrument was calibrated based on the height of the participant as well as distance from the screen. During the task, the registration was monitored using an additional screen in order to detect potential problems. At the end of the data collection, the raw data were processed in order to classify eye movements into different types (e.g. fixation, saccade, etc.) using the velocity-threshold fixation identification (I-VT) algorithm implemented in Tobii Studio v. 3.3.1. This algorithm classifies eye movements based on the velocity of the directional shifts of the eye. In the current analysis we mainly focused on fixations, which are considered the most interesting metric in previous studies (van der Lans et al., 2011) based on their ability to indicate the moment in which an information is most probably registered by the brain (Rayner, 1998).

1.4.4 Comparison of performances between high school and university students

A total of 56 high school (Group 1) and 66 university students (Group 2) were included and were all able to complete the assigned task. After exclusion of 5 participants from Group 1 and 3 from Group 2 due to technical problems occurred during the task, 51 and 63 participants from Group 1 and Group 2 were included in the analyses, respectively (Table 1.3).

Table 1.3: Demographic characteristics of the sample

	Group 1	**Group 2**	**Statistics**	**p**
Number	51	63	-	-
Female (%)	41.2	50.8	1.05^1	0.306
Age (mean ± SD)	17.88 (± 0.86)	23.41 (± 3.11)	28.00^2	<0.001
Previous use of the website (%)	43.1	73.0	10.45^1	0.001

[1]Pearson's χ^2; [2]Mann-Whitney U. Abbreviations: SD, standard deviation
Group 1: high school students, Group 2: university students

Student's t test showed that, as expected, the two groups differed in terms of age, while Pearson's chi-squared test showed no significant gender difference between high school and university students (Table 1.3). A higher proportion of students from the group of university students, as expected, had previously used the UniCa website (Table 1.3). We computed three metrics to evaluate performances of participants: time to completion of the task (defined as the time passed between the last and the first fixation for each participant), the total number of fixations and a difficulty ratio dR$_i$ calculated as:

$$\mathrm{dR}_i = \frac{n_V}{n_T} \tag{1.1}$$

where i = 1, ..., m is the task identifier, n_V is the number of pages visited to complete a task and n_T is the minimum number of pages required to be visited to complete a task (Goldberg et al., 2002). All analyses were conducted using R v. 3.6.1 (R Core Team, 2019).

All participants successfully completed the task within six minutes (mean ± standard deviation [SD], Group 1: 1.98 ± 1.33 minutes, Group 2: 1.63 ± 1.27 minutes). Times to completion across the tasks ranged from 15 seconds to 5.54 minutes for Group 1 and from 21 seconds to 5.43 minutes for Group 2. We observed a significant positive correlation between time to completion and the other two metrics (fixations:

Spearman's rho = 0.81, $p < 0.0001$; difficulty ratio: Spearman's rho = 0.83, $p < 0.0001$). Tasks for which the metrics were found to be above two SD from the mean were considered to be particularly complex. While no task showed such complexity based on time to completion or number of fixations, for Task 4 the mean difficulty ratio in both groups of participants was above two SD from the mean (Group 1: Task 4: 6.88, mean of all tasks = 3.38; SD = 1.61; Group 2: Task 4: 6.42, mean of all tasks = 3.10; SD = 1.60).

For each task, normality of distribution for the three metrics was tested with the Shapiro-Wilk test (Shapiro and Wilk, 1965). Performances between the two groups of students were compared using Student's t test (with or without Welch's correction for unequal variances according to results of Levene's test) in case of normal distribution. Otherwise, Mann–Whitney U test was used. The Bonferroni correction was used to adjust results for multiple testing based on ten tasks. A p-value < 0.005 (i.e. 0.05/10) was considered significant. For the majority of the tasks, no significant differences were observed in the performance of the two groups of students after multiple testing correction (Table 1.4).

For Task 8 (in which participants were asked to retrieve the page of the Human Sciences library) university students showed a better performance based on all three metrics. However, only the number of fixations was significantly different between the two groups after multiple testing correction.

Tasks completed in a less efficient way by high school compared to university students might allow to identify specific pages that could be improved to make them easier to browse for users with scarce previous knowledge of the website. The worse performance shown by high school students in Task 8 might have been prompted by a different viewing behavior in the home page. In order to explore this hypothesis, analyses with Markov chain were conducted, as will be explained in more details in the next section.

In addition, high school students showed a higher difficulty ratio for Task 5 (Table 1.4). In other words, students included in this group opened a much higher number of pages compared to the minimum number required to complete the task. This could be explained by a lack of understanding of the meaning of specific links (and therefore an increased number of pages opened to reach the right one) or by an increased difficulty in the retrieval of the information even after reaching the correct page. This aspect was investigated through a qualitative analysis on heat maps and gaze plots, which were produced using fixations coordinates.

The gaze plot of the destination page of Task 5 (Figure 1.6) showed a reduced number of fixations located in the area containing the needed information, while

Table 1.4: Comparison of the viewing behavior between the two groups of high school (Group 1) and university students (Group 2)

			Time to completion				**Number of fixations**				**Difficulty ratio**			
Task	$\mathbf{n}_{G1}$	$\mathbf{n}_{G2}$	$\bar{T}_{G1}$	$\bar{T}_{G2}$	**Stat**	**p**	$\bar{N}_{G1}$	$\bar{N}_{G2}$	**Stat**	**p**	$\bar{D}_{G1}$	$\bar{D}_{G2}$	**Stat**	**p**
Task 1	8	7	66.61	75.49	-0.50^{1}	0.626	176.13	177.71	-0.03^{1}	0.978	2.00	2.71	-1.42^{1}	0.182
Task 2	2	6	82.57	72.21	0.26^{1}	0.834	145.00	188.67	-0.51^{1}	0.628	2.50	3.08	-0.91^{1}	0.402
Task 3	9	5	134.29	53.41	37.0^{2}	0.060	275.78	118.40	39.00^{2}	0.029	3.55	1.60	37.0^{2}	0.060
Task 4	4	6	170.54	140.82	0.44^{1}	0.681	397.25	439.67	-0.18^{1}	0.860	6.88	6.42	0.15^{1}	0.884
Task 5	3	6	218.40	172.69	1.74^{1}	0.134	519.00	225.83	2.38^{1}	0.049	4.33	3.17	6.67^{1}	0.0003
Task 6	3	9	198.02	119.06	21.00^{2}	0.209	443.67	147.22	23.00^{2}	0.100	4.22	3.96	16.50^{2}	0.641
Task 7	5	6	139.36	182.42	-0.98^{1}	0.355	347.60	391.83	-0.27^{1}	0.794	4.22	4.06	-2.07^{1}	0.071
Task 8	5	7	140.77	46.86	4.02^{1}	0.007	348.60	86.29	3.68^{1}	0.004	2.40	1.36	32.5^{2}	0.016
Task 9	8	6	69.83	47.95	26.00^{2}	0.852	175.75	124.00	25.00^{2}	0.950	1.50	1.44	14.5^{2}	0.212
Task 10	4	5	68.46	56.67	0.56^{1}	0.591	160.75	115.60	0.72^{1}	0.498	2.17	1.53	1.27^{1}	0.283

[1] Student's t; [2] Mann-Whitney U. Significant results after multiple testing correction (using a threshold $p < 0.005$) are reported in bold. Abbreviations: G1, Group 1; G2, Group 2; n, number; Stat, statistics, $\bar{T}$, mean time to completion; $\bar{N}$, mean number of fixations; $\bar{D}$, mean difficulty ratio.

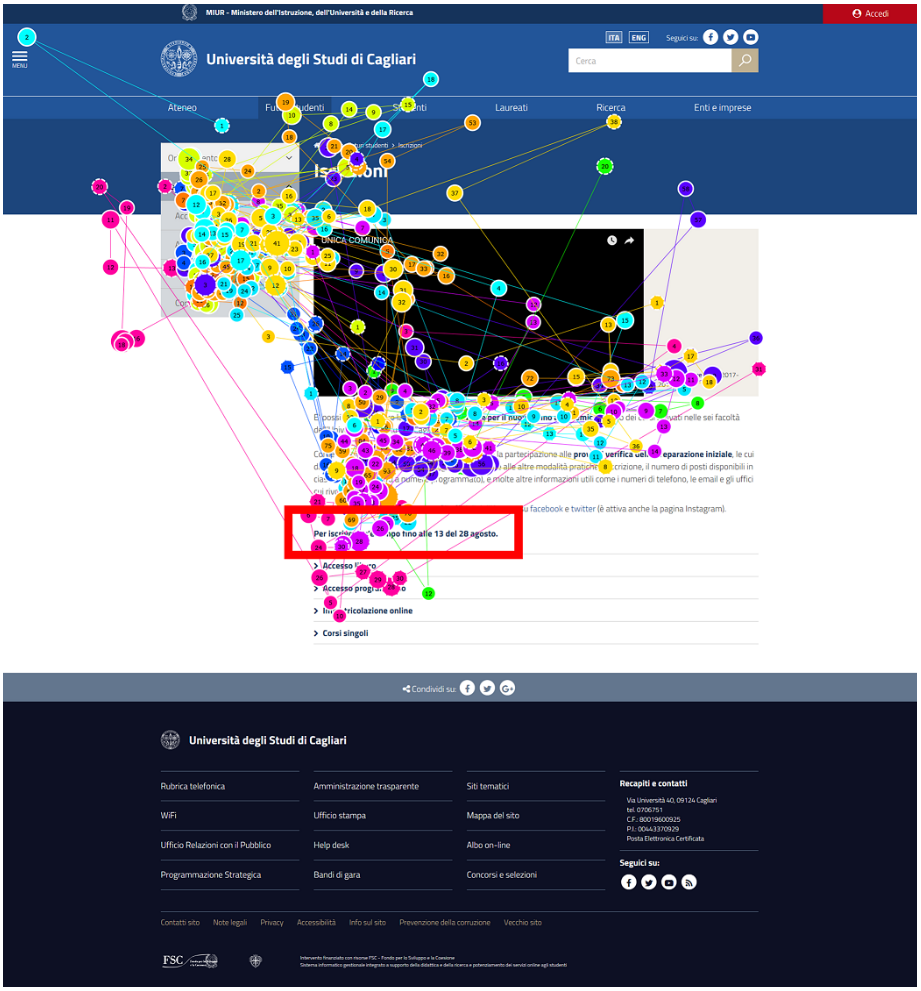

MIUR - Ministero dell'Istruzione, dell'Università e della Ricerca
Accedi
MENU
Università degli Studi di Cagliari
ITA
ENG
Seguici su:
Cerca
Ateneo
Futuri studenti
Studenti
Laureati
Ricerca
Enti e imprese
Futuri studenti > Iscrizioni
Iscrizioni
UNICA COMUNICA
Per iscriversi c'è tempo fino alle 13 del 28 agosto.
Accesso libero
Accesso programmato
Immatricolazione online
Corsi singoli
Condividi su:
Università degli Studi di Cagliari
Rubrica telefonica
WiFi
Ufficio Relazioni con il Pubblico
Programmazione Strategica
Amministrazione trasparente
Ufficio stampa
Help desk
Bandi di gara
Siti tematici
Mappa del sito
Albo on-line
Concorsi e selezioni
Recapiti e contatti
Via Università 40, 09124 Cagliari
tel. 0706751
C.F.: 80019600925
P.I.: 00443370929
Posta Elettronica Certificata
Seguici su:
Contatti sito
Note legali
Privacy
Accessibilità
Info sul sito
Prevenzione della corruzione
Vecchio sito
FSC
Intervento finanziato con risorse FSC - Fondo per lo Sviluppo e la Coesione
Sistema informatico gestionale integrato a supporto della didattica e della ricerca e potenziamento dei servizi online agli studenti

Figure 1.6: Gaze plot of the destination page of Task 5

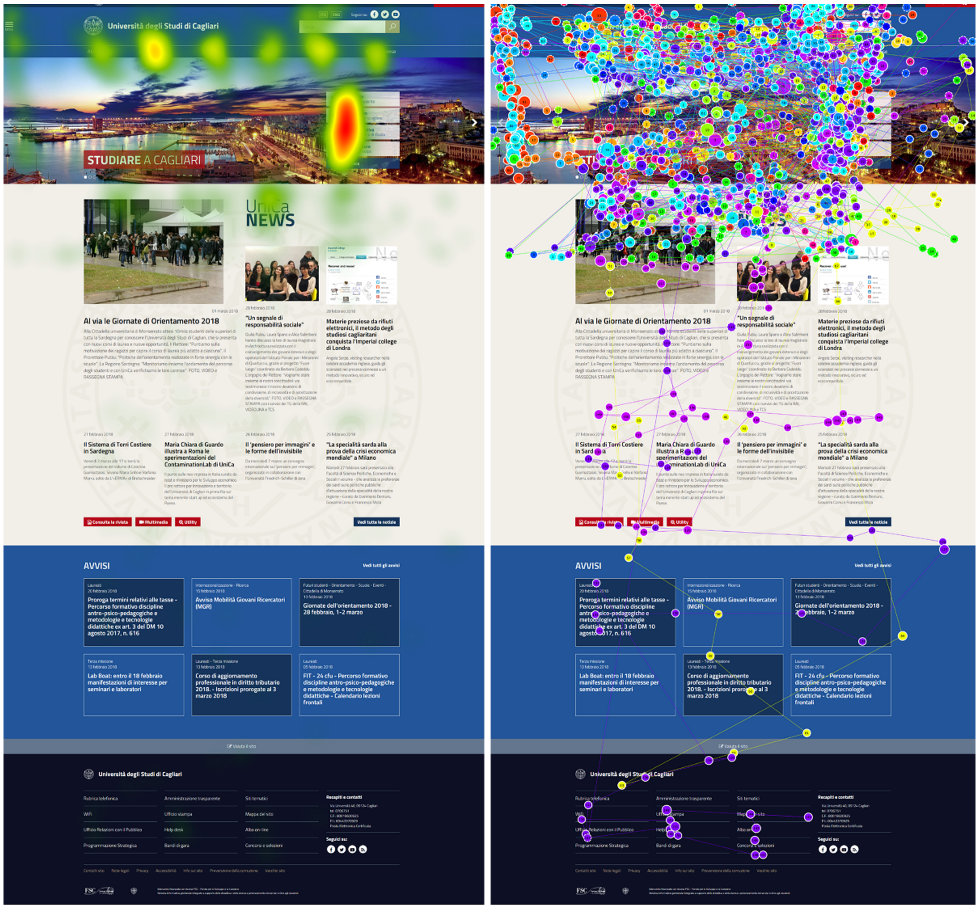

Università degli Studi di Cagliari
STUDIARE A CAGLIARI
UniCa NEWS
Al via le Giornate di Orientamento 2018
AVVISI
Università degli Studi di Cagliari

Figure 1.7: Heat map and gaze plot of the UniCa homepage

the large majority of participants focused on the navigation bar. Therefore, most participants failed to retrieve the information and did not realize they had reached the correct page. This observation is useful to plan design of pages containing relevant information, such as the deadline to register to admission tests, in order to make them more visible. In addition, such information would be better placed in the upper part of the screen, as we found that the majority of users do not scroll a page but merely observe the visible part (Figure 1.7). The left part of the figure shows the heat map of the home page for high school students. In the heat map, clusters of fixations are represented with colors ranging from red (areas with a higher number of fixations) to green (areas with less fixations). The gaze plot (on the right) shows the sequence of fixations in the order in which they occurred. Overall, the fact that all participants were able to complete assigned tasks, as well as the lack of differences in the performances of two groups of students for the majority of the tasks, suggest that the website offers a good web usability even to users with a low level of previous knowledge of the site.

1.4.5 Analysis of the viewing behavior using Markov chain

A Markov chain is a discrete time stochastic process $X_0,\ X_1,\ \ldots,\ X_t$ where the following property holds: the distribution of X_t only depends upon X_{t-1}, while it is independent from all previous values $X_0,\ X_1,\ \ldots,\ X_{t-2}$, exception made for the last one X_{t-1} (Gilks et al., 1996). This can also be written as a conditional probability,

$$\begin{aligned} & P\left[X_{t=x} \mid X_0 =\ x_0,\ X_1 =\ x_1,\ \ldots,\ X_{t-1} =\ x_{t-1}\right] = \\ & P[Xt = x|X_{t-1} = x_{t-1}] = \\ & p_{ij} \end{aligned} \tag{1.2}$$

where p_{ij} in Equation (1.2) is the probability that the chain jumps from state i to state j and $i,\ j$ are states of a countable set A. Let a transition probability be the probability of transitioning from state i to state j in one jump. For the transition probabilities, the following property holds: $\sum_{j\in S} p_{ij} = 1$ with $i \in$ S. The matrix $P =\ (p_{ij})$ represents the transition matrix of the chain (Serfozo, 2009). The stationary distribution of a Markov chain with transition matrix P is a vector of probabilities π that follows the property: $\pi P\ =\ \pi$. To converge to a stationary distribution, a Markov chain needs to be irreducible, recurrent and aperiodic (Gilks et al., 1996). Let X be defined as a Markov chain $P\left[X_t = x \mid X_{t-1} = x_{t-1}\right]$ then, X is called irreducible if:

$$P_{ij}\left(t\right) > 0 \quad \text{for some } t > 0 \qquad \forall i, j \tag{1.3}$$

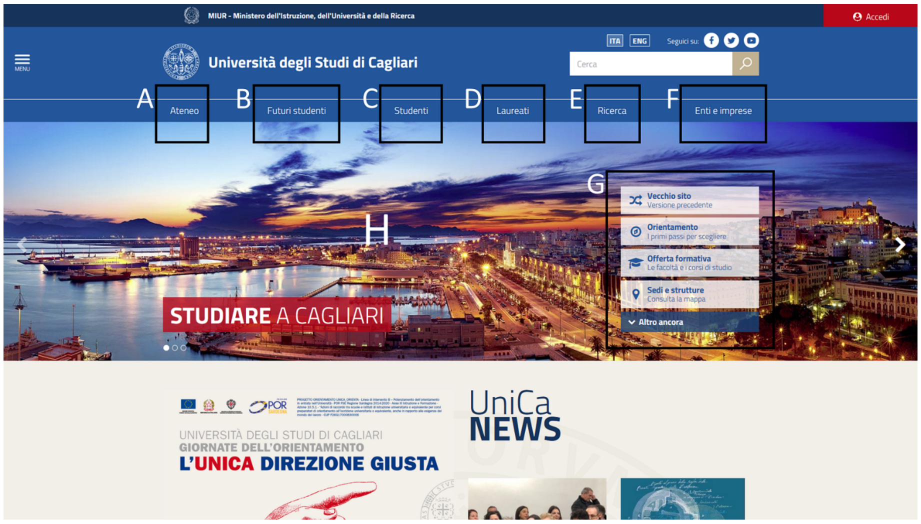


Figure 1.8: Definition of the areas of interest (AOI) in the UniCa home page. Definition of the AOIs drawn across the main buttons of the menu in the home page. AOIs are indicated with letters from A to G, while the area outside any AOI is indicated with the letter H.

An irreducible chain X is recurrent if:

$$\sum P_{ij}\left(t\right) = \infty \qquad \forall i, j \tag{1.4}$$

An irreducible chain X is called aperiodic if:

$$\text{greatest common divider}\left\{t > 0:\ P_{ii}\left(t\right) > 0\right\} = \ 1 \qquad \forall i. \tag{1.5}$$

If a chain X satisfies properties of Equation (1.3), (1.4) and (1.5), the stationary distribution represents the limiting distribution of successive iterates from the chain, regardless of the starting probabilities of each state (Gilks et al., 1996). We applied Markov chain to the analysis of the viewing behavior on the home page of the UniCa website. The states of the Markov chain were represented by AOIs on the page. To this aim, we defined seven AOIs around the main buttons of the menus in the UniCa home page and named them with capital letters from A to G (Figure 1.8). A padding of 30 pixels was applied around every button in order to be able to also capture fixations in proximity of each menu.

For each participant, fixations in the home page were extracted and assigned to the corresponding AOI according to their (x, y) coordinates. Fixations outside any AOI were assigned to an eighth region (named with the capital letter H). For each participant $k \in K$, let D_k be a vector of variable size containing the sequence of states representing the AOIs in which each fixation was made. Transitions between states contained in D_k were used to define an n × n with n = 8 transition matrix T_k. For each group of participants g ∈ [1, 2], let $M^{(g)}$ be the matrix defined as:

$$M^{(g)} = \sum_{k \in K} T_k \tag{1.6}$$

that is, Equation (1.6) is the matrix containing all transitions between states. Finally, the transition probability matrices $P^{(g)}$ were obtained by computing probabilities of transitions for each $M^{(g)}_{i,j}$.

Markov chain analysis was conducted on the home page as this page was the starting point of all tasks and is commonly considered the most important page of a website. Data for a total of 48 participants from Group 1 and 53 from Group 2 having at least two fixations with known (x, y) coordinates in the home page, visible without doing any scrolling, were included. Figure 1.9 shows the fixations made in the area of the home page visible without scrolling by high school students (a) or university students (b) and how they were attributed to defined AOIs (panels c and d).

Table 1.5: Transitions between the AOIs of the home page in high school (Group 1) and university students (Group 2)

Task	**n**$_{G1}$	**n**$_{G2}$	**Mean**$_{G1}$	**SD**$_{G1}$	**Mean**$_{G2}$	**SD**$_{G2}$
Task 1	8	6	9.00	8.38	9.33	9.24
Task 2	1	6	16.00	0.00	17.33	14.32
Task 3	9	5	8.89	8.19	4.60	3.65
Task 4	4	6	6.00	7.57	14.17	9.39
Task 5	2	4	32.50	34.65	1.00	1.41
Task 6	2	5	7.00	0.00	2.80	1.64
Task 7	5	6	18.20	12.19	6.17	6.52
Task 8	5	5	16.20	11.71	5.20	4.32
Task 9	8	6	15.50	7.76	15.83	4.79
Task 10	4	4	4.75	4.43	5.75	3.96

Only participants with at least two fixations with known (x,y) coordinates in the home page are included. Abbreviations: AOI, areas of interest; G1, group 1; G2, group 2; SD, standard deviation.

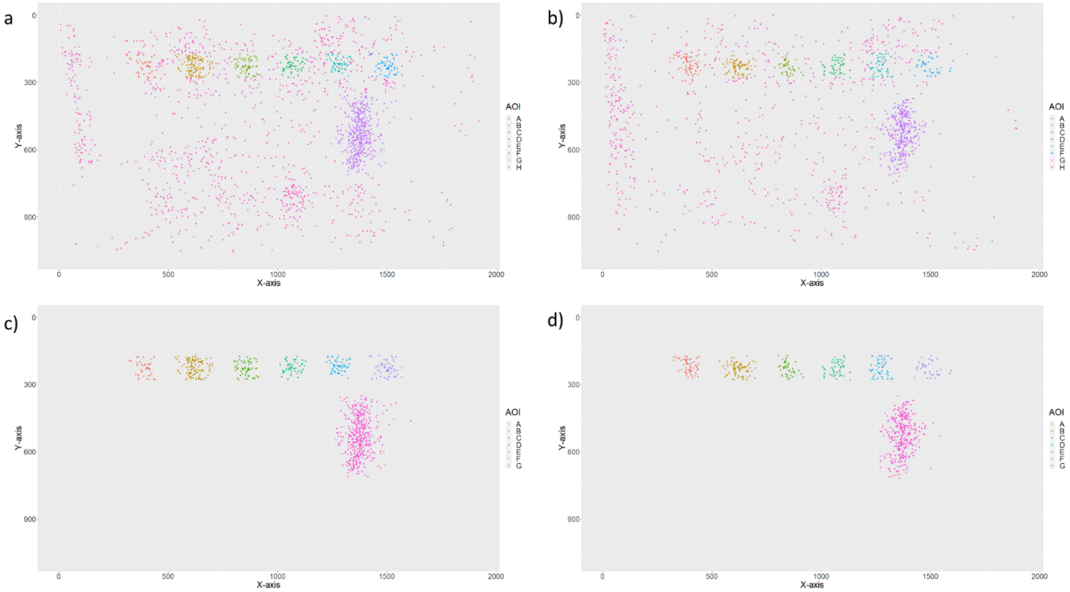


Figure 1.9: Assignments of the fixations to the areas of interest (AOI) defined in the UniCA home page. This figure shows: (a) fixations made by high school students on the home page in the area visible without scrolling; (b) fixations made by high school students on the defined AOIs; (c) fixations made by university students on the area of the home page visible without scrolling; and (d) fixations made by university students on the defined AOIs

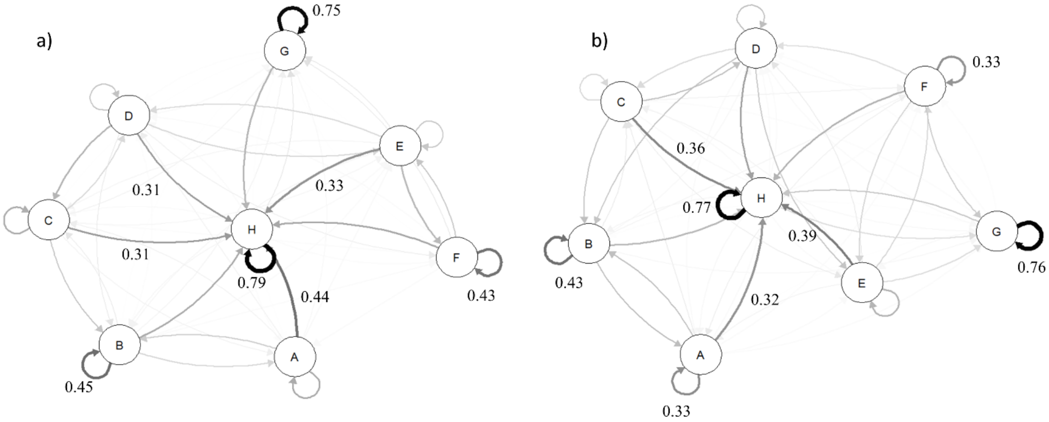


Figure 1.10: Oriented graphs constructed using the transition probabilities between AOIs high school (a) and university (b) students. The graph shows the transitions between the defined AOIs (letters from A to G) and the area outside any AOI (letter H) for high school students (a) and university students (b). The width of the arcs is proportional to the value of the corresponding transition probabilities. Only transition probabilities ≥ 0.30 are reported.

Table 1.5 reports the mean number of transitions between AOIs stratified based on each task and each group of participants. The oriented graphs constructed using the transition probabilities between AOIs for the two groups of participants and generated using the qgraph R package (Epskam et al., 2012) are shown in Figure 1.10. The width of the arcs is proportional to the value of the corresponding transition probabilities. For better clarity, only values of transition probabilities ≥ 0.30 are reported. For both high school (a) and university students (b), it can be observed that the highest transition probabilities are either self-loops or are directed to the H region (the area of the page outside any defined AOI). The matrices $P^{(g)}$ with g $\in$ [1, 2] of these transition probabilities were used as input for the steadyStates function of the markovchain R package (Spedicato, 2017) to obtain the stationary distributions (Table 1.6).

In this function, eigenvectors corresponding to identity eigenvalues are identified and then normalized to sum up to one. The verifyHomogeneity function of the same package, which uses a chi-square-based test, was used to verify if transition matrices of the two groups of participants belonged to the same Markov chain. No significant differences between the two transition matrices were found (χ^2 = 76.32, p = 0.12), supporting the hypothesis that, on a global level, users with different

Table 1.6: Markov chain stationary distributions for high school (Group 1) and university students (Group 2)

Group	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
All tasks								
Group 1	0.022	0.060	0.032	0.027	0.030	0.032	0.247	0.550
Group 2	0.048	0.068	0.032	0.034	0.036	0.026	0.258	0.499
Task 8								
Group 1	0.007	0.030	0.016	0.011	0.028	0.016	0.250	0.640
Group 2	0.185	0.220	0.033	0.026	0.013	0.007	0.337	0.179

Areas of interests (AOI) are indicated with letters from A to G. The area outside any AOI is indicated with the letter H.

levels of knowledge of the portal show similar behavior when searching information on the home page. We then focused on Task 8, as for this task all three metrics showed significant differences in the performance of the two groups of participants. For this task, significant differences were found in the transition matrices of the two groups of participants ($\chi^2 = 115.29$, $p < 0.001$). This observation is in line with the observed difference in performance between the two groups of students reported in the previous section, as well as with differences observed in the qualitative analysis of scanpaths which were plotted using the scanpath R package (von der Malsburg and Vasishth, 2011). As shown in Figure 1.11, the majority of high school students, compared to university students, went back and forth between different AOIs and made several observations in the area outside any AOI.

For university students, the areas with the highest probabilities in the stationary distribution were the buttons of the menu corresponding to AOIs A, B and G, while for high school students it was the area of the page outside any AOI (H). This might be related to the absence of a specific item of the menu labelled “libraries”. While this probably did not represent a problem for university students, as they might be more aware that libraries can be found among university buildings or services, this might not have been the case for prospective students.

The use of eye tracking to evaluate web usability presents some technical limitations. The eye tracker is able to identify the area in which the gaze is focused but does not take into consideration peripheral vision, which plays an important role during activities such as reading or exploration of a website. For instance, peripheral vision is involved in the so-called “banner blindness”, i.e. ignoring the presence of an advertising banner in a web page (Benway and Lane, 1998). In addition, while the eye tracker is able to identify fixation coordinates with high precision, it does not

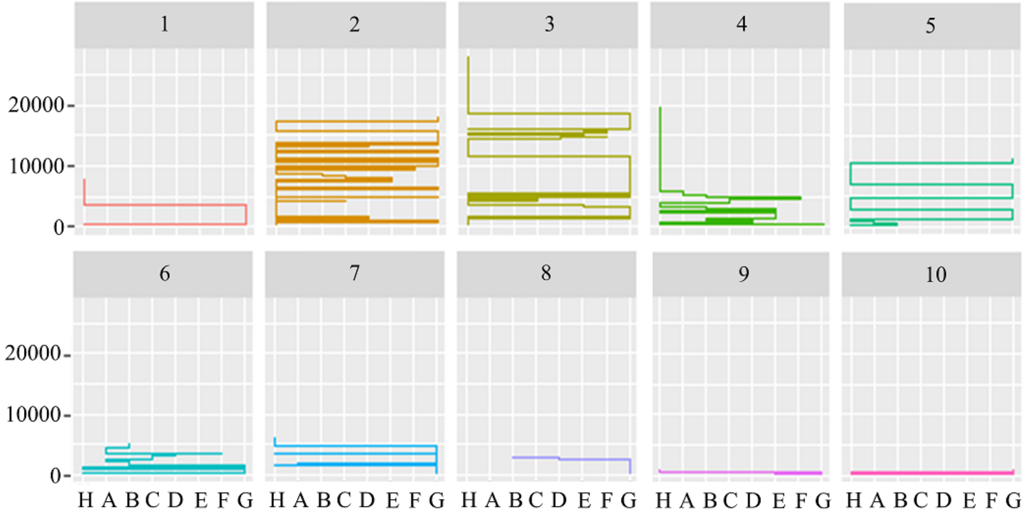


Figure 1.11: Scanpath for high school (from 1 to 5) and university (from 6 to 10) students for Task 8. The figure shows scanpaths for participants with at least two fixations with known (x,y) coordinates in the home page. The letters from A to G indicate areas of interest (AOI) as defined in Figure 1.8, while the letter H indicate the area outside any AOI. The plots show the sequence of fixations in the different areas of the home page (X axis) in milliseconds (Y axis).

allow to gain information regarding the reasons underlying viewing behavior. While studies investigating this aspect can be designed, they are however much more expensive and sometime invasive, as they usually need additional instruments able to measure brain activity or reactions based on the object a participant is looking at. Additional limitations of the current study are related to the relatively limited number of participants performing each task (due to the fact that a high number of tasks was designed to have a more comprehensive evaluation of relevant pages).

In conclusion, while we observed a high efficiency for most analyzed pages, the combination of qualitative and quantitative analysis allowed to suggest some changes to improve the web usability of the University of Cagliari's website, with a specific focus on users with low levels of previous knowledge of the website.

1.5 Application II: Digital flyers

Nowadays, almost every company exploits the potential of Internet using digital marketing techniques in order to reach a higher number of people and decrease costs. However, traditional approaches such as paper flyers (also known as door drop flyers) are still used by a large number of retailers to advertise their products and limited-time offers. Digital flyers are able to offer the strengths of both techniques, combining traditional and innovative approaches. In fact, while they picture a higher number of products compared to classic internet ads or banners, they are still able to reach a much higher number of people compared to paper flyers. While the eye tracking technology has successfully been used to evaluate how people interact with a web page, only few studies have applied it to evaluate how visual information is processed by the consumers as they are observing a flyer (Pentus et al., 2018). Studying the viewing behavior of consumers might allow to measure differences in the effectiveness of two digital flyers, as well as to identify clusters of consumers based on their viewing behavior.

In this study, we applied the eye tracking technology to evaluate viewing behavior on digital flyers from two different retailers. To this aim, we designed three tasks and applied the analytical approaches described in the following sections.

1.5.1 Design of the tasks

In February 2020, we selected the latest flyers from two major retailers promoting technology products. We designed three tasks that required to find a specific product (unique for each task) (Table 1.7). Features of this product were communicated to the participant right before the test. In order to successfully complete the task, each participant had to search for the object satisfying the features on a specific page of one flyer and then on the other one (the two flyers were presented in random order). The task ended when the subject communicated the name of the right object. This procedure was conducted for both flyers. In case of error in the identification of the object, the participant was told to keep searching and the number of errors was registered. During the test, participants had to use only their eyes (no mouse, no keyboard) to locate the object.

Table 1.7: Description of the tasks

Task	**Description of the Task**
Task 1	Find the washer machine with the lowest price
Task 2	Find the smartphone with the best percentage of discount
Task 3	Find a TV screen measuring at least 40 inches, with the lowest original price

1.5.2 Participants

Participants were recruited among university students of the University of Cagliari, mainly from Economics and Law departments, randomly selected in group study rooms (in different days of the week and different times of the day). Each student performed the three tasks. For each participant, we collected information about age, gender, residency and university course. A total of 25 university students completed the three tasks. Three students were excluded due to technical problems occurred during the task, leading to a final sample of 22 participants. 41% of the participants were women and mean age (± SD) was 23.86 ± 3.56.

1.5.3 Data collection

For each participant, eye movements during the task were gathered with a screen-based Tobii X2-60 Compact eye tracker applied to a 25-inch monitor as in the study described in Section 1.4. Before each participant started the tasks, the instrument was calibrated according to the specific height and distance from the screen. As in the previous study, eye movements were classified into different types using the I-VT algorithm implemented in Tobii Studio v. 3.3.1, which classifies eye movements based on the velocity of the directional shifts of the eye. Among different types of eye movements, we chose to focus on fixations, which are considered the metric of most interest in similar studies (van der Lans et al., 2011).

1.5.4 Comparison of the effectiveness of the two flyers

The effectiveness of the two flyers was compared using time to completion for each task as well as the total number of fixations. Normality of distribution of time to completion and number of fixations during each task was evaluated using the Shapiro-Wilk test (Shapiro and Wilk, 1965). Distribution of time to completion and number of fixations was normal for Task 2 and 3 but not for Task 1. For the latter task, non-parametric tests were used. For all tasks, a strong correlation between time and number of fixations was observed (Task 1: Spearman's rho: 0.84, $p < 0.001$; Task 2: Pearson's correlation coefficient: 0.96, $p < 0.001$; Task 3: Pearson's correlation coefficient: 0.71, $p < 0.001$). For each task, time to completion and fixations between the flyers of the two retailers were compared using the paired samples t-test (Tasks 2 and 3) or Wilcoxon test (Task 1). The Bonferroni correction was used to adjust results for multiple testing based on three tasks. A p-value < 0.017 (i.e. 0.05/3) was considered significant.

A better performance of Flyer 2 was observed during Task 1. Specifically, participants completed this task in a faster way and with a lower number of fixations when looking at Flyer 2 compared to Flyer 1 ($p < 0.001$ for both). Additionally, participants completed Task 3 in a faster way when looking at Flyer 2 compared to Flyer 1, although this difference was not significant after adjusting for multiple testing ($p = 0.03$). No significant difference was observed for Task 2.

These findings might be explained by differences in the layout of the two flyers. In fact, while both flyers showed a similar number of objects of the same type, the second flyer represented them using a more orderly design (the pictures of the promoted objects had similar dimensions and were evenly distributed across the page). Conversely, as regard to the task for which no significant differences were detected, the design of the pages selected for this task was similar among the two flyers.

1.5.5 Analysis of gaze transitions using Markov chain

For Task 3, we further analyzed gaze transitions among AOIs defined in each flyer using Markov chain. This analytical approach and its application to eye tracking data has been discussed in Section 1.4.5. To this aim, seven AOIs were drawn around the main objects pictured in each flyer. Capital letters from A to G were assigned to the defined AOIs and considered to be the states of the Markov Chain. Fixations made by each participant during the task were assigned to corresponding AOIs based on their coordinates on the screen (x, y). This allowed to obtain a sequence of states needed for the computation of transitions. The presence of differences in

viewing behavior between the two flyers was then analyzed by comparison of the two stationary distributions using the verify Homogeneity function in the Markov Chain R package (Spedicato, 2017) in R. 3.6.1 (R Core Team, 2019).
The two stationary distributions were significantly different ($\chi^2 = 380.95$, p < 0.001). This finding is in accordance with the differences in performances observed in time to completion of the task and total number of fixations between the two flyers, supporting the hypothesis that, on a global level, participants looked at the flyers in a different way while searching information.

1.5.6 Scanpath analysis

A qualitative analysis of the scanpaths for each participant was also conducted. To this aim, scanpaths for each participant were created using filled circles to plot the sequence of the different fixations. The dimension of each circle was proportional to the duration of the corresponding fixation (for an example of scanpaths see Figure 1.12).

Scanpaths showed a lower number of deviations and changes of direction during the observation of Flyer 2 compared to Flyer 1, in line with the quantitative results showing a better performance of Flyer 2 as regard to time and number of fixations required to complete the task in the whole sample. In conclusion, our results showed significant differences in the performance of digital flyers from two retailers. Future steps of this research might involve analyses aimed at identifying clusters of participants with similar viewing behavior, as well as characteristics of areas of a page with higher density of fixations during free observation. Such analyses might allow to further characterize which aspects of a digital flyer might be associated with higher effectiveness.

1.6 Application III: Tourist landscapes

Digital media exert a relevant influence on tourism management due to the fact that several people use Internet as a primary source of information when choosing their travel destination (Garín-Muñoz and Amaral, 2011). For this reason, cultural sites or leisure destinations as well as hotels and travel agencies, advertise their location using websites, social media accounts, or pages on travel fare aggregators/search engines. Images are the main media used to promote the attractiveness of a destination (Ruhanen et al., 2013). Accordingly, it has been shown that images can affect travel choices (Wang and Sparks, 2016). The analysis of user viewing behavior measured with the eye-tracking technology - allowing to assess the exact position of the eyes

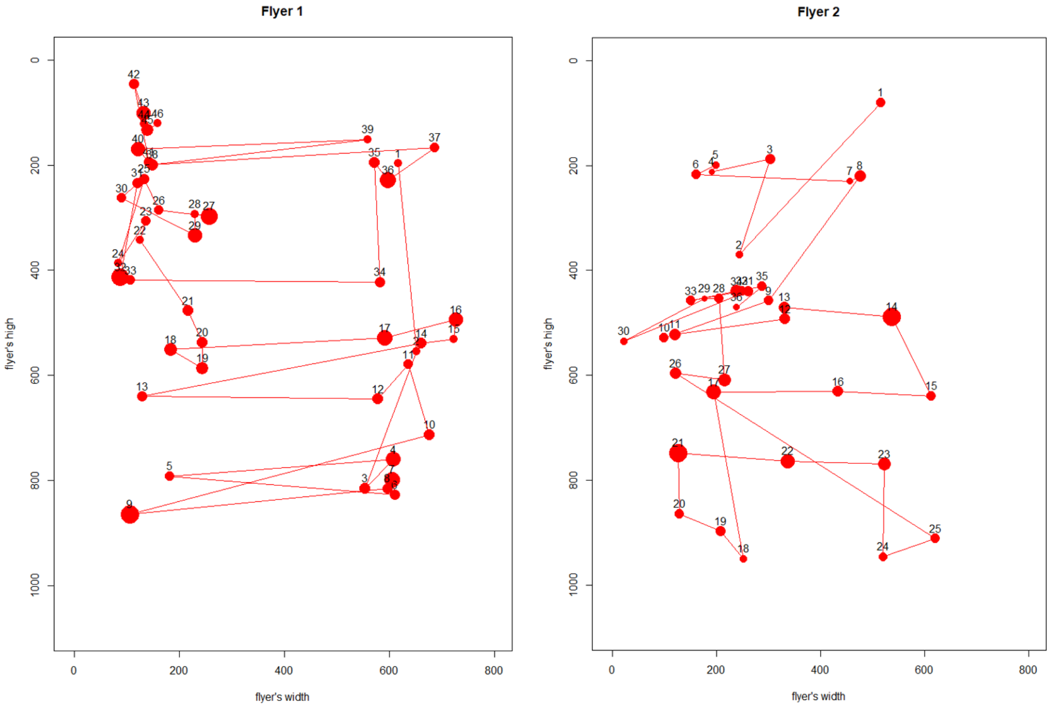


Figure 1.12: Example of representation of scanpaths for a participant for Flyer 1 and 2 during Task 1

during the visualization of images, texts, or other stimuli - can increase the effectiveness of digital media (Scott et al., 2019). Specifically, metrics allowing to conduct a quantitative analysis of viewing behavior (e.g. the number of fixations) can allow to improve the effectiveness of a website or to conduct consumer segmentation. For the current study we focused on landscapes, which are considered to be an element of great importance for attraction and development of tourism (Jiménez-García et al., 2020).

In this study, we leveraged the eye tracking technology to analyze viewing behavior on a large publicly available data set of images depicting natural and city landscapes. In addition, we assessed whether classification models can be used to separate these two classes of images based on two metrics calculated using eye tracking data: number of fixations and path length covered by the eye gaze of each participant.

1.6.1 Data set

For the current analysis we used the publicly available MIT saliency benchmark repository data set, including 1003 images, mostly depicting natural indoor or outdoor scenes (Judd et al., 2009). In this data set, eye tracking data were obtained for 15 participants (age: 18-35 years) who looked at each image for 3 seconds in free viewing with a 1 second pause (gray screen) between images. The experimental conditions were the following: participants were seated in a dark room two feet apart from a 19" screen (1280x1024 resolution), and their range of motion was limited using a chin rest to stabilize the head. Eye tracking measurements were conducted using an ETL 400 ISCAN 240Hz model. In this data set, images were collected from two online repositories: Flickr and LabelMe.

Included images depict different objects, such as indoor or outdoor scenes, people, animals, buildings and so on. Each image was manually reviewed and assigned to one of three possible classes: (i) natural landscapes, (ii) city landscapes, (iii) other, based on the main element depicted in the image. Only images in which the main element was a natural landscape or a city landscape were assigned to these two categories. For example, images showing a valley or a desert were classified as “natural landscape”. However, images depicting a natural object, but not a natural landscape (e.g. a single flower) were classified as “other”. Analyses were conducted on 187 images classified as “city landscapes” and 225 classified as “natural landscapes”, while images classified as “other” were removed. Figure 1.13 shows some examples of images from these two classes.

We hypothesized that natural landscapes might represent more homogenous pictures with fewer different stimuli to focus on, while city landscapes might induce users (e.g.,

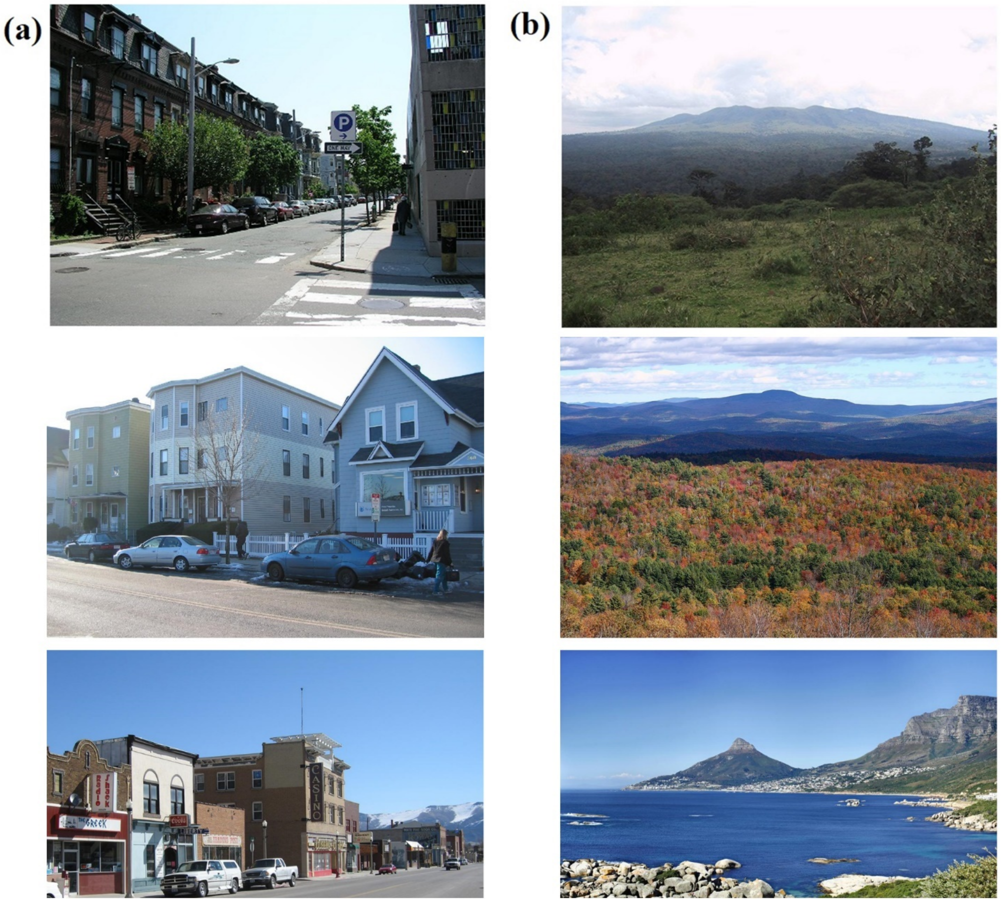


Figure 1.13: Examples of (a) city landscapes and (b) natural landscapes

visitors browsing a touristic website) to shift from one object to another (e.g., from a car to a building to a road sign). In accordance, we hypothesized that the path followed by the observer's eye on a picture might be longer in images depicting city compared to natural landscapes.

1.6.2 Analysis of viewing behavior based on fixations and path length

For each image, two metrics quantifying the viewing behavior of participants were computed: total number of fixations and path length covered by the gaze of each participant during the observation. For each image, this metric was calculated as the sum of the Euclidean distances between fixations, based on (x, y) coordinates of each fixation. For both variables, normality of distribution was assessed using the Shapiro-Wilk test (Shapiro and Wilk, 1965), while homogeneity of variance with Levene's test. Based on results from these tests, the number of fixations and the path length were compared between the two classes of images using Mann Whitney's U test and Welch's t-test, respectively. Both path length and number of fixations showed a significant difference when comparing images depicting natural and city landscapes. Specifically, we observed shorter path length ($p < 0.001$) and number of fixations ($p < 0.001$) in natural compared to city landscapes (Table 1.8).

Table 1.8: Summary statistics for path length and number of fixations

	Path length (pixel)		**Number of fixations**	
	Natural	**City**	**Natural**	**City**
Min	4,668	8,011	70	79
Q1	14,267	18,522	103	116
Median	17,504	21,317	112	123
Mean	17,766	21,431	111.2	123
(± SD)	(± 4,942)	(± 4,322)	(± 12.84)	(± 12.67)
Q3	21,287	24,298	120	131
Max	31,938	32,020	148	160

For natural landscapes, n = 187, for city landscapes, n = 225.
Abbreviation: SD, standard deviation

These findings are in accordance with our hypothesis that natural landscapes, due to their homogenous nature, might be easier to visually explore compared to city landscapes. Previous studies have also highlighted that nature images are easier to comprehend (Wang and Sparks, 2016), as well as to recognize and memorize (Dupont et al., 2013). In a subsequent step, we tested whether path length and the number of fixations can be used as predictors to separate natural from city landscapes using different supervised classification models. These analyses and production of relative plots were carried out with R v. 3.6.3, (R Core Team, 2020) using the packages mclust (Scrucca et al., 2016), MASS (Venables and Ripley, 2002), class (Venables and Ripley, 2002), factoextra, and ggplot2 (Wickham, 2016). We selected four widely used models: logistic regression (LR) with a decision rule, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and K-nearest neighbours (KNN). In the following sections, the theory underlying these models will be illustrated in detail.

1.6.3 Supervised classification models: logistic regression

The logistic function was developed in the 19th century to describe the growth of populations and autocatalytic chemical reactions. In such a case we consider the temporal evolution of a quantity W(t) and its growth rate

$$W^{*}(t) = dW(t)/dt. \tag{1.7}$$

A simple assumption is that W*(t) is proportional to W(t) and a constant β

$$W^{*}(t) = \beta W(t) \tag{1.8}$$

and therefore, β = W*(t) / W(t), where β is the constant rate of growth. This leads to exponential growth

$$W(t) = W(0)e^{\beta t}, \tag{1.9}$$

where W(0) is the value of W at t = 0. This model might be reasonable if the growth it is not opposed in any way, but that's not always the case. Nonetheless, there are some real-world situations that might be approximated by this model. For example, as Malthus wrote in 1798: *The power of population is indefinitely greater than the power in the earth to produce subsistence for man. Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio. A slight acquaintance with numbers will shew the immensity of the first power in comparison to the second. By that law of our nature which makes food necessary to the life of man, the effects of these two unequal powers must be kept equal. This implies a strong and constantly operating check on population from the difficulty of subsistence. This difficulty must fall somewhere and must necessarily be severely felt be a large portion of mankind* [Thomas Robert Malthus - An Essay on the Principle of Population, as It Affects the Future Improvement of Society, With Remarks on the Speculations of Mr Godwin, Mr Condorcet and Other Writers].

If it is true that the human population, without restrictions, follows a geometric progression, it was also argued that, to be more realistic, we should consider some constraint factor to the Equation (1.9) which becomes

$$W^{*}(t) = \beta W(t) – \varphi(W(t)) \tag{1.10}$$

This addition is due to Pierre-François Verhulst, a Belgian mathematician and statistician, who later expanded Equation (1.10) into

$$W^{*}(t) = \beta W(t)(\omega – W(t)) \tag{1.11}$$

where ω is the upper limit for W. As we can see in Equation (1.11), the population growth $W^{*}(t)$ is a function of both the population at time t and the level of saturation already attained, which can be also written as a proportion between current level and maximum achievable level

$$P(t) = \beta P(t)\{1 – P(t)\}. \tag{1.12}$$

The solution of this differential equation is

$$P(t) = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}} \tag{1.13}$$

which was later named logistic function. As sometimes happen, the logistic function was discovered again by Raymond Pearl and Lowell Reed in their study on the United States' population growth in 1920. This is probably due to the fact that the idea of logistic growth is simple but still effective. It has been used in the past to model, for example, population growth and is now used in fields such as marketing to study how a new product is able to penetrate a new market (Cramer, 2002).
An important property of the logistic function is that it always returns values between 0 and 1. This is particularly useful when we want to apply a *logistic regression model* to our data. The main types of logistic regression are binomial, multinomial, or ordinal. Binomial (or binary) logistic regression is used when the outcome of the response variable is binary or dichotomous in nature, i.e. when it has only two possible values (e.g. success, failure or win, loss) often represented or encoded as "0" and "1". A binomial logistic regression model is based on the Bernoulli distribution, whose probability distribution function (PDF) is a member of the exponential family and can be written as $f(y;\pi) = \pi^{y}(1-\pi)^{1-y}$ where π is the probability of success (and (1 - π) the probability of failure). Multinomial logistic regression extends

the binomial regression, in the sense that it deals with outcomes that might have three or more possible values (for example in medicine we might observe "disease A", "disease B", "disease C"). In this model, order is not important. The last type of logistic regression is the ordinal logistic regression, which is similar to the multinomial case, but the outcome values can be ordered. For example, the dependent variable might be the score in a Math test evaluated as "very poor" (if the score is below 25), "poor" (from 26 to 50), "good" (from 51 to 75), and "very good" (from 76 to 100). The binary logistic regression model has a binary response variable (or dependent variable), and usually one or more predictors (models with no predictors are also possible) (Hilbe, 2009). In general, we can denote independent variables as $X_1, X_2, ..., X_k$ where k is the number of variables being considered. Each of these X might be a continuous, categorical, or binary independent variable or even an interaction between two or more predictors (Kleinbaum and Klein, 2010). In case we have more than one predictor we have a *multiple logistic regression* and we can write this model as:

$$P(X) = \frac{e^{\alpha\ +\ \sum \beta_i X_i}}{1 +\ e^{\alpha\ +\ \sum \beta_i X_i}} = \frac{1}{1 +\ e^{-(\alpha + \sum \beta_i X_i)}} \tag{1.14}$$

where i = 1, ..., k and P(X) is a shortening notation for $P(Y = 1|X_i)$. This is the probability that our dependent variable Y $\in$ 0, 1 assumes a value equal to 1, given all i predictors. The true coefficients α, β_1, β_2, ..., β_k in Equation (1.14) are unknown, but we can compute the estimators $\hat{\alpha}$, $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ using our data. The method of estimation of these parameters is maximum likelihood (ML) which is based on the likelihood function. When we use ML in a logistic regression model we compute the estimates $\hat{\alpha}$, $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ such that the probability for each observation to be correctly classified is maximum. Using this method we are able, once we have entered the values of the parameters into Equation (1.14), to obtain a value for P(X) that approaches one for all those observations that have "success/win" in the response variable and a number close to zero for all observations in the "failure/loss" category. More formally, given θ the vector of parameters such as $\theta = (\theta_1, \theta_2, ..., \theta_k)$, we can express the unconditional formula for likelihood function as:

$$L\ (\theta) = \prod_{j=1}^{m} P\left(X_j\right) \prod_{j=m+1}^{n} [\{1 - P\left(X_j\right)]\} \tag{1.15}$$

with j = 1, ..., m, ..., n and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}, ..., \hat{\theta}_k)$ representing the set of k estimated parameters that maximize the likelihood function. In practice, however, since attempting to maximize the likelihood function $L(\theta)$ is computationally intensive, we maximize the natural logarithm of $L(\theta)$. The probability of obtaining the data for the case j can be defined as P(X_j), where P(X) represents the formula for a given observation (Kleinbaum and Klein, 2010). It is possible to rewrite Equation (1.14) in terms of log odds

$$\frac{P\left(x\right)}{1-P\left(x\right)}=e^{\alpha\ +\ \sum\beta_i X_i} \tag{1.16}$$

where the quantity on the left-hand side of the equation is called odds. A small probability of "success/win" yields an odds close to zero, while an high value of P(X) let the odds goes to ∞. If we apply the logarithm to Equation (1.16) we obtain

$$\log\left(\frac{P\left(x\right)}{1-P\left(x\right)}\right)=\alpha\ +\ \sum\beta_i X_i \tag{1.17}$$

which is called log-odds or *logit*. If we compare Equation (1.17) with multiple linear regression, we can see that they are very similar. The difference is that when we have a different value for one of the predictors, the change (expressed by the corresponding β) will affect the whole logit (but not in a linear way) and we should make further calculation (mainly using the exponential function) to obtain the change for P(X). This leads to the conclusion that the same change in one of the predictors may have a different effect on the logit, and therefore on P(X), depending on the current value of the predictor. Anyhow, the value obtained will be a continuous value and not a 0/1 value. Therefore when we are in a classification setting, we need to apply a decision rule to convert this value in one of the two possible values of the response variable. Usually a cutoff is used to ensure that all values above the threshold are assigned to one and those below the threshold to zero (James et al., 2013).

1.6.4 Supervised classification models: LDA

The main problem with logistic regression arises when the response variable has a number of classes higher than two. In this case we can use several other methods. One of these methods is *Linear Discriminant Analysis* (LDA), which can be exploited when the response variable has a number of classes equal or higher than two. Furthermore, LDA can be used to obtain more stability or better performance with respect to the logistic regression when the number of observations is small or when the two classes are well-separated, which may cause problems with the estimation of parameters. On the other hand, if we assume that the independent variables are not normally distributed, then we might prefer to use logistic regression or other methods (James et al., 2013). LDA combines the distribution (assumed normal) of the p predictors $\boldsymbol{x} = x_1, x_2, ..., x_p$ and the Bayes' theorem, to compute the posterior probability that an observation belongs to a specific class of the response variable. With the Bayes' theorem

$$P\left(y = k|x\right) = \frac{\pi_k f_k\left(x\right)}{\sum_{l=1}^{C} \pi_l f_l\left(x\right)} \tag{1.18}$$

we are able to obtain the probability that an observation belongs to one of the classes of the response variable, given the predictors values. To make the explanation clearer, we can also write Equation (1.18) as:

$$P\left(y = k|x\right) = \frac{P\left(x|y = k\right) P\left(y = k\right)}{P\left(x\right)} = \frac{P\left(x|y = k\right) P\left(y = k\right)}{\sum_l P\left(x|y = l\right) \cdot P\left(y = l\right)} \tag{1.19}$$

Here, the posterior probability $P\left(y = k|x\right)$ is computed using the prior probability $P\left(y = k\right)$ and the density function. To build a classification model for a response variable with a number z of (unordered) classes, where K = 2, 3, ..., z, we need to assume some density function for our predictors and to compute an estimate of μ_k, σ^2 and π_k using our data. In case we have only one predictor, the density function (assumed Gaussian) has the following form:

$$P\left(x|y = k\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2}\frac{\left(x - \mu_k\right)^2}{\sigma^2}\right) \tag{1.20}$$

where μ_k is the mean and σ^2 is the variance for the k-th class. Let also assume that all the K classes have the same variance σ^2. Now we compute the estimates for $\hat{\mu}_k, \hat{\sigma}^2, \hat{\pi}_k$ as:

$$\hat{\mu}_k = \frac{1}{n_k} \sum x_i \tag{1.21}$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum \left(x_i - \hat{\mu}_k\right)^2 \tag{1.22}$$

$$\hat{\pi}_k = \frac{n_k}{n} \tag{1.23}$$

where Equation (1.21) is the mean computed separately in each class, Equation (1.22) represents the variance computed in each class and averaged according to its weight, and Equation (1.23) is simply the number of observations in the k-th class divided by the total number of observations. Plugging these estimates in Equation (1.19) and rearranging its terms, we obtain the linear discriminant as:

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k) \tag{1.24}$$

which will assign each observation to the class for which $\hat{\delta}_k(x)$ is largest. A similar formula is used when p > 1, i.e. the number of predictors is more than one (James et al., 2013). In that case the multivariate Gaussian density will become:

$$P\left(x|y=k\right) = \frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right) \tag{1.25}$$

and the discriminant function will be written as:

$$\hat{\delta}_k\left(x\right) = x^T\Sigma^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\Sigma^{-1}\hat{\mu}_k \ + log\left(\hat{\pi}_k\right) \tag{1.26}$$

1.6.5 Supervised classification models: QDA

While an assumption of the LDA model is that, for each class, observations come from a multivariate Gaussian distribution, with a specific mean for each class and a shared covariance matrix, QDA assumes a specific covariance matrix for each class. Therefore, the distribution of the observations from the k-th class can be indicated as X $\sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the k-th class. Under this assumption, like in LDA, an observation X = x is assigned to the class for which the quantity computed in Equation (1.27) is largest:

$$\hat{\delta}_k(x) = -\frac{1}{2}\log\left|\Sigma_k\right| - \frac{1}{2}\left(x-\mu_k\right)^t \Sigma_k^{-1}\left(x-\mu_k\right) + log(\hat{\pi}_k) + C \tag{1.27}$$

where C is a constant term (James et al., 2013). The decision boundary created by QDA is quadratic, so it can be applied to several types of non-linear problems and lead to better results. QDA can also be useful when the number of observations is not enough to get accurate estimates using non-parametric methods (e.g. KNN). QDA is therefore a more flexible method than LDA, but less flexible than the KNN method, that will be discussed in the next section. The name QDA derives from the fact that the quantity in Equation (1.27) is computed as a quadratic function. Based on the difference in the assumptions, LDA or QDA might be preferred in different situations. Specifically, the choice between these two models is related to the bias-variance trade-off. When there are *p* predictors, p(p+1)/2 parameters have to be estimated to define a covariance matrix. QDA estimates a specific covariance matrix for each class, leading to Kp(p+1)/2 parameters. It is therefore easy to understand how the number of parameters to estimate can greatly increase based on the number of predictors. Conversely, as LDA assumes that the K classes share a common covariance matrix, Kp linear coefficients should be estimated, leading to a classifier with substantially lower variance. While this might allow to obtain a better performance, the bias could be high in case the assumption that the K classes share a common covariance matrix is not met. In general, LDA usually shows a better performance in case the training data set includes a limited number of observations, while QDA might be preferred in cases in which the training set is large or in case it is clear that the assumption that the K classes share a common covariance matrix is not met (James et al., 2013).

1.6.6 Supervised classification models: KNN

The KNN classifier attempts to estimate the conditional distribution of Y given X, in order to assign a given observation to the class with highest estimated probability. Let K be a positive integer and x_0 a test observation. First, the K points nearest to x_0 in the training data are identified and defined as N_0. Next, the conditional probability for class j is estimated as the fraction of points in N_0 whose response values are equal to j as in:

$$P\left(Y=j|X=x_0\right)=\frac{1}{K}\sum I\left(y_i=j\right) \tag{1.28}$$

The Bayes rule is then used to assign x_0 to the class for which the probability is higher. While the KNN can be considered a simple approach, its performance can approach the optimal Bayes classifier (James et al., 2013).

1.6.7 Performances of supervised models in the classification of natural and city landscapes

The four models described in the previous sections were trained using 80% of the data set (n = 330 images). Performance was tested using the remaining 20% of the data set (n = 82 images) using k-fold cross-validation with k = 5. Results of comparisons of these models are shown in Table 1.9.

As shown in Table 1.9, the four classification methods showed similar performances. In particular, sensitivity ranged from slightly above 66% to 74%, with the highest performance shown by logistic regression. Specificity was slightly lower and

Table 1.9: Performance of four models (LR, LDA, QDA, and KNN) in the classification of landscapes

	LR	**LDA**	**QDA**	**KNN**
Sensitivity	0.743	0.724	0.719	0.662
Specificity	0.608	0.616	0.642	0.662
Accuracy	0.680	0.672	0.680	0.621
F1-score	0.714	0.704	0.707	0.653

ranged from 61% to 66%, with the best performance achieved by KNN. Based on how the images were coded, this means that most misclassification errors are made when we try to predict the "city landscapes" class. Accuracy ranged from 62% to 68%, with best performances achieved by logistic regression and QDA. Finally, F1-score ranged from 65% to 71%, with the best performance achieved by logistic regression. Overall, among the four tested models logistic regression showed the best performance and proved to be the best classification method for this task.

1.6.8 Analysis of viewing behavior using unsupervised classification models

Finally, we compared how the two classes of images are visualized using two unsupervised classification methods: the hard clustering performed using K-Means Clustering algorithm (K-means) and the soft clustering performed using Gaussian Mixture Model clustering method (GMM). K-means and GMM are two popular clustering methods which work following an iterative procedure. K-means is non-probabilistic and performs hard assignments, i.e. each point can only belong to one class. Conversely, GMM is a probabilistic algorithm based on multivariate Gaussian distributions as in Equation (1.29)

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\Big\{-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\Big\} \tag{1.29}$$

so that, when the expectation-maximization (EM) algorithm converges, each point is assigned to a class with a certain probability. GMM is more flexible than K-means because it allows decision boundaries to assume an elliptical shape while K-means only a circular shape. The number of clusters was set equal to two based on the presence of two classes of images. In addition, we confirmed this number to be

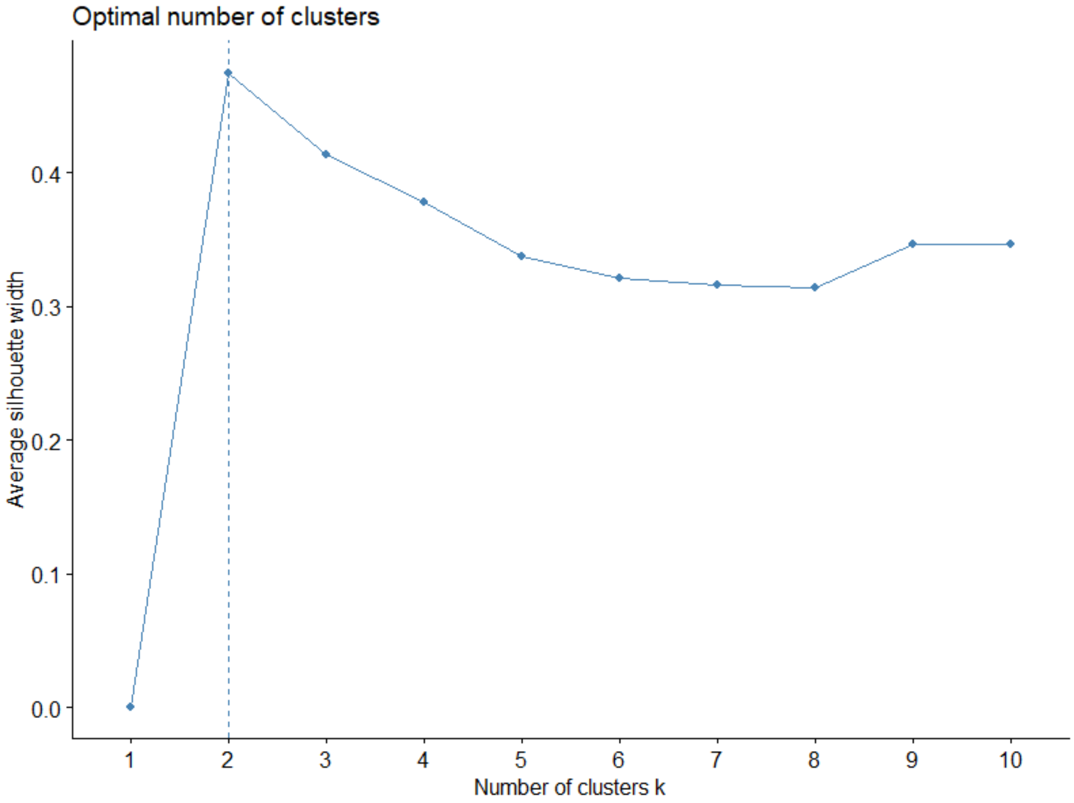


Figure 1.14: Optimum number of clusters based on the silhouette method

optimal using the silhouette method as shown in Figure 1.14.

Visualization of the two classes was similar using both clustering methods, as shown in Figure 1.15. Both in K-means clustering (panel a) and GMM (panel b) plots, the “city landscapes” class is coloured in blue and the “natural landscapes” class in red. Different symbols were used to display correctly classified points (empty circles or squares for city and natural landscapes, respectively) and misclassified points (black filled circles and squares for city and natural landscapes, respectively). Comparison of the two plots shows very similar results as regards to misclassification errors.

These findings can be of interest to stakeholders who have to decide whether to insert images depicting “city landscapes” or “natural landscapes” in touristic web portals. Our results would suggest that images of “natural landscapes” might be preferred, as these can be observed with a lower number of fixations and be perceived as easier to visually explore (therefore allowing a user to visualize a higher

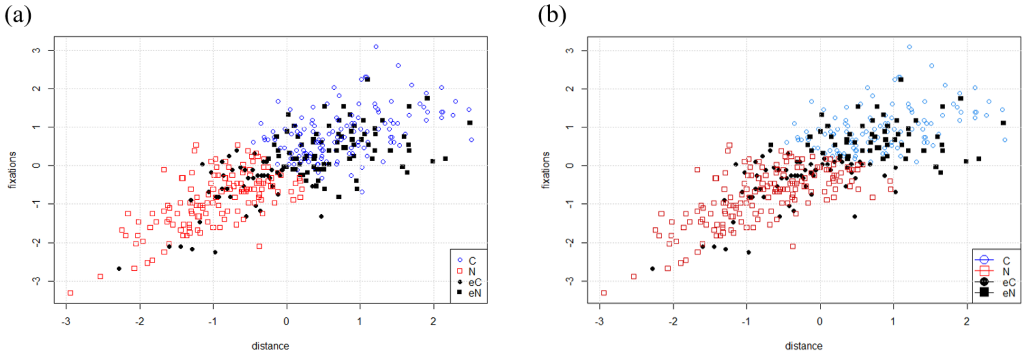


Figure 1.15: Comparison of clustering using (a) K-means and (b) GMM. Abbreviations: C: city landscapes, N: natural landscapes, eC: fixations erroneously classified as city landscapes, eN: fixations erroneously classified as natural landscapes

number of pictures or explore other pages of the website). When choosing city landscapes, images with a reduced number of elements should be privileged, in order to make their perception easier for a typical user. Overall, images chosen to communicate a message should be simple and depict a reduced number of elements, as these images might be more easily processed by the brain, thus being potentially more effective in the engagement of viewers. Our results support the utility of the analysis of eye-tracking data to increase available knowledge on the use of images in tourism promotion. The main limitations of this study include the small number of participants for which viewing behavior data were available. In addition, as time of observation was fixed to 3 seconds for each image, it was not possible to analyze potential differences in this metric between city and natural landscapes. Finally, while the number of images depicting city or natural landscapes was relatively limited, we partially addressed this limit using a k-fold cross-validation approach. Nonetheless, the analysis of additional data sets will allow to further explore these findings as well as to assess the potential role of other variables (e.g., time of observation) to reduce the high number of misclassification errors we observed.

Chapter 2

Sentiment Analysis

2.1 Sentiment analysis

2.1.1 Definition and characteristics of sentiment analysis

Sentiment analysis is a natural language processing technique that aims to computationally treat opinions or sentiments of a document, such a text or a speech (Medhat et al., 2014). It also focuses on identifying and quantifying expressions that reflect an opinion's holder sentiment (i.e., positive, negative or neutral) toward entities (e.g., products) or specific aspects (e.g., products' price) (Cambria et al., 2017). However, sentiment analysis is much more than that, since it involves many facets and multiple sub-problems (e.g., topic mining, sentiment summarization, emotion detection). Sentiment analysis, also sometimes referred to as *opinion mining* or *polarity classification*, is aimed at analyzing and classifying text into sentiments with a polarity or specific emotions using different approaches (Pang and Lee, 2008). Although some scholars began working on textual analysis well before, the term sentiment analysis first appeared in the article from (Nasukawa and Yi, 2003), and the term opinion mining in the one from (Dave et al., 2003). Nowadays, communication between e.g. a company and consumers has become more and more bidirectional, as not only the first entity sends messages/information to the latter part, but also vice versa. The communication between companies and consumers has been facilitated by the diffusion of Internet. In particular, social networks allow people from all around the world to have a direct interaction with companies. Of course, this is not only true for companies but also for e.g. politicians, athletes, actors and so on, just to provide a few examples. There are several occasions in which the availability of a tool that can help to dissect the meaning of a text without needing human interpretation is of great interest and value, e.g. to filter e-mails, letters, messages or any other type of communication daily received by companies. Social media can be a gold mine of information and provide a great quantity of useful data to a company, but the huge number of posts published every day makes the task of exploiting them highly challenging. In fact, in order to use this information, reliable pipelines for data download, cleaning, organization and interpretation need to be defined. Ideally, these operations should require the lowest possible amount of human work and allow to gain up-to-date knowledge. For instance, it would be useful to know in a prompt way that, due to a given event, many people in a specific area of the world need a certain product or service. This knowledge could be exploited by different stakeholders in different situations, such as a company trying to increase the number of its customers, but also a humanitarian organization that aims to help people in areas struck by war or natural disasters. While these two conditions are very different from each other, in both of them we can underline how crucial it is to interpret a message in a prompt and correct way. Nonetheless, the automatic interpretation of the meaning of written text is not a trivial task and sentiment analysis is only able to address a part of it, i.e. the one concerning its polarity. Sentiment analysis can classify polarity using different categories (e.g. positive/negative or positive/neutral/negative) or using quantitative metrics to better underline the strength of the identified polarity. In addition, sentiment analysis can be performed at three different levels: document level, sentence level and aspect level. At a document level, the text (for instance a review or an article) can be thought as a series of sentences that are evaluated separately and then aggregated to define the overall sentiment. At a sentence level, we define the polarity of each sentence, thus leading to a more detailed analysis compared to the previous setting. The analysis at the aspect level is even more refined, as it can be conducted for each defined aspect of interest. For instance, we can define the polarity of different aspects represented in a product review, such as customers' opinions regarding functionality, price, aesthetics, delivery and so on. In fact, the opinions of consumers on these different aspects do not necessarily coincide, and in the same product review we might find expression of a positive sentiment for functionality and a negative sentiment for price (Liu, 2015).

One of the most relevant concepts in sentiment analysis is the definition of *opinion*. Usually, an opinion is a broad concept that encompasses sentiment, evaluation, target, and opinion holder. An opinion orientation is also called sentiment, which is what we analyze more often and is usually included in one of the three classic categories: positive, negative or neutral. We can express the concept of opinion as:

$$O = e, a, s, h, t(2.1) \tag{2.1}$$

where e is the entity on which the opinion is expressed, a is a specific aspect of the entity discussed in the opinion, s is the sentiment elicited, h is the person who hold the opinion, and t is the timestamp. An opinion can be expressed on multiple entities at the same time, and a single entity can be considered as a collection of several aspects, so in this case $e = a_1, a_2, ..., a_n$. For example, the review of a smartphone can concern aspects such as video camera, memory, price, and so on. Opinions can be expressed in several ways, but the main categories are regular opinions (Liu 2006; Liu 2010) or comparative opinions (Jindal and Liu, 2016). A regular opinion can be expressed directly or indirectly. Examples of direct expressions can be found in sentences such as "the view was beautiful" or "the room was clean". An indirect opinion is probably less frequent and although is expressed on a target, it also gives important information about another entity. For example, the sentence "After taking that medicine, my sight was no longer clear" is an indirect opinion since, even though the sight is the subject of this sentence, we also have an indirect (negative) opinion about the medicine. The second type of opinion is comparative. In this case, the sentence is expressed using the comparative or superlative form of an adjective or adverb, but that's not always the case (e.g., "I prefer Android"). There are many ways to compare entities or aspects of an entity, but we will not go into detail on this topic. An example of comparative opinion is: "Apple is better than Android" or "Android is the best" (Liu, 2015).

The two main approaches that can be applied to perform sentiment analysis are the lexicon-based approach and the machine learning approach. In the lexicon-based approach, words contained in a list, defined as lexicon or dictionary, are associated with a specific sentiment polarity. The sentiment of a document text in calculated based on the number and strength of positively/negatively associated words contained in the text (Turney, 2002). Based on the characteristics and aims of the study, the experimenter can choose whether to use an existing lexicon or to build a new one. On the other hand, the machine learning approach consists in building a model to accomplish a supervised classification task. Several models or classifiers can be used, e.g. Naïve Bayes (NB), Support Vector Machine (SVM), logistic regression, decision trees, neural networks, etc. While the conditions in which they might work best are different, they all require to collect, label and process the data, split the data set into training and test sets, build the model on the training set and finally assess model performance.

Importantly, in the last few years, a growing number of studies applied deep learning methods to perform sentiment analysis. Deep learning is based on artificial neural networks (ANNs), which have been designed similarly to the human biological neural networks. To continue with the analogy, ANNs are composed of a vast number of information processing units (called neurons) that are structured in layers and work similarly to how the human brain works. They can mimic the learning process of a biological brain by altering the connection weights between neurons to learn to accomplish tasks (such as classification). The modern neural networks are very different from those created at the end of the last century. We refer to that type of neural network with the term "shallow" to indicate that they were composed of only one or two layers, while today's ones are composed of many layers connected to each other and are therefore defined as "deep". The increased computing power and the availability of a large amount of training data greatly increased the interest in this field of study. In fact, today these neural networks are used for various purposes, such as computer vision, speech recognition, and sentiment analysis tasks. Specifically, with regard to sentiment analysis, deep learning models often employ word embeddings as input features (Yadav and Vishwakarma 2020; Cambria 2018; Sailunaz and Alhajj 2019). Word embedding is an approach that converts words into numerical vectors, a necessary step to proceed with the computerized analysis of the text. Basically, it entails the mathematical embedding of a multi-dimensional space into a space with a considerably smaller number of dimensions (e.g. Word2Vec). Word embeddings have some drawbacks, one of which is that words with various meanings are grouped into a single representation. Polysemy and homonymy, in other words, are not effectively managed. To overcome this problem, researchers have recently proposed embeddings that exploit the context of a word to distinguish between various possible meanings (e.g. BERT). While a detailed description goes beyond the scope of this thesis, more details on these methods can be found in (Devlin et al., 2018).

2.2 Lexicon-based approach

The most important words that help to correctly classify a sentence, document or speech into positive or negative texts are sentiment or opinion words. Many of these words are adjectives such as good, wonderful, amazing (positive) or bad, poor, terrible (negative). A list of these words is called a lexicon (or sentiment lexicon). We can also add several other expressions such as "break a leg", "to kill two birds with one stone", and other idioms that help the researchers to overcome situations when the meaning of a sentence taken word by word is different from the overall meaning of the sentence. Usually each of the entries of the lexicon is associated with a polarity that defines whether that word or expression can be considered as positive, negative or neutral. However, there are several problems in using a lexicon-based approach to classify a sentence or a text as positive or negative (or neutral).

Domain specificity: many words may have a completely different meaning in different domains. An example might be the word “sick” that is used as a negative word in many contexts, e.g. “the patient is sick” but has an opposite meaning if used in slang expressions like “that trick was sick” which is a compliment. For this reason, the polarity or sentiment orientation can be different depending on the reference domain or context.

Once more without sentiment: the fact that a sentence or text contains a word does not necessarily mean that a feeling is being expressed. This can happen in interrogative sentences (e.g., “Can you tell me which camera is good?”) or conditional sentences (e.g., “I wonder if I can find a good camera in this shop”). Both of these sentences use the word “good” but not to express a positive sentiment. The former is a simple request for information, while the latter is more of an ironic comment. This does not mean that it is not possible to express sentiments using interrogative or conditional sentences. For example, the sentence “Does anyone know how to repair this terrible printer?” is saying that the printer is really bad.

Sarcasm: sarcasm detection is one of the main challenges in sentiment analysis. For example, a sentence like “What a great car!” can be a real compliment or a sarcastic comment depending on the context. As human beings we are able to understand, even using non-textual information such as the tone of voice or images, if someone is being sarcastic, but it’s not an easy task if we only have access to a written text. Luckily, sarcasm is not so common in reviews, but it can be very common in other domains such as in politics.

Sentiment without sentiment words: in many cases it is possible to express a positive or negative sentiment without any sentiment words. For example, “The washer uses a lot of water” can be considered as a negative opinion about the washer because it wastes too much water, even though it is written as a simple description of the user’s experience. This opinion can be considered more objective than something like “The washer was ugly” because it is based on an objective evaluation. Another example of this type of a positive/negative comment expressed without any sentiment words is: “After sleeping on the mattress for two days, a valley has formed in the middle". Even if valley is probably exaggerated, the sentence is pretty objective but still very negative in relation to the characteristics of the mattress (Liu, 2015).

2.3 Machine learning approach

In many practical cases, when we try to classify a text, the main question is whether it is a positive or a negative text. Product reviews are texts in which customers point out which are the best and the worst features of an object. Since many of these reviews are based on a score or rating system (e.g. stars), it is possible to set a decision rule to decide which reviews are good and which are bad. For example, using the 5-stars rating system, we could assign a "positive" label to all reviews with 4 or 5 stars and a "negative" label if the number of stars is 1 or 2. In this case, reviews with 3 stars are excluded because they are midway between positive and negative and they are probably a mix of elements belonging to the two classes. In other cases, they can be considered as "neutral" and form a new class, if that's how we want to build our data set. Supervised machine learning methods are widely used in sentiment analysis. One of the first studies to apply this approach was that from Pang and colleagues (2002) who reported good performance when trying to classify positive or negative movie reviews using NB and SVM (see Section 2.5.5 and Section 2.5.6 for details). The key element when using a supervised machine learning approach is to build a set of features which, once plugged into the classification model, provide the best possible result. These features can be based on counting words frequency, the labeling of each word of the text according to grammar rules, modifiers of sentiment orientation and so on.

2.3.1 Data cleaning

Preprocessing or data cleaning is a necessary step at the beginning of each sentiment analysis. Text data are very noisy and need to be cleaned and prepared for the analysis. Additionally, some words are not really helpful and can be eliminated. Reducing the number of words reduces the dimensionality of the problem and hence can improve and speed up the classification. The main steps involved in the process are text cleaning (e.g., remove HTML tags, white spaces, and punctuation; lowercase conversion), expanding abbreviations, stemming, tokenization, lemmatization, stop words removal, handling of negation, emoji and other non-textual elements. These are the fundamental steps to be taken before the feature extraction phase.

2.3.2 Feature extraction

Term frequency: in sentiment analysis, the frequency-inverse document frequency (tf-idf) allows to build a matrix where each word is counted within each document of a collection or corpus. Using tf-idf it is possible to have an idea of how important each word is within a specific document of the corpus. It assigns a weight to words based both on the number of times they appear in a document as well as the number of documents they appear in (essentially the weight changes according to whether the word is specific to a document or if it is a commonly used word). The term frequency part *tf(t, d)* is computed as:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2.2}$$

where *tf(t, d)* is the number of times the term t appears in document d divided by the number of words in document d. There are several variations to this computation. For example, it is possible to apply logarithm, normalization, or a binary count (1 if term t is in document d, 0 otherwise) instead of raw count. The inverse document frequency is a weighting function that takes into account the rarity of the term t with respect to the whole corpus. The IDF part is computed as:

$$\mathrm{idf}\,(t,\ D) = log\frac{N}{n_t} \tag{2.3}$$

where N is the number of documents of a corpus D and $n_t = |d \in D : t \in d|$, which is sometimes written as $1 + |d \in D : t \in d|$ to avoid division by zero, is the number of documents of the corpus D where the term t appears. Equations (2.2) and (2.3) can be modified using different ways to compute weights (e.g., using log normalization $log\,(1 + f_{t,d})$ for tf weights and $log\,\left(1 + \frac{N}{n_t}\right)$ for idf weights) (Manning et al., 2008). The tf-idf is therefore computed as the product of the two parts:

$$\text{tf-idf}\,(t,\ d,\ D) = \mathrm{tf}\,(t,\ d)\ \times\ \mathrm{idf}\,(t,\ D) \tag{2.4}$$

Part of speech (POS): in the POS tagging step, every part of a speech is labelled according to the Penn Treebank POS tags (Santorini, 1990). This is an important step in feature engineering, since adjectives, for example, are one of the biggest drivers in sentiment analysis. Table 2.1 shows the parts of speech according to Penn Treebank POS tags.

Table 2.1: Alphabetical list of part-of-speech tags used in the Penn Treebank Project

Number	**Tag**	**Description**
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Source: Penn Treebank Project

2.3.3 Other aspects to take into account

Sentiment words: These words are particularly relevant in representing sentiment even in short sentences. For example, the sentence “the book is great” is very short, but also quickly recognizable as a positive comment. Other examples of positive words are: good, beautiful, amazing. The same thing applies to negative words such as: bad, awful, poor. Besides single words, there are also small sentences, mostly idiomatic expression, such as “it’s not rocket science”. Sometimes the sentiment is the opposite of what we would assign with a word-by-word evaluation, e.g. “break a leg” is positive.

Sentiment shifters: sentiment shifters are words or expressions used to revert or modify (e.g. to increase or decrease) the sentiment orientation of a sentence. The three main sentiment shifters are negations, amplifiers and downtoners. For example, the sentence “I like it” can be negated using “not”, i.e. “I do not like it”. Negations are very important because they completely reverse the sentiment orientation.
An example of amplifier is “really”, i.e. “I really like it” is somehow more positive than “I like it”. Conversely, downtoners de-amplify the impact of positive or negative words. For example, “I hardly like it” is less positive than “I like it”. Among these three types of sentiment shifters, negations are probably those to handle with more caution. For example, the sentence “Not only I like it, I also...” is positive even if it starts with “not”.

Syntactic dependency: several researchers have studied features based on the words’ dependency. One of the best-known types is probably the n-gram. An n-gram (unigram, bigram, trigram, and so on) is a sequence of n words that can be found in a given sentence or text. Models based on n-grams are widely used in linguistics, biology, and for plagiarism detection.

2.4 Analysis of the literature

In order to summarize studies applying sentiment analysis to commonly investigated fields such as finance, politics and reaction to events, we conducted a bibliometric analysis. To this aim, we conducted a search on SCOPUS updated to July 11 2021, with no language and date restrictions, including articles and conference papers. The search strategy is reported in the Appendix. The bibliometric analysis was conducted using the Bibliometrix R package (Aria and Cuccurullo, 2017). In the following sections we report the main results of this search for each field, as well as characteristics of the most relevant studies.

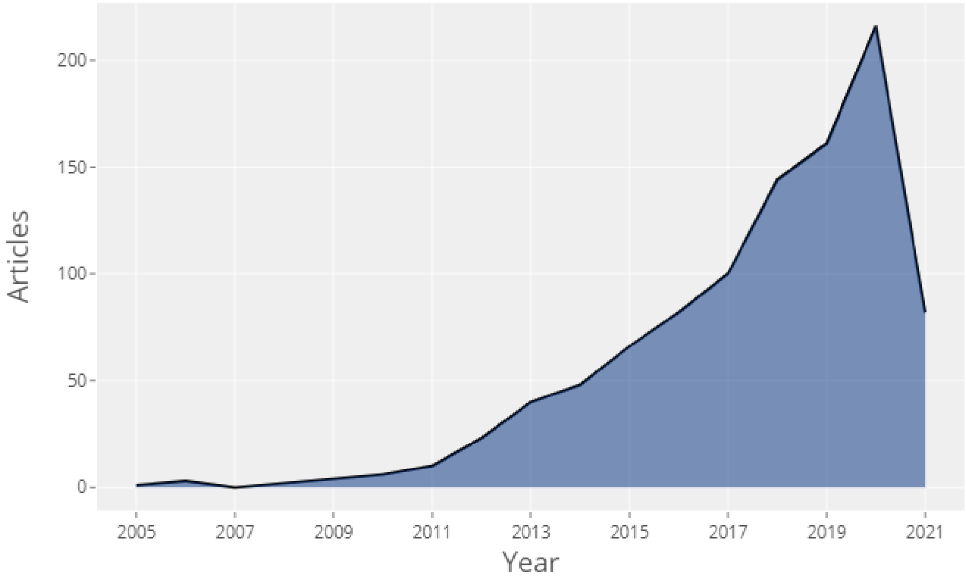


Figure 2.1: Annual scientific production regarding articles applying sentiment analysis to finance and economics related topics

2.4.1 Finance

Sentiment analysis has been increasingly applied to finance and economics in the last few years (Figure 2.1). In total, our search identified 988 documents (404 articles and 584 conference papers) published from 2005 to 2021 that applied sentiment analysis to finance-related topics. In this field, articles on sentiment analysis observed an annual growth rate of 34.15%. Figure 2.2 shows the distribution of retrieved articles among countries, based on affiliations of authors.

The analysis of the scientific production showed China to be the most represented country among affiliations of authors of the retrieved articles (Table 2.2), while US was the country with most cited articles (Table 2.3). Overall, the average number of citations per document was 11.26.

In the financial domain, sentiment analysis is widely used in three main fields: financial forecasting, banking and corporate finance, as outlined in a recent review (Gupta et al., 2020). Among authors exploring the first topic, Das and Chen (2007) collected messages from the Yahoo's message board to predict the Morgan Stanley High-Tech Index trend. They used five classifiers and a majority vote (best of five) to predict trends for a range of metrics. Another relevant work in this field involved Twitter data to predict Dow Jones, S&P 500, and NASDAQ indices (Zhang et al.,

Table 2.2: Top 10 countries for number of articles or conference papers related to sentiment analysis in finance-related topics

Country	**Number of studies**
China	308
India	260
US	256
UK	81
Italy	66
Germany	61
Australia	48
South Korea	47
Canada	41
Turkey	36

Table 2.3: Top 10 countries for citations on articles or conference papers related to sentiment analysis in finance-related topics

Country	**Total Citations**	**Average article citations**
US	1,563	30.06
China	593	5.99
UK	545	22.71
India	378	6.41
Hong Kong	371	37.10
Japan	291	41.57
Italy	267	14.05
Netherlands	265	44.17
Singapore	236	39.33
Slovenia	204	40.80

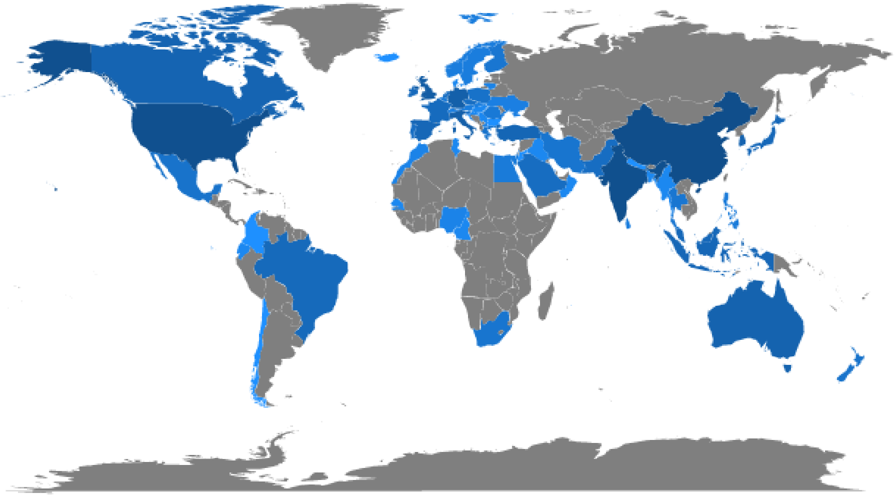


Figure 2.2: Distribution of articles applying sentiment analysis to finance and economics related topics based on affiliations of authors

2021). The authors showed an inverse relationship between emotions and indices (positive emotions made the Dow Jones go down and vice versa). Other authors performed a similar analysis restricted to experts' tweets (Bar-Heim et al., 2011).
In another study, Nassirtoussi and colleagues (2015) showed that financial news-headlines can predict with high accuracy (up to 83.33%) the directional movement of a currency pair in the foreign exchange market. To this aim, the authors used a multi-layer algorithm consisting in a Semantic Abstraction Layer (addressing the problem of multiple words in a text referring to the same concept), a Sentiment Integration Layer (proposing a sentiment weight that reflects investors' sentiment) and an algorithm called Synchronous Targeted Feature Reduction (aimed at dimensionality reduction) (Nassirtoussi et al., 2015).
An alternative approach was used by Nguyen and colleagues (2015) who built a model to predict stock price movements based on messages on social media. The authors collected posts on 18 companies published on Yahoo Finance Message Board for a period of one year. This message board was used to discuss news or write comments on the companies or company events. The authors showed that the incorporation of data from sentiment analysis on these posts increased accuracy compared to the model using historical prices only, and this increase was more relevant for stocks difficult to predict (Nguyen et al., 2015). Finally, several other authors used sentiment analysis on non-financial sources of information to predict market trends (Zhang and Skiena 2010; Si et al. 2013).

Sentiment analysis in the banking field can be useful for different applications related to e.g. detection of money laundering, reporting quality, risk management or customer relationship management (Gupta et al., 2020). Among the most interesting contributions, Fritz and Tows (2018) used sentiment analysis to assess the quality of 343 banks' annual risk reports from 30 German banks between 2002 and 2013. Quality of the reports was evaluated in terms of their fulfillment of regulatory requirements. The authors found text mining measures to explain a great proportion of the variance in reporting quality and identified discrepancies between the reports of distressed and non-distressed banks (Fritz and Tows, 2018).

Applications of sentiment analysis in corporate finance can include, among others, analysis of reports, sustainability analysis, analysis of competitors (Gemar and Jiménez-Quintero, 2015) or fraud detection (Gupta et al., 2020). Among the most interesting applications, documents such as the annual report of a company can provide useful elements to predict the future of the company. However, this information are often hidden and text-mining techniques can be applied to extract them. An example is the study of Lee and colleagues (2018), who analyzed annual reports of 54 different US companies in the information technology sector to correlate text patterns with sales performances of the companies. The authors found text length and patterns of frequently appearing words, but not the positive or negative tone of a report, to be correlated with the sales performance (Lee et al., 2018).

2.4.2 Politics

Sentiment analysis is widely used to evaluate trends in voters' opinion during, for instance, elections or presidential campaigns. Our search identified 2152 documents (829 articles and 1323 conference papers) published from 2006 to 2021 applying sentiment analysis to politics-related topics (Figure 2.3). In this field, articles on sentiment analysis observed an annual growth rate of 32.83%. Figure 2.4 shows the distribution of retrieved articles among countries, based on affiliations of authors.

The analysis of the scientific production showed India to be the most represented country among affiliations of authors of the retrieved articles (Table 2.4), while US was the country with most cited articles (Table 2.5). Overall, the average number of citations per document was 13.34.

In politics, sentiment analysis has been used as a tool to forecast political results, to assess the level of polarization in the electorate or to identify political topics. Ceron and colleagues (2014), for example, focused on the Italian and French political elections. They found that even though Internet users are not representative of the whole country population, it is possible to use opinions expressed through social

Table 2.4: Top 10 countries for number of articles or conference papers related to sentiment analysis in politics-related topics

Country	Number of studies
India	643
US	611
China	438
Indonesia	151
Italy	147
UK	145
Spain	143
Brazil	104
Germany	92
Pakistan	86

Table 2.5: Top 10 countries for citations on articles or conference papers related to sentiment analysis in politics-related topics

Country	Total Citations	Average article citations
US	7,970	66.98
India	1,965	13.10
China	1,094	7.44
Korea	877	21.93
Italy	861	21.53
Spain	804	13.63
Singapore	733	66.64
UK	425	12.50
Hong Kong	424	28.27
Canada	382	23.88

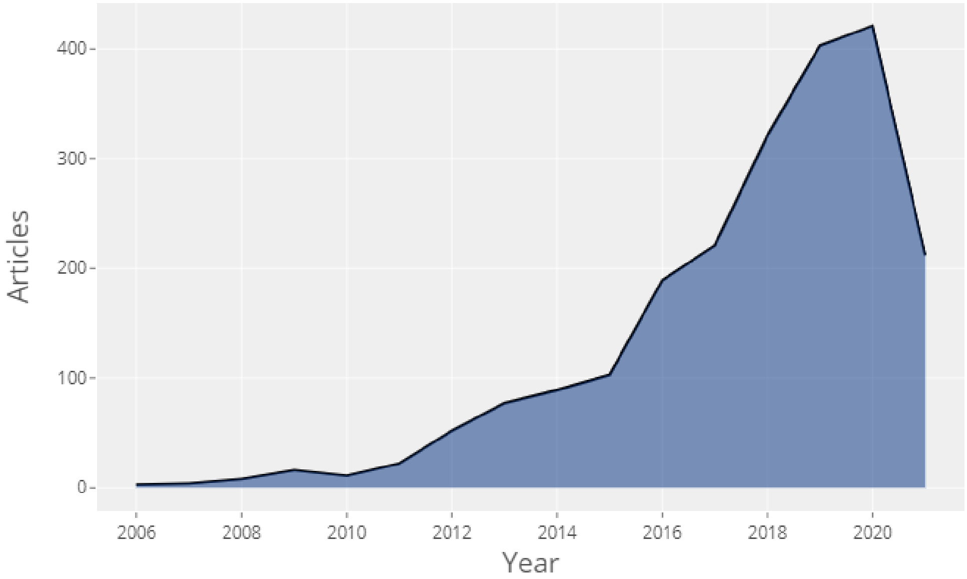


Figure 2.3: Annual scientific production regarding articles applying sentiment analysis to politics related topics

media to forecast electoral results. Additionally, they showed a correlation between social media preferences and the results of traditional mass surveys. In another paper, Burnap and colleagues used tweets to predict the UK election outcome (Burnap et al., 2016). They were able to correctly predict the top three parties in terms of voting shares, including some shifts in votes between parties that had gained consensus and others that had lost it, but they were unable to assess whether the presence of geo-located data would improve prediction or not (Burnap et al., 2016). Conversely, an older study from Metaxas and colleagues (2011) argued that predicting the outcome of an election using sentiment analysis on tweet data is not a viable technique in most cases. They tested the predictive power of social media metrics to predict the outcome of different Senate races of recent US elections, showing poor performance. They also highlighted challenges that limit the predictability of election results through Twitter data and proposed a set of standards to improve results (Metaxas et al., 2011).

Other authors suggested that the creation of a personalized lexicon might also be useful to improve analyses on political texts. Specifically, Young and Soroka (2012) designed the Lexicoder Sentiment Dictionary, a new lexicon to assess specific emotions expressed in political tweets. The authors showed classification of news content performed by this dictionary to be more similar to human coding, compared with

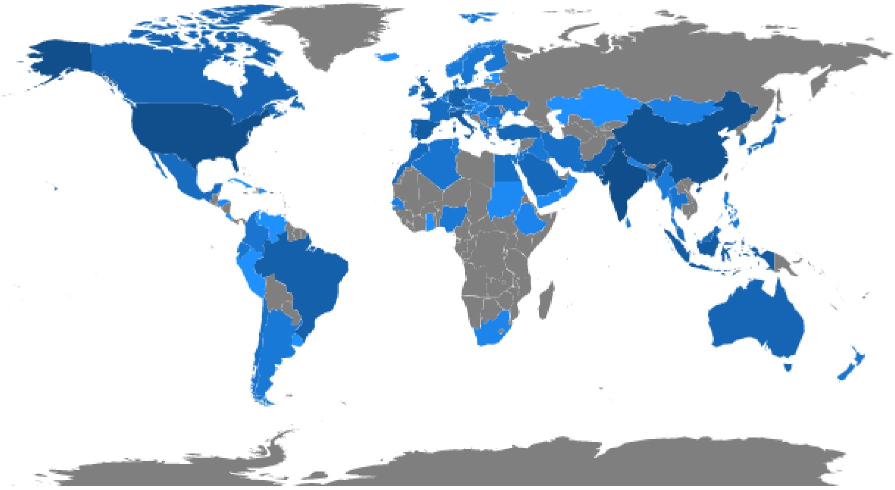


Figure 2.4: Distribution of articles applying sentiment analysis to politics related topics based on affiliations of authors

other available dictionaries. The authors showed that disgust and trust were the main emotions identified in the news contents, and also that negativity was expressed much more than positivity. These findings support the importance of the use of dictionaries specifically designed based on the investigated field.
Conover and colleagues (2011) analyzed more than 250,000 tweets posted from users with different political orientation in the six weeks before the 2010 US congressional elections. Through the application of a network clustering algorithm, the authors showed that retweets from users with different political beliefs exhibit a well-separated structure, while the user-to-user mention network forms a highly heterogeneous cluster in which users interact at a higher rate compared to the network of retweets (Conover et al., 2011). This suggests that the two main interaction mechanisms (retweets and mentions) on Twitter are characterized by different network topologies. Finally, Rill and colleagues (2014) collected about 4 million tweets posted before and during the election in Germany to develop a tool to detect emerging political topics in Twitter before other media. The authors showed that the topics emerged earlier in Twitter than in Google Trends, supporting the utility of Twitter as a source to detect new trends in political topics.

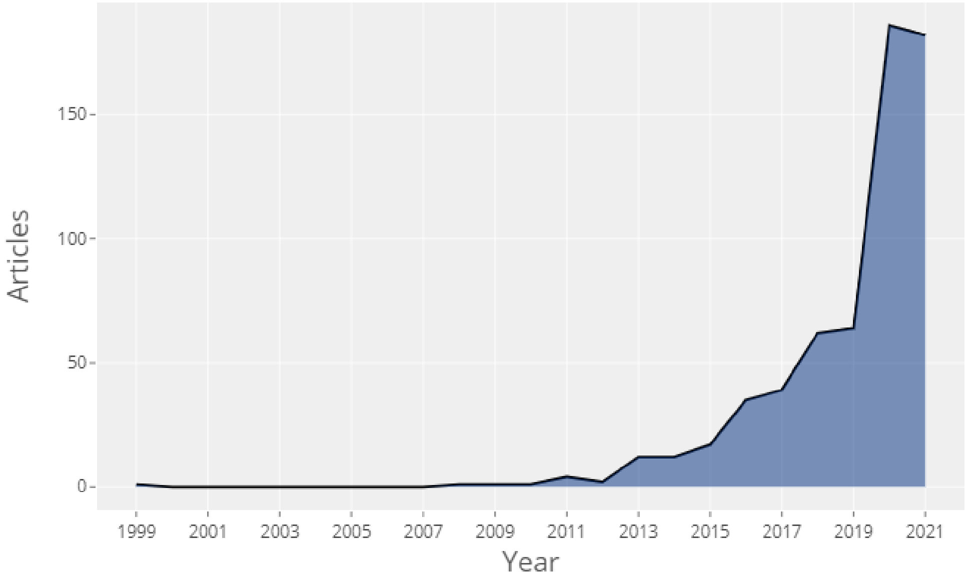


Figure 2.5: Annual scientific production regarding articles applying sentiment analysis to reaction to events

2.4.3 Reaction to events

Finally, sentiment analysis has been increasingly applied to analyze reactions to events (Figure 2.5). In total, our search identified 619 documents (341 articles and 278 conference papers) published from 1999 to 2021 that applied sentiment analysis to the analysis of reaction to events. While the number of documents retrieved for this field was lower compared to finance or politics related studies, this field is characterized by a substantial growth as we observed an annual growth rate of 45.02%. This observation can be explained by a steep increase in the number of studies in the last two years, due to the large body of research on the COVID-19 pandemic. Figure 2.6 shows the distribution of retrieved articles among countries, based on affiliations of authors.

The analysis of the scientific production showed US to be the most represented country as regard to affiliations of authors of the retrieved articles (Table 2.6), while China as regard to most cited articles (Table 2.7). Overall, the average number of citations per document was 6.06.

Several studies have exploited the potential of sentiment analysis using data from Twitter, i.e. data from common users expressing their opinion on the most heterogeneous topics. An interesting subgroup of these studies analyzes tweets in response

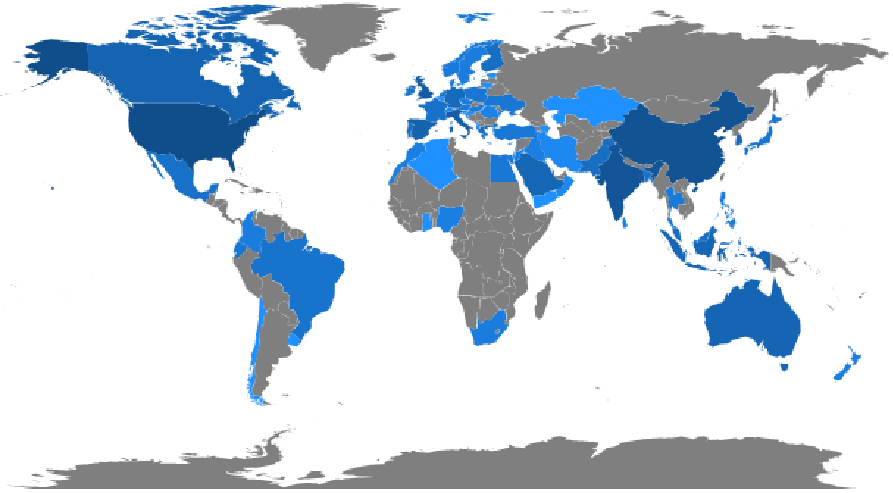


Figure 2.6: Distribution of articles applying sentiment analysis to reaction to events based on affiliations of authors

Table 2.6: Top 10 countries for number of articles or conference papers related to sentiment analysis and reaction to events

Country	**Number of studies**
US	283
China	186
India	135
UK	80
Italy	67
Spain	46
Canada	39
Australia	36
Saudi Arabia	35
Indonesia	34

Table 2.7: Top 10 countries for citations on articles or conference papers related to sentiment analysis and reaction to events

Country	**Total Citations**	**Average article citations**
China	764	11.58
US	641	10.17
UK	211	13.19
Australia	204	18.55
Quatar	169	33.80
India	155	3.69
Turkey	85	28.33
Italy	83	10.38
Portugal	78	19.50
Spain	69	3.63

to important events. Given the sheer number of users who post their opinion on the social network on average, it is possible to carry out interesting analyses that investigate how people react to certain events or facts (see, for example, Paul and Dredze 2017). In this section, we focus on articles that analyzed events such as migrant crisis, terrorist attacks and disease outbreaks.

About one million refugees and migrants came to Europe in 2015 due to negative events such as wars, diseases, terrorism or natural disasters (Vollmer and Karakayali, 2018). Many of these people have made desperate journeys and in very risky conditions, which brought them from Africa to the Mediterranean countries, to then continue towards their final destination (in many cases, northern European countries). With the increase in the number and media coverage of arrivals, public opinion in countries such as Germany has shifted from an initial "welcome culture" to a less warm reception (Backfried and Shalunts, 2016). Pope and Griffith (2016) collected data in English and German to assess if any difference in sentiment could be found comparing tweets from people speaking the two languages. A few years after, Öztürk and Ayvaz (2018) used a similar approach for the Syrian refugee crisis, comparing approximately 350,000 English and Turkish tweets. Their findings included the discovery of a different point of view among those who commented on the event in English (more concerned about politics) compared to those who commented in Turkish (worried about what was happening in a nearby country).

Terrorism attacks is another hot topic in sentiment analysis on Twitter data. The study from Cheong and Lee (2011) was one of the first to analyze, using machine learning classifiers, tweets in a terrorism-related scenario. Many people share their reactions to text, images or video of people involved in a terroristic attack. Twitter was also a primary source of information during attacks such as the ones that took place in Mumbai (India) in 2008 and in Jakarta (Indonesia) in 2009 (Cheong and Lee, 2011). In this study, the authors proposed a framework to identify possible threats combining the scanning of important topics using the latest trends in Twitter, the data collection for selected topics, the analysis of such data and finally a summary or visual report. The importance of an alarm system of this type is somehow underlined by two studies related to the attacks on Westgate Mall (Kenya) and Woolwich (UK). The former emphasized the importance of a filter to extract important information from redundant data (Simon et al., 2014). The latter built a negative binomial regression model on about 400,000 tweets to study the survival time of information related to the attack, showing that sentiment analysis is able to predict survival time of the information flow (Burnap et al., 2014). Since many terroristic groups (e.g. ISIS) use Twitter to spread fake news or propaganda announcements, Ashcroft et al. (2015) studied the possibility to stop these kinds of messages using a machine learning approach to assess if a tweet is somehow supporting a terrorist organization. To address this challenge, they used stylometric, time-based and sentiment-based features as well as machine learning classifiers such as SVM, NB and Adaboost, getting good results for such a delicate task. The paper of Güneyli and colleagues (2017) slightly shifts the point of view as it is focused on how Turkish political leaders reacted or talked about terrorism during election campaign in 2015. Another example is the analysis of the Uri (India) terroristic attack in 2016 (Garg et al., 2017) where the authors used SVM and NB to analyze about 60,000 tweets. Interestingly, the authors also showed that the survival time for negative tweets was longer than the time for the positive ones. Some authors focused on the emotional part of tweets (Harb et al., 2019). In this case, the events were two terrorist attacks occurred in the UK (Manchester and London) in 2017. Two deep learning models [Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM)] trained with different data sets (e.g. an existing pre-labeled data set, a data set automatically labeled using hashtags, etc.) were used to classify tweets according to the (mainly negative) emotion conveyed (anger, fear, sadness, disgust, surprise and a residual category). They showed a similar accuracy for both deep learning models with an F1-score of 63%. Conde-Cespedes et al. (2018) assessed whether general account violations can be useful in detecting potential threats. They collected about 200,000 tweets in different languages (but mostly Arabic tweets translated in English) before and after the Paris terroristic attacks in 2015. These tweets are classified into two categories: pro-ISIS or neutral. They used SVM trained with features derived from keywords and sentiment analysis scores and obtained a 90% of accuracy.

The last type of event covered in this section are disease outbreaks. Chunara et al. (2012) analyzed data from different sources (news, tweets and official reports by the government) during the first 100 days of the 2010 Haitian cholera outbreak. They showed how Twitter and other sources can be an informal but useful source of information (in addition to government or health institutions announcements) to predict infection trends a few weeks in advance and therefore estimate the outbreaks dynamics. Nowadays, people heavily rely on Internet to stay informed (as, for example, during the H1N1 influenza virus outbreak) (Jones and Salathe, 2009). In 2010, Chew and Eysenbach (2010) collected about two million tweets related to the H1N1 influenza virus outbreak and showed that sentiment analysis performed on tweets is a useful tool to measure public perception, allowing authorities to address real as well as perceived concerns. This is of great importance if we consider that fake news help a disease to spread more quickly, with increased costs and higher loss of human lives. On the other hand, correct information might help to take countermeasures (e.g. social distancing) that, especially in the initial stage of an outbreak (when a vaccine is not available), are of vital importance. Signorini et al. (2011) showed a different type of insight in which tweets analysis might be useful during an outbreak: the extraction of information from a live stream of tweets as a hot spots' early detection system. As said, Twitter has proven itself as a valuable source of information to predict future trends of infection spreading (Szomszor et al., 2010). The authors analyzed about three million tweets to predict results from official reports one week in advance. In the study from Smith et al. (2016) the tweets analyzed are related to the flu season in US in 2012/2013. The authors used machine learning classifiers to separate tweets about flu awareness from tweets about the infection, showing two different trends in the data. For example, they showed how levels of awareness dropped after the flu peak, while infection levels were still high. Broniatowski et al. (2013) used the same data (Twitter data posted during the same flu season) to build a model able to distinguish between tweets reporting an infection from tweets mentioning flu. Using this tool, they were able to predict changes in influenza prevalence with an accuracy of 85%.

Ebola outbreaks are another phenomenon under study, especially since 2014. A lot of authors mostly focused on the perception of the disease among people living in Western countries (and not in Africa, where the outbreak took place). For example, Fung et al. (2014) showed that, despite the disease outbreak being very far away from the US (only few cases hit the US, while the large majority of cases were in Guinea, Sierra Leone and Liberia), people were very concerned about their own safety. This was shown through the analysis of the levels of anxiety, anger and other negative emotions of people talking about this topic (significantly higher than those observed during the influenza outbreak). Lazard et al. (2015) exploited a Center for Disease Control (CDC) live Twitter event to collect data and extrapolate the main topics people were talking about. They highlighted eight topics like the ones discussed nowadays by authorities and everyday users during the COVID-19 outbreak (e.g. ways of transmission, symptoms, survival of the virus outside the body and protections). Towers et al. (2015) extract Twitter data and web searches from the first days in which the media reported some Ebola cases in the United States (September 2014) until the end of October. They compared this information with the number of searches on Google (e.g. Ebola symptoms), video and news related to Ebola, finding a strong relationship between the two variables.

So far, we've talked about disease outbreaks from an external point of view, now let's move on to an internal point of view. Oyeyemi et al. (2014) collected tweets using keywords such as "Ebola" and "prevention" or "cure" and found that many of them are carrying misleading information. The authors split data into three categories: correct information, medical misinformation and other (a generic category for tweets not in the first two categories). Medical disinformation can generate big issues for those who fall victim to it (and sometimes also for the people who come into contact with them) since it leads to defend against concrete threats with useless or even harmful actions. Therefore, the Nigerian government decided to use Twitter to respond to misinformation. Guidry et al. (2017) analyzed how Twitter and Instagram have been used by important organizations [CDC, World Health Organization (WHO) and Médecins Sans Frontières (MSF)], to educate the public about Ebola. They found that Instagram posts were significantly more likely than tweets to feature contents related to risk perception, e.g. information about adverse outcomes. Similarly, Liang et al. (2019) compared the relevance of dissemination of information on Twitter from important actors (broadcasting) compared to word of mouth (viral spreading). The analysis was performed on all tweets posted about Ebola in a 14-month period (March 2014 – May 2015) focusing on the retweeting patterns. The authors identified four types of users: influential, hidden influential, disseminator and common user. They highlighted the relevance of a broadcasting-type communication, concluding that it would be useful for health authorities to establish a partnership, particularly with influential users, to communicate more efficiently.
During the Zika virus outbreak in Central and South America, Fu et al. (2016) col-

lected a sample of more than one billion tweets (in English, Spanish and Portuguese) using the keyword “Zika”, posted from May 2015 to April 2016. They identified 20 topics that were grouped into five themes (impact and reaction to Zika virus, concern for pregnancy and microcephaly, transmission routes and case reports), and found that government authorities had a less relevant impact than user-generated contents in the dissemination of information, highlighting the fact that it was necessary to prevent the proliferation of disinformation.

2.4.4 Other fields

Sentiment analysis has been applied in several other fields, such as tourism or criminology. In tourism, Hu and Chen (2016) focused on how the tourism industry has been influenced by electronic word-of-mouth. They tried to refuse some assumptions such as that all reviews are equally visible to online users or that review rating and hotel star class do not interact. Using variables such as review content, sentiment, author, and visibility, they assessed the presence of an interaction between review rating and hotel star class and showed that visibility has a strong effect on review helpfulness. In a study conducted in 2017, Kim and colleagues (2017) applied a hybrid text mining methodology to study tourists’ behavior patterns. Specifically, they analyzed roughly 20,000 online reviews of Paris classified according to 14 categories (e.g., restaurants, sightseeing, hotels, shopping and so on) and further analyzed some of these categories to better understand why they convey negativity or positivity (Kim et al., 2017). In another study, Cheng and Jin (2019) aimed at investigating which attributes influence the experience of Airbnb users the most. They reported that people are heavily influenced by past hotel experiences and that the main attributes they pay attention to are location, amenities and host, but not price, and that the main downside was noise. In another paper, Liu and colleagues (Liu et al. 2019) focused on the Chinese tourists’ evaluations of destinations in Australia and compared these reviews with those of non-Chinese tourists. They analyzed more than 36,000 online reviews (posted on domestic online social media and travel agencies), finding that the market features and preferences of Chinese tourists are very different from those of international tourists. Finally, an interesting work from Moro and colleagues (2019) exploited the features of gamification on more than 67,000 reviews from TripAdvisor. The authors trained four neural networks to understand how 12 gamification features helped to explain review length and sentiment score, finding three badge features to be the most relevant: total number of badges, the passport badges, and the explorer badges.

In criminology, Dadvar and De Jong (2012) explored the field of cyberbullying. They argue that it is possible to increase the accuracy of cyberbullying detection using data such as users' information, their characteristics (e.g., age, gender), and post-harassing behavior. They concluded that author information can be exploited to improve the identification of misbehavior in online social networks. Other researchers on cyberbullying have focused on: 1) the creation of an aggressiveness score, using a regression model instead of a classic binary classification to map documents according to their level of aggressiveness (Bosque and Garza, 2014); 2) how to use sentiment analysis of posts on hacking forums to predict malicious cyber events (Deb et al., 2018); 3) how to leverage social data to predict cyber-security attacks (Hernandez-Suarez et al., 2018); and 4) how to use in-game chat data and other information from a popular multiplayer game to build a scoring scheme to identify and reduce cyberbullying (Murnion et al., 2018).

Li and colleagues (2014) focused on the underground economy which represents a huge source of money for cyber criminals. For example, malware can be used to skim credit/debit card information and then this information can be sold to third parties to commit further crimes. In their study, the authors developed a deep learning-based framework and assessed its performance on a Russian forum showing that the framework was able to identify top malware sellers. Chen et al. (2015) exploited data from Twitter and weather to predict crime. Specifically, aim of this study was to accurately predict time and location where a particular type of crime would occur. They chose a lexicon-based approach combined with weather data and historical crime incidents. Using a kernel density estimation, they were able to get better results than the benchmark model.

2.5 Application: Impact of the COVID-19 outbreak on sentiment towards Italy and implications for stock market

During the recent Coronavirus disease 2019 (COVID-19) outbreak, Twitter has been widely used to share opinions and reactions to events. Italy was one of the first European countries to be struck by the virus and to organize countermeasures (lockdown, stay-at-home orders). As a result of these events, the country has suffered a reputation damage. Based on previous literature suggesting that Twitter is a widely used platform to share reaction to events, we hypothesized that the analysis of tweets collected before and after the COVID-19 outbreak might be useful to investigate changes in opinions about Italy. Analyzing the sentiment trend with different lexicons-based methods, we found a breakpoint corresponding to the date of the first Italian case of COVID-19. This event caused a relevant change in sentiment scores that we used as proxy for the country reputation. We also demonstrated that sentiment scores about Italy are strongly related with the levels of the Italian Stock Exchange main index (FTSE-MIB index), and they can therefore be seen as early detection signals of changes in the values of FTSE-MIB. Finally, we exploited the machine learning approach (comparing the performance of two classifiers) to make a content-based classification of tweets posted before and after the outbreak into a positive, neutral or negative class.

In order to provide a comprehensive picture of sentiment towards Italy, we compared different lexicons and also added an evaluation of text polarity performed with machine learning classifiers. Besides general sentiment, we also analyzed individual emotions to better understand how they might behave in reactions to events. Importantly, we expanded the analysis using economic data from the main Italian stock exchange index to assess whether changes in sentiment towards Italy might serve as early detection signals of stock market changes. Previous studies have shown how sentiment analysis is able to predict stock market movements (see, for example, Pagolu et al. 2016) and suggested that the polarity of a sentiment expressed on Twitter can be reflected on different aspects of the life of a country, including the performance of the stock market which can be influenced by exogenous factors. Indeed, we observed a strong relationship between this proxy for country reputation and stock market performance. Specifically, the change in sentiment was correlated with changes in stock exchange index values up to eight days earlier.
To the best of our knowledge, the present study is the first one using social media opinions to consider the effects of the COVID-19 outbreak on both the reputation of a country and its economy.

2.5.1 Context of the analysis: Twitter and COVID-19

Twitter is a microblogging and social networking service used by millions of people from all over the world to interact and publish contents in response to events. During the COVID-19 outbreak - which was first identified in December 2019 in Wuhan (Hubei, China) and resulted in a serious threat to public health worldwide (Hui et al., 2020) - Twitter is largely used to share information and express sentiment and concerns. Italy was one of the first European countries to be severely affected by COVID-19 as well as to implement extraordinary measures to limit viral transmission, such as lockdown and stay-at-home orders (Remuzzi and Remuzzi, 2020). This situation might have led to extensive concerns towards Italy, leading to potential damage to the country's reputation, as well as loss of investments and tourism flows. In particular, since country reputation is usually studied in terms of strategic public diplomacy, effective nation building and nation branding (see, for example, Yang et al. 2008), it is very likely for a serious health threat like the COVID-19 outbreak to negatively affect all the three dimensions of the reputation of a country. In the following, we focus specifically on the third dimension (national branding) through the assessment of the sentiment towards Italy via sentiment analysis of Twitter data.

2.5.2 Lexicon-based methods

In order to evaluate the sentiment of tweets, six different lexicons were used: sentimentR (Rinker, 2019), vader (Hutto and Gilbert, 2014), nrc, afinn, bing and syuzhet (Syuzhet R package, Jockers 2015) (Figure 2.7). In the following paragraphs, these lexicons will be described in detail.

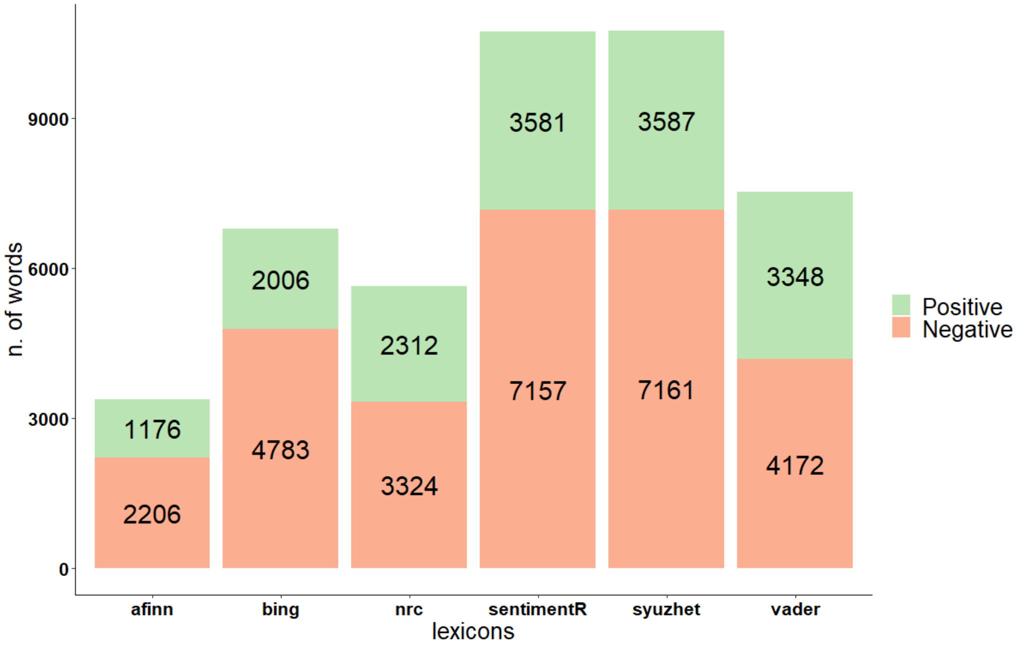


Figure 2.7: Number of positive and negative words included in the six lexicons

Syuzhet The name Syuzhet comes from the Russian formalists Victor Shklovsky and Vladimir Propp who divided narrative into two components, the "fabula" and the "syuzhet". Basically, syuzhet is the technique of a narrative, the way the elements of a story are organized. The Syuzhet lexicon was developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers. This lexicon is implemented in the Syuzhet package and includes 10,748 words (3,587 positive and 7,161 negative). The values associated with these words range from –1 to +1.

Afinn The afinn lexicon was developed by Finn Årup Nielsen to create the AFINN Word Database. Initially, words were retrieved from tweets downloaded for online sentiment analysis in relation to the United Nation Climate Conference (COP15) in 2009. The word list was originally derived from a set of obscene words as well as a few positive words. It was then extended using tweets posted for COP15, particularly those which scored high on sentiment. Later additions include words from the Original Balanced Affective Word List (Greg Siegle), Internet slang from the Urban Dictionary (including acronyms such as WTF, LOL and ROFL) and the Compass DeRose Guide to Emotion Words. The first version (AFINN-96) to be distributed comprised 1,468 unique words, including some sentences. A later version was extended to include 2,477 different words. The afinn lexicon implemented in the Syuzhet package has 3,382 words (1,176 positive and 2,206 negative) and values associated with these words range from –5 to +5 (curiously, only "kind of" and "some kind" are scored as zero) (Figure 2.8). The words were scored manually by the author.

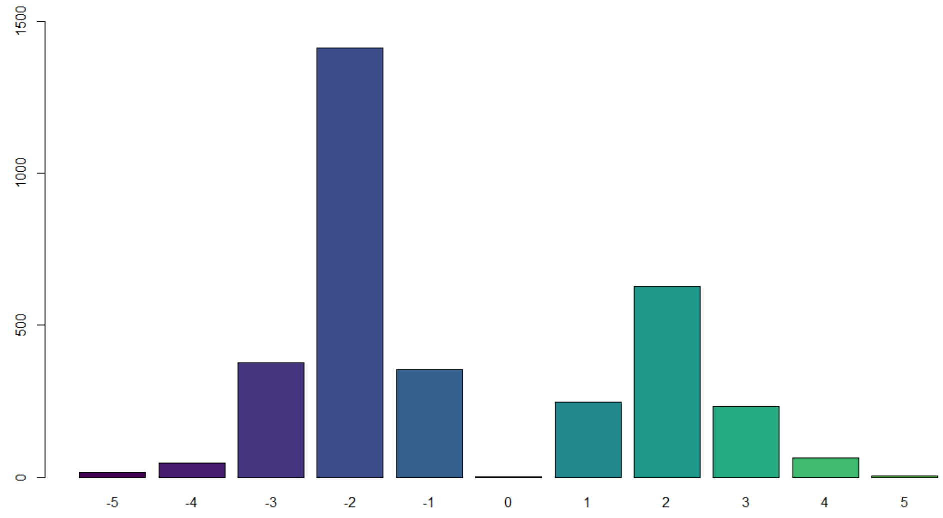


Figure 2.8: Distribution of scores of words included in the afinn lexicon

Bing The bing lexicon was first developed by Minqing Hu and Bing Liu as the Opinion Lexicon. The bing lexicon implemented in the Syuzhet package has 6,789 words (2,006 positive and 4,783 negative) and the only two values associated with these words are −1 if negative and +1 if positive.

NRC The NRC lexicon was developed in 2010 by Saif M. Mohammad and Peter D. Turney to detect emotions in texts using the Amazon Mechanical Turk, a popular crowdsourcing platform. The authors focused on emotions such as: joy, sadness, anger, fear, trust, disgust, surprise and anticipation, because these are the basic and prototypical emotions (Plutchik, 1980) and more complex emotions can be viewed as combinations of these basic emotions. Terms in the lexicon were chosen to include some of the most frequent nouns, verbs, adjectives and adverbs. Each word has been manually annotated by many users, trying to make sure that users annotated only words they were familiar with. To reduce the number of errors or cheating users, the authors asked which word was closest in meaning to the target (4 options). In this way, there is a 75% chance that an answer given at random would be eliminated in the quality control phase. The NRC lexicon implemented in the Syuzhet package has 13,901 words (6,468 unique words) representing 8 emotions and 2 sentiments (see Figure 2.9 and Appendix for details), each one associated with the value 1.

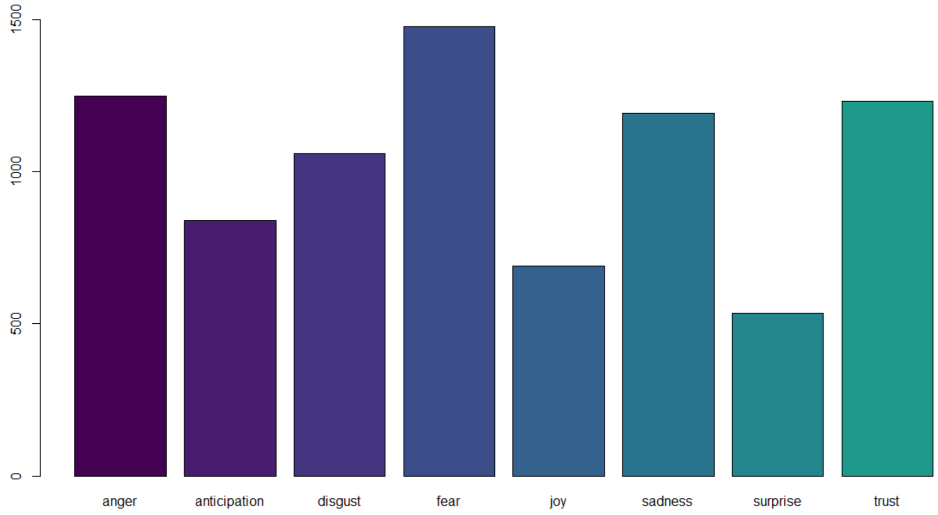


Figure 2.9: Distribution of words for each emotion included in the NRC lexicon

SentimentR The SentimentR package was developed by Tyler Rinker and exploits a dictionary lookup approach enhanced by valence shifters such as amplifiers (intensifiers), de-amplifiers (downtoners), negators and adversative conjunctions, which respectively increase, decrease, reverse the valence of polarized words. The goal is to try to get better results in terms of accuracy compared with just using a lexicon to evaluate text. For this reason, the analysis might be slower on large data sets. Since valence shifters affect the polarized words, in the case of negators and adversative conjunctions, for example, the overall sentiment of a sentence may be reversed. In the case of amplifiers/de-amplifiers, the evaluation might be over or underestimated. The author analyzed several data sets (e.g., products reviews, political debates, books) and showed that these valence shifters might affect the analysis (Rinker, 2019). In fact, up to 26% of times a polarized word appears in a text, the valence shifter also appears (amplifiers usually appear less frequently and de-amplifiers are much rarer). The package retrieves the needed sentiment from another R package (lexicon) and the dictionary represents a modified version of the Syuzhet lexicon, including 11,710 words (3,891 positive, 13 scored neutral and 7,819 negative). Values associated with these words range from –2 to +1. Valence shifters are integers values ranging from 1 to 4. The package also allows for custom dictionaries where words can be added or removed from an existing dictionary. The package has been employed in several contexts, particularly in the analysis of sentiment expressed by energy company consumers on Twitter (Ikoro et al., 2018).

Vader The Vader package was developed by C.J. Hutto in 2014 and is freely available under the MIT License. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed to analyze social media. This lexicon was built using ratings from ten independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability) and also include emoticons, acronyms and other elements. For example, positive terms such as “okay”, “good”, and “great” have values of 0.9, 1.9, and 3.1 (respectively), while negative terms such as “horrible”, the frowning emoticon :(, and “suck” (or its slang derivative “sux”) have values of –2.5, –2.2, and –1.5 (respectively). The compound score is a normalized, weighted composite score which is obtained as the sum of the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). The compound score can be used to classify sentences setting a threshold. For example, positive (compound score $\geq$ 0.05), neutral (compound score between –0.05 and 0.05), or negative (compound score $\leq$ –0.05). The Vader package has a lexicon with 7,520 words (3,348 positive and 4,172 negative) rated on a scale from –4 (extremely negative) to 4 (extremely positive), also

including zeros for neutral words. An extract of the six lexicons is available in the Appendix.
Using a Sankey diagram, we can graphically represent how many words each dictionary shares with the others and how many are present only in that dictionary (connection to itself). For example, it's easy to see that syuzhet and sentimentR have the highest number of words in common and that Vader is the lexicon with the highest number of unshared words (Figure 2.10).

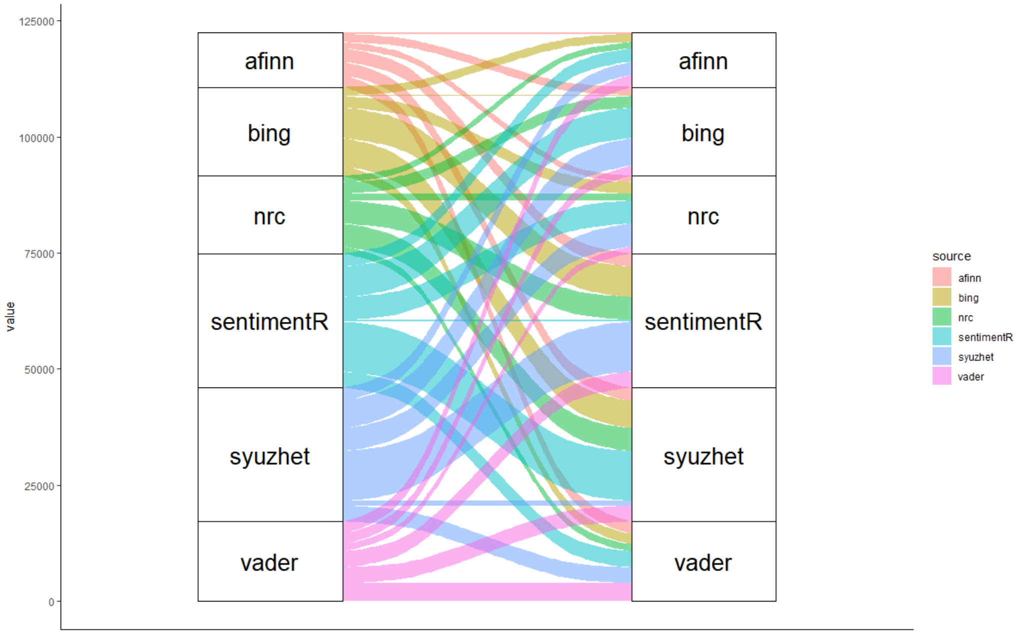


Figure 2.10: Sankey diagram showing words shared among the six lexicons or unique for a lexicon

2.5.3 Analysis of the evolution of sentiment towards Italy using the lexicon-based approach

In order to conduct a sentiment analysis to evaluate the temporal evolution of the sentiment towards Italy before and during the COVID-19 outbreak, all tweets posted in the period October 2019 – May 2020, in English language and reporting the keyword "Italy" were downloaded. In total, 4,481,104 tweets were collected. Initial data cleaning consisted in quality control and removal of tweets for which incomplete content was downloaded. Finally, 4,480,788 tweets were retained and used for further

Table 2.8: Sample of standardized mean lexicons' score

Date	**Afinn**	**Bing**	**NRC**	**Syuzhet**	**SentimentR**	**Vader**
2019-10-01	0,17	0,14	0,08	0,13	0,12	0,11
2019-10-02	0,18	0,16	0,14	0,17	0,17	0,15
2019-10-03	0,23	0,22	0,13	0,19	0,18	0,19
2019-10-04	0,24	0,23	0,10	0,19	0,16	0,18
...	...	...	...	...	...	...
2020-05-28	0,07	0,03	0,04	0,06	0,07	0,03
2020-05-29	0,11	0,09	0,07	0,11	0,11	0,08
2020-05-30	0,09	0,04	0,04	0,08	0,08	0,06
2020-05-31	0,04	0,00	0,04	0,03	0,05	0,01

analysis. During preprocessing of tweets, punctuation marks, hashtags, mentions and links were removed. In addition, tweets were converted in lower case letters. Preprocessing was conducted in R (R Core Team, 2020) version 3.6.3.
For each day and each lexicon we computed a sentiment score, calculated as the mean score for tweets collected in a single day. These values were then plotted to observe the temporal evolution of the sentiment towards Italy (Figure 2.11a). The mean scores were then standardized to better compare the different methods as they use different scales to evaluate the sentiment of a text. (Figure 2.11b). Standardization has been performed using the following formula (Equation 2.5):

$$z = \frac{x - \bar{x}}{sd} \tag{2.5}$$

where $\bar{x}$ is the mean of the sample, and sd is the standard deviation of the sample. These z-scores have been plotted in Figure 2.11b and an extract is reported in Table 2.8. Standardized sentiment scores allowed to observe a high concordance among different methods. As can be observed in Figure 2.11, for all lexicons the most positive scores were observed on New Year's Eve. Notably, while neutral or slightly positive values can be steadily observed in the previous months, from February 21, 2020 there is a drop in sentiment towards extremely negative values. This date corresponds to the reporting of the first Italian case of COVID-19 in Codogno (Romagnani et al., 2020). From this date, sentiment scores remain negative up to May, when a slow increase towards less negative / neutral values can start to be observed.

The empirical observation of a sharp decrease in sentiment scores at February 21, 2020 was verified trough an analysis aimed at finding breakpoints, i.e. structural breaks constituting unexpected changes in a time series. These changes can occur at single or multiple points and their detection can allow to increase knowledge on

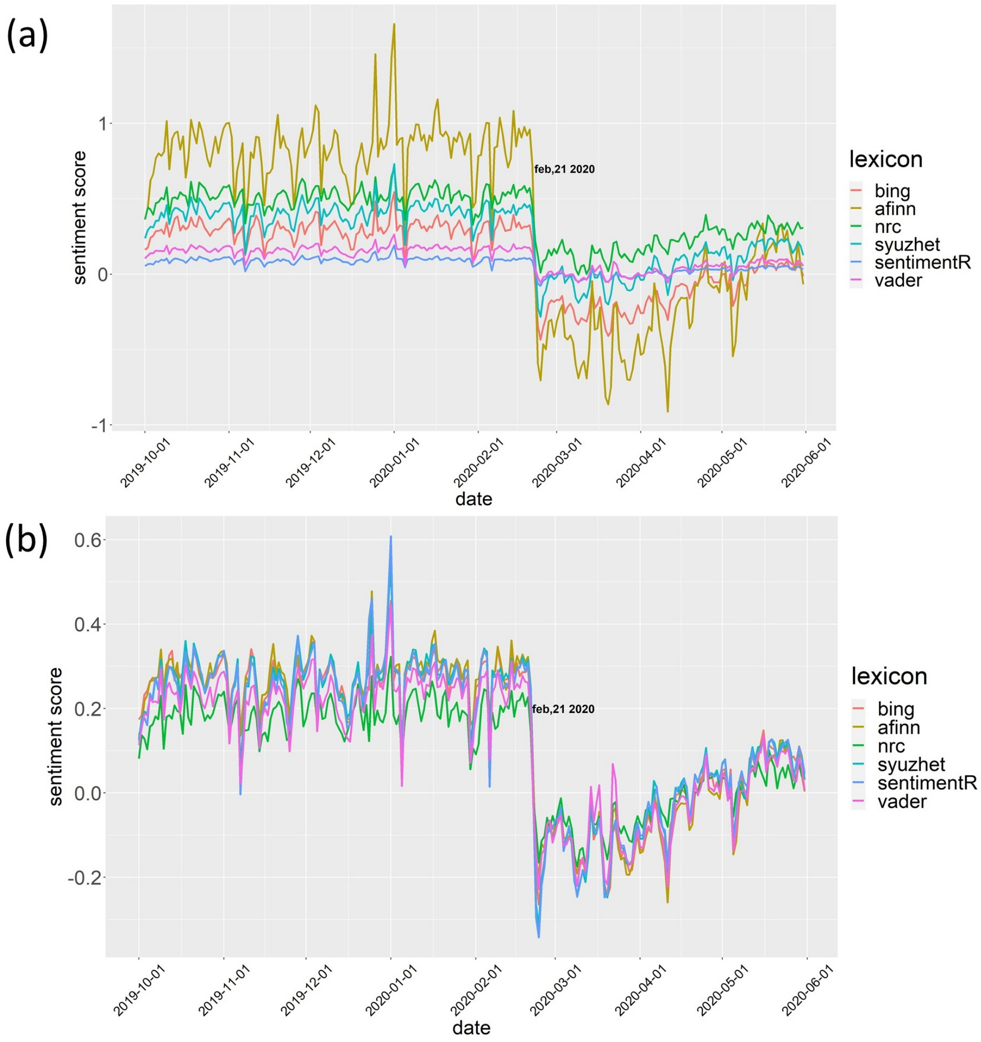


Figure 2.11: Sentiment score (a) and standardized sentiment score (b) of collected tweets including the keyword "Italy" from October 2019 to May 2020

Table 2.9: Identified break dates using different lexicons

Lexicon	**Break dates**	
Afinn	February 21	April 20
Nrc	February 21	April 11
bing	February 21	April 13
syuzhet	February 21	April 12
vader	February 21	April 13
sentimentR	February 21	April 12

the phenomenon object of study. This analysis was conducted with the strucchange R package (Zeileis et al., 2002) for sentiment scores computed with all six lexicons. Basically, this method is aimed at assessing deviations from stability in the classical linear regression model (Equation 2.6):

$$y_i = x_i^T \beta + u_i \tag{2.6}$$

In many applications, it is reasonable to believe that there might be m breakpoints (especially if some exogenous event occurs) in which a different regression might be applied. For example, in the case of a single breakpoint, we might better explain the data using a regression (let's call it Regression A) for the first part and another regression (regression B) for the second part. In general, there might be *m + 1* segments in which regression coefficients are fitted, and therefore the model can be rewritten as in Equation 2.7

$$y_i = x_i^T \beta_j + u_i \qquad i = i_{j-1} + 1, \ldots, i_j;\ j = 1, \ldots, m+1 \tag{2.7}$$

where j represents the segment index. The breakpoints function estimates the breakpoints by minimizing the residual sum of squares (RSS) or Bayesian Information Criterion (BIC) of the Equation 2.7. For sentiment scores computed with all lexicons, we were able to confirm the existence of a breakpoint on 21 February 2020 (Table 2.9). This finding confirms that, regardless of the method used to evaluate sentiment scores, there is high concordance in the identification of a sudden change of sentiment on this date.
The optimal number of breakpoints was two for all lexicons (Figure 2.12).

However, we observed a small variability between lexicons as regard to the date of the second breakpoint (Table 2.9). A graphical representation of the breakpoints obtained from all lexicons is reported in the Appendix. Based on the consistent observation for all lexicons of a structural breakpoint at February 21, 2020, we defined

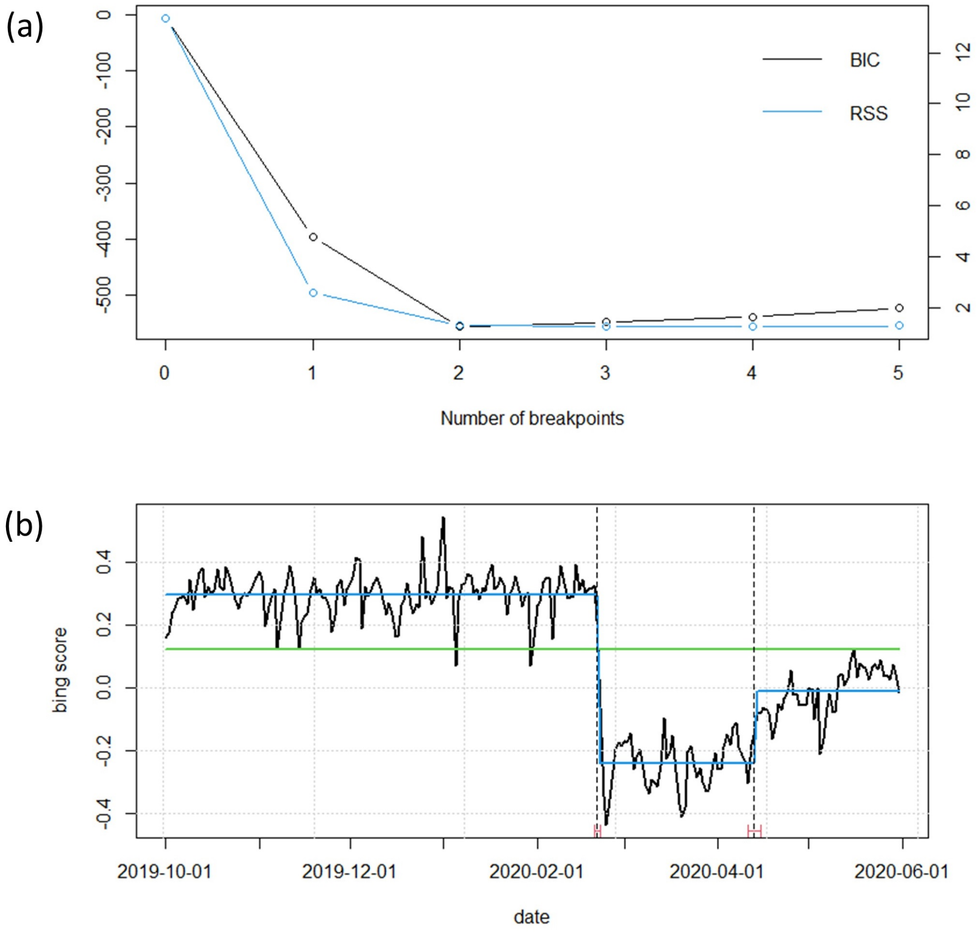


Figure 2.12: (a) BIC and Residual Sum of Squares using the bing lexicon, (b) Breakpoints in sentiment score using the bing lexicon

two different periods for the subsequent analyses: Period A from October 1, 2019 to February 20, 2020; and Period B from February 21, 2020 to May 31, 2020. We conducted more detailed analyses on tweets included in these two periods using the nrc lexicon, as this method allows to conduct analyses on specific emotions. Specifically, the nrc lexicon allows to evaluate a text in terms of eight basic emotions including four negative (anger, disgust, fear and sadness) and four positive emotions (anticipation, joy, surprise and trust) by computing the mean of the number of words associated with each emotion for each day. These values have been used to plot graphs in Figure 2.13 and Figure 2.14. The scores obtained for sentiment as well as for specific emotions show non-normal distribution according to the Shapiro-Wilk test. The Mann-Whitney U test was used to compare sentiment and emotions between Period A and Period B. As shown in Figure 2.13, general *positive* sentiment is not significantly decreased (p = 0.84) in Period B compared with Period A, whilst *negative* sentiment is increased (p < 0.001).

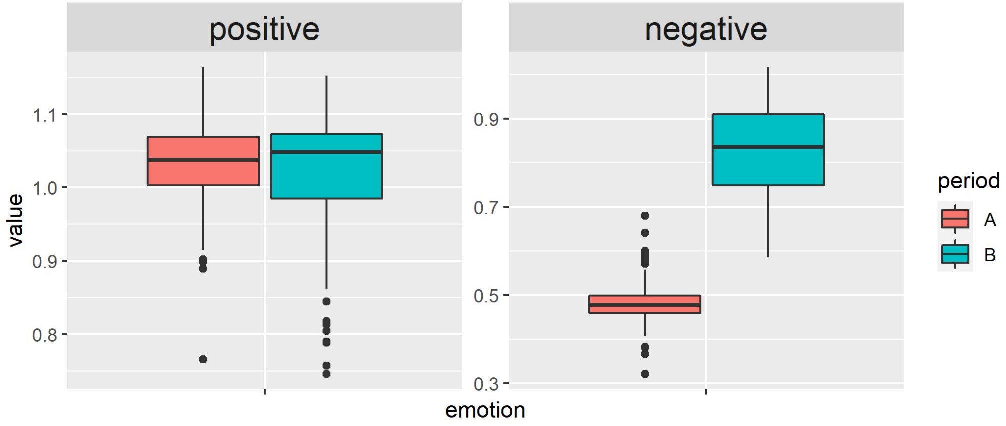


Figure 2.13: Boxplots showing positive and negative sentiment in Period A (October 1 2019 – 20 February 2020) and Period B (21 February – 31 May 2020)

Although a rise of negative emotions in Period B is somewhat expected and in accordance with our hypothesis, positive emotions overall remain stable. This finding is further explored, and partly confirmed, through the analysis of specific emotions. As shown in Figure 2.14, all negative emotions show higher values in Period B compared with Period A (p < 0.001). Conversely, positive emotions show a more heterogenous course. Specifically, joy show lower values in Period B (p < 0.001), whilst anticipation, surprise and trust show higher values in Period B compared with Period A (p < 0.001). A decrease of joy can be expected during such a hard time,

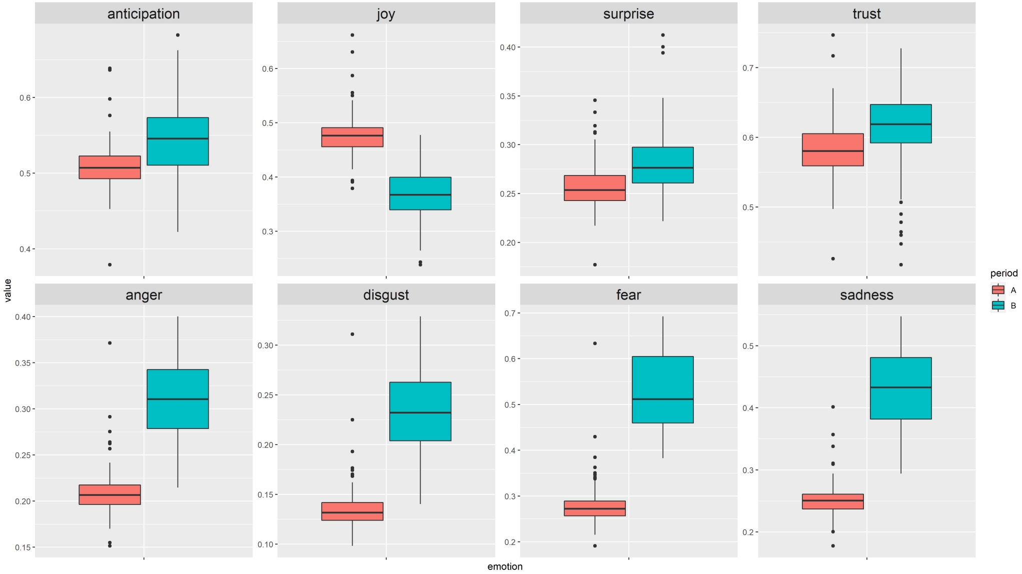


Figure 2.14: Boxplots showing positive and negative emotions in Period A (October 1 2019 – 20 February 2020) and Period B (21 February – 31 May 2020)

whilst the other three emotions might increase for different reasons. Specifically, the increase in values of anticipation and surprise might be interpreted as follows: even if the COVID-19 outbreak is a negative event, it has the power to generate surprise and to increase the desire to know what will happen in the near future. On the other hand, the increase in trust might be related to the willingness to believe in a speedy recovery as well as to encouragements towards Italy from other countries.

Figure 2.15 shows the day-to-day detailed temporal evolution for the eight emotions from October 1 to May 31. It is possible to observe that, from February 21, 2020, the negative emotions (right panel) rise, whilst positive emotions (left panel) do not all act in the same way. Specifically, anticipation and trust first decrease and subsequently show a progressive increase in the last days, joy first decreases and then remains stable, and surprise rises. Figure 2.15 gives us the opportunity to grasp subtle nuances that do not emerge by observing only the general trend. The detail of the various emotions allows us to observe an increase in the levels of negative emotions, which is expected since an event such as the outbreak of an epidemic disease certainly has the potential to increase anger, disgust, fear and sadness both for those who experience the event firsthand and for those who are not directly affected by it. Positive emotions have more diversified trends, which in any case can be easily interpreted taking into consideration the individual emotion. Joy, for example, un-

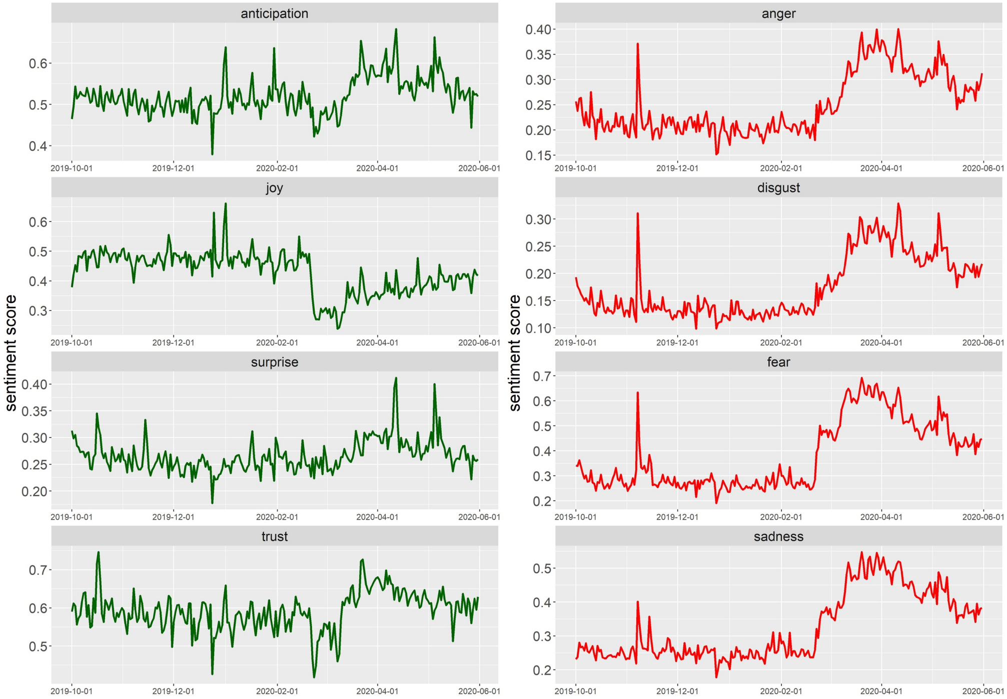


Figure 2.15: Temporal evolution of the positive (on the left: anticipation, joy, surprise and trust) and negative (on the right: anger, disgust, fear and sadness) emotions from October 1 to May 31

dergoes a sharp decrease which brings its value in period B well below that of period A and can be explained by the worsening of the situation that was occurring in Italy day after day. Trust, on the other hand, has a similar sharp decrease but it recovers in a very short time (about a month). This can be explained by the fact that many people, at that time, wanted to express their trust that things would get better for Italy using, for example, slogans such as *"Andrà tutto bene"* (it will be okay) which were very popular in Italy during the spread of the pandemic.

2.5.4 Sentiment as an early detection signal of stock market performance

We collected data from the main Italian stock exchange index (FTSE-MIB) to assess if it was possible to identify a trend like the one observed with sentiment analysis. Specifically, we collected closing values of FTSE-MIB from October 1, 2019 to

May 31, 2020 (using linear interpolation we also added the values corresponding to weekends and holidays, in order not to break the time series and make the period compatible with that of the sentiment analysis). Based on these data, we have highlighted the presence of three breakpoints: (i) November 6, 2019, (ii) March 7, 2020 and (iii) April 26, 2020 (Figure 2.16). As shown by the figure, the first and most relevant breakpoint after the COVID-19 outbreak period is on March 7, i.e. 15 days after the drastic drop in sentiment (February 21, 2020). It is interesting to note that the day after the observed breakpoint, the Lombardy region was set into lockdown.

Since various studies have applied sentiment analysis on different types of data to predict share values (see Section 2.4.1), we wanted to verify whether an association between sentiment scores evalaued before and after the COVID-19 outbreak and stock market index values might be identified. Considering the time lag (15 days) between the main breakpoint in sentiment scores and the corresponding breakpoint in FTSE-MIB values, our hypothesis is that an association exists when we apply a time lag (i.e. past sentiment scores from today associated with stock exchange index values in the next days).

In order to analyze the relationship between the change of sentiment and FTSE-MIB index values over time, we constructed a vector autoregressive (VAR) model using the vars package in R (Pfaff, 2008). To evaluate if the association between sentiment scores and FTSE-MIB prices depends from the specific method used to perform sentiment analysis, we estimate the model separately for each method and each lexicon used to compute sentiment scores. First, differences of log transformed data are taken to make the time series stationary. Stationarity was checked via inspection of the time series plot as well as with the Augmented Dickey-Fuller test implemented in the tseries R package (Trapletti and Hornik, 2021). The optimal lag length for the sentiment computed using each lexicon was estimated using the VARselect function of the vars R package, based on the Akaike Information Criterion (AIC) criterion. Stability of the models was checked based on eigenvalues of the companion coefficient matrix using the vars package. The bidirectional association between sentiment and FTSE-MIB index values was analyzed using the Granger-causality test.

Using the VARselect function, we observed different optimal lags in the relationships between FTSE-MIB index values and sentiment scores computed using the six dictionaries (Table 2.10).

We then conducted Granger-causality test on the VAR models estimated using these time lags. The null hypothesis for this test is that lagged values of a variable X that evolves over time (i.e. sentiment scores) do not explain the variation in variable Y (i.e. FTSE-MIB index values). In other words, variable X Granger-causes variable

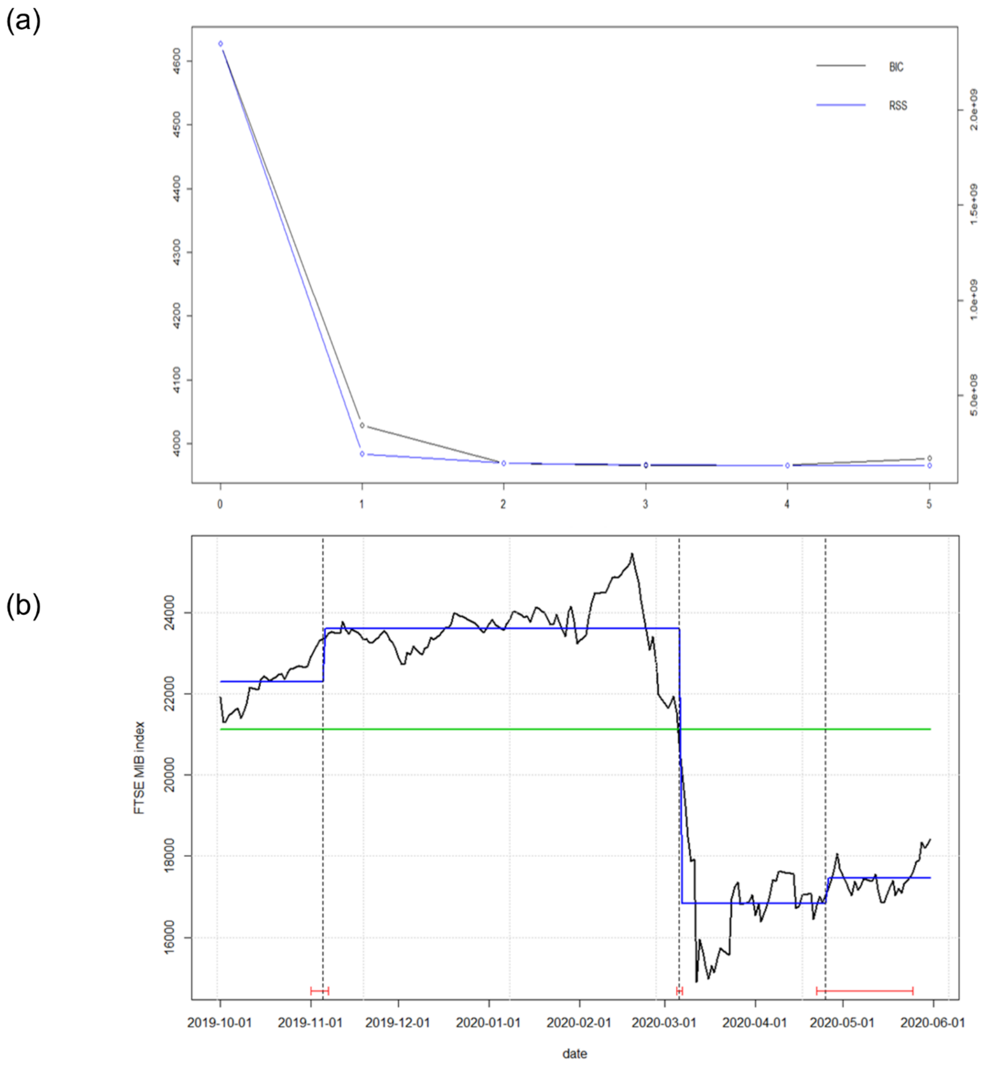


Figure 2.16: (a) BIC and Residual Sum of Squares using the FTSE-MIB index values; (b) Breakpoints in FTSE-MIB index values

Table 2.10: Optimal time lags and results of the Granger-causality test

Lexicon	**Optimal lag (days)**	**Sentiment G-causes FTSE-MIB**	**FTSE-MIB G-causes sentiment**
Afinn	8	0.025	0.439
Bing	4	0.135	0.171
NRC	6	0.051	0.463
Syuzhet	8	0.085	0.919
SentimentR	4	0.224	0.640
Vader	5	0.034	0.427

Abbreviations: G-causes, Granger-causes

Y in case predictions of the values of Y, based on its own past values and on the past values of X, are better than predictions of the values of Y based only on past values of Y. We found that FTSE-MIB index values do not Granger-cause sentiment computed using any dictionary. Conversely, sentiment scores computed with afinn and vader are found to Granger-cause FTSE-MIB index values and a trend is observed for NRC (Table 2.10). While both the drop in sentiment scores as well as in FTSE-MIB values are caused by the pandemic, our results suggest that sentiment scores computed with afinn and vader are useful to predict FTSE-MIB index values with an optimal lag of 8 and 5 days, respectively.

Overall, these findings enforce the idea that a change in sentiment scores can be considered as an early detection signal (up to eight days earlier) for potential effects on the stock market values.

2.5.5 Machine learning approaches: Naïve Bayes

Many classification problems can be tackled using supervised machine learning algorithms such as NB. While one of the historical and more famous uses of these classifiers is spam filtering, text classification is another application that has gained popularity in the last two decades (Sahami et al., 1998). This classifier is based on the Bayes theorem which says that, given two events A and B, we can compute the conditional probability of A given B as the probability of B given A times the probability of A, divided by the probability of B, as shown in Equation 2.8:

$$P\left(A|B\right) = \frac{P(B|A)\ P(A)}{P(B)} \tag{2.8}$$

The assumption of the NB classifier is that features x_j $(j = 1, \ldots, p)$ are conditionally independent: $P\,(x_j|\,X = \;x_1, \ldots, x_p) = P(x_j) \quad \forall\; j \in [1, p]$. For any dependent categorical variable y with k classes (k $\geq$ 2) the Bayes theorem can be written as:

$$P(C = C_k \mid X = x_1, \ldots, x_p) = \frac{P(C_k)\prod_{j=1}^{p} P(x_j \mid C_k)}{P(x_1, \ldots, x_p)} \tag{2.9}$$

where C is one of the C_k classes of y, and X is a vector of random variables x_j. In the context of sentiment analysis, the challenge is to correctly assign each tweet $t_i\,(i = 1, \ldots, n)$ to one of the k classes of y. For instance, tweets can be classified either in positive or negative (k = 2). Since the denominator in Equation 2.9 is the same for each class, we can compute an approximated version of Equation 2.9 eliminating the term in the denominator, as:

$$P(C_k \mid x_1, \ldots, x_p) \propto P(C_k)\prod_{i=1}^{p} P(x_j \mid C_k) \tag{2.10}$$

so that the assigned class $\hat{C}$ is chosen according to:

$$\hat{C} = argmax\; P(C_k)\prod_{i=1}^{p} P(x_j \mid C_k) \tag{2.11}$$

Therefore, any tweet t_i is assigned to one, and only one, of the C_k classes.

2.5.6 Machine learning approaches: Support Vector Machine

SVM is a non-probabilistic binary linear classifier (Suykens and Vandewalle, 1999), as it is not based on any specific probability distribution. The goal of SVM is to find a line (hyperplane) which best separates data. In sentiment analysis context, the task is to assign tweets $t_i\,(i = 1, \ldots., n)$ to one of the possible classes (for instance, positive or negative) of the categorical dependent variable y. The hyperplane that best separates should be placed as far as possible from the nearest points of the two classes. The distance between a point and the hyperplane is called margin (hard or soft). The distinction between hard and soft margin is related to the possibility of the soft margin to “making exceptions” and allows for misclassified data if they do not exceed a certain threshold. In case of hard margins, if they are large, the probability of incorrectly classifying that observation is smaller. Using training data $(\overrightarrow{x}_1, y_1), \ldots, (\overrightarrow{x}_n, y_n)$, where $\overrightarrow{x}_i$ are vectors of feature values of observation

i $(i = 1, \ldots, n)$, and y_i are labels (+1 or –1, depending on the class we are considering) is possible to train the linear SVM classifier. The goal is to find a hyperplane

$$\vec{w} \cdot \vec{x}_i + b = 0 \tag{2.12}$$

that best separates the two classes ($\vec{w}$ is the normal vector to the hyperplane). It is possible to add constraints about the margin to avoid having any observation within the margins. These constraints are usually specified as in Equation 2.13:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \qquad \forall \quad 1 \leq i \leq n \tag{2.13}$$

One of the biggest problems that may arise when we are strictly looking for the best hyperplane is that a single new observation might cause a great shift of the hyperplane, and therefore reduce the margin by a great amount. For this reason, it would be better to have a classifier able to tolerate a sub-optimal separation of the two classes, but capable of carrying out the classification with an overall greater precision (soft margins). Basically, with a soft margin we allow some observation to fall on the wrong side of the margin, or even of the hyper-plane. Since each margin is distant $1/|| \vec{w} ||$ from the hyper-plane that separates the classes, the distance between the two margins is equal to $2/|| \vec{w} ||$. Therefore, to maximize this distance we should minimize $|| \vec{w} ||$. To achieve this goal, we should solve the quadratic programming problem in Equation 2.14:

$$\begin{cases} min\ f : \ \frac{1}{2} \ ||W||^2 \\ s.t.g : y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \end{cases} \tag{2.14}$$

Equation 2.14 is a constrained optimization problem solvable with the Lagrange multipliers method. To apply this method, we make use of a slack variable ξ_i such that the constraint can be expressed as

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq B(1 - \xi_i) \tag{2.15}$$

with $B = 1/|| \vec{w} ||$, $\xi_i \geq 0$ and $\sum_{i=1}^{n} \xi_i \leq K$. In this way, some points can be on the wrong side, but only up to a given distance (less than or equal to the constant K). The optimization problem reported in Equation 2.14, subject to the constraint specified in Equation 2.15, finally becomes

$$\begin{cases} y_i(\vec{w} \cdot \vec{x}_i + b) \geq B(1 - \xi_i) \\ min||w||^2 \quad s.t. \quad \xi_i \geq 0 \sum_{i=1}^{n} \xi_i \leq K \end{cases} \tag{2.16}$$

2.5.7 Performance of machine learning approaches

We apply the widely used machine learning approaches described in the previous sections to analyze the polarity of tweets and compare their performances in the prediction of tweet polarity before and after the COVID-19 outbreak. Considering that, after the outbreak, the topic was largely discussed on Twitter, we aimed to evaluate whether the two classifiers showed differences in performance when analyzing texts pertaining to more diverse topics (whole data set and Period A) or a more homogenous topic (Period B). This analysis was conducted on 10,000 tweets randomly sampled (5,000 tweets from Period A and 5,000 from Period B). Tweets were manually labeled into "positive", "negative" or "neutral" based on their content. Tweets consisting only of hashtags, urls, emoticons or single words were excluded. A total of 6,409 tweets were labelled as positive, neutral or negative (3,227 from Period A and 3,182 from Period B) and used for subsequent analysis. Preprocessing and data cleaning operations included removal of punctuation, numbers and stop words. After data cleaning, tweets were used to form a document text matrix using the tm R package (Feinerer and Hornik, 2019). The analysis were conducted on the whole data set of 6,409 positive, negative or neutral tweets, as well as on the two subsets. NB and SVM models were estimated using the e1071 (Meyer, 2019) and caret (Kuhn, 2008) R packages.

As the specific content of the tweets is more homogeneous in Period B, being closely related to COVID-19, we hypothesize the two classifiers to perform better for tweets posted in this period compared to those posted in the pre-outbreak period (Period A). Results about the k-fold cross-validated (k = 5) performance are reported in Table 2.11.

Respect to the whole test set (Period A + Period B), NB reaches a good accuracy for negative and positive tweets (0.77 and 0.78, respectively) and a lower accuracy for neutral tweets (0.66). A similar difference is observed for sensitivity and positive predictive value (PPV), for which a good performance is only observed for positive and negative tweets. Conversely, specificity is high for all examined classes. Restricting the analysis to Period A, the classifier shows a general worse performance, especially in the classification of positive and negative tweets. An opposite result is observed when restricting to Period B. For negative tweets, all metrics show an improvement compared to either Period A or the whole data set (Table 2.11), while more variable results can be observed for neutral and positive tweets. Therefore, NB shows an improvement in the classification accuracy when analyzing a data set in which a specific topic, i.e. the COVID-19 pandemic, characterizes the content of a large part of negative tweets (Period B) compared to periods in which more diverse topics are discussed by users (whole data set and Period A).

Table 2.11: Performance of Naïve Bayes and Support Vector Machine in the classification of tweets

	NB			**SVM**		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Whole data set						
Sensitivity	0.72	0.53	0.69	0.67	0.66	0.67
Specificity	0.83	0.79	0.86	0.88	0.74	0.89
PPV	0.69	0.52	0.73	0.73	0.53	0.78
NPV	0.86	0.79	0.83	0.84	0.83	0.83
Accuracy	0.78	0.66	0.77	0.77	0.70	0.78
Period A						
Sensitivity	0.63	0.58	0.63	0.58	0.74	0.56
Specificity	0.84	0.76	0.82	0.90	0.65	0.89
PPV	0.68	0.52	0.65	0.76	0.49	0.73
NPV	0.81	0.81	0.81	0.80	0.85	0.79
Accuracy	0.74	0.67	0.72	0.74	0.70	0.73
Period B						
Sensitivity	0.73	0.46	0.73	0.69	0.59	0.69
Specificity	0.81	0.81	0.86	0.86	0.74	0.90
PPV	0.65	0.51	0.76	0.70	0.50	0.81
NPV	0.86	0.78	0.84	0.85	0.81	0.83
Accuracy	0.77	0.64	0.79	0.77	0.66	0.79

Abbreviation: NPV, negative predictive value; PPV, positive predictive value

Next, SVM with a radial kernel is estimated in the same way as NB. Again, when analyzing the whole data set (Period A + Period B) the classifier shows a good accuracy, specificity and PPV for positive and negative tweets, but a worse performance for neutral tweets (Table 2.11). When analyzing Period A, consistently with the results obtained with NB, SVM shows lower accuracy in the classification of positive and negative tweets. Conversely, higher sensitivity and negative predictive value (NPV) are obtained for neutral tweets. When analyzing Period B, as observed with NB, the classification of negative tweets improves based on all metrics, while the classification of neutral and positive tweets generally shows a lower performance compared with the whole data set. Results obtained for the whole data set (Period A + B, Table 2.11) evidence that the two classifiers show similar performance based on all metrics except sensitivity, for which the best results for the neutral class are obtained with SVM. Results obtained for the tweets collected during Period A reveal that, while SVM shows a higher sensitivity in the classification of neutral compared with positive and negative tweets, the opposite is observed with NB. Accuracy and NPV are similar for all classes and for both models. When analyzing the tweets collected during Period B it can be observed that, for both models, the classification of negative tweets shows better performances compared with other classes as well as with negative tweets related to Period A or the whole data set. Overall, these results are consistent with our initial assumption that classifiers' performance is improving for negative tweets following the beginning of the outbreak, as the content of tweets is more specifically focused on the pandemic. This was especially true for negative tweets, as we can speculate their content to be particularly homogeneous and mostly related to the pandemic. The class for which we observed a general worse performance is the neutral class. This is probably due to the fact that neutral tweets can contain mixed feelings about the object of study or ambiguous statements hard to interpret, that can therefore lead to a higher number of classification errors.

The results of this research must be interpreted in light of some limitations. First, data was collected from a single social network (i.e., Twitter). However, as also shown in the literature review, this platform is widely used to evaluate reactions to important events. In addition, our findings might not be applicable to different countries as results could vary due to the different country's reputation or other factors (e.g. cultural or socioeconomic factors). Despite these limitations, our findings contribute to elucidate how the COVID-19 outbreak affected sentiment towards Italy, one of the first countries to be severely affected, and can also help to shed light on the relationship between country reputation and the possible economic repercussions of an event of this magnitude. In conclusion, we applied lexicon-based sentiment analysis and machine learning methods to evaluate the sentiment towards Italy before and after the COVID-19 outbreak using real data collected from Twitter. We observed a rise in negative emotions towards Italy in correspondence of the first Italian case of COVID-19, followed by a change towards more neutral or slightly positive values. We showed that sentiment scores can be also used as early detection signals of changes in stock exchange values. Future developments of this work will include the evaluation of the sentiment towards different countries as well as the potential impact of government restrictions such as social distancing, lockdown or travel restrictions.

Chapter 3

Building a new dictionary: the Eye dictionary

3.1 Lexicon-based approach: limitations of existing lexicons and studies aimed at creating new dictionaries for sentiment analysis

As discussed in detail in Section 2.2, the sentiment analysis lexicon-based approach is based on a dictionary, i.e. a tool where hundreds or thousands of words are associated with a polarity (e.g. words such as “bad” or “ugly” have a negative polarity, while “good” or “beautiful” a positive polarity). Each word of the text we want to classify is searched in the dictionary. If the word is present, the value assigned to that word will contribute to the overall text sentiment (in combination with the other words present both in the text and in the dictionary). A summarizing function (e.g. average or sum) is then applied in order to obtain a value representative of the whole text. While a number of dictionaries are already available, the main reasons that can lead to the development of novel dictionaries can be related to the need to create an instrument adapted to the language or the specific field under study, or to improve the performance, addressing limitations of currently used dictionaries, as will be the focus of our work. The majority of dictionaries for sentiment analysis is available in English. Nonetheless, a number of authors focused on the development of dictionaries in other languages. For instance, some authors used cross-lingual approaches to create dictionaries for sentiment analysis in widely used languages such as Chinese, starting from English lexicon-based systems via translation of English text (Yao et al. 2006; Wan 2008).

The creation of a new dictionary poses specific challenges in the case of minor languages for which a substantial work, as well as a deep knowledge of the language, might be needed to translate labels or assign new polarities. Approaches that have proven to be helpful to address these challenges include use of language translation tools such as Google Translate API, manual checking and automatic annotation tools. Mikula and colleagues (2017) used an innovative approach to adapt sentiment dictionaries to the Slovak language using automatic as well as human annotation. Specifically, they used an optimization algorithm called Particle Swarm Optimization (PSO) to assign polarities and showed that this approach outperformed a dictionary annotated by a human. Beside providing a tool to classify texts in different languages, some authors focused on improvement of existing approaches as in the case of Viet-SentiLex, a dictionary developed in Vietnamese and able to consider the polarity of ambiguous sentiment words (Viet et al., 2018). Momtazi (2012) developed a German opinion dictionary including 1,864 words for the analysis of sentiment strengths of positive and negative texts published on social media. The study extended previous work in which the authors developed sentiment analysis dictionaries for the German language. In this study, however, the lists of positive and negative words, but no strength degree, were provided (Waltinger, 2010).

Other authors focused on the creation of dictionaries aimed at improving classification performance in specific fields such as online news (Rao et al., 2013), software engineering (Islam and Zibran, 2018) or the financial domain (Tabari and Hadzikadic, 2019). Among these, the study from Rao et al. (2013) describes a method to automatically build word-level and topic-level dictionaries to detect social emotions. Differently from previously available dictionaries for the classification of general domain texts, this approach was specifically designed to allow the identification of entities (e.g. products or brands) or events able to evoke different social emotions (Rao et al., 2013).

While different dictionaries for sentiment analysis are available, the variability in their performance in the classification of different types of texts calls for the development of alternative dictionaries. In particular, a high number of dictionaries simply assign binary values to positive (e.g. +1) or negative (e.g. -1) words, without taking into consideration potential differences in the relevance of assessed words (Hansen et al. 2011; Hu and Liu 2004). However, a non-trivial task in sentiment analysis is the identification of words that should be assigned greater importance. In other terms, specific weights to attribute to words should be defined based on some metrics reflecting their potential relevance. Dictionaries which assign weights often rely on subjective evaluation from the authors or from crowdsourcing platforms to assess the degree of positive or negative sentiment associated with a word (Mohammad and Turney, 2013). While of high value, this kind of evaluation does not take into consideration potential differences related to how much a word is able to attract the attention of a reader. Words able to attract more attention such as unusual or catching words might be looked at for a longer time and be more easily remembered. Therefore, we can speculate that these words might play a more relevant role in the determination of the polarity of a text compared with words that attract less attention. We hypothesize that the eye tracking technology, which allows to assess the exact position of the eyes during the visualization of texts, images or other stimuli, can contribute significantly to the identification of words that may be able to gain more attention from a reader and are thus potentially more relevant. Eye tracking allows to analyze eye position and movements with a high degree of accuracy and is therefore used for several applications in which we need to draw conclusions about the cognitive processes that takes place in our mind when we are observing something. Indeed, this technology allows to gain a better understanding of the user's visual attention as well as information processing during either the observation of an object, the execution of tasks (Zammarchi et al., 2021) or reading. The latter application can be useful to gain insights into diseases characterized by abnormal reading patterns, such as dyslexia (Nilsson Benfatto et al., 2016), neurodegenerative disorders (Fernandez et al., 2013) or psychiatric disorders (Zammarchi and Conversano 2021; Rubin et al. 2021. Besides the study of medical conditions, another field in which the application of eye tracking to reading can provide a substantial contribution is the assessment of human emotions (Lim et al., 2020). To this regard, most of the available studies focused on processing of sentences characterized by sarcasm (Olkoniemi et al., 2019) or contents associated with different emotions such as fear (Kaakinen and Simola, 2020). In addition, other studies explored whether eye tracking might be useful to investigate predictability of words during reading (Luke and Christianson, 2018) or to assess naturalistic reading patterns (Cop et al., 2017). To our knowledge no previous study has leveraged eye tracking data to improve the classification of texts using a lexicon-based sentiment analysis approach. In the following section we present the Eye dictionary, a new dictionary for sentiment analysis developed using eye-tracking data as weights to determine the relevance of words based on the attention they might receive. We evaluate the performance of the developed dictionary in the classification of the polarity of positive and negative texts compared with other widely used dictionaries. Finally, we discuss obtained results, strengths and limitations of this work as well as future expansions.

3.2 Development of a new lexicon: the Eye dictionary

To develop a dictionary based on eye tracking data, two main aspects have to be defined: weights and polarities. In Section 3.2.1 we describe the data sets used to compute weights and polarities, while in Section 3.2.2 we describe how these components of the dictionary were defined.

3.2.1 Data sets for the definition of weights and polarities

To compute weights, we used Provo Corpus and the Ghent Eye-Tracking Corpus (GECO), two large corpora of eye tracking data.
The Provo Corpus was originally developed to investigate predictability effects in reading (Luke and Christianson, 2018). This corpus includes 55 paragraphs taken from various sources (e.g. news articles, science magazines and works of fiction). The length of the paragraphs ranged from 39 to 62 words (average: 50 words). Each paragraph contained 2.5 sentences on average (range: 1-5) and sentences were on average 13.3 words long (range: 3-52). Overall, the Provo Corpus included 1,197 unique words. Participants included 84 students enrolled in Brigham Young University. All participants had 20/20 corrected or uncorrected vision. Eye movements were recorded using an SR Research EyeLink 1000 Plus eye-tracker (spatial resolution of 0.01°), sampling at 1000 Hz. Participants were seated at 60 cm from a monitor (resolution: 1,600 × 900). A chin and forehead rest were used to minimize head movements. Although viewing was binocular, eye movements were only recorded from the right eye. Each participant was presented texts in random order. Each text was presented only after a stable fixation was detected. Data cleaning procedure included removal of fixations shorter than 80 ms and longer than 800 ms (about 4% of the data).

The GECO corpus consisted in the English version of the novel “The Mysterious Affair at Styles” by Agatha Christie, including 5,031 sentences, for a total of 5,013 unique words (Cop et al., 2017). Participants included 14 students enrolled in a bachelor’s or master’s program of psychology at the University of Southampton (age average: 21.8 years, range: 18–36 years). All participants had normal or corrected-to-normal vision, and none reported having any language and/or reading impairments. Eye tracking movements were recorded with a desktop-mounted EyeLink 1000 system (SR Research, Canada), sampling at 1000 Hz. A chinrest was used to minimize head movements. As in the case of the Provo Corpus, although viewing was binocular, eye movements were only recorded from the right eye.

Polarities of the words included in the Eye dictionary were computed using two strategies. In a first version of the dictionary, polarities were computed based on a large data set of movie reviews including 50,000 texts from the Internet Movie Database (IMDB), labeled as positive or negative reviews (Maas et al., 2011). Briefly, this collection of reviews was constructed allowing no more than 30 reviews per movie and included an even number of highly polarized positive (score ≥ 7 out of 10) and negative (score ≤ 4 out of 10) reviews. In a second version of the dictionary, polarities were manually computed rating the words as positive, negative and neutral. For neutral words, the polarity computed as described for the first version of the dictionary was retained.

3.2.2 Computation of weights and polarities

Across both corpora, we extracted eye tracking data in the form of dwell time for each word (i.e. total reading time calculated as the sum of the duration across all fixations on a given word). For each word w included in each corpus of eye tracking data, the average dwell time $\bar{d^w}$ based on the total number of occurrences of the word in the corpus is calculated as in Equation (3.1)

$$\bar{d^w} = \frac{1}{n}\sum_{i=1}^{n} d_i^w \tag{3.1}$$

where n is the number of occurrences of a word w in the data set and d^w is the dwell time for the word w. The average global dwell time $\bar{d}$ for the whole data set is computed as in Equation (3.2)

$$\bar{d} = \frac{1}{m}\sum_{i=1}^{m} d_i \tag{3.2}$$

where m is the number of all occurrences of all words observed in the data set and d_i is the dwell time for the occurrence i of a word in the data set. Each weight v for each word w is then calculated as the ratio in Equation (3.3)

$$v^w = \frac{\frac{1}{n}\sum_{i=1}^{n} d_i^w}{\frac{1}{m}\sum_{i=1}^{m} d_i} \tag{3.3}$$

and these values have been normalized using the min-max normalization. In case words were present in both corpora, the weight was defined as the average of the two normalized values.

Polarities are computed using a large data set of movie reviews including 50,000 texts, labeled as positive and negative reviews (Maas et al., 2011). To assess if a word has a positive or negative polarity, we compute a probability in the form of Equation (3.4):

$$P\left(w_{pos}\right) = \frac{N_{w_{pos}}}{N_w} \qquad P(w_{neg}) = \frac{N_{w_{neg}}}{N_w} \tag{3.4}$$

where $P(w_{pos})$ is the probability that the word w is positive, $N_{w_{pos}}$ is the number of occurrences of the word w in positive labeled texts and N_w is the number of occurrences of the word w. The same computation is made for negatives. Given the probabilities in Equation (3.4) we assign a polarity p to each word w as in Equation (3.5)

$$p^w = \begin{cases} 1 & if\ P(w_{pos}) > P(w_{neg}) \\ 0 & if\ P(w_{pos}) = P(w_{neg}) \\ -1 & if\ P(w_{pos}) < P(w_{neg}) \end{cases} \tag{3.5}$$

Therefore, we assign the word w a positive (+1) or negative value (-1) in case $P(w_{pos})$ is greater or lower than 0.5, respectively. If the probability is exactly 0.5 the word w is assigned 0 (neutral). For each word, a final value s is then computed as the product of weights and polarities as in Equation (3.6)

$$s^w = v^w \cdot p^w \tag{3.6}$$

3.2.3 Comparison with other lexicons

We compared the performance of the Eye dictionary in the classification of sentiment polarity with four dictionaries included in the Syuzhet R package (Jockers, 2015): nrc, afinn, bing and syuzhet. As these lexicons have been introduced in Section 2.5.2, only the main characteristics of the four lexicons will be reported here. The number of positive and negative words included in these dictionaries is shown in Figure 3.1.

The NRC dictionary was developed in 2010 to detect emotion in texts (Mohammad and Turney, 2013). Each word was manually annotated by different users on the Amazon Mechanical Turk crowdsourcing platform. The NRC lexicon implemented in the Syuzhet R package includes 5,636 words associated with a positive or negative sentiment. The afinn lexicon was initially developed using words retrieved from tweets related to the United Nation Climate Conference (COP15) in 2009 (Hansen et al., 2011). It was then extended to include words from the Original Balanced Affective Word List, Internet slang from the Urban Dictionary (including acronyms

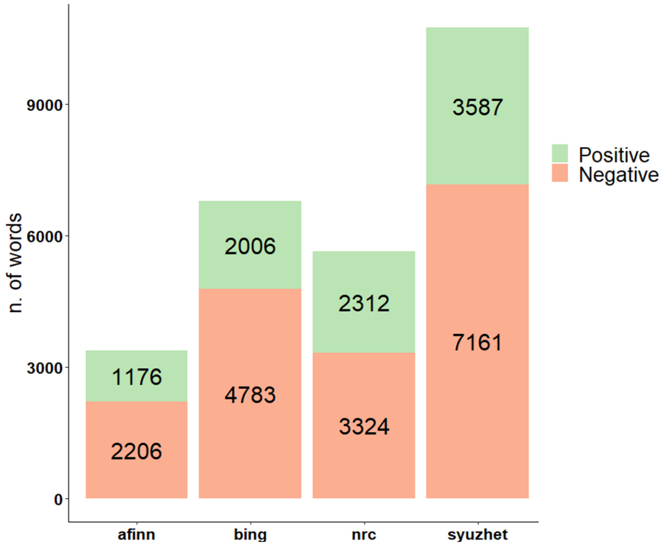


Figure 3.1: Number of positive and negative words included in the four dictionaries

such as WTF, LOL and ROFL) and the Compass DeRose Guide to Emotion Words. The afinn lexicon implemented in the Syuzhet R package includes 3,382 words scored manually by the author (1,176 positive and 2,206 negative). Values associated with these words range from –5 to +5. The bing lexicon was developed by Minqing Hu and Bing Liu as the Opinion Lexicon (Hu and Liu, 2004). The bing lexicon implemented in the Syuzhet package includes 6,789 words (2,006 positive and 4,783 negative). Words are assigned the values –1 if negative and +1 if positive. Finally, the Syuzhet lexicon was developed in the Nebraska Literary Lab (Jockers, 2015). This dictionary includes 10,748 words (3,587 positive and 7,161 negative) with values ranging from -1 to +1.

3.2.4 Description of the Eye dictionary

The Eye dictionary includes a total of 5,329 unique words for which weights and polarities were calculated as described in the previous sections. Of these, 1,113 words were present in the Provo Corpus, 4,814 in the GECO corpus and 598 in both corpora. Of the 5,329 words included in the Eye dictionary, 3,119 were assigned a positive polarity and 2,210 a negative polarity based on the automatic computation of polarity. Words included in the Eye dictionary were characterized based on parts of speech (POS) tagging (also known as grammatical tagging) using the latest version

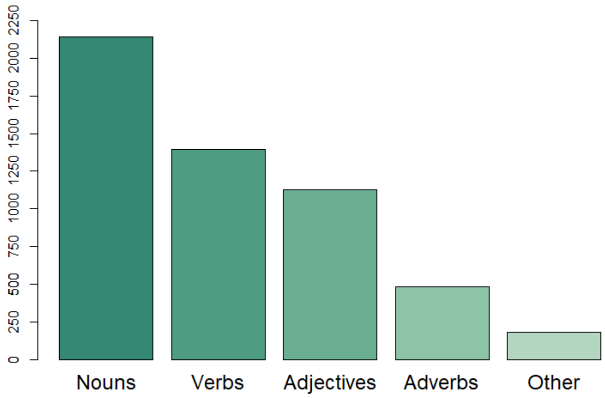


Figure 3.2: Parts of speech (POS) tagging of the 5,329 words included in the Eye dictionary

(C7) of the Constituent Likelihood Automatic Word-Tagging System (CLAWS). This tagger was first developed by the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University in 1987 (Garside, 1987) and has ben subsequently improved and developed (Garside and Smith, 1997). CLAWS shows 96-97% accuracy and its latest version was used to tag 100 million words of the British National Corpus (BNC). Using this system, words included in the Eye dictionary were divided into 10 classes based on their tags. Based on the assigned tags, we found that the Eye dictionary contains 1,128 adjectives, 484 adverbs, 2,142 nouns, 1,393 verbs and 182 words included in minor categories (25 conjunctions, 23 determiners, 11 interjections, 33 prepositions, 48 pronouns and 42 other) (Figure 3.2).

In order to obtain a version of the Eye dictionary excluding stopwords, we excluded from the 5,329 unique words a list of 124 stopwords defined based on the stopwords R package (`https://github.com/quanteda/stopwords`), leading to a total of 5,205 unique words for the Eye dictionary without stopwords (ws) version.

Finally, we developed an alternative version of the Eye dictionary in which the polarity of words was manually revised by a single rater. For words rated as positive and negative the manual classification was retained, while for words rated as neutral the polarity originally computed as in the first version of the dictionary was retained. This version of the Eye dictionary included 2,950 words with a positive polarity and 2,379 with a negative polarity.

3.3 Application: Using the Eye dictionary for the classification of reviews and comparison with other dictionaries

First, we conducted a qualitative analysis aimed at verifying whether texts rated with the Eye dictionary show similar scores compared with other dictionaries. In order to do this, we downloaded all tweets posted between January 1 2022 and February 18 2022 containing the keyword "Chinese New Year" in the English language. We chose this topic in order to be able to show a growing positive sentiment towards the date of Chinese New Year 2022 (February 1). We downloaded a total of 185,335 tweets that were preprocessed as described in Section 2.5.3. As shown in Figure 3.3, we observed high concordance in the evolution of the daily standardized average sentiment scores between the Eye Dictionary and the other four dictionaries.

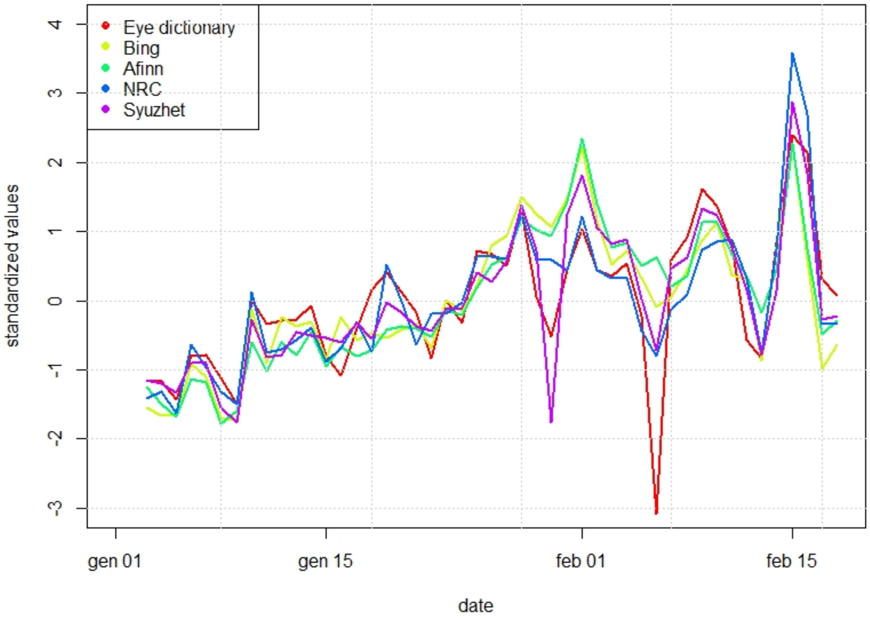


Figure 3.3: Evolution of sentiment of tweets on Chinese New Year 2022

Next, we assessed the performance of the dictionary based on eye tracking data in the classification of sentiment polarity using a collection of labeled texts consisting in 1,000 consumer reviews (500 positive and 500 negative) from Yelp (Kotzias et al., 2015) and compared the performance with that of the other four dictionaries previously described. For each of the 1,000 texts, the overall sentiment is defined based on the algebraic sum of signed values assigned to each word by a dictionary. Specifically, a text was classified as positive in case the algebraic sum of word values was greater than zero, negative if the algebraic sum was less than zero and neutral in case it was exactly zero.

Table 3.1 shows the performance of the Eye dictionary and the four other dictionaries in terms of precision, recall, F1-score and accuracy.

Table 3.1: Comparison between Eye dictionary and four other dictionaries

Metric	**Eye dict.**	**Eye dict.(ws)**	**bing**	**afinn**	**nrc**	**syuzhet**
Precision positive	0.71	0.62	0.58	0.56	0.45	0.61
Precision negative	0.60	0.62	0.67	0.64	0.40	0.76
Recall positive	0.46	0.63	0.80	0.80	0.59	0.86
Recall negative	0.82	0.61	0.42	0.37	0.27	0.45
F1-score positive	0.56	0.62	0.67	0.66	0.51	0.71
F1-score negative	0.69	0.62	0.52	0.47	0.32	0.57
Accuracy	0.64	0.62	0.61	0.58	0.43	0.66

Abbreviations: Eye dict., Eye dictionary; ws, without stopwords

The best performance in the classification of positive texts as regards to precision was obtained by the Eye dictionary (0.71) followed by syuzhet (0.61), while a lower performance was shown by the Eye dictionary as regards to precision in the classification of negative texts. The Eye dictionary showed the best performance based on recall and F1 score for negative texts, while Syuzhet accomplished the best results for these metrics in the classification of positive texts. The Eye dictionary also showed the second-best accuracy after the syuzhet dictionary. After removal of stopwords, the Eye dictionary showed improved performance in the recall and F1-score of positive texts but precision for positive texts, recall and F1-score for negative texts and accuracy worsened.

Next, we evaluated the performance of the version of the Eye dictionary in which we integrated manual rating of polarity (for negative and positive words) and automatic definition of polarity (for neutral words), based on the assumption that the polarity of neutral words might be more affected by a subjective rating and could therefore benefit of an automatic definition according to the recurrence of these words in a large data set of previously labelled texts. The performance of the version of the Eye dictionary with the revised version of polarity is shown in Table 3.2.

Table 3.2: Performance of the Eye dictionary with revised rating of polarity

Metric	**Eye dictionary**	**bing**	**afinn**	**nrc**	**syuzhet**
Precision positive	0.68	0.58	0.56	0.45	0.61
Precision negative	0.66	0.67	0.64	0.40	0.76
Recall positive	0.65	0.80	0.80	0.59	0.86
Recall negative	0.69	0.42	0.37	0.27	0.45
F1-score positive	0.66	0.67	0.66	0.51	0.71
F1-score negative	0.68	0.52	0.47	0.32	0.57
Accuracy	0.67	0.61	0.58	0.43	0.66

The change in the classification of polarity led to a general improvement of the performance based on all metrics except precision positive, which showed a slight decrease from 0.71 to 0.68. The Eye dictionary still achieved the best performance based on this metric as well as on recall and F1 score for negative texts. In addition, the Eye dictionary achieved the best accuracy.

Analysis of errors

We examined the false-positive and false-negative errors for the Eye dictionary based on the performance shown in the Yelp data set of consumer reviews. We identified the following main sources of errors:

- Negations

Examples: *Honestly it didn't taste that fresh*

- Indirect expressions of satisfaction or dissatisfaction.

Example: *The service here is fair at best*

- Idiomatic expressions, i.e. an expression, word, or phrase that has a figurative meaning different from the literal meaning of the idiom's individual elements

Example: *Omelets are to die for*

Presence of mixed emotions

Example: *The place was fairly clean, but the food simply wasn't worth it*

Comparison with other products

Example: *I've had better bagels from the grocery store*

To conclude, we have described a new sentiment analysis dictionary, called the Eye dictionary, built by leveraging eye tracking data to compute weights, based on the hypothesis that dwell time can serve as a measure of relevance of a word (i.e. its ability to gain attention from a reader). The performance of this dictionary in the classification of different types of texts has been compared with four existing and widely used dictionaries. Notably, the Eye dictionary was able to achieve a performance similar or better compared to most of the other dictionaries even if it includes a much lower number of words. Specifically, the Eye dictionary showed a better accuracy compared with the syuzhet dictionary, which includes more than twice the number of words (10,748 vs 5,329 words). We can therefore hypothesize that increasing the number of words of the Eye dictionary through collection of additional eye-tracking data might allow to further improve the performance. Future developments will therefore include the expansion of the number of words included in the dictionary as well as definition of rules to handle cases in which the classification is particularly challenging, such as sentences including negations, amplifiers, downtoners and multi-word or idiomatic expressions.

Chapter 4

Side project

4.1 Education using games and online resources

Mobile applications are novel and useful tools that can be used to spread education and increase engagement in a topic. They can be used in a variety of settings, thanks to the wide diffusion of smartphones, tablets, smartwatches and similar devices. Among the most appreciated features of mobile applications, we can list the fact that their contents can usually be viewed in a fast way (for instances a few minutes per day) and they are therefore able to engage the user in the long-term. In the case of applications focused on medicine, based on the contents they offer, these applications can be targeted towards different type of users such as patients, citizens, health care professionals or students. For instance, mobile applications can be used to convey information to patients as well as to enhance adherence during treatment with different medications such as antiretroviral therapy (Horvath et al., 2019), immunosuppressive medication (Levine et al., 2019), antipsychotics (Pozza et al., 2020), vitamin D (Goodman et al., 2016) or smoking cessation (Chu et al. 2019; Pifarre et al. 2017). In the field of pharmacology, mobile applications are currently being tested in pharmacovigilance as a tool to report suspected adverse drug reactions and/or share safety information (Ahn et al. 2019; Egbring et al. 2016; Pierce et al. 2019).

Mobile applications can also implement games, which often represent additional tools to increase the motivation of the users and let them actively learn concepts or acquire skills. To this regard, gamification can be defined as the application of elements typical of game playing to nongaming activities. Being able to increase interactivity, rewards and motivation (Sera and Wheeler, 2017), gamification has been increasingly used in several fields, such as education, health and sport (Dell and Chudow 2019; Jones and Wisniewski 2019; Lam et al. 2019; Shannon 2019; Lopez and Tucker 2018). Mobile applications can complement and extend traditional teaching methods and tools, allowing students to improve their preparation in a fun way and even when they are not in class (for instance during travelling). In addition, mobile applications or games can include contents which generally do not find enough space in course programs due to paucity of time. Finally, mobile applications and games can increase engagement towards a specific field, thus potentially allowing the student to consider it for higher-level studies or as a professional career.

4.2 Literature review of studies using mobile apps to improve education in medicine and pharmacology

Mobile applications are increasingly being used to convey information to healthcare students / professionals in a simple and fast way. One of the most recent examples during the recent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) outbreak, is represented by a mobile health platform developed to disseminate up-to-date and validated information to the medical staff of the Children's Hospital at the University Hospitals of Geneva (Zamberg et al., 2020).

Medicine, and specifically pharmacology, are among the fields in which mobile applications and games might be extremely useful due to the fact that they might be able to engage the user and covey concepts that might otherwise appear to be too difficult or complex. These approaches might be applied in the classroom settings, in special contexts such as laboratories, or be available for the students to apply at home or during travelling. In this sense, recent studies described the use of interactive trivia (Jones and Wisniewski, 2019) and quest games (Lam et al., 2019) in the classroom setting for first year pharmacy students. In another recent study, Ameri and colleagues showed that LabSafety, an educational application aimed at teaching students about safety measures in laboratories, was able to increase the knowledge and engagement of pharmacy students (Ameri et al., 2020). Mobile applications might be particularly useful in fields in which a number of studies is suggesting that the knowledge gained during university courses should be expanded, as in the case of pharmacogenetics (Just et al. 2017; Pisanu et al. 2014). This discipline is aimed at identifying the contribution of genetic variants to interindividual variability in the efficacy and safety of drugs (Hockings et al., 2020).

While pharmacogenetics has become part of the clinical decision making in some medical fields, such as oncology (Al-Mahayri et al. 2020; Chan et al. 2020; Joshi et al.

2019; Morganti et al. 2019; Wang et al. 2020) and cardiology (Davila-Fajardo et al. 2019; Zhu et al. 2020), in other settings, such as in psychiatry, it is more utilized as a research and discovery tool (Amare et al. 2017; Koromina et al. 2020; van Westrhenen et al. 2020). Among the challenges that limit the application of pharmacogenetics in clinical practice, one of the most relevant is the limited knowledge of a substantial part of physicians and pharmacists regarding the interpretation of pharmacogenetics tests (Just et al. 2017; Karuna et al. 2020; Kim et al. 2020; Petit et al. 2020; Pisanu et al. 2014). The pharmacogenetics educational background shows substantial differences among states and even among different regions in the same state (Pisanu et al., 2014). While it will be important to try to harmonize pharmacogenetics education among training programs (Just et al., 2017), alternative tools such as mobile applications might also help to increase learning and engagement in this topic. Dodson and Baker (2020) evaluated the perceptions on a mobile clinical decision support tool on pharmacogenetics among nurse practitioners and students. The application provided a case study on a patient requiring pharmacogenetic testing and illustrated clinical evidence-based guidelines (Dodson and Baker, 2020). While this prototype is still under development, it represents a good example of how mobile applications can be used to increase engagement in pharmacogenetics using a user-friendly mechanism.

As nowadays several providers directly propose pharmacogenetic testing to physicians or, in some cases, directly to patients, physicians need to be able to evaluate quality and clinical utility of different pharmacogenetics tests, as well as to provide correct information on their interpretation. The Food and Drug Administration (FDA) recently issued a safety communication warning against the use of unapproved pharmacogenetic testing being marketed directly to patients or offered through health care providers (FDA Safety Communication, 2018). In this scenario, the development of additional educational tools, such as mobile platforms and applications, appears to be of great importance to complement and integrate knowledge provided by university courses, in order for physicians and other health care providers to be able to critically evaluate scientific evidence supporting different pharmacogenetic test as well as their clinical impact.

Existing mobile applications dedicated to pharmacology mostly provide lists of concepts or flash cards. While some applications provide trivia, it is often not possible to choose the topic to exercise on. In addition, specialized topics such as pharmacogenetics, as well as latest research findings, are not included. We developed PharmacoloGenius, the first pharmacology mobile application designed to freely offer a curated list of resources and games aimed at increasing students' engagement in basic concepts of pharmacology as well as research and clinical applications (Zammarchi et al., 2020). This free Android application is organized in different sections (study resources for pharmacology, games and news) to provide the user with both resources to increase knowledge on theoretical topics as well as novel research findings and opportunities for researchers in pharmacology.

4.3 Application: the PharmacoloGenius mobile app

4.3.1 Development of the application and debugging

The application was developed in Java, using the multi-language and multi-platform Android Studio integrated development environment (IDE) and designed in accordance with the target application program interface (API) level requirements for an application to be available on Google Play Store [minimum software development kit (SDK) version 21, corresponding to Android Lollipop, minimum SDK target 28 corresponding to Android Pie]. Internet access is required if the user wish to be able to visualize real-time updates of the News section. No access to camera, contacts, location or storage is required.

To ensure compatibility with different Android devices, testing and debugging was conducted using both real Android smartphones and tablets, as well as different emulated Android Virtual Devices. A testing session of the pilot version of the application was conducted during the "Innovation in Pharmacology Education" workshop held at Pharmacology 2019. Testers included undergraduate students, PhD students, researchers and full-time pharmacology professors who were able to provide comments and suggestions on how to improve available sections as well as which additional contents might be more useful. The PharmacoloGenius Android application can be downloaded from Google Play Store (PharmacoloGenius download link).

4.3.2 Sections of the PharmacoloGenius mobile application

Three main sections were developed: Pharmacology Study Resources, Games and News (Figure 4.1).

Pharmacology Study Resources

The first section offers a curated list of resources pertaining to Basic Pharmacology, Applied Pharmacology and Pharmacogenetics such as online courses, video and additional educational tools (Figure 4.1 b). Target users include: 1) pre- or post-graduate students; 2) healthcare professionals desiring to expand their knowledge in specific fields of pharmacology via online educational tools; and 3) educators that

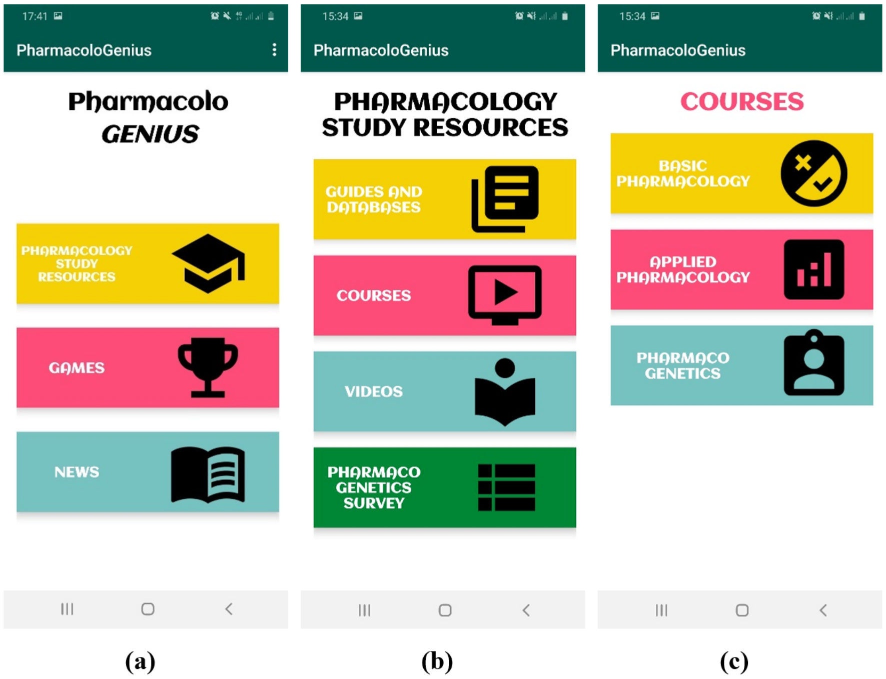
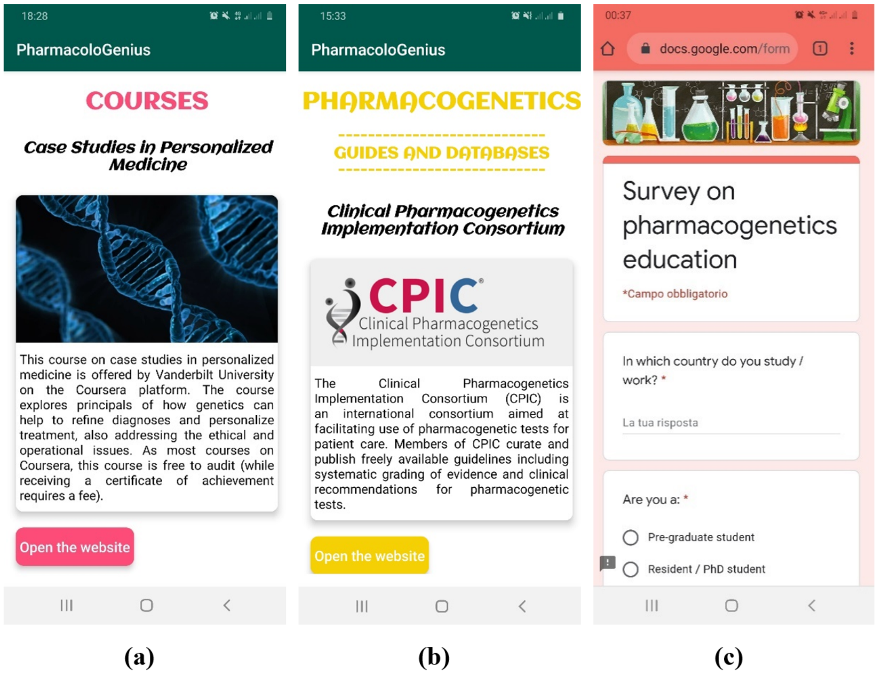


Figure 4.1: Graphical user interface of the PharmacoloGenius app. (a) Frist screen of the app; (b) Pharmacology study resource screen; (c) Courses screen

need to retrieve tools that can be complementary to traditional materials used during lessons.

Selected resources pertain to three main topics: Basic Pharmacology, Applied Pharmacology and Pharmacogenetics (Figure 4.1 c). Basic and Applied pharmacology resources included online guides, courses and videos (Figure 4.2 a). Pharmacogenetics resources include links to evidence-based guidelines and a list of databases that can be used to evaluate frequency of genetic variants in different populations or assess the functional and clinical impact of genetic variants (Figure 4.2 b). Criteria followed to select these resources included: 1) being freely available; 2) available in English language; and 3) developed or endorsed by scientific societies, universities, international research initiative or other reliable sources.

Finally, this section includes a link to an anonymous Pharmacogenetics Survey (Figure 4.2 c) that was designed to collect information regarding the level of confidence of users in pharmacogenetics, assess whether this topic was covered during their graduate course and collect users' opinions regarding which specific topics they would need the most to be represented on educational tools.



COURSES

Case Studies in Personalized Medicine

This course on case studies in personalized medicine is offered by Vanderbilt University on the Coursera platform. The course explores principals of how genetics can help to refine diagnoses and personalize treatment, also addressing the ethical and operational issues. As most courses on Coursera, this course is free to audit (while receiving a certificate of achievement requires a fee).

PHARMACOGENETICS

GUIDES AND DATABASES

Clinical Pharmacogenetics Implementation Consortium

The Clinical Pharmacogenetics Implementation Consortium (CPIC) is an international consortium aimed at facilitating use of pharmacogenetic tests for patient care. Members of CPIC curate and publish freely available guidelines including systematic grading of evidence and clinical recommendations for pharmacogenetic tests.

Survey on pharmacogenetics education

*Campo obbligatorio

In which country do you study / work? *

La tua risposta

Are you a: *

○ Pre-graduate student

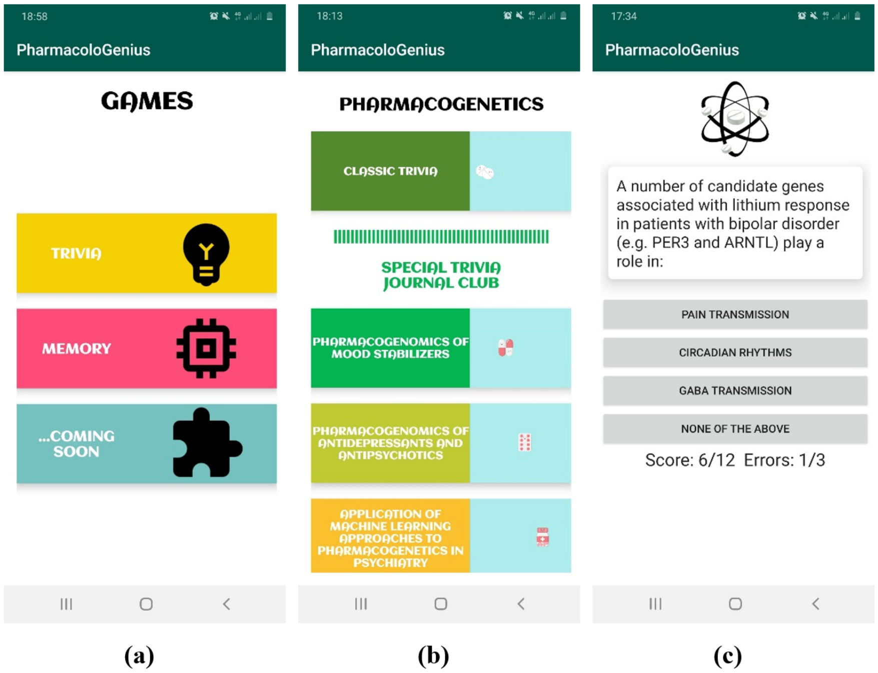
○ Resident / PhD student

Figure 4.2: Graphical user interface of the PharmacoloGenius app. (a) Pharmacogenetics courses screen; (b) Pharmacogenetics databases screen; (c) Survey on pharmacogenetics education

Games

The second section includes a memory game and a series of original multiple-choice trivia (Figure 4.3 a). The memory game requires coupling pharmacology-related pictures with a number of moves as small as possible. This game is available in three versions characterized by different difficulty levels (4x3, 4x4 or 5x4). The user wins the match in case he/she is able to provide the correct answer to a final question on pharmacological topics.

Trivia include a series of questions developed using different sources as references, such as The Concise Guide to Pharmacology 2019/2020 (Alexander et al., 2019), the IUPHAR/BPS Guide to Pharmacology database (www.guidetopharmacology.org) as well as textbooks such as Goodman and Gilman's The Pharmacological Basis of Therapeutics 13th Edition (Brunton et al., 2018). Undergraduate students represent the main target for trivia, which are focused on different theoretical concepts such as pharmacokinetics, pharmacodynamics or pharmacogenetics. At the start of each game, ten questions are randomly extracted from a larger set and presented to the



A number of candidate genes associated with lithium response in patients with bipolar disorder (e.g. PER3 and ARNTL) play a role in:

Score: 6/12 Errors: 1/3

Figure 4.3: Graphical user interface of the PharmacoloGenius app. (a) Game screen; (b) Pharmacogenetics trivia screen; (c) Example of a game session

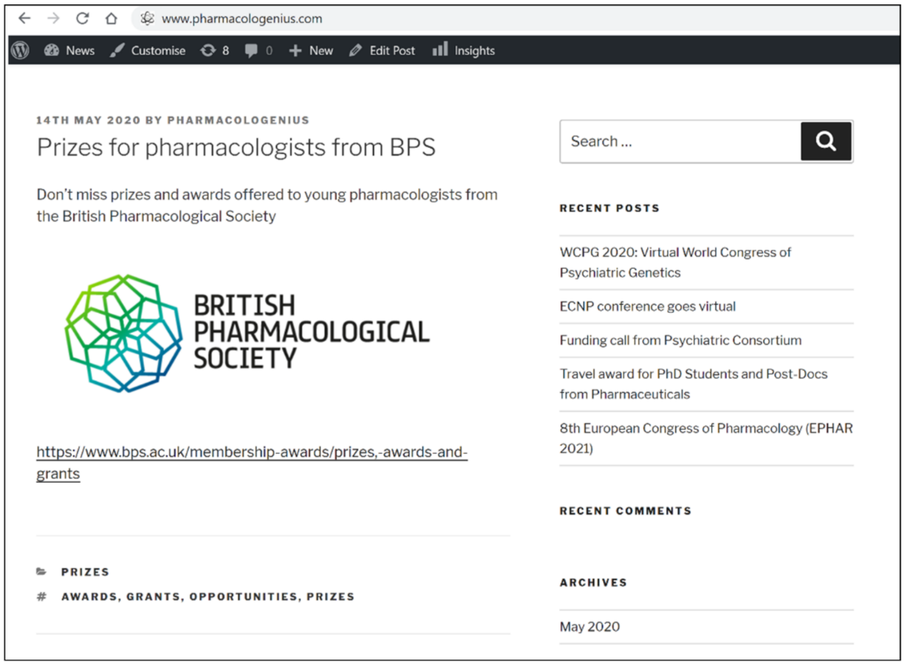
user. Besides trivia on basic pharmacological concepts, for the topics "Pharmacogenetics" a second database of special "Journal club" trivia was developed based on research articles of particular interest (Figure 4.3). These trivia are targeted towards early career researchers or healthcare professionals who desire to increase knowledge on the state-of-the-art of specific relevant research or clinical applications.

These articles were selected according to the following criteria: 1) published in peer-reviewed journals; 2) published in journals listed on PubMed; 3) being open access; 4) relevant in the field; and 5) published in the last five years.
Before the trivia is started, the user can open and read the article on which the trivia is based through the link provided on the screen. The special trivia are currently focused on pharmacogenetics of different psychotropic drugs such as mood stabilizers, antipsychotics and antidepressants, as well as applications of machine learning methods to this field. Whenever the user provides a wrong answer, a short paragraph is shown to explain which specific textbook or internet resource / article the user can check to find the correct answer and revise that topic. The user gains points for each correct answer and visualizes the number of errors and correct answers. Additional

trivia on other pharmacological topics are currently under development and will be regularly added to the application.

News

Finally, the News section offers timely updates regarding opportunities for students and researchers working in the pharmacology field (Figure 4.4). It is developed as a self-hosted blog implementing the open source WordPress platform (version 5.4) and can be accessed via any web browser on either mobile devices or personal computers at www.pharmacologenius.com.



14TH MAY 2020 BY PHARMACOLOGENIUS

Prizes for pharmacologists from BPS

Don't miss prizes and awards offered to young pharmacologists from the British Pharmacological Society

BRITISH PHARMACOLOGICAL SOCIETY

https://www.bps.ac.uk/membership-awards/prizes,-awards-and-grants

PRIZES

AWARDS, GRANTS, OPPORTUNITIES, PRIZES

RECENT POSTS

- WCPG 2020: Virtual World Congress of Psychiatric Genetics
- ECNP conference goes virtual
- Funding call from Psychiatric Consortium
- Travel award for PhD Students and Post-Docs from Pharmaceuticals
- 8th European Congress of Pharmacology (EPHAR 2021)

RECENT COMMENTS

ARCHIVES

May 2020

Figure 4.4: Graphical user interface of the PharmacoloGenius News section, providing updates on funding calls, bursaries, conferences, relevant research articles and other opportunities for pharmacologists

The News section provides frequent updates on opportunities for researchers in pharmacology such as travel grants, conferences, bursaries and funding calls. Short posts on different types of opportunities and updates are organized in different categories, in order for the user to be able to quickly visualize the type of posts of interest (e.g. Conferences, Grants, Prizes, Travel Awards, News and so on). The user can also search posts of interest based on tags, which are used to group news with similar contents across different categories.

PharmacoloGenius offers a wide range of resources, including contents and news, to pharmacology students or health care professionals. While social networks are currently more used than mobile applications to convey information to students or researchers, mobile applications can offer some important advantages as they can provide a less distracting environment as well as more carefully curated and reliable information. Limitations of the current version of PharmacoloGenius include the relatively low number of pharmacological topics covered by trivia to date, although this section is currently being expanded. Moreover, additional testing sessions with students and healthcare professionals will be carried out, in order to further improve the quality of contents and their presentation.

In conclusion, PharmacoloGenius represents a useful resource to increase knowledge and engagement on pharmacology concepts, as well as to keep updated on clinical and research applications and opportunities.

Appendix

Table A.1: Research strings used for bibliometrics analysis in Section 2.4

Search strategy for finance related articles:
TITLE-ABS-KEY(“sentiment analysis” AND (“finance” OR “financial” OR “stock market*” OR “bank*” OR “economics” OR “financial forecasting” OR “corporate finance”)) AND (LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “ar”))
Search strategy for politics related articles:
TITLE-ABS-KEY (“sentiment analysis” AND (“politic*” OR “elect*” OR “campaign” OR “vote*”)) AND (LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “ar”))
Search strategy for events related articles:
TITLE-ABS-KEY (“sentiment analysis” AND (“terroris*” OR “outbreak*” OR “pandemy” OR “attack*” OR “crisis” OR “migra*” OR “infection”)) AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “cp”))

Table A.2: Number of words in the six lexicons used in the analysis

Lexicon	**N. of words**	**N. of positive words**	**N. of negative words**
Syuzhet	10,748	3,587	7,161
Afinn	3,382	1,176	2,206
Bing	6,789	2,006	4,783
Nrc[1]	5,636	2,312	3,324
Jockers-Rinker[2]	11,738	3,581	7,157
VADER[3]	7,520	3,348	4,172

[1] Nrc lexicon has 13,901 words but since a word can appear more than once, the table only shows words classified as positive or negative (for details see Table A.3)

[2] The Jockers-Rinker lexicon is the default sentiment used in the “sentimentR” package and is a modified version of the syuzhet lexicon.

[3] The VADER lexicon has been retrieved from the GitHub page of the author (`https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt`). Last access: 19/07/2021

Table A.3: Frequencies of words in the “nrc” lexicon for each sentiment and emotion

Category	**N. of words**
Anger	1,247
Anticipation	839
Disgust	1,058
Fear	1,476
Joy	689
Sadness	1,191
Surprise	534
Trust	1,231
Positive	2,312
Negative	3,324

Note: words can be included in more than one category

Table A.4: Examples of words in the “Afinn” lexicon

Word	**Value**
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
abductions	-2
...	...
youthful	2
yucky	-2
yummy	3
zealot	-2
zealots	-2
zealous	2

Table A.5: Examples of words in the “Bing” lexicon

Word	**Value**
a+	1
abound	1
abounds	1
abundance	1
abundant	1
accessable	1
. . .	. . .
zapped	-1
zaps	-1
zealot	-1
zealous	-1
zealously	-1
zombie	-1

Table A.6: Examples of words in the “nrc” lexicon

Language	**Word**	**Sentiment**	**Value**
english	abba	positive	1
english	ability	positive	1
english	abovementioned	positive	1
english	absolute	positive	1
english	absolution	positive	1
english	absorbed	positive	1
. . .	. . .	. . .	. . .
english	worthy	trust	1
english	wot	trust	1
english	yearning	trust	1
english	zeal	trust	1
english	zealous	trust	1
english	zest	trust	1

Table A.7: Examples of words in the “Syuzhet” lexicon

Word	**Value**
abandon	-0.75
abandoned	-0.50
abandoner	-0.25
abandonment	-0.25
abandons	-1.00
abducted	-1.00
...	...
zenith	0.40
zest	0.50
zombie	-0.25
zombies	-0.25
false	-0.60
true	0.50

Table A.8: Examples of words in the “SentimentR” lexicon

Word	**Value**
a plus	1.00
abandon	-0.75
abandoned	-0.50
abandoner	-0.25
abandonment	-0.25
abandons	-1.00
...	...
zealously	-1.00
zenith	0.40
zest	0.50
zippy	1.00
zombie	-0.25
zombies	-0.25

Table A.9: Examples of words in the “Vader” lexicon

Word	**Value**
abandon	-1.9
abandoned	-2.0
abandoner	-1.9
abandoners	-1.9
abandoning	-1.6
abandonment	-2.4
...	...
youthful	1.3
yucky	-1.8
yummy	2.4
zealot	-1.9
zealots	-0.8
zealous	0.5

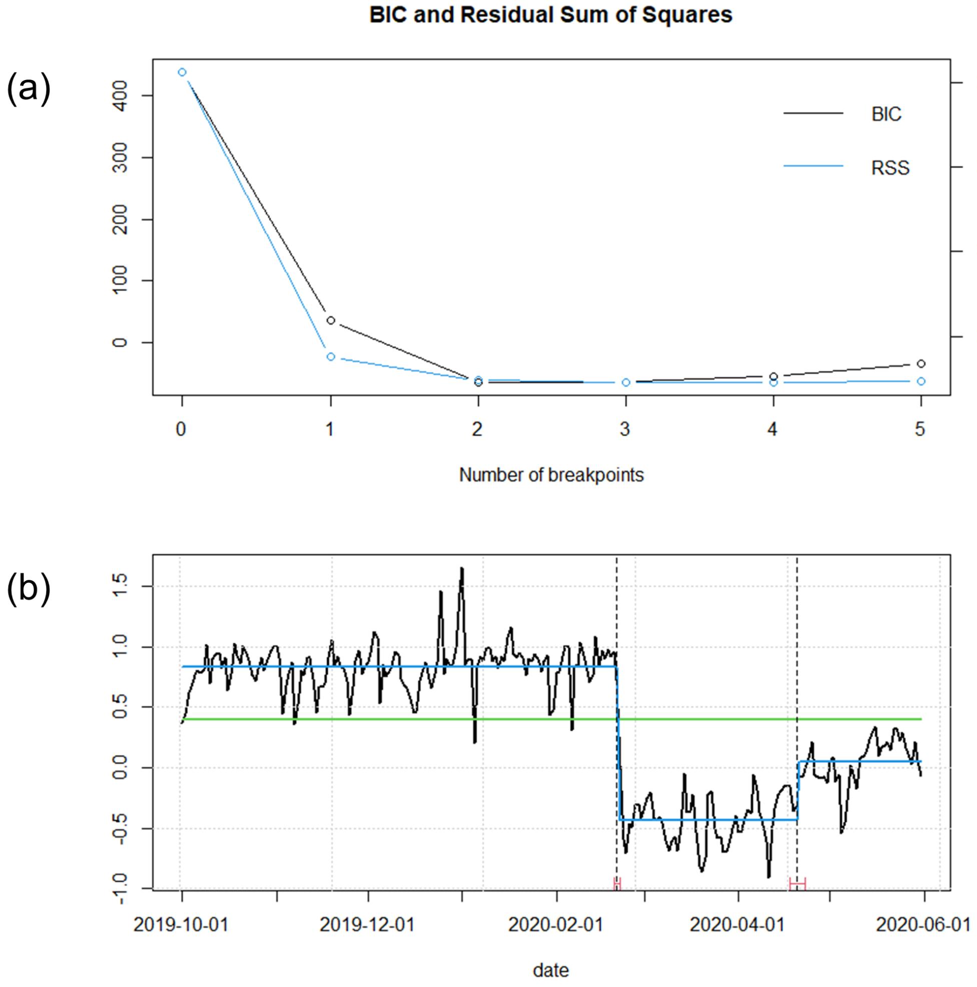


Figure A.1: (a) BIC and Residual Sum of Squares using the afinn lexicon, (b) Breakpoints in sentiment score using the afinn lexicon

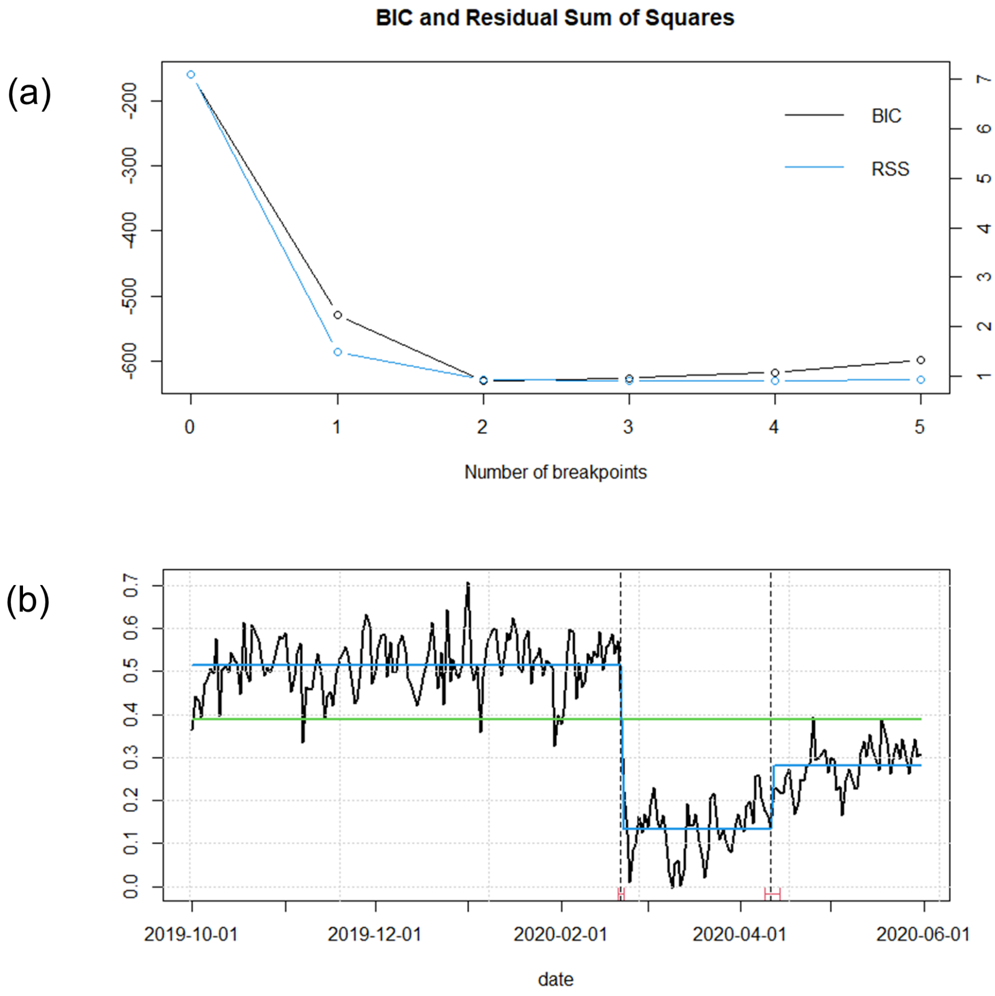


Figure A.2: (a) BIC and Residual Sum of Squares using the nrc lexicon, (b) Breakpoints in sentiment score using the nrc lexicon

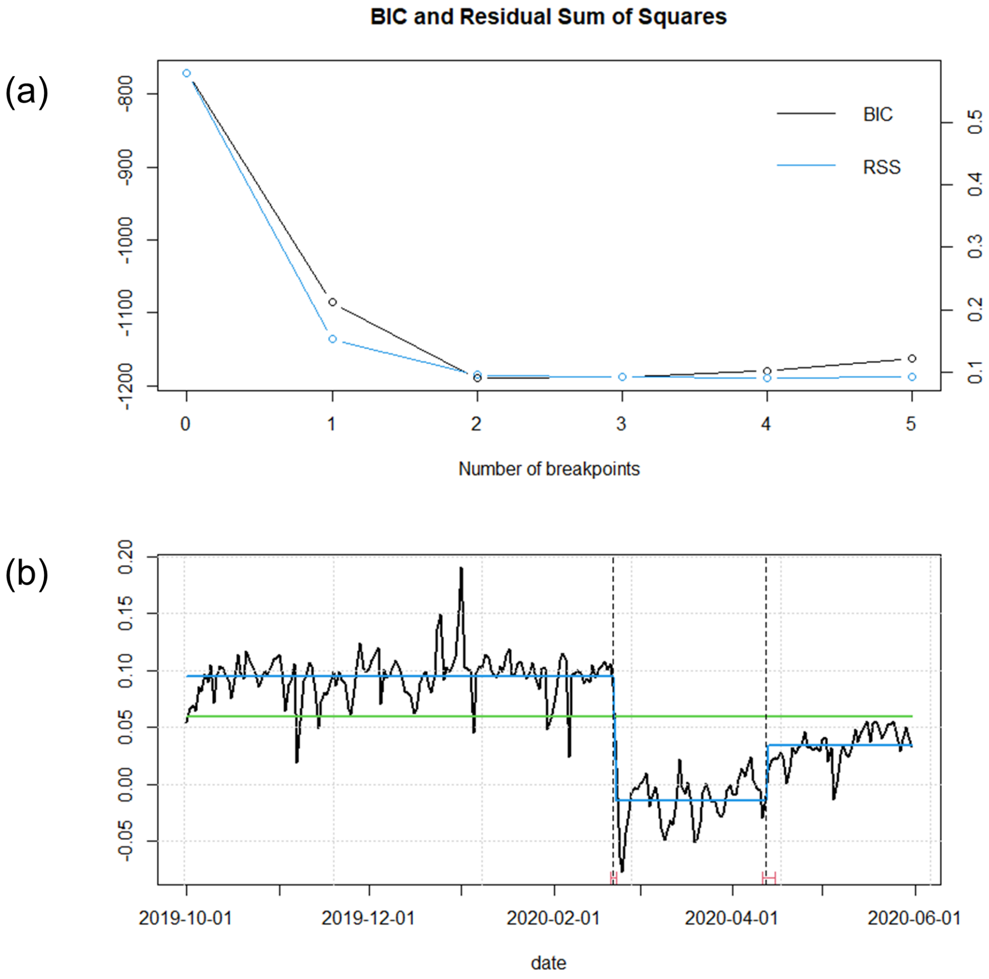


Figure A.3: (a) BIC and Residual Sum of Squares using the syuzhet lexicon, (b) Breakpoints in sentiment score using the syuzhet lexicon

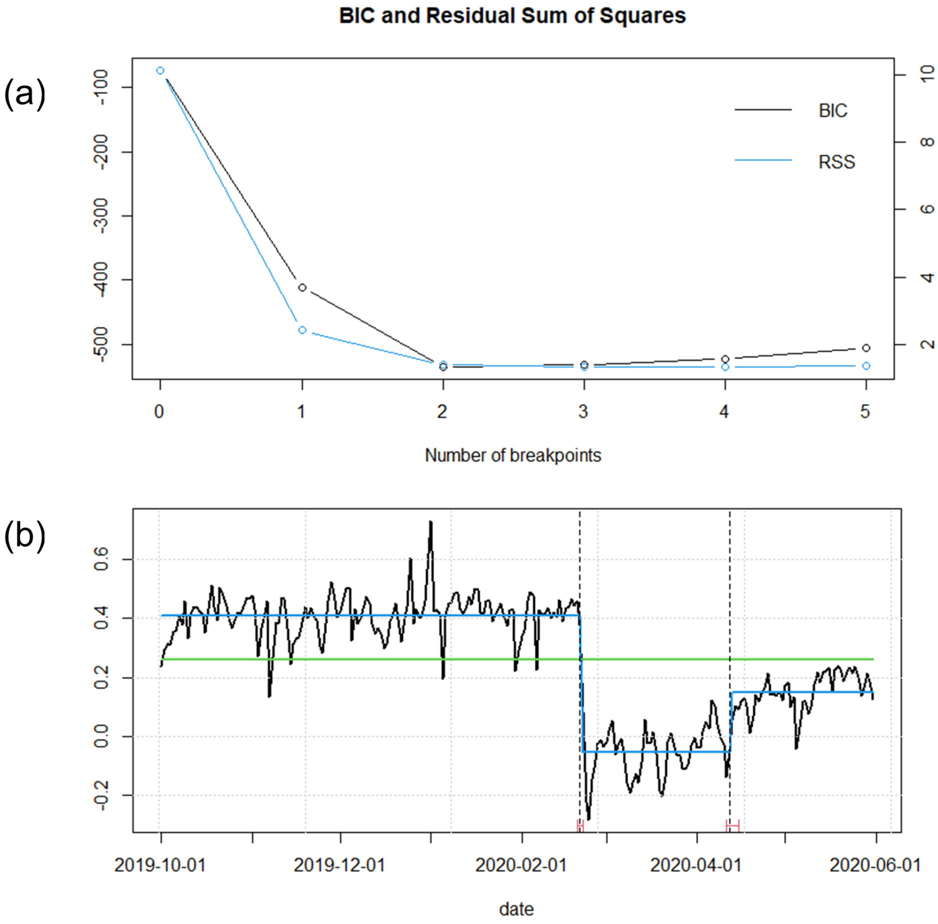


Figure A.4: (a) BIC and Residual Sum of Squares using the sentimentR lexicon, (b) Breakpoints in sentiment score using the sentimentR lexicon

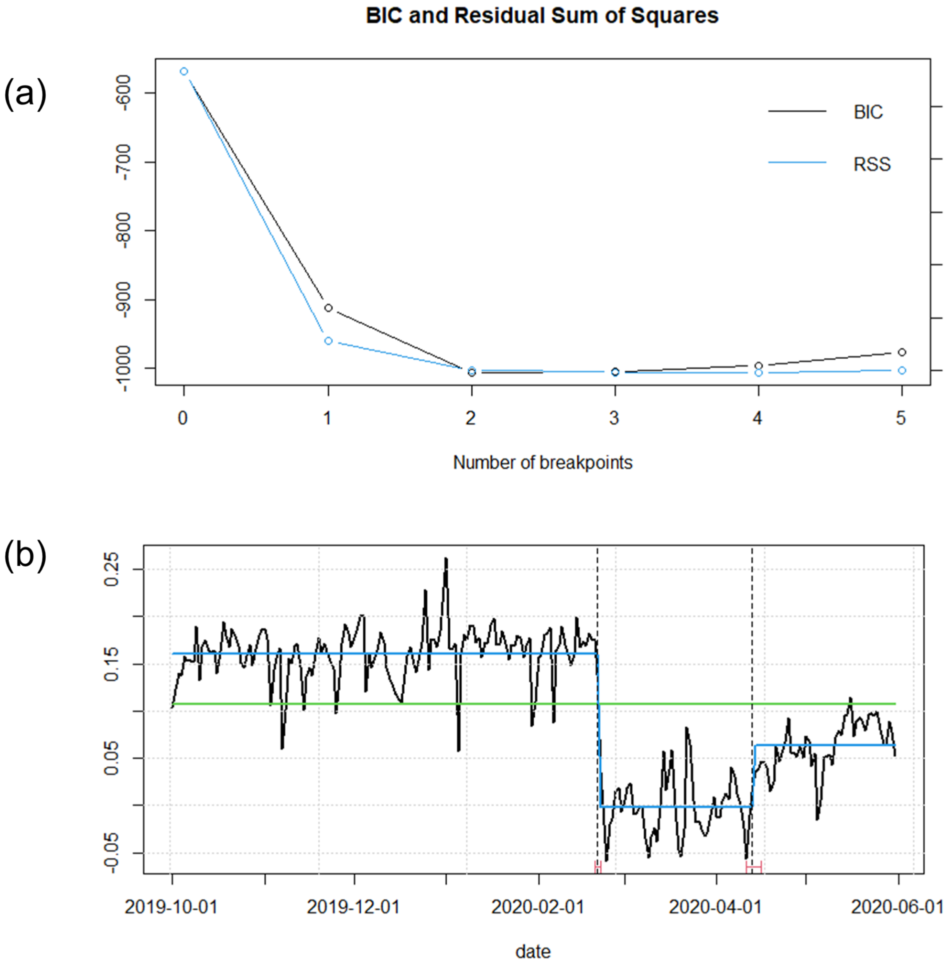


Figure A.5: (a) BIC and Residual Sum of Squares using the vader lexicon, (b) Breakpoints in sentiment score using the vader lexicon

Bibliography

S. H. Ahn, J. Zhiang, H. Kim, S. Chang, J. Shin, M. Kim, Y. Lee, J. H. Lee, and Y. R. Park. Postvaccination fever response rates in children derived using the fever coach mobile app: A retrospective observational study. *JMIR Mhealth Uhealth*, 7 (4):e12223, 2019.

Z. N. Al-Mahayri, G. P. Patrinos, and B. R. Ali. Toxicity and pharmacogenomic biomarkers in breast cancer chemotherapy. *Front Pharmacol*, 11:445, 2020.

S. P. H. Alexander, E. Kelly, A. Mathie, J. A. Peters, E. L. Veale, J. F. Armstrong, E. Faccenda, D. H. Simon, A. J. Pawson, J. L. Sharman, C. Southan, J. A. Davies, and C. Collaborators. The concise guide to pharmacology 2019/20: Introduction and other protein targets. *Br J Pharmacol*, 176(Suppl 1):S1–S20, 2019.

A. T. Amare, K. O. Schubert, and B. T. Baune. Pharmacogenomics in the treatment of mood disorders: Strategies and opportunities for personalized psychiatry. *EPMA J*, 8(3):211–227, 2017.

A. Ameri, R. Khajouei, A. Ameri, and Y. Jahani. Labsafety, the pharmaceutical laboratory android application, for improving the knowledge of pharmacy students. *Biochem Mol Biol Educ;*, 48(1):44–53, 2020.

C. Andrà, P. Lindstrom, F. Arzarello, K. Holmqvist, O. Robutti, and C. Sabena. Reading mathematics representations: an eye tracking study. *International Journal of Science and Mathematics Education;*, 13:237–259, 2015.

L. Antunez, L. Vidal, A. Sapolinski, A. Giménez, A. Maiche, and G. Ares. How do design features influence consumer attention when looking for nutritional information on food labels? results from an eye tracking study on pan bread labels. *International Journal of Food Sciences and Nutrition;*, 64:515–527, 2013.

M. Aria and C. Cuccurullo. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4):959–975, 2017.

M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha. Detecting jihadist messages on twitter. *In: European Intelligence and Security Informatics Conference. IEEE*, pages 161–164, 2015. doi: 10.1109/EISIC.2015.27.

O. Augereau, K. Kunze, H. Fujiyoshi, and K. Kise. Estimation of english skill with a mobile eye tracker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. , Germany*, 2016.

M. P. Bach. Usage of social neuroscience in e-commerce research - current research and future opportunities. *Journal of Theoretical and Applied Electronic Commerce Research*, 13(1):I–IX, 2018.

G. Backfried and G. Shalunts. Sentiment analysis of media in german on the refugee crisis in europe. In *Paloma Díaz*, pages 234–241. Narjès Bellamine Ben Saoud, Julie Dugdale and Chihab Hanachi, eds, Information Systems for Crisis Response and Management in Mediterranean Countries. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-47093-1.

R. Bar-Heim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein. Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)*, 2011.

J. Beltràn, M. S. Garcìa-Vàzquez, J. Benois-Pineau, L. M. Gutierrez-Robledo, and J. F. Dartigues. Computational techniques for eye movements analysis towards supporting early diagnosis of alzheimer's disease: A review. *Comput Math Methods Med*, page 2676409, 2018.

J. P. Benway and D. M. Lane. *Banner Blindness: Web Searchers Often Miss 'Obvious' Links*. Internet Technical Group, 1998.

L. Bix, R. P. Sundar, N. M. Bello, C. Peltier, L. J. Weatherspoon, and M. W. Becker. To see or not to see: Do front of pack nutrition labels affect attention to overall nutrition information? *PLoS One*, 10:e0139732, 2015.

M. H. Black, N. T. M. Chen, K. K. Iyer, O. V. Lipp, S. Bolte, M. Falkmer, T. Tan, and S. Girdler. Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography. *Neuroscience & Biobehavioral Reviews;*, 80:488–515, 2017.

L. P. Bosque and S. E. Garza. Aggressive text detection for cyberbullying. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8856:221–232, 2014.

L. Brady and C. Phillips. Aesthetics and usability: A look at color and balance. *Usability News, 5.1*, 2003.

D. A. Broniatowski, M. J. Paul, and M. Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS ONE*, 8(12):e83672, 2013.

L. L. Brunton, B. Chabner, and B. C. Knollmann. *Goodman & Gilman's pharmacological basis of therapeutics*. McGraw-Hill 13th edition; New York, US, 2018.

P. Burnap, M. L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):206, 2014.

P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233, 2016.

E. Cambria, D. Das, S. Bandyopadhyay, and A. E. Feraco. *A practical guide to sentiment analysis*. Springer International Publishing, Cham, Switzerland, 2017.

P. S. H. D. . K. K. Cambria, E. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *In Proceedings of the AAAI conference on artificial intelligence*, 32(1), 2018.

A. Ceron, L. Curini, S. M. Iacus, and G. Porro. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New Media and Society*, 16(2):340–358, 2014.

C. W. H. Chan, B. M. H. Law, W. K. W. So, K. M. Chow, and M. M. Y. Waye. Pharmacogenomics of breast cancer: highlighting cyp2d6 and tamoxifen. *J Cancer Res Clin Oncol*, 146(6):1395–1404, 2020.

N. T. M. Chen and P. J. F. Clarke. Gaze-based assessments of vigilance and avoidance in social anxiety: a review. *Current Psychiatry Reports*, 19:59, 2017.

X. Chen, Y. Cho, and S. Y. Jang. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium*, page 63–68. SIEDS 2015, 2015.

M. Cheng and X. Jin. What do airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70, 2019.

M. Cheong and V. C. Lee. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, 13(1):45–59, 2011.

C. Chew and G. Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PLOS ONE*, 5(11):e14118, 2010.

K. H. Chu, C. G. Escobar-Viera, S. J. Matheny, E. M. Davis, and B. A. Primack. Tobacco cessation mobile app intervention (just kwit! study): protocol for a pilot randomized controlled pragmatic trial. *Trials*, 20(1):147, 2019.

R. Chunara, J. R. Andrews, and J. S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Am J Trop Med Hyg*, 86(1):39–45, 2012.

P. Conde-Cespedes, J. Chavando, and E. Deberry. Detection of suspicious accounts on twitter using word2vec and sentiment analysis. In S. Cham, editor, *International Conference on Multimedia and Network Information System*, pages 362–371, 2018.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

U. Cop, N. Dirix, D. Drieghe, and W. Duyck. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615, 2017.

J. S. Cramer. The origins of logistic regression (december 2002). Working paper no. 2002-119/4, Tinbergen Institute, 2002. URL `https://ssrn.com/abstract=360300`.

M. Dadvar and F. De Jong. Cyberbullying detection: A step toward a safer internet yard. In *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*, pages 121–125, 2012.

G. Dagnelie. *Visual Prosthetics: Physiology, Bioengineering, Rehabilitation.* Springer Science & Business Media, 2011. ISBN 978-1-4419-0754-7.

M. Das, S. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528. WWW 2003, 2003.

C. L. Davila-Fajardo, X. Diaz-Villamarin, A. Antunez-Rodriguez, A. E. Fernandez-Gomez, P. Garcia-Navas, L. J. Martinez-Gonzalez, J. A. Dávila-Fajardo, and J. C. Barrera. Pharmacogenetics in the treatment of cardiovascular diseases and its current progress regarding implementation in the clinical routine. *Genes*, 10:4, 2019.

A. Deb, K. Lerman, and E. Ferrara. Predicting cyber-events by leveraging hacker sentiment. *Information*, 9(11):280, 2018.

K. A. Dell and M. B. Chudow. A web-based review game as a measure of overall course knowledge in pharmacotherapeutics. *Curr Pharm Teach Learn*, 11(8):838–842, 2019.

J. Devlin, M. Chang, K. Lee, and T. K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018.

S. Djamasbi, M. Siegel, and T. Tullis. Generation y, web design, and eye tracking. *International Journal of Human-Computer Studies*, 68:307–323, 2010.

C. H. Dodson and E. Baker. Focus group testing of a mobile app for pharmacogenetic-guided dosing. *J Am Assoc Nurse Pract*, 33(3):205–210, 2020.

W. Dong, H. Liao, R. Roth, and S. Wang. Eye tracking to explore the potential of enhanced imagery basemaps, web mapping. *The Cartographic Journal*, 51:313–329, 2014.

A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer International Publishing, 2017.

H. B. L. Duh, G. C. B. Tan, and V. H. Chen. Usability evaluation for mobile device: A comparison of laboratory and field tests. In M. Nieminen and M. Roykkee, editors, *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, page 181–186. 2006.

L. Dupont, M. Antrop, and V. Van Eetvelde. Eye-tracking analysis in landscape perception research: influence of photograph properties and landscape characteristics. *Landscape Research*, 39(4):1–18, 2013.

M. Egbring, E. Far, M. Roos, M. Dietrich, M. Brauchbar, G. A. Kullak-Ublick, and A. Trojan. A mobile app to stabilize daily functional activity of breast cancer patients in collaboration with the physician: A randomized controlled clinical trial. *J Med Internet Res*, 18(9):e238, 2016.

S. Ehmke, C. Wilson. Identifying web usability problems from eyetracking data. In *Proceedings of the 21st British HCI Group Annual Conference on HCI 2007: HCI...but not as we know it*, University of Lancaster, United Kingdom, 2007.

S. Epskam, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom. qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48:1–18, 2012.

I. Feinerer and K. Hornik. tm: Text mining package. r package version 0.7-7. 2019. URL `https://CRAN.R-project.org/package=tm`.

G. Fernandez, P. Mandolesi, N. P. Rotstein, O. Colombo, O. Agamennoni, and L. E. Politi. Eye movement alterations during reading in patients with early alzheimer disease. *Invest Ophthalmol Vis Sci*, 54:8345–8352, 2013.

T. W. Frazier, M. Strauss, E. W. Klingemier, E. E. Zetzer, A. Y. Hardan, C. Eng, and E. A. Youngstrom. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, 56:546–555, 2017.

D. Fritz and E. Tows. Text mining and reporting quality in German banks—a cooccurrence and sentiment analysis. *Univers J Account Finance*, 6(2):54–81, 2018.

K. W. Fu, H. Liang, N. Saroha, Z. T. H. Tse, P. Ip, and I. C. H. Fung. How people react to zika virus outbreaks on twitter? a computational content analysis. *Am J Infect Control*, 44(12):1700–1702, 2016.

I. C. H. Fung, Z. T. H. Tse, C. N. Cheung, A. S. Miu, and K. W. Fu. Ebola and the social media. *Lancet*, 384(9961):2207, 2014.

P. Garg, H. Garg, and V. Ranga. Sentiment analysis of the uri terror attack using twitter. In *2017 International conference on computing, communication and* *automation (ICCCA)*, pages 17–20, IEEE, 2017. communication and automation (ICCCA).

T. Garín-Muñoz and T. Amaral. Internet usage for travel and tourism. the case of spain. *Tourism Economics*, 17:1071–1085, 2011.

R. Garside. *The CLAWS Word-tagging System*, volume The Computational Analysis of English: A Corpus-based Approach. Longman, London, 1987.

R. Garside and N. Smith. A hybrid grammatical tagger: Claws4. In L. G. Garside, R. and A. e. McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121. Longman, London, 1997.

G. Gemar and J. A. Jiménez-Quintero. Text mining social media for competitive analysis. *Tour Manag Stud*, 11(1):84–90, 2015.

J. J. Gibson. Acritical review of the concept of set in contemporary experimental psychology. *Psychological Bulletin*, 38:781–817, 1941.

W. R. Gilks, S. Richardson, and D. E. Spiegelhalter. *Markov chain Monte Carlo in practice*. Springer, Chapman & Hall/CRC Interdisciplinary Statistics (Book 2), Boca Raton, FL, 1996.

A. Güneyli, M. Ersoy, and S. Kiralp. Terrorism in the 2015 election period in turkey: Content analysis of political leaders' social media activity. *J. UCS*, 23(3):256–279, 2017.

J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. Eye tracking in web search tasks: Design implications. *In: Eye Tracking Research & Applications (ETRA) Symposium, ACM*, pages 51–58, 2002.

X. P. Goldberg, J. H. Kotval. Computer interface evaluation using eye movements: Methods and constructs. *In: International Journal of Industrial Ergonomics*, 24: 631–645, 1999.

S. Goodman, B. Morrongiello, and K. Meckling. A randomized, controlled trial evaluating the efficacy of an online intervention targeting vitamin d intake, knowledge and status among young adults. *Int J Behav Nutr Phys Act*, 13(1):116, 2016.

R. W. Graham, D. J. Jeffery. Location, location, location: eye tracking evidence that consumers preferentially view prominently positioned nutrition information. *Journal of The American Dietetic Association*, 111:1704–1711, 2011.

R. W. Graham, D. J. Jeffery. Predictors of nutrition label viewing during food purchase decision making: an eye tracking investigation. *Public Health Nutrition*, 15:189–197, 2012.

R. L. Gregory. *Eye and brain: The psychology of seeing*. Princeton University Press, Princeton, NJ, 1990.

J. P. Guidry, Y. Jin, C. A. Orr, M. Messner, and S. Meganck. Ebola on instagram and twitter: How health organizations address the health crisis in their social media engagement. *Public relations review*, 43(3):477–486, 2017.

A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah. Comprehensive review of text-mining applications in finance. *Financ Innov*, 6:39, 2020.

L. Hansen, A. Arvidsson, F. Nielsen, E. Colleoni, and M. Etter. *Good Friends, Bad News - Affect and Virality in Twitter*, volume Future Information Technology. Communications in Computer and Information Science. Springer, Berlin, Heidelberg, 2011.

J. G. Harb, R. Ebeling, and K. Becker. Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. *Journal of Information and Data Management*, 10(2):97–115, 2019.

A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, V. Martinez-Hernandez, H. Perez-Meana, and V. Olivares-Mercado, J. Sanchez. Social sentiment sensor in twitter for predicting cyber-attacks using l1 regularization. *Sensors*, 18(5):1380, 2018.

J. M. Hilbe. *Logistic Regression Models (Chapman & Hall/CRC Texts in Statistical Science)*, 2009.

J. K. Hockings, A. L. Pasternak, A. L. Erwin, N. T. Mason, C. Eng, and J. K. Hicks. Pharmacogenomics: An evolving clinical tool for precision medicine. *Cleve Clin J Med*, 87(2):91–99, 2020.

J. Holsanova, H. Rahm, and K. Holmqvist. Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Personality and Individual Differences*, 2006.

K. J. Horvath, S. Lammert, R. F. MacLehose, T. Danh, J. V. Baker, and A. W. Carrico. A pilot study of a mobile app to support hiv antiretroviral therapy adherence among men who have sex with men who use stimulants. *AIDS Behav*, 23(11):3184–3198, 2019.

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, 2004.

Y. H. Hu and K. Chen. Predicting hotel review helpfulness: The impact of review visibility. *and interaction between hotel stars and review ratings International Journal of Information Management*, 36(6):929–944, 2016.

D. S. Hui, E. I. Azhar, T. A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T. D. Mchugh, Z. A. Memish, C. Drosten, A. Zumla, and E. Petersen. The continuing 2019-ncov epidemic threat of novel coronaviruses to global health - the latest 2019 novel coronavirus outbreak in wuhan, china. *Int J Infect Dis*, 91:264–266, 2020.

M. B. Hunter and R. F. M. Chin. *Impaired social attention detected through eye movements in children with early-onset epilepsy.* Epilepsia, 2021. doi: 10.1111/epi.16962.

C. J. Hutto and E. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, MI, June 2014, 2014. Ann Arbor.

K. C. Hwang, Y. M. Lee. Using an eye tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human–Computer Interaction*, 34:15–24, 2017.

V. Ikoro, M. Sharmina, K. Malik, and R. Batista-Navarro. Analyzing sentiments expressed on twitter by uk energy company consumers. In *2018 Fifth International Conference on Social Networks Analysis*, pages 95–98. Management and Security (SNAMS), 2018.

M. R. Islam and M. Zibran. Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text. *J. Syst. Softw*, 145: 125–146, 2018.

R. J. K. Jacob and K. S. . Karn. Eye tracking in human computer interaction and usability research: Ready to deliver the promises. In R. R. Hyona and H. Deubel, editors, *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements.* Elsevier, Oxford, 2003. URL `http://www.eecs.tufts.edu/`.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, New York, 2013.

W. James. *The principles of psychology.* Harvard University Press, Cambridge, MA, 1981.

M. Jiménez-García, J. Ruiz-Chico, and A. R. Peña-Sánchez. Landscape and tourism: Evolution of research topics. *Land*, 9(12):1–17, 2020.

N. Jindal and B. Liu. Mining comparative sentences and relations. In *Proceedings of National Conference on Artificial Intelligence (AAAI-2006)*, 2016.

M. L. Jockers. Extract Sentiment and Plot Arcs from Text, Syuzhet, 2015. URL `https://github.com/mjockers/syuzhet`.

E. P. Jones and C. S. Wisniewski. Gamification of a mobile applications lecture in a pharmacy course. *Med Ref Serv Q*, 38(4):339–346, 2019.

J. H. Jones and M. Salathe. Early assessment of anxiety and behavioral response to novel swine-origin influenza a (h1n1). *PLOS ONE*, 4(12):e8032, 2009.

R. P. Joshi, D. F. Steiner, E. Q. Konnick, and C. J. Suarez. Pharma-oncogenomics in the era of personal genomics: A quick guide to online resources and tools. *Adv Exp Med Biol*, 1168:103–115, 2019.

T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *ICCV*, 2009.

K. S. Just, M. Steffens, J. J. Swen, G. P. Patrinos, H. J. Guchelaar, and J. C. Stingl. Medical education in pharmacogenomics-results from a survey on pharmacogenetic knowledge in healthcare professionals within the european pharmacogenomics clinical implementation project ubiquitous pharmacogenomics (u-pgx). *Eur J Clin Pharmacol*, 73(10):1247–1252, 2017.

J. K. Kaakinen and J. Simola. Fluctuation in pupil size and spontaneous blinks reflect story transportation. *J Eye Mov Res*, 13:1212–1223, 2020.

B. Kapitaniak, M. Walczak, M. Kosobudzki, Z. Jozwiak, and A. Bortkiewicz. Application of eye tracking in the testing of drivers: A review of research. *International Journal of Occupational Medicine and Environmental Health*, 28(6):941–954, 2015.

N. Karuna, P. Tragulpiankit, S. Mahasirimongkol, and S. Chumnumwat. Knowledge, attitude, and practice towards pharmacogenomics among hospital pharmacists in thailand. *Pharmacogenet Genomics*, 30(4):73–80, 2020.

K. Kim, O. J. Park, S. Yun, and H. Yun. What makes tourists feel negatively about tourism destinations? application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change*, 123: 362–369, 2017.

W. Y. Kim, H. S. Kim, M. Oh, and J. G. Shin. Survey of physicians' views on the clinical implementation of pharmacogenomics-based personalized therapy. *Transl Clin Pharmacol*, 28(1):34–42, 2020.

D. G. Kleinbaum and M. Klein. *Logistic regression.* Springer-Verlag, New York, 2010.

M. Koromina, S. Koutsilieri, and G. P. Patrinos. Delineating significant genome-wide associations of variants with antipsychotic and antidepressant treatment response: implications for clinical pharmacogenomics. *Hum Genomics*, 14(1):4, 2020.

S. M. Kosslyn. *Image and brain.* MIT Press, Cambridge, MA, 1994.

D. Kotzias, M. Denil, N. De Freitas, and P. Smyth. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 597–606, 2015. ACM, USA.

M. Kuhn. Caret package. *Journal of Statistical Software*, 28(5), 2008.

M. Kuhnel, L. Seiler, A. Honal, and D. Ifenthaler. Mobile learning analytics in higher education: Usability testing and evaluation of an app prototype. In *Paper presented at the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age*, Algarve, Portugal, 2017. Vilamoura.

K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise. Towards inferring language expertise using eye tracking. *CHI 13 Extended Abstracts on Human Factors in Computing Systems*, 13, 2013a.

K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling. I know what you are reading: recognition of document types using mobile eye tracking. *ISWC '13: Proceedings of the 2013 International Symposium on Wearable Computers*, pages 113–116, 2013b.

J. T. Lam, M. A. Gutierrez, J. A. Goad, L. Odessky, and J. Bock. Use of virtual games for interactive learning in a pharmacy curriculum. *Curr Pharm Teach Learn*, 11(1):51–57, 2019.

J. Lamberz, T. Litfin, O. Teckert, and G. Meeh-Bunse. Still searching or have you found it already? – usability and web design of an educational website. *Business Systems Research*, 9(1):19–30, 2018.

A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran. Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's ebola live twitter chat. *Am J Infect Control*, 43(10):1109–1111, 2015.

B. Lee, J. H. Park, L. Kwon, Y. H. Moon, Y. H. Shin, G. Kim, and H. J. Kim. About relationship between business text patterns and financial performance in corporate data. *J. open innov*, 4:3, 2018.

D. Levine, J. Torabi, K. Choinski, J. P. Rocca, and J. A. Graham. Transplant surgery enters a new era: Increasing immunosuppressive medication adherence through mobile apps and smart watches. *Am J Surg*, 218(1):18–20, 2019.

M. N. Levy, B. M. Koeppen, and B. A. Stanton. *Berne & Levy Physiology*. Elsevier, 2017.

W. Li and H. Chen. Identifying top sellers in underground economy using deep learning-based sentiment analysis. *Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014*, 2014.

H. Liang, I. C. H. Fung, Z. T. H. Tse, J. Yin, C. H. Chan, L. E. Pechta, J. Smith, B., R. D. Marquez-Lameda, M. I. Meltzer, K. M. Lubell, and K. W. Fu. How did ebola information spread on twitter: broadcasting or viral spreading? *BMC public health*, 19(1):438, 2019.

H. Liao, W. Dong, C. Peng, and H. Liu. Exploring differences of visual attention in pedestrian navigation when using 2d maps and 3d geo-browsers. *Cartography and Geographic Information Science*, 44:474–490, 2017.

H. Liao, X. Wang, W. Dong, and L. Meng. Measuring the influence of map label density on perceived complexity: a user study using eye tracking. *Cartography and Geographic Information Science*, 46(3):210–227, 2019.

J. Z. Lim, J. Mountstephens, and J. Teo. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*, 20(8):210–227, 2020.

B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer, 2006.

B. Liu. Sentiment analysis and subjectivity. In D. F. Indurkhya, N., editor, *Handbook of Natural Language Processing.* Chapman and Hall, 2010.

B. Liu. *Sentiment analysis: mining opinions, sentiments, and emotions.* Cambridge University Press, Cambridge, 2015.

Y. Liu, K. Huang, J. Bao, and K. Chen. Listen to the voices from home: An analysis of chinese tourists' sentiments regarding australian destinations. *Tourism Management*, 71:337–347, 2019.

P. N. Lopes, P. Solovey, and R. Straus. Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences*, 35:641–658, 2003.

C. Lopez and C. Tucker. *Towards personalized adaptive gamification: a machine learning model for predicting performance.* IEEE Trans. Games, 2018.

S. G. Luke and K. Christianson. The provo corpus: A large eye-tracking corpus with predictability norms. *Behav Res*, 50:826–833, 2018.

A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*, pages 142–150. ACL, USA, 2011.

L. Maffei. Enciclopedia della scienza e della tecnica. 2007.

C. D. Manning, P. Raghavan, and H. Schutze. Scoring, term weighting, and the vector space model. *Introduction to Information Retrieval*, pages 100–123, 2008.

S. M. Manson, L. Kne, K. R. Dyke, J. Shannon, and S. Eria. Using eye tracking and mouse metrics to test usability of web mapping. *Cartography and Geographic Information Science;*, 39(1):48–60, 2012.

S. Martinez-Conde and S. L. Macknik. From exploration to fixation: An integrative view of yarbus's vision. *Perception*, 44:884–899, 2015.

C. C. McDonald, A. H. Goodwin, A. K. Pradhan, M. R. Romoser, and A. F. Williams. A review of hazard anticipation training programs for young drivers. *Journal of Adolescent Health*, 57(1):S15–23, 2015.

A. McNeill, S. Gravely, S. C. Hitchman, L. Bauld, D. Hammond, and J. Hartmann-Boyce. Tobacco packaging design for reducing tobacco use. *Cochrane Database Syst Rev*, 4:CD011244, 2017.

W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 165–171, 2011.

D. Meyer. e1071 r package. 2019. URL `https://cran.r-project.org/web/packages/e1071/index.html`.

M. Mikula and K. Gao, X. Machová. Adapting sentiment analysis system from english to slovak. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, 2017.

L. M. Miller and D. L. Cassady. Making healthy food choices using nutrition facts panels. the roles of knowledge, motivation, dietary modifications goals, and age. *Appetite*, 59(1):129–139, 2012.

S. Mohammad and P. Turney. Crowdsourcing a word-emotion association lexicon,. *Computational Intelligence*, 29:436–465, 2013.

S. Momtazi. Fine-grained German sentiment analysis on social media. In *Proc. of the 9th Intl. Conf. on Language Resources and Evaluation, LREC 2012.*, 2012.

S. Morganti, P. Tarantino, E. Ferraro, P. D'Amico, B. A. Duso, and G. Curigliano. Next generation sequencing (ngs): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Adv Exp Med Biol*, 1168:9–30, 2019.

K. Morita, K. Miura, K. Kasai, and R. Hashimoto. Eye movement characteristics in schizophrenia: A recent update with clinical implications. *Neuropsychopharmacol Rep*, 40(1):2–9, 2020.

S. Moro, P. Ramos, J. Esmerado, and S. M. J. Jalali. Can we trace back hotel online reviews' characteristics using gamification features? *International Journal of Information Management*, 44:88–95, 2019.

F. Muñoz-Leiva, J. Hernández-Méndez, and D. Gómez-Carmona. Measuring advertising effectiveness in travel 2.0 websites through eye-tracking technology. *Physiology & Behavior*, 200:83–95, 2019.

S. Murnion, W. J. Buchanan, A. Smales, and G. Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers and Security*, 76: 197–213, 2018.

A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. Ngo. Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Syst Appl*, 42(1):306–324, 2015.

T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, Florida, 2003. 70-77.

T. H. Nguyen, K. Shirai, and J. Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Syst Appl*, 42(24):9603–9611, 2015.

M. Nilsson Benfatto, G. Oqvist Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson. Screening for dyslexia using eye tracking during reading. *PLoS One*, 11: e0165508, 2016.

H. Olkoniemi, V. Stromberg, and J. K. Kaakinen. The ability to recognise emotions predicts the time-course of sarcasm processing: Evidence from eye movements. *Q J Exp Psychol (Hove)*, 72, 2019.

R. Oyekunle, O. Bello, Q. Jubril, I. Sikiru, and A. Balogun. Usability evaluation using eye-tracking on e-commerce and education domains. *Journal of Information Technology and Computing*, 1(1):1–13, 2020.

S. O. Oyeyemi, E. Gabarron, and R. Wynn. Ebola, twitter, and misinformation: a dangerous combination? *BMJ*, 349:g6178, 2014.

N. Oztürk and S. Ayvaz. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147, 2018.

V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi*, pages 1345–1350, 2016.

B. Pan, L. Zhang, and R. Law. The complex matter of online hotel choice. *Cornell Hospitality Quarterly*, 54(1):74–83, 2013.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, 2002.

K. Park, M. Jeong, and M. Kim. Usability evaluation of menu interfaces for smartwatches. *Journal of Computer Information Systems*, 60(2):156–165, 2020.

M. J. Paul and M. Dredze. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183, 2017.

Penn Treebank Project. URL `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`.

K. Pentus, K. Ploom, A. Kuusik, and T. Mehine. How to optimize sales flyers – novel experiment design. *Baltic J. of Manag*, pages 1746–5265, 2018.

C. Petit, A. Croisetiere, F. Chen, and I. Laverdiere. Are pharmacists from the province of quebec ready to integrate pharmacogenetics into their practice. *Pharmacogenomics*, 21(4):247–256, 2020.

B. Pfaff. Var, svar and svec models: Implementation within r package vars. *Journal of Statistical Software*, 27(4), 2008.

PharmacoloGenius download link. URL `https://play.google.com/store/apps/details?id=com.gzcp.pharmacologenius&hl=en`.

C. E. Pierce, S. T. de Vries, S. Bodin-Parssinen, L. Harmark, P. Tregunno, D. J. Lewis, S. Maskell, R. Van Eemeren, A. Ptaszynska-Neophytou, V. Newbould, N. Dasgupta, A. F. Z. Wisniewski, S. Gama, and P. Mol. Recommendations on the use of mobile applications for the collection and communication of pharmaceutical product safety information: Lessons from imi web-radr. *Drug Saf*, 42(4):477–489, 2019.

M. Pifarre, A. Carrera, J. Vilaplana, J. Cuadrado, S. Solsona, F. Abella, F. Solsona, and R. Alves. Tcontrol: A mobile app to follow up tobacco-quitting patients. *Comput Methods Programs Biomed*, 142:81–89, 2017.

C. Pisanu, E. E. Tsermpini, E. Mavroidi, T. Katsila, G. P. Patrinos, and A. Squassina. Assessment of the pharmacogenomics educational environment in southeast europe. *Public Health Genomics*, 17(5-6):272–279, 2014.

R. Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33, 1980.

A. Poole, L. J. Ball, and P. Phillips. In search of salience: A response time and eye movement analysis of bookmark recognition. In S. Fincher, P. Markopolous, D. Moore, and R. Ruddle, editors, *People and Computers XVIII-Design for Life: Proceedings of HCI 2004*. Springer-Verlag, London, 2004.

D. Pope and J. Griffith. An analysis of online twitter sentiment surrounding the european refugee crisis. In *In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, (IC3K 2016)*, pages 299–306, 2016.

L. Popova, J. Nonnemaker, N. Taylor, B. Bradfield, and A. Kim. Warning labels on sugar-sweetened beverages: An eye tracking approach. *Am J Health Behav*, 1(43): 2, 2019.

M. I. Posner, C. R. R. Snyder, and B. J. Davidson. Attention and the detection of signals. *Experimental Psychology: General*, 109(2):160–174, 1980.

A. Pozza, A. Coluccia, G. Gualtieri, and F. Ferretti. nhancing adherence to antipsychotic treatment for bipolar disorders. comparison of mobile app-based psychoeducation, group psychoeducation, and the combination of both: protocol of a three-arm single-blinded parallel-group multi-centre randomised trial. *Clin Ter*, 171(2):e7–e93, 2020.

R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.

R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL, 2020. URL `http://www.R-project.org/`.

Y. Rao, J. Lei, W. Liu, Q. Li, and M. Chen. Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17:723–742, 2013.

K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422, 1998.

A. Remuzzi and G. Remuzzi. Covid-19 and italy: what next? *Lancet*, 395(10231): 1225–1228, 2020.

S. Rill, D. Reinel, J. Scheidt, and R. Zicari. Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69, 2014.

T. W. Rinker. sentimentr: Calculate text polarity sentiment version 2.7.1. 2019. URL `http://github.com/trinker/sentimentr`.

S. Ritthiron and A. Jiamsanguanwong. Usability evaluation of the university library network's website using an eye tracking. In *ICAIP 2017 Proceedings of the International Conference on Advances in Image Processing*, pages 184–188, 2017.

D. A. Robinson. The oculomotor control system: A review. In *Proceedings of the IEEE*, 1032–1049, 1968.

P. Romagnani, G. Gnone, F. Guzzi, S. Negrini, A. Guastalla, F. Annunziato, S. Romagnani, and R. De Palma. The covid-19 infection: lessons from the italian experience. *J Public Health Policy*, 41(3):238–244, 2020.

S. P. Roth, A. N. Tuch, E. D. Mekler, J. A. Bargas-Avila, and K. Opwis. Location matters, especially for non-salient features–an eye tracking study on the effects of web object placement on different types of websites. *International Journal of Human-Computer Studies*, 71(3):228–235, 2013.

M. Rubin, N. Bhattacharya, J. Gwizdka, Z. Griffin, and M. Telch. The influence of ptsd symptoms on selective visual attention while reading. *Cogn Emot*, 19:1–8, 2021.

L. M. Ruhanen, C.-L. J. McLennan, and B. D. Moyle. Strategic issues in the australian tourism industry: A 10-year analysis of national strategies and plans. *Asia Pacific Journal of Tourism Research*, 18(3):220–240, 2013.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. *In: Learning for Text Categorization: Papers from the 1998 workshop*, pages 98–105, 1998.

K. Sailunaz and R. Alhajj. Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, 36:101003, 2019.

B. Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* University of Pennsylvania, School of Engineering and Applied Science, Department of Computer and Information Science, 1990.

K. Schag, E. J. Leehr, P. Meneguzzo, P. Martus, S. Zipfel, and K. E. Giel. Food-related impulsivity assessed by longitudinal laboratory tasks is reduced in patients with binge eating disorder in a randomized controlled trial. *Sci Rep*, 11(1):8225, 2021.

R. E. Scherr, K. D. Laugero, D. J. Graham, B. T. Cunningham, L. Jahns, K. R. Lora, M. Reicks, and A. R. Mobley. Innovative techniques for evaluating behavioral nutrition interventions. *Advances in Nutrition*, 8(1):113–125, 2017.

N. Scott, R. Zhang, D. Le, and B. Moyle. A review of eye-tracking research in tourism. *Current Issues in Tourism*, 22(10):1244–1261, 2019.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016.

L. Sera and E. Wheeler. Game on: The gamification of the pharmacy classroom. *Curr Pharm Teach Learn*, 9(1):155–159, 2017.

R. Serfozo. *Basics of applied stochastic processes.* Springer (Berlin), Berlin, Heidelberg, 2009.

C. Shannon. Engaging students in searching the literature. *Med Ref Serv Q*, 38(4): 326–338, 2019.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3):591–611, 1965.

D. F. Shebilske, W. L. Fisher. Understanding extended discourse through the eyes: How and why. In M. Groner, C. Menz, D. F. Fisher, and R. A. Monty, editors, *Eye movements and psychological functions: International views (pp. 303–314). Hillsdale, NJ: Lawrence Erlbaum.* 1983.

J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. *Exploiting Topic Based Twitter Sentiment for Stock Prediction.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013), 2013.

A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PLOS ONE*, 6(5):e19467, 2011.

T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, and B. Adini. Twitter in the cross fire—the use of social media in the westgate mall terror attack in kenya. *PLOS ONE*, 9(8):e104136, 2014.

M. Smith, D. A. Broniatowski, M. J. Paul, and M. Dredze. Towards real-time measurement of public epidemic awareness: Monitoring influenza awareness through twitter. In C. Stanford, editor, *AAAI spring symposium on observational studies through social media and other human-generated content*, 2016.

G. Spedicato. Discrete time Markov chains with r. *The R Journal*, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-036/index.html`.

M. Stojek, L. M. Shank, A. Vannucci, D. M. Bongiorno, E. E. Nelson, A. J. Waters, S. G. Engel, K. N. Boutelle, D. S. Pine, J. A. Yanovski, and M. Tanofsky-Kraff. A systematic review of attentional biases in disorders involving binge eating. *Appetite*, 123:367–389, 2018.

N. Stott, J. R. E. Fox, and M. O. Williams. Attentional bias in eating disorders: A meta-review. *Int J Eat Disord*, 2021. doi: 10.1002/eat.23560.

J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

M. Szomszor, P. Kostkova, and E. De Quincey. Swineflu: Twitter predicts swine flu outbreak in 2009. In *International conference on electronic healthcare*, pages 18–26, Berlin, Heidelberg, 2010. Springer.

N. Tabari and M. Hadzikadic. Context sensitive sentiment analysis of financial tweets: A new dictionary. In R. Bembenik, S. Ł., G. Protaziuk, M. Kryszkiewicz, and H. Rybinski, editors, *Intelligent Methods and Big Data in Industrial Applications. Studies in Big Data*. Springer, Cham, 2019.

P. Thier and U. J. Ilg. The neural basis of smooth-pursuit eye movements. *Current Opinion in Neurobiology*, 15(6):645–52, 2005.

Tobii Studio User Manual V. 3.3.1. URL `https://www.tobiipro.com`.

S. Towers, S. Afzal, G. Bernal, N. Bliss, S. Brown, B. Espinoza, J. Jackson, J. Judson-Garcia, M. Khan, M. Lin, R. Mamada, V. M. Moreno, F. Nazari, K. Okuneye, M. L. Ross, C. Rodriguez, J. Medlock, D. Ebert, and C. Castillo-Chavez. Mass media and the contagion of fear: the case of ebola in america. *PLOS ONE*, 10(6): e0129179, 2015.

A. Trapletti and K. Hornik. tseries: Time series analysis and computational finance. 2021. URL `https://CRAN.R-project.org/package=tseries`.

A. Treisman. Features and objects in visual processing. *Scientific American*, 255(5): 114B–125, 1986.

A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics, PA*, page 417–424. 2002.

R. van der Lans, M. Wedel, and R. Pieters. Defining eye-fixation sequences across individuals and tasks: the binocular-individual threshold (bit) algorithm. *Behavior Research Methods*, 43:239–257, 2011.

R. van Westrhenen, K. J. Aitchison, M. Ingelman-Sundberg, and M. M. Jukic. Pharmacogenomics of antidepressant and antipsychotic treatment: How far have we got and where are we going? *Front Psychiatry*, 11:94, 2020.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S, Fourth edition*. Springer, New York, (US), 2002. ISBN 0-387-95457-0.

V. Viet, Q. Huyn, and K. Yamamoto. Vietsentilex: a sentiment dictionary that considers the polarity of ambiguous sentiment words. In S. Politzer-Ahles, Y. Y. Hsu, C. H. Huang, and Y. Yao, editors, *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, Hong Kong*. 2018.

B. Vollmer and S. Karakayali. The volatility of the discourse on refugees in germany. *Journal of immigrant & refugee studies*, 16(1-2):118–139, 2018.

T. von der Malsburg and S. Vasishth. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65:109–127, 2011.

H. Von Helmholtz. Handbuch der physiologischen optik (treatise on physiological optics). The Optical Society of America, Rochester, NY, 1925.

T. Vuori, M. Olkkonen, M. Polonen, A. Siren, and J. Hakkinen. Can eye movements be quantitatively applied to image quality studies? *Proceedings NordiCHI 04, ACM Press*, 4:335–333, 2004.

U. Waltinger. Germanporalityclues: A lexical resource for German sentiment analysis. In *Proceedings of Language Resources and Evaluation Conference (LREC)*. 2010.

X. Wan. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), Stroudsburg, Association for Computational Linguistics*, page 553–561, 2008.

J. Wang, P. Antonenko, M. Celepkolu, Y. Jimenez, E. Fieldman, and A. Fieldman. Exploring relationships between eye tracking and traditional usability testing data. *International Journal of Human-Computer Interaction*, 36(6):483–494, 2018.

Y. Wang and B. Sparks. An eye tracking study of tourism photo stimuli: Image characteristics and ethnicity. *Journal of Travel Research*, 55(5):588–602, 2016.

Y. T. Wang, M. Y. Merl, J. Yang, Z. X. Zhu, and G. H. Li. Opportunities for pharmacists to integrate pharmacogenomics into clinical practice. *Pharmacogenomics J*, 20(2):169–178, 2020.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016. ISBN 978-3-319-24277-4. URL `http://ggplot2.org`.

A. Yadav and D. K. Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.

F. Y. Yang, M. J. Tsai, G. L. Chiou, S. W. Y. Lee, C. C. Chang, and L. L. Chen. Instructional suggestions supporting science learning in digital environments based on a review of eye tracking studies. *Journal of Educational Technology & Society*, 21(2):28–45, 2018.

S. U. Yang, H. Shin, J. H. Lee, and B. Wrigley. Country reputation in multidimensions: predictors, effects, and communication channels. *Journal of Public Relations Research*, 20(4):421–440, 2008.

J. Yao, G. Wu, J. Liu, and Y. Zheng. Using bilingual lexicon to judge sentiment orientation of chinese words. In K. Seoul, editor, *Proceedings of 6th International Conference on Computer and Information Technology (CIT'06)*, 2006.

A. L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967.

L. Young and S. Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29:205–231, 2012.

L. R. Young and D. Sheena. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7(5):397–439, 1975.

I. Zamberg, S. Manzano, K. Posfay-Barbe, O. Windisch, T. Agoritsas, and E. Schiffer. A mobile health platform to disseminate validated institutional measurements during the covid-19 outbreak: Utilization-focused evaluation study. *JMIR Public Health Surveill*, 6(2):e18668, 2020.

G. Zammarchi and C. Conversano. Application of eye tracking technology in medicine: A bibliometric analysis. *Vision*, 4(5):56, 2021.

G. Zammarchi, M. Del Zompo, A. Squassina, and C. Pisanu. Increasing engagement in pharmacology and pharmacogenetics education using games and online resources: The pharmacologenius mobile app. *Drug Development Research*, 2020.

G. Zammarchi, L. Frigau, and F. Mola. Markov chain to analyze web usability of a university website using eye tracking data. *Statistical Analysis and Data Mining*, 14(4):331–341, 2021.

B. A. Zardari, Z. Hussain, A. A. Arain, W. H. Rizvi, and M. S. Vighio. *QUEST e-learning portal: applying heuristic evaluation, usability testing and eye tracking.* Univ Access Inf Soc, 2020. URL `https://doi.org/10.1007/s10209-020-00774-z`.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. strucchange: An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002.

S. Zhang, D. Chen, Y. Tang, and L. Zhang. Children asd evaluation through joint analysis of eeg and eye-tracking recordings with graph convolution network. *Front Hum Neurosci*, 15(65134):9, 2021.

W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)*, 2010.

Y. Zhu, K. M. Swanson, R. L. Rojas, Z. Wang, J. L. St Sauver, S. L. Visscher, L. J. Prokop, S. J. Bielinski, L. Wang, R. Weinshilboum, and B. J. Borah. Systematic review of the evidence on the cost-effectiveness of pharmacogenomics-guided treatment for cardiovascular diseases. *Genet Med*, 22(3):475–486, 2020.