Department of Business and Management

Master's Degree in Marketing Analytics and Metrics

Chair of Customer Intelligence & Big Data

# Using sentiment analysis to measure the outcome of marketing campaigns on Twitter, an empirical approach

SUPERVISOR

Prof. Giuseppe Francesco Italiano

CANDIDATE

Felice Cassone

N. 703541

CO-SUPERVISOR

Prof. Alessio Maria Braccini

ACADEMIC YEAR 2019 – 2020

*A Guglielmo, Virginia e Florianne,*
*il cui supporto ha rappresentato*
*le fondamenta di ogni mio traguardo.*

# Index

## Introduction

Every day we produce an immense amount of data, research in facts shows that there are 2.5 *quintillion bytes* of data created each day at our pace, which is only accelerating in recent years (Marr, 2018).

An important part of this data, however, is difficult to analyze from standard methods and yet it retains a key part of the consumer decision journey: the opinions of people publicly expressed on the internet. The part of the consumer journey at play is the active evaluation, that happens after an initial consideration from the consumer and consist of a general search of others' opinions and reviews about the relevant product category. (Court and Elzinga, 2009)

This type of opinions and subjective information can be referred to as *sentiment* and the difficulties in analyzing it derive from its expression in a natural language instead of a machine readable one. The best way to analyze them is still rooted in human analysis, but there are methods that are slowly approaching our precision and understanding through the use of characteristics that machines share with us, such as natural language processing and machine learning. These allow them to comprehend the meaning behind words or entire phrases and to learn from what they did in the past to improve future analysis.

Sentiment analysis is thus defined as the measurement of an author's opinion about specific entities (Feldman, 2013). And can represent a key asset for firms that seek to measure the reception of a marketing campaign, a new product, the opinions of their users or the overall thought of any large pool of unstructured human-created information.

In recent times there has been an "explosion" of sentiment shared online through social media like *Twitter* and *Facebook*. These snippets of text can be a goldmine for companies that are looking to monitor their public reputation, in fact sentiment analysis allows them to see in real time what the word-of -mouth is like for their average customer.

Since the technology behind text analysis is still behind human recognition it is difficult to analyze a full text with multiple paragraphs. A source like Twitter, however, does not run the risk of overcoming this machines' limit, thanks to its characters limit, set at 280 per tweet, with only 12% of tweets longer than 140 characters (Perez, 2018).

Twitter also offers the opportunity to aggregate different content with a keyword, under the form of *hashtags*, so that it can be even easier for a company to monitor certain topics.

The flexibility and ease of creating such publics and communities when they are needed and without restrictions, is what gave Twitter recognition as the preferred platform for events discussion. This recognition is evident in the common use of the platform by media organizations, politicians, and, most importantly, industries and firms, that choose to carry out part of their marketing strategies or their public interactions on it.

The aim of this work is to show that, through sentiment analysis applied to tweets, it is possible to measure and evaluate the opinions of thousands of customers in a single glance. In particular, hashtags will be used to aggregate different tweets revolving around particular marketing campaigns and brands (such as *#MyCalvins*, *#NintendoDirect*, and *#ShareaCoke*) in order to determine the reaction of customers to those and their general sentiment. The success of the hashtag incorporation in various promotional channels has led brand-related hashtags to become extremely popular, with 70% of the most frequently used hashtags in 2015 being brand related (Simply Measured, 2015).

The final outcome of the developed algorithm will create an evaluation of the overall reception and sentiment of the aggregated tweets, while also returning the most used words from the customers and their geolocations, so that the firm will be able to make an in-depth real time assessment of the campaign.

The algorithm will use the Twitter API to scrape the tweets, it will be based on a supervised learning approach for their analysis and it will use a Naïve Bayes classification model to tag their words into sentiment classes.

Similar works have been accomplished in the past in different ways: among others, Wang at al. (2011) have analyzed tweets using a graph-based sentiment analysis that as able to classify the texts by their polarity, while Agarwal et al. (2011) used a handmade emoticons dictionary to classify the tweets. On a different note, Liu (2010) created a way to exclude objective opinions in his sentiment analysis of movie reviews in order to obtain better and more useful results.

This work sets itself to improve on past literature by merging the many different methods used and incorporating new differentiators, hence the analysis will be carried by employing both a machine learning classifier and an emoticon dictionary and the results will be differentiated also based on their subjectivity, excluding objective tweets.

# 1. Research analysis

## 1.1. Qualitative and quantitative methods for research

In the marketing research field, a distinction can be made between the quantitative and the qualitative method (Neumann, 2011). Quantitative research focuses on the collection of data from a large sample of respondents from a defined population and uses statistical, mathematical and computational techniques for data analysis in order to generalize the result to the entire population. Qualitative research instead focuses on the exploration of a phenomenon in a more unstructured way, with the objective of obtaining valuable insights about attitudes, beliefs and emotions of the narrow group of subjects on which the research is carried. (Newman, 1998)

The type of information that the researchers obtain using one or the other greatly differs: the output of quantitative analysis is data in a raw form, the rigid and objective research approaches that this method uses yield a formalized and mathematic solution to the problem, basing on a standardized sequence of instructions. On the contrary, the output of qualitative research are the subjective opinions that the researcher is able to create through a personal reasoning after interpreting the data gathered from the subjects.

The most significant difference between these two types of analysis relies in their measurement process: what links the data to the concepts.

Three features separate qualitative from quantitative approaches to measurement.

Timing is the first difference. While in qualitative research the measure phase and the data collection phase are simultaneous, in quantitative research at first the actions to measure must be converted in variables, then the data is gathered and analyzed.

The data itself represent the second difference. In a qualitative study the information gathered is in the form of numbers, written or spoken words, actions, symbols and even images. The data in these cases is left diverse and unstandardized. A quantitative study, on the other hand, uses techniques that produce data in the form of numbers. It raises the level of abstraction from the initial ideas to their specific numerical representation. This way it creates a compact, uniform and consistent way to empirically represent abstract

ideas. While numerical data convert information into a standard and condensed format, qualitative data are voluminous, diverse, and nonstandard. (Neumann, 2011)

The last difference in the approaches to measurement is how the concepts are connected with the data. In qualitative researches many of the concepts used are developed and refined while the data is being gathered. The researchers can reexamine and reflect on the data simultaneously and interactively. This way the new ideas are able to steer the reasoning and suggest new ways and phenomena to measure. Conversely, in quantitative researches the reflection and definition of concept is done before the data gathering, then a measuring technique is chosen in order to bridge the abstract concepts with the empirical data.

Even with their differences, both of the methods of measurements intimately connect how we perceive and think about the social world with what we find in it (Neumann, 2011). Nevertheless either of the two received their share of criticism: quantitative research is criticized as a rigid approach that ignores the inherent subjectivity of human social interactions (Holstein and Gubrium, 1995) while qualitative research is described as a subjective and non-scientific method that lacks structural coherence (Poggenpoel and Myburgh, 2005).

Despite the contradictory methodologies used, recent development in research methodologies (Kelle, 2006; Olsen, 2004; Srnka, 2007) suggest that a new approach can be conceived by integrating the two, in order to improve both the rigor and the connection to the data at hand.

### 1.2. The debate between quantitative and quantitative methods

The number of differences between qualitative and quantitative research listed in the previous paragraph have created, starting from the 1980s, an antipathy between the two methods, that created full-fledged “paradigm wars”.

The paradigm wars saw researchers with different doctrines and methods of research strongly claiming that their line of thought was the most appropriate. In 1989,

Gage described the situation as if he was narrating it two decades into the future (hence in our past in 2009), stating that “it was in 1989 that the "Paradigm Wars" had come to a sanguinary climax” (Gage, 1989).

He argued that there were three hypothetical futures at hand:

1. The positivist, establishment, mainstream, standard, objectivity-seeking and quantitative approach had succumbed to their critiques.
2. Nothing had really changed, and the research wars were still going on.
3. Peace had broken between the two approaches out and a dialogue was created, lifting the discussion to a new level of insight, making progress toward the generation of a new theory that fitted together the previous ones. (Gage, 1989)

By 2009, paraphrasing the author, peace has indeed broken out, but not in the productive way that was anticipated. Rather than being settled or resolved in favor of a clear winner, the paradigm of research in the social sciences embedded the distinction between quantitative and qualitative methods in a way that often implies that they are incommensurable approaches.

This distance between the two methods has also been defined as an *epistemological chasm*, as described in Walby’s (2001) work, who noted that such chasms are usual between similar disciplines, despite being hard to justify on a philosophical standpoint.

The literature is composed of three different positions on the matter of this division: empiricist, realist and constructionist (Olsen, 2004).

The empiricist position (Silverman, 2001) claims that qualitative and quantitative techniques are completely antithetical: it assumes a dualism between the school of qualitative epistemology versus the school of quantitative epistemology. This line of thought has been disregarded by Olsen since in her opinion a more integrated epistemology is in need for social science, not two competing epistemological schools.

The realist position (Sayer, 1992) is an alternative to the empiricists’ because it argues that “social objects are often affected by the way they are construed, but that they also have an ongoing real existence that is not constituted entirely by how today’s researchers construe them” (Sayer, 2000). This opinion is thus plural in respect to methodology and to theories and offers a starting ground for embarking on integrated mixed-methods research. However, this theory is unlikely to spawn a more parsimonious

approach and Olsen has found its weakness in its inbred difficulty to simplify the research topics at hand, making it instead more elaborated and time-consuming.

The third position is the constructionist viewpoint (Andrews, 2012), that idealizes that all social objects are merely socially constructed. Meaning that their definition and even existence depend upon their observers' minds. Being a complete opposite of the realist line of reasoning, it does not come as a surprise that constructionism itself appears too simplistic to serve as a methodology by itself (Olsen, 2004).

After outlining these tree viewpoints in the literature, Olsen presents an innovative approach that proposes either of the three as a possible philosophical starting point of research but considers all of them at the same time. This approach puts empiricism, realism, and constructionism at different edges of a triangle of viewpoints, giving to researchers the possibility to start from either of the three, but advising the realist approach because of its capability to mix qualitative and quantitative research.

According to such thinking, triangulation across the qualitative-quantitative divide (shown in Figure 1) is only consistent with a pluralist theoretical viewpoint.

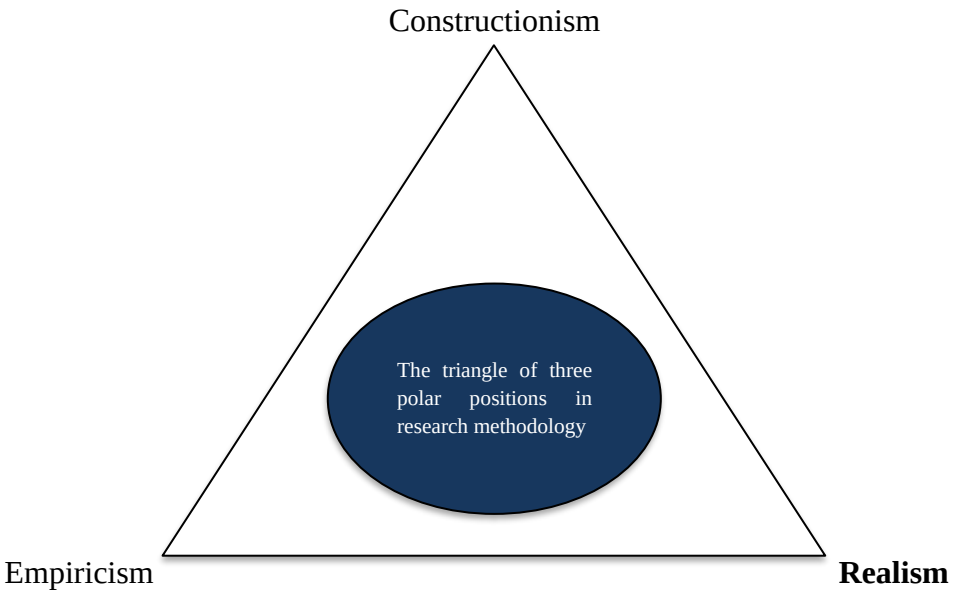


*Figure 1: The triangulation across the qualitative-quantitative divide*

According to Bryman (2001), the combination of different methodologies resulting from a mixed approach will generally tend to have a leading strategy that is actually used to start out the research, and a follow-up strategy to widen the enquiry and further

investigate the data at hand. This theory applies especially well to this strategy, where, for example, starting with a realist approach, a researcher can begin analyzing the data available and directly measurable and only after may move to more empirical inquiries, this double process helps explore and improve his knowledge of the real world.

### 1.2.1. The rise of mixed methods

After addressing the theoretical base of the possible merging between quantitative and qualitative methods for research, it's important to also show how this mixed method can be used to cope with the weaknesses of qualitative and quantitative methods using the complementary strengths of each.

Kelle, after studying the evolution of the aforementioned "paradigm wars" from the 1980s, came to the conclusion that, if the arguments developed by followers of one school in order to highlight the methodological problems of the competing school are just disregarded as obsolete, important issues could be overlooked. (Kelle, 2006)

The problem in the past literature has been a chauvinism on the researchers' own doctrines. This mindset has led to a lack of answers regarding arguments stressing methodological limitations of both qualitative and quantitative research. A researcher siding for a particular doctrine tended to answer a problem of his tradition by emphasizing problems of the other tradition. In this way problems that could have been solved using a dialectic approach were simply neglected. This has led to a situation where the potential of quantitative and qualitative methods to cope with problems of the competing method has not been utilized. (Kelle, 2006)

The first problem of this kind that can be analyzed is one that was already mentioned by renowned sociologists such as Herbert Blumer: quantitative researchers are blamed for their alienation from the world they are investigating (Blumer, 1940).

To put it in Filstead's terms (1970), this kind of research would lack a "firsthand involvement with the social world" indispensable for any adequate understanding of social phenomena.

To understand the fundaments of this problematic we must analyze the methodological concept underlying more sophisticated forms of quantitative research: the Hypothetico-Deductive (HD) method. As the name suggests, it consists of a use of two operations: the formulation of hypotheses and the deduction of consequences from them in order to arrive at conclusions which - though hypothetical - are well supported. (Føllesdal, 1979)

A pivotal argument against this approach lies in its definition: the hypotheses in these systems are not justified "from above", as they are in an axiomatic system, instead, they are justified from below, through their consequences (Føllesdal, 1979). Therefore, in order to create a legitimate hypothesis a researcher must have an understanding of social phenomena, that requires knowledge about patterns, structures and rules characteristic of the particular social life world that is under his lens. Such information usually forms an integral part of culturally specific stocks of knowledge.

However, social researchers that are using an HD approach often employ their personal commonsense knowledge (Kelle & Lüdemann, 1998). This *heuristic of commonsense knowledge* is particularly dangerous if the background of the researcher and the research itself differ to a certain degree. For example if the researcher belongs to a different gender, social class or ethnicity he will not have access to culture-specific stocks of knowledge to formulate hypotheses and define variables; this in turn may result in problems of theory building and hypothesis construction, leading to mis-specification of statistical models (Kelle, 2006).

This shortcoming of quantitative methods can be easily dealt with by applying qualitative techniques in conjunction with the quantitative ones already in use. Kelle applied this reasoning in a panel study carried out to investigate the status passage from school to the labor market in Germany in the 1990s (Heinz et al., 1998; Kelle & Zinn, 1998). The researchers administered a survey to all school leavers in two cities that had just started vocational training in two defined occupations to collect data about their future occupational careers. From the large quantitative sample, a smaller one (*n=120*) was drawn to conduct qualitative interviews on, focusing on work experiences, aspirations and reflections on careers.

By only analyzing the quantitative data, it was apparent that there was a lack of understanding about causality and relevant information about the subjects' occupational

life. Kelle however found that by using the qualitative data gathered, they were able to obtain a clearer understanding about the processes underlying the association between a certain occupation and a level of occupational aspirations. Furthermore, by using the access they obtained to the local knowledge of occupational cultures provided by the qualitative interview data, the researchers were able to add two additional context variables to the initial model, making it more realistic and further improving its explanatory power.

Another potential flaw in quantitative research methods is that the lack of what makes qualitative interviews "complex social and organizational phenomena rather than just a research method" (Qu & Dumay, 2011): a more direct human contact. Even the best constructed questionnaire may generate an invalid or misleading result if the research subjects understand a question in a different way than the researchers want them to, or if they don't consider the topics treated relevant enough.

The interview cannot be regarded as a mere exchange of information between the interviewer and an interviewee seen as a "passive provider of data" (Kelle, 2006). It is indeed a complex process of social interaction where the interviewee can act according to his own motives and goals and can hide his intentions, hold back information, invent it and so forth. Furthermore, an interview is highly dependent on the skill of the interviewer to interpret the action and motives of his counterpart.

It is indeed clear that misunderstandings and accidental mistakes as well as willful omissions or falsities can pose a significant threat to the validity and quality of data coming from interviews (Lillis, 1999).

Kelle treated this weakness of the qualitative methods in another one of his studies, where data and results taken from a research about the satisfaction of care homes' residents show how these threats for validity can be identified and treated with mixed methods designs (Kelle & Niggemann, 2002, 2003). The purpose of this study was to measure the satisfaction and needs of care home residents. It was composed of semi-structured interviews ($n$=40) and standardized questionnaires ($n=244$) administered to different subjects in care homes in the whole of Germany of varying size and ownership.

The researchers noticed that there were strong divergences between the data collected in quantitative and qualitative manner: data collected with the questionnaires showed in fact a very positive image about the residents' satisfaction in opposition to the

qualitative data collected during interviews. The findings of the research determined that it was difficult for respondents to express negatively or judge badly the actors of the institution they occupy, however qualitative methods were able to bring interviewees out of their initial reserve. If a relationship that involved mutual trust and understanding developed between the interview partners, even respondents who were extremely cautious in the beginning were willing to report about negative experiences, thus crucially enriching the data quality.

This study can lead to the conclusion that certain quantitative instruments are not improvable to the point that they will perfectly explain human behavior. Consequently, the use of qualitative methods may yield useful 'negative information' about the validity of the data gathered.

Having already showed two improvements that mixed methods bring over the use of a single one, Kelle moved to show the possible solutions for the traditional complaint that statisticians and sociologist like Lundberg (1941) always posed to qualitative studies: they do not provide a basis for sound generalizations because of the lack of representativeness of the small *n* studies that characterize them.

The argument against the generalizability of qualitative studies started in the 1930s when Znaniecki (1934) made a distinction between "analytic induction" and "enumerative induction", which is the ordinary statistical way of studying relationships with correlation. He argued that in qualitative research a form of generalization superior to statistical inference had to be used and its dependence was not on the number of cases, "but on the strength of the theoretical reasoning" (Seale, 1999).

This concept is called "theoretical generalization" (Tsang, 2014) and the idea behind it is that there is a social process or structure at work in every single case that can be proved with a detailed and incisive theoretical analysis.

The term "transferability" is used in literature (Polit, 2010) to address questions rising from the concerns relative to the possible scope of such qualitative studies: its critics suggest that this kind of reasoning is too close to the idea of generalizability.

The contribution of mixed studies to this problem has been suggested even in classical writings (Barton, 1955) and it consists in testing and examining through large-scale quantitative surveys the findings from qualitative studies with small numbers of

observations in a limited domain. This qualitative inquiry, however, would only yield valuable results if effected in a very thorough and not resource-efficient way.

In the subsequent decades other researchers have tried to address this problem through mixed methods. Glaser and Strauss (1967) presented the concept of "theoretical sampling" as "a process of data collection for generating theory whereby the analyst jointly collects codes and analyses data and decides what data to collect next and where to find them, in order to develop a theory as it emerges" (Glaser and Strauss, 1967).

According to these theorists, quantitative research could be used to maximize or minimize the differences in a sample or provide an overview about the existence or the distribution of certain types of social patterns in the investigated domain.

Consequently, a quantitative study can provide a sampling that allows the researchers to manually select the cases they want to expand the research on, be they typical, deviant or extreme.

Mollenkopf and Baas (2002) applied this method in their study about the mobility and health situation of elderly people. They obtained standardized data by administering questionnaires to a large sample of respondents and, on the basis of these data, they were able to identify four types of elderly persons based on their health and mobility levels.

The researchers were able to carry qualitative in-depth interviews with a small number of respondents in each group in order to identify how they experienced their health and mobility situation. Having previously classified the different types of subjects, they had the possibility to scrutinize more intensively the people with considerable health problems and high mobility, as members of this group had developed successful coping strategies that the researchers were interested in.

After having showed what different researchers have accomplished through mixed methods studies, we can appreciate the various technical approaches that this research methods offer.

The first is a sequential qualitative-quantitative design (*qual* > *quan*), where a qualitative study has the aim to identify the core issues at hand and to develop the hypotheses and the theoretical bases, which are to be further examined in a successive quantitative study, carried out in order to check if the concept that are deemed relevant in a smaller number of cases are able to explain social phenomena in a greater domain.

This design helps to cope with two of the flaws of single methods mentioned above: the limited transferability of conclusions from qualitative researches with a small *n* as well as the initially mentioned risks of the heuristic of commonsense knowledge, plaguing quantitative Hypothetico-Deductive research. In fact, by starting with a qualitative study, researchers can obtain access to local knowledge that is key to develop the relevant hypotheses and concept to test on a greater scale.

A second way to combine the methods is the opposite of the first: it starts with a quantitative study, followed by a qualitative one. In such a sequential quantitative-qualitative design (*quan* > *qual*) a quantitative research is carried in order to better narrow the problem areas and research questions to further investigate with the help of qualitative methods and data. This design helps to cope with two problems of quantitative research: the difficulty to understand the quantitative data without the proper sociocultural knowledge and the doubt that the research is focusing only on remote or marginal cases.

A third and last design can fulfil similar function to the aforementioned two, but with its series of restrictions and benefits: the parallel qualitative-quantitative design (*qual* + *quan*). Its qualitative part can yield information that explains statistical associations, identify new variables and develop additional clarifications. However, since the two studies are conducted in parallel, the quantitative study cannot benefit from the information already retrieved in the qualitative one and vice versa. A great benefit of this design is that, since it interviews the same subjects with two different techniques, it can help the researchers to identify measurements problems and artifacts of both quantitative and qualitative data. (Kelle, 2006)

In conclusion, quantitative and qualitative methods can fulfill distinct yet complementary purposes within mixed-method designs. Quantitative methods are suited to give an overview about the matter under study and can describe its subjects on a macro level, whereas qualitative methods can be utilized to tap into the local knowledge in order to develop grounded hypotheses that cover relevant phenomena.

Qualitative and quantitative methods can be used to answer completely different questions: the results of statistical analyses show what kinds of actions subjects typically perform, while the analysis of qualitative data helps to answer why the subjects perform those actions and what the motivations behind them are.

The best use of these two methods is not as substitutes but in conjunction, to overcome each other's shortcomings.

## 1.3. Marketing research in the web context

With the advent of the Internet and the World Wide Web, the collection of data analyzable with traditional methods of research, both quantitative and qualitative, has changed. Access to user-generated content is allowed in an immediate and spontaneous way: this results in the possibility to gather information about the opinion of a population and the expression of subjective and personal ideas of many subjects. (Miller, 2006)

The web and especially the social networking platforms offer to researchers large amounts of data. Even for qualitative research, traditional tools such as focus groups and questionnaires, when carried out online, allow to collect considerably larger volumes of data in a considerably shorter time. (Kaden, Linda & Prince, 2011)

However this evolution of research is not without its drawbacks: as Kotler (2010) explains: the people that have access to the internet and actually use it to express their opinions and thoughts about a relevant subject for research are not certainly representative on the entire population the researcher should refer to. A marketing research carried out online is not suitable for every kind of product: its representation is closely linked to variables such as the degree of computerization of the population, the type of consumers who use social media to communicate and the type of digital platform that is taken into consideration in the survey.

Another aspect to consider when approaching a web analysis is that, unlike offline survey, messages posted online (referred to as *User Generated Content – UGC*) are written spontaneously by users and received to the researcher in a dirty and unordered manner (referred to as *online chatter*). It is nonetheless important to highlight the value and at the same time the limitations of this type of data: on the one hand, their unstructured nature requires a greater effort than the typical offline survey that follows a standardized script; on the other hand, the spontaneity of the data received by the user and not

addressed by the researcher allows to collect free and unguided opinions, possibly revealing links and information not initially considered. (Tirunillai & Tellis, 2014)

In this context, we will show sentiment analysis as a synthesis between the quantitative method and the qualitative method in an online search context, that is, the extraction of information from the Internet, with the awareness of the limits, but also of the advantages that a survey of an online-only population obviously presents.

## 1.4. Sentiment analysis as a synthesis between qualitative and quantitative

The objective of this research work is to present sentiment analysis as a method capable of addressing the limits of quantitative and qualitative research methods, leveraging of the advantages of the aforementioned online marketing research and, in particular, the content spontaneously produced by users on social networking platforms.

The analysis of the sentiment of the Internet population collects types of information that are typical of qualitative research, namely feelings, impressions and consumer opinions about a brand, product or message; on the other hand, it removes the limits of those same methods by minimizing subjectivity and taking advantage of the rigor of statistical tools that are typical of quantitative methods (Rambocas, 2013).

While still keeping in mind the issues in representativity of only considering an online population, the application of this type of analysis to the web context allows the access to a massive amount of data, that allows to analyze a broader sample than traditional methods.

A key advantage of sentiment analysis is that, while it can access vast amount of data as already mentioned, it also allows a "human factor" to access it. As Ceron, Curini, and Iacus (2014) explained in their publication, sentiment analysis does not work by using the brute force of calculators in order to extract information from a text, it doesn't just count the number of times that a word appears, the number of likes and so on. This type of analysis uses the data coming from different sources in a qualitative way, through which a researcher can further explore a text to really extract the meaning (or *sentiment*),

in the same way he would in an analysis of focus groups on the merits and defects of a product, as opposed to a mere sales count analysis.

Sentiment analysis is in fact defined as the analytical measurement of an author's opinion about specific entities (Feldman, 2013).

This type of analysis is closely linked to the concept of *opinion mining*, a term first introduced by Dave et al. (2003) to indicate a technique that can process a keyword search and identify attributes (positive, neutral, negative) for each term, so that once the distributions of these terms are aggregated, it becomes possible to extract the opinion associated with each key term.

It is however crucial to understand that any purely quantitative linguistic model is wrong at the start, because of the basis of how language itself works. Every single sentence, however thoughtful and well-constructed, can drastically change its meaning with the inclusion of even the slightest variation. (Ceron, Curini, and Iacus, 2014). For example including a sarcastic hashtag at the end of a tweet can dramatically change its meaning, or the position of a comma in a phrase can completely revert its message, like in the famous quote attributed to a Latin sybil: *"Ibis, redibis numquam peribis in bello"*, that can be translated in either *"you will go, you will return, never in war will you perish"* or *"you will go, you will never return, in war you will perish"* based on the position of the second comma (after *redibis* or after *numquam*).

Fundamentally, the complexity of language is so high that any completely automatic method to interpret it is deemed to fail, nonetheless these methods are the key to find and interpret possible recurrences in themes and words that, with a final human scrutiny, can allow large scale analysis on potentially thousands of texts (and millions of words) at once.

## 1.5. Natural language processing

When talking about sentiment analysis we enter the field of *Natural Language Processing* (*NLP*), the sector of Artificial Intelligence and computer sciences that deals with the relationship of computers with human language, or "natural language". Natural

language understanding is, in fact, the main challenge of AI and deep learning techniques, through which it is possible to teach machines to recognize words, understand texts and communicate with humans (Bates, 1995).

The core of any NLP work is the important topic of natural language understanding. There are three major problems involved in the process of creating computer programs with the objective of understanding human language: the first pertains to the thought processes in use, the second to the representation and meaning of the linguistic input, and the third to the world knowledge necessary to understand it. As a result, an NLP system may begin at the single word level in order to establish the morphological structure and nature of the word; and then may move on to the sentence level to determine the word order, grammar, and meaning of the entire sentence; and finally to the context and the overall context or domain. A given word or sentence may have a specific meaning or connotation in each environment and may be related to many other words or sentences in the given context. (Chowdhury, 2003)

In order to understand natural languages, it is important to comprehend that there are seven interdependent levels (Liddy, 1998, 2001) that people use to extract meaning from language:

- *Phonetic* or phonological level that deals with the pronunciation of words. It specializes in the interpretation of sounds within and across words and in doing so it uses three types of rules in its analysis: 1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence.
- *Morphological* level that deals with the smallest components of words that carry meaning. It analyzes morphemes, the smallest units of meaning; since the meaning of each morpheme (like prefixes and suffixes) remains the same across words, humans, and consequently an NLP system, can recognize the meaning of each in order to represent the meaning of the word.
- *Lexical* level that deals with the lexical meaning of individual words. It is composed of several types of processing in order to arrive to a word-level understanding. The first of these is the assignment of a single part-of-speech tag to each word so that words that have more than one lexical function

are given the most probable part-of-speech tag based on the context in which they occur. At the second processing, the words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning based on the semantic theory utilized in the NLP system. The lexical level many times requires a proper lexicon to work, but its nature and the extent of information included in it may vary. Lexicons can fluctuate between a simple version composed only by words and their parts-of-speech, and increasingly complex version containing information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments.

- *Syntactic* level that deals with the grammar of words and structure of sentences. It focuses on uncovering the overall grammatical structure of a sentence. The output of this level of processing is a schematized representation of the sentence that reveals the structural dependency relationships between the words.

- *Semantic* level that deals with the meaning of both words and sentences. It centers on the interactions between word-level meanings in the sentence and it allows to disambiguate words with multiple senses in order to allow one and only one meaning of polysemous words to be selected and included in the semantic representation of the sentence.

- *Discourse* level that deals with units of text longer than a sentence. Instead of interpreting multi-sentence texts individually as separated blocks, it rather focuses on the properties of the text as a whole to convey meaning throughout the existing connections between the sentences that compose it. One of its most common uses it the anaphora resolution, meaning the replacing of semantically vacant words such as pronouns with the appropriate entity to which they refer to (Mitkov, 2014). The discourse level is also used to recognize the discourse or text structure in order to determine the function of different sentences for the purpose of better representing the whole text.

- *Pragmatic* level that deals with the knowledge that comes from the outside world, namely, from outside the content of the document. Its goal is to explain how extra meanings are included in a text without actually being physically encoded into them.

Since humans use all of these level in their communication, a natural language processing system will be more capable the more of these seven levels it will include in its analysis.

# 2. Sentiment analysis, theory and methodology

## 2.1. **Definition of sentiment analysis**

After having briefly defined sentiment analysis as the measurement of the opinion of an author in a determined text (Feldman, 2013), we can deepen its study with the aim of better comprehending the basis behind it.

In order to understand sentiment analysis, we must begin from the analysis of textual information; it can be broadly divided into two main types: *facts* and *opinions*, where *facts* are objective expressions and information about entities and their characteristics and *opinions* are subjective expressions that describe people's sentiments and ideas about entities and their characteristics (Yu, Hatzivassiloglou, 2003).

Before the *World Wide Web*, the study of sentiment was mainly focused on information mining instead of opinion mining, for the simple fact that opinionated text was much more difficult to come by prior to the advent of the internet. The Web has in fact dramatically changed the way people express themselves: now it's easy and practically free to express a personal view about anything by posting it on forums, brand sites or social media (the *User Generated Content* already mentioned). It is safe to say that if a consumer wants to buy a new product one of his main sources of opinions will be product reviews and judgments on the Web (Liu, 2010).

However, since the volume of opinionated text is ever expanding, it is increasingly difficult to monitor opinions and views on the Web without any automated aid. It would be a formidable task for a human to find relevant sources online, extract the opinions in them, read them and, more importantly, give structure to them and summarize their overall result. Sentiment analysis grows out of this necessity.

The commercial value of its practical applications can be seen by making a simple query on a search engine for "*sentiment analysis services"*: dozens of companies offer these services to the enterprise level, including leading firms such as *Microsoft* and *Hitachi*.

To give a proper definition to sentiment analysis we must start from is components and what it aims to study: the opinion mining.

In particular, "an *object* (*o*) is an entity which can be a product, person, event, organization, or topic. It is associated with a pair, *o*: (*T*, *A*), where *T* is a hierarchy of *components* (or *parts*), *sub-components*, and so on, and *A* is a set of *attributes* of *o*. Each component has its own set of sub-components and attributes." (Liu, 2010)

In literature the term *feature* is commonly used to represent both components and attributes in order to simplify the analysis (Ding & Liu, 2007); in this view the object itself is also seen as a feature and an opinionated comment on it is called *general opinion*, while an opinionated comment on any of its specific features is called a *specific opinion*.

Any feature in a text can be *explicit* or *implicit*: "if a feature *f* or any of its synonyms appears in a sentence *s*, *f* is called an *explicit feature* in *s*. If neither *f* nor any of its synonyms appear in *s* but *f* is implied, then *f* is called an *implicit feature* in *s*." (Liu, 2010)

Continuing to follow Liu's (2010) definitions, an *opinion* on a feature *f* is a positive or negative view, sentiment or attitude on *f* expressed by an *opinion holder* – the person or organization expressing the opinion. The opinion has an *orientation*, that indicates whether it conveys a positive, negative or neutral message.

Piecing the abovementioned definitions and concepts together, Liu, jointly with other researchers, was able to define a model of an object, an opinionated text and the overall mining objective, which are collectively referred to as the *feature-based sentiment analysis model* (Wang, Liu, Song, & Lu, 2014).

"Model of an object:

An object *o* is represented with a finite set of features, $F = \{f_1, f_2, \ldots, f_n\}$, which includes the object itself as a special feature. Each feature $f_i \in F$ can be expressed with any one of a finite set of words or phrases $W_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$, which are synonyms of the feature, or indicated by any one of a finite set of feature indicators $I_i = \{i_{i1}, i_{i2}, \ldots, i_{iq}\}$ of the feature.

Model of an opinionated document:

A general opinionated document *d* contains opinions on a set of objects $\{o_1, o_2, \ldots, o_q\}$ from a set of opinion holders $\{h_1, h_2, \ldots, h_p\}$. The opinions on each object $o_j$ are expressed on a subset $F_j$ of features of $o_j$. An opinion can belong to any one of the following two types:

1. Direct opinion:

A direct opinion is a quintuple ($o_j$, $f_{jk}$, $oo_{ijkl}$, $h_i$, $t_l$), where $o_j$ is an object, $f_{jk}$ is a feature of the object $o_j$, $oo_{ijkl}$ is the orientation or polarity of the opinion on feature $f_{jk}$ of object $o_j$, $h_i$ is the opinion holder and $t_l$ is the time when the opinion is expressed by $h_i$. The opinion orientation $oo_{ijkl}$ can be positive, negative or neutral. […]

2. Comparative opinion:

A comparative opinion expresses a relation of similarities or differences between two or more objects, and/or object preferences of the opinion holder based on some of the shared features of the objects.

Objective of mining direct opinions:

Given an opinionated document *d*,

1. discover all opinion quintuples ($o_j$, $f_{jk}$, $oo_{ijkl}$, $h_i$, $t_l$) in *d*, and
2. identify all the synonyms ($W_{jk}$) and feature indicators $I_{jk}$ of each feature $f_{jk}$ in *d*." (Liu, 2010)

As the definitions show, the process of sentiment analysis mainly consists in deriving structured, qualitative data from unstructured, raw texts. In fact, the data coming from the comments is divided in quintuplets where each variable is numerically measurable and capable of being stored in databases and tables; in turn, the data coming from these tables can be visualized to gain insights on opinions and sentiments.

For example, among the possible outputs of a query from a sentiment analysis database there is a feature buzz summary, that shows the frequency of feature mentions, telling a company what are the most talked-about and mentioned features; an object buzz summary, that works similarly, but it instead focuses on objects in order to possibly compare different products, and finally trend tracking, that creates reports of variables' trends based on time.

One of the most important research topics in sentiment analysis is *sentiment classification*. Since the early 2000s (Pang, Lee, & Vaithyanathan, 2002) the research has started applying machine learning techniques in order to classify an opinionated document - mostly product reviews (Cui, Mittal, & Datar, 2006), as expressing a positive or negative opinion.

The operation of classifying a sentence as expressing a positive or negative opinion is called *sentence level sentiment classification* and it can be defined as following:

"Given an opinionated document *d* which comments on an object *o*, determine the orientation *oo* of the opinion expressed on *o*, i.e., discover the opinion orientation *oo* on feature *f* in the quintuple (*o, f, so, h, t*), where *f = o* and *h, t, o* are assumed to be known or irrelevant." (Liu, 2010)

Before measuring the sentiment of a sentence, it is also important to determine if what is under analysis is subjective or objective, in order to filter out phrases that contain no opinion from the measure. It can be done using traditional supervised learning methods (Wiebe, Bruce, & O'Hara, 1999) to classify single parts-of-speech inside sentences.

The existing techniques for sentiment classification are based mostly on supervised learning, although there are some that use unsupervised methods.

## 2.2. Automated Learning

### 2.2.1 Machine Learning

Machine learning is the set of techniques that allows a machine to learn and perfect itself in a certain skill. The *machine* is defined as a computer or software that uses the algorithms to whom the skill is taught. Machine learning is also called *automated learning* as algorithms become able to perform the task automatically and independently of the instructions of a human researcher, learning instead from the data itself. (Bishop, 2006)

The field of machine learning is characterized by two main types of tasks: supervised, and unsupervised. The key difference between the two types is that supervised learning is accomplished using a *ground truth*, i.e. having prior human-made knowledge of what the output values for the samples should be. Consequently, its goal is to learn a function that, given a sample of data and intended outputs, best approximates the relationship between input and output observable in the data. Unsupervised learning, in contrast, does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points. It then looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision. (Hinton, 1999)

### 2.2.2 **Supervised and unsupervised learning methods**

Since most of the existing techniques for sentiment classification are based on supervised learning (Liu, 2010), it will be explored first in this elaborate.

The goal of supervised machine learning is to build a brief model of the distribution of classification labels in terms of predictor features. Its output is a classifier used to assign class labels (positive or negative) to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. (Kotsiantis, Zaharakis, & Pintelas, 2007)

The main task of sentiment classification is to construct an appropriate set of features among which the most used (Pang & Lee, 2008) are:

- *Terms and their frequency*: these features are the individual words or word n-grams and the number of times they appear; in some cases, the position of single words is also taken into account.
- *Part-of-speech tags*: these features, often referred to as POS tags, are labels assigned to each token (word) in a text corpus to indicate its part of speech and often also other grammatical categories such as tense, number, case, etc.
- *Opinion words and phrases*: *Opinion words* are tokens generally employed to express positive or negative sentiments. For example, "*nice"*, "*beautiful"* and "*great"* are positive opinion words, whereas "*terrible"*, "*bad"* and "*nasty"* are negative opinion words. Verbs (like "*love"* and "*hate"*) and nouns (like "*trash"*), although less used, are still a way to convey opinions. In addition to individual words, phrases and idioms are also commonly used to express opinions (like "*dead in water"*).
- *Negation*: the appearance of these feature can completely change the opinion of the text, for example the phrase "*I do not like this product"* is negative just because of the introduction of *not*. However, the inclusion of a *Negation* does not necessarily deem the phrase as having a negative opinion; for example, the phrase "*I could not be happier with this product"* is extremely positive.

Unsupervised learning, on the other hand, works without giving any previous assumptions and definitions to the model about the outcome of variables it analyzes, it processes the data on its own trying to find a model behind it. It is extremely useful in cases when there is no pre-labeled data, or the structure of the data is not certain, and the researcher wants to learn more about the nature of process under analysis, without making any previous assumptions about its outcome. (Brody & Elhadad, 2010)

An established method of this kind was proposed by Turney (2002) and the algorithm he presented is split into three steps.

The first step consists of the extraction of phrases containing adjectives or adverbs. It does so because research (Rittman et al, 2004) has shown that adjectives and adverbs are generally good indicators of subjectivity and opinions in phrases. Nevertheless, even if a single adjective has the power to indicate subjectivity, it may not be enough to determine the opinion orientation of its phrase, therefore the algorithm extracts two consecutive words, where one is an adjective and the other is a context word.

The second step consist in the estimation of the orientation of the extracted phrase using the pointwise mutual information (PMI) measure given in the following equation:

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

In the formula, $Pr(term_1 \wedge term_2)$ is the co-occurrence probability of $term_1$ and $term_2$, and $Pr(term_1)Pr(term_2)$ gives the probability that the two terms co-occur if they are statistically independent. The ratio between $Pr(term_1 \wedge term_2)$ and $Pr(term_1)Pr(term_2)$ is in consequence a measure of the degree of statistical dependence between them. The *log* of this ratio is the amount of information that is acquired about the presence of one of the words when the other is observed.

The opinion orientation (*oo*) of the phrase is then constructed measuring its association with the positive reference word "*excellent*" and with the negative reference word "*poor*":

$$oo(phrase) = PMI(phrase, \text{“}excellent\text{”}) - PMI(phrase, \text{“}poor\text{”}).$$

The probabilities are finally calculated by making queries to a search engine with the objective of collecting the number of *hits* (number of relevant documents to the query). This way, by computing a search of two terms both together and separately, it is possible to estimate the probabilities in the first equation.

Turney used a search engine that could compute the *NEAR* operator, which is able to constrain its search only to documents that contain the words within ten words of one another. Considering *hits*(*query*) as the number of hits returned, the final equation can be reworked as:

$$oo(phrase) = \log_2\left(\frac{hits(\text{phrase } NEAR \text{ "excellent"})hits(\text{"poor"})}{hits(\text{phrase } NEAR \text{ "poor"})hits(\text{"excellent"})}\right).$$

Thus, the final step of Turney's methods is to compute the average *oo* of all the phrases in the given text and classify the overall document as positive if the average *oo* is positive or negative otherwise. (Turney, 2002)

This work will however adopt the supervised learning method, since it has been established as one of the most successful techniques, when combined with machine learning, to accomplish sentiment analysis (Pang, Lee, & Vaithyanathan, 2002).

Supervised classification methods generally require more control by the hand of the researcher, as they require the final classes to which the elements are assigned (the sentiment level in this particular case) to be known before proceeding in the analysis. Specifically, for this sentiment analysis, texts that have previously been labeled will be used, that is, parts-of-speech to which a sentiment value has already been assigned (positive, negative or neutral, following a basic classification).

When conducting supervised learning, the main problems to keep in check are *model complexity* and the *bias-variance tradeoff*. Both of these are interrelated.

The *model complexity* is referred to the complexity of the function that the algorithm is attempting to learn. The appropriate level of model complexity is determined by the nature of the training data at hand. If the amount of data is modest, or if it is not uniformly spread throughout the various possible classes, then the model of choice should be a low-complexity one. This is because a high-complexity model would *overfit* if applied to a dataset with a small number of data points. (Briscoe, & Feldman, 2011)

*Overfitting* refers to learning a function that corresponds too closely or exactly to the data used to train it, but does not generalize to other data points (Everitt, Skrondal, 2010), in other words, the algorithm is strictly learning to produce the training data without learning the actual trend or structure in the data that leads to this final output.

The *bias-variance tradeoff* is also referred to model generalization. In fact, any model is balanced between *bias*, which is the constant error term, and *variance*, which is the amount by which the error could vary between different training sets. For example, an high bias and low variance model would be systematically wrong 25% of the time, whereas a low bias and high variance model would be wrong anywhere from 10% to 50% of the time, depending on the dataset that was used to train it. Typically, bias and variance move in opposite directions of each other; an increase in bias will usually lead to a decrease in variance, and vice versa. (Briscoe, & Feldman, 2011)

As with all machine learning techniques, the process can be substantially broken down into 3 steps:

1. *The training phase:*

   A defined set of textual elements, subset of the entire population (i.e. the total sum of texts that will be analyzed), is used to teach the classification model to infer the rules that allow each element to be given the correct label.

2. *The test phase:*

   The rules created in the previous phase are used to classify items that are not part of the previous subset (i.e. those that are not yet classified).

3. *The implementation phase:*

   The resulting model is finally used to classify the new elements, that is, the future textual content that the researcher would like to analyze. The model will then use the rules generated during the training phase and validated during the test phase. (Gama and de Carvalho, 2009)

When applying this model to sentiment analysis the classification becomes about assigning a label (class) to a part-of-speech that is able to describe its sentiment (either positive, negative or the in-between). The final objective of the use of machine learning in this case is to build a model that is able to append said label in an automatic way, without the aid of a researcher.

The generated model can be considered predictive, in the sense that it can correctly predict the sentiment of the text (the predefined class it belongs to) from a series of characteristics possessed by the text itself (i.e. its features, or explanatory variables).

The classification is thus defined as supervised because the possible classes are decided ex ante by the researcher, who governs the investigation by defining at the outset what dimensions the analysis is going to use. Using a supervised approach therefore requires that, in the training phase, the model uses on a previously classified dataset (i.e. text with a sentiment label already assigned) to train on.

### 2.2.3 **Classification techniques**

Predictive classification modeling can be defined as the task of approximating a mapping function (*f*) from input variables (*x*) to discrete output variables (*y*) (Kotsiantis et al, 2007).

There is a wide variety of machine learning techniques that are commonly used in supervised classification tasks. Using a supervised leaning approach in sentiment analysis in fact requires a data corpus, which serves as a preparation document for classification learning. The classification, in turn, can be executed in different ways based on the theorems applied.

The basic functions available for classification include: *Naïve Bayes*, *Support Vector Machines*, *Decision Trees* and *Maximum Entropy*.

A *Naïve Bayes* classifier is a probabilistic classifier based on applying Bayes' theorem that assumes that attributes are conditionally independent. In fact, its key difference from other classifiers is that Naïve Bayes assumes that the features are independent of each other and have no type of correlation between them. However, as easily imagined, this is not the case in real life. This naïve assumption of features being uncorrelated is thus the reason why this algorithm is named "naïve".

We can describe how the classifier operates starting by defining *P(x)* as the probability that an event *x* occurs: it is calculated as the number of the desired outcome divided by the total number of outcomes. Conditional probability, on the other hand, is

the likelihood that an event *x* occurs given that another event *(y)* that has a relation with event *x* has already occurred. The probability of event *x* given that event *y* has occurred is denoted as *P(x|y).* Finally, a joint probability is the probability of two events occurring together and is denoted as *P(x) × P(y).*

Bayes' Theorem can be thus defined as:

$$P(x|y) = \frac{P(x) \times P(y)}{P(y)}$$

Or the probability of event *x,* given that event *y* occurs equals to the probability that *x* and *y* occur together divided by the probability of *y*.

This classifier is constructed based on the frequency of occurrence of each feature per class in the training data set. And under the assumption of features being independent, *P(x1, x2 , … , xn | yi)* it can be written as:

$$P(x1, x2, \ldots, xn|y_i) = P(x1|y_i) \times P(x2|y_i) \times \ldots \times P(xn|y_i)$$

The classification is conducted by deriving the maximum posterior which is the maximal *P(x1, x2,…xn|X),* applying the Bayes theorem assumption. This assumption greatly reduces the required calculations by only counting the class distribution and not their interdependence. Even though the assumption is not valid in most real cases since the attributes are dependent, Naïve Bayes is able to perform impressively in a number of different contexts (Lewis, 1998).

This classifier has some advantages and disadvantages over its substitutes: firstly, the assumption that all features are independent makes its algorithm very fast compared to others, therefore it is prone to works with high-dimensional data such as text classification or spam detection. On the downside, because of the aforementioned assumption it is less accurate than other algorithms (Rish, 2001).

The *Support Vector Machines* classifier is based on the statistical learning theory (Vapnik, 1995): in short, it creates a line or *hyperplane*[1] which separates the data into different classes.

Binary classifiers like SVMs attempt to find a separating line or a hyperplane between different types of data so that it maximizes the separation margin between observations from different classes.

The objective of an SVM algorithm, after analyzing the data and drawing a number of possible separator lines, is to find the data points closest to the lines from both the classes. These points are called *support vectors*. After this process, it computes the distance between the lines and the support vectors. This distance is called *margin*, the algorithm's goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane. Thus, SVM try to make a decision boundary in order for the separation between the two data classes to be as wide as possible.

When the data is bidimensional, the separating line adopted is a hyperplane, whereas for three-dimensional data a plane with two dimensions divides the 3d space into two parts and thus act as a hyperplane. This rule can be expanded, stating that for a space of n dimensions a hyperplane of n-1 dimensions is able to separate it into two parts.

SVMs have two tuning parameters: *C* and *Gamma*.

*C* controls the trade-off between having a smooth decision boundary and classifying the training points correctly. When the value of *C* is high it means the algorithm will get more training points correctly.

*Gamma* defines how much the influence of a single training example is considered. If it has a low value it means that every data point has a far reach and conversely high value of gamma means that every point has close reach (Vapnik, 1995).

This classifier has its share of advantages and disadvantages: it can work well when the datapoints have a clearer margin of separation or when the number of dimensions is greater than the number of objects; on the other side, since its nature requires clear separation between data, it can find difficulties when operating in large or noisy datasets (Karamizadeh et al, 2014).

[1] A *hyperplane* in an n-dimensional Euclidean space "is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts" (Curtis, 1968).

The *Decision Tree* classifier builds classification models in the form of a tree structure: it creates a diagram-like structure in which each internal node represents a test on a feature (e.g. whether a word is subjective or objective) , each leaf node represents a class label (the decision taken after computing all features) and the single branches represent the features that have led to those class labels. The paths from the root to the single leaf represent classification rules.

A DT utilizes an *if-then* rule set which is mutually exclusive and comprehensive for the classification at hand. The rules that sequentially split the data are learned in succession using the training data. Each time that a rule is added, the tuples that it covers are removed and this process goes on until no more data is left. (Quinlan, 1986)

The tree is constructed in a top-down recursive manner, it requires all attributes to be categorical or at least discretized in advance. The attributes at the top of the tree have more impact towards in the classification.

Being a simpler model to represent makes the DT classifiers easier to understand and their nature makes them resistant to outliers, so they require less data preprocessing. On the negative side, these classifiers are prone to overfitting and thus to generating a high number of useless branches, so they require the researcher to *pre-prune*, which halts tree construction early, or *post-prune*, which removes branches from the fully grown tree in the end (Bramer, 2007).

Finally, *Maximum Entropy* classifiers are used in response to a weakness of the Naïve Bayes classifiers: their performance depends on the degree to which the features are independent, but the datasets rarely are, hence they may result in poor performances.

The Maximum Entropy classifier does instead use mutually dependent features to positively classify data. This classifier is based on the idea that all that is known should be modeled and nothing should be assumed about the unknown. To accomplish this goal, this method considers all the classifiers that are consistent with a training dataset (i.e all the models that fit the data), then chooses the classifier that maximizes entropy out of them. (Nigam et al, 1999)

Due to its nature, this classifier is useful when no prior information is known about the data that will be analyzed, but on in parallel it requires more time to train comparing to other classifiers (Marouli, 2014).

## 2.3 **Opinion lexicon generation**

Opinion words, also referred to as polar words, opinion-bearing words, and sentiment words, are one of the main protagonists of the sentiment classification task. Apart from individual words there are also opinion phrases and idioms. Together they are called opinion lexicon and can be divided into two types, the base type and the comparative type. Opinion lexicon belonging to the comparative type is used to express comparative and superlative opinions and uses words like *better* or *worse*. Unlike opinion lexicon of the base type, the words of the comparative type do not express a generic sentiment on an object, but a comparative opinion on more than one object, so they can be more difficult to measure. (Liu, 2010)

The collection of opinion lexicon and generation of opinion list can be accomplished using three different approaches: *manual approach*, *dictionary-based approach*, and *corpus-based approach*.

The *manual approach* is the simplest and most time consuming: for a set of opinions belonging to a specific category, the researchers give a score to each, with a high score indicating a high possibility of its being a typical opinion for the category. This gives the user a simple overview of tendencies in the original opinion data (Morinaga, 2002).

This method is mostly used in combination with the other two because of its laborious nature and in order to compensate for eventual errors in the other methods.

The second approach, defined as *corpus-based approach*, relies on both syntactic or co-occurrence patterns and a seed list of opinion lexicon to find other opinion words in a large corpus.

One of the major ways of operating it was proposed by Hatzivassiloglou and McKeown (1997): it uses a list of seed opinion adjective words and employs them in combination with a set of linguistic constraints or conventions on connectives with the objective of identifying additional adjective opinion words and their orientations. One of its constraints is the particle "*and"* which suggests that conjoined adjectives usually have the same orientation; for example, in the phrase "*This product is beautiful and cheap*" if "*beautiful*" is established to be positive, it can be inferred that "*cheap*" is also positive,

since people tend to express the same opinion on the same side of a conjunction. Other constraints are also designed for other particles, even if they can work in different ways: "*or"*, "*but"*, "*either-or"*, and "*neither-nor"* have all different meanings for the overall method. This rule is called *sentiment consistency*.

The method was subsequently revised by Kanayama and Nasukawa (2006), that expanded the approach by proposing the idea of *intra-sentential* (within a sentence) and *inter-sentential* (between sentences) sentiment consistency, called *coherency*. In short, it implies that the same opinion orientation (positive or negative) is retained in a number of consecutive sentences and eventual opinion changes are indicated by adversative expressions such as "*but"* and "*however"*.

The third method used is called *dictionary-based approach* and it consists of collecting an initial set of opinion words and lexicon manually with known orientations and sentiments, and then to grow it by either searching the net for their synonyms and antonyms or expanding it using machine learning (Cruz, 2016).

This approach has one main shortcoming: it has difficulties to find the sentiment of lexicon with domain specific orientations. However, an unsupervised approach can be utilized to overcome the limits of the supervised one in use. In fact, as Andreevskaia and Bergler (2006) showed, it is possible to create a lexicon via the unsupervised labeling of words or phrases with their sentiment polarity (also referred to as *semantic orientation*). In their work, seed words for which the polarity is already known are assumed to be provided beforehand, in which case the labels can be determined by propagating the existing ones belonging to the seed words to synonyms, or to terms that co-occur with them in general text or in dictionaries.

In this thesis the dictionary-based approach has been adopted, since it represents the most resource-efficient way to create a sentiment set. In fact, by starting from a sentiment dictionary that contains human-tagged text, such as those provided by universities online and expanding it using machine learning it is possible to analyze through Natural Language Processing almost any kind of text.

The merits of human encoding are that, based on the researcher's ability, each text can be encoded with a very low margin of error, regardless of the language used, the context of discussion, the use of metaphors or rhetorical figures.

### 2.3.1. **From text to data: the stemming process**

Having already established the complexity of language, it is easy to imagine that an automated process would encounter many difficulties in analyzing it, but fortunately not all this complexity is indispensable for a textual analysis.

The initial, but fundamental process is the reduction of the text in quantitative data, in order to make a statistical model able to deal with it. A text contains in fact many words or auxiliary symbols that can be filtered through preliminary analysis. In general, a text or document are part of a set of texts called *corpus* and a collection of them is called *corpora*. There are algorithms that are more efficient on short texts and others that work better with longer texts but, regardless of length, all methods involve a similar process of reducing texts in data matrices. One of the first procedures that are performed is the so-called preprocessing phase of the texts, and it consists in deleting the information about the order in which the words appear in the text. (Jurafsky, & Martin, 2009)

Even if the previous operation would seem counterintuitive, it is done with the objective of avoiding to deal with the complex superstructures of language. The final result of the process is defined as a "*bag-of-words*", meaning the sum of the terms, without taking their order into account. Text can be further reduced to a decreased set of terms called "*stems*". A stem can be a single word (*unigram*) or a pair of words (*bigram*) can also be extrapolated if their order is meaningful (i.e. the words that compose "*White House*" taken together carry a different meaning than "*white*" and "*house*" taken apart: the first is a name while the second tells a characteristic about a particular type of house) or even a triad of words (trigram) and so on. In general, considering groups of three or more words does not add much to the analysis in terms of precision and quality, so most researchers stop at bigrams in their analysis. (Zhang, 2010)

Stems do not have to be complete words for the system to work, in fact it's preferred to reduce the term to their basic root: "*argue*", "*argues*" and "*arguing*" can be described from the stem "*argu*". All conjunctions, punctuation, articles, prepositions, suffixes and prefixes, verbal deficiencies, etc. can also be removed, together with words that appear

too frequently within a corpus (e.g. in 90% or more of texts) or too rarely (less than 5% of the texts).

All the steps described until now, when considered together, are described as the stemming phase.

Counting the number of terms in a language, one might think that the stem matrix contains an exorbitant number of lines. The Oxford English Dictionary in fact counts more than 170.000 words currently in use for English only.

In practice, however, any empirical analysis can prove that a typical stem matrix has no more than 300 or 500 stems and very often even less. What presents the computational challenge is instead the number of rows that the matrix will contain, meaning the number of texts to be analyzed, that can be even several million for each analysis.

## 2.4. Opinion Search and Retrieval

Considering the importance that web search acquired for scientific research since its inception, it is not hard to imagine what tools for opinion search might generate in terms of usage and recognition.

Through a single tool it is possible to fetch user generated content around a specific argument selected from the user. Examples of research queries that can be made are: finding public opinions around a particular subject or one of its features, e.g. finding online opinions about the battery life of a smartphone, or finding all the opinion expressed by a single figure (be it a person or an organization) on a certain topic in a given time, e.g. a world leader's yearly opinions about economy.

Similarly to how search engines work, opinion search also has the task of retrieving documents/phrases relevant to the user's query and it does so using similar techniques. However, it has one major difference: it has to determine first if the sentence found expresses a subjective opinion and then if that opinion is positive or negative (Zhang, 2008). This is where the sentiment analysis part comes into play, in fact traditional search does not perform these important sub-tasks that make opinion search more challenging.

Moving to ranking, which is one of the most important features of search engines, we can appreciate another great difference between the two types of search. Traditional search's basic premise is that the top-ranking results contain all the information to satisfy the user's needs (Liu, 2006).

This model is valid for factual information because it is not supposed to change between different results, in truth one fact is equivalent to any number of the same fact. So, if the first page contains all the relevant information that the user requested, there is no need to continue the research further. For opinion search this model is not valid: since opinions can be divergent, only considering a part of them can lead to a misrepresentation in the proportion between negatives and positives. In fact, one opinion is not equivalent to multiple opinions. In this case all the data available needs to be analyzed; if however the amount of documents gets impractical to examine, a reduction can be made in the time period in which the data is gathered, so that only the opinions expressed during a certain time will be taken into account.

What an opinion search engine can do that a standard search engine cannot is also to provide a feature-based summary for each search. It can be difficult to operate, but such engine can associate objects to opinions and sentiments and thus provide a description of the sentiment of the single feature. This would give a user the possibility to consult the web in search of the opinions of others on many features of the same object at the same time, before making a final buying decision.

The program that this thesis revolves around operates as an opinion search engine and it is composed of two separate tasks:

1. *The data retrieval:*

   This part is operated through a scraping algorithm that explores online posts on the *Twitter* platform (referred to as *tweets*) tagged with a specific hashtag. After the input of a word as a query, any tweet tagged with the hashtag corresponding to that word will be saved onto a file in order to analyze it.

2. *The data analysis:*

   This part consists of two sub-tasks: first the tweets are analyzed to determine their level of subjectivity so that the objective ones (or *non-opinionated*) can be removed, then each tweet is run through classifiers in order to divide them

between groups expressing a positive, negative or mixed opinion. The system uses supervised learning for both tasks, obtaining the opinionated training data and the non-opinionated training data from sentiment dictionaries available online. After this analysis, a keyword assessment is carried, in order to determine which words are more connected to negative opinions, representing the possible downsides and shortcomings of a product, and which are connected to positive ones, representing the recognized positive sides of the product.

### 2.4.1. **Twitter**

Twitter is known for its massive spreading of instant messages and comments (i.e. tweets) and their varied nature. Anything can be posted by anyone: individuals, companies and politicians can publish comments on world news, entertainment gossips about celebrities, and discussions over any products. In addition to displaying news and reports, Twitter itself is also one of the bigger platforms where different opinions are shown or exchanged. No matter where people come from, what religious belief they hold, rich or poor, civilized or uneducated, they comment, discuss, compliment, argue and complain over topics they are interested in, sharing their own feelings freely.

As of 2020, over half a billion tweets are shared per day every day[2]. It is known that user-generated content with rich sentiment information can be very precious for analysis purposes, thus all this data lends itself very well to an opinion search engine.

While single-tweet sentiment analysis results can still provide useful information, the overall or general sentiment tendency towards topics are more can give businesses the opportunity to see the bigger picture behind their or their competitor's actions.

For example, a technology firm is curious about how people feel about Apple's new *iPhone* and it can be of great convenience for them if major opinions are collected from a massive number of tweets. A political candidate would crave to get an overview about the support and opposition for other candidates in Twitter at the same time. In all these

[2] Data available at https://www.internetlivestats.com/twitter-statistics/

scenarios, a sentiment analysis of the topic during a set time period is in the best option, but the social network does not offer one.

This thesis taps into this demand by using the unique characteristics of *hashtags* in Twitter. In the platform, hashtags are a mean of aggregating tweets, in addition to a convention for adding additional context and metadata to them. They are created organically by Twitter users, or with a specific interest by companies as a way to categorize messages and to highlight topics. They are simply composed by prefixing a word or a phrase with a hash symbol, such as "*#hashtag*".

To be precise, hashtags can be categorized into three types (Wang, 2011). Most hashtags (*topic hashtags*) serve as user annotated topics, like "*#Trump"*. In other cases, hashtags (*sentiment hashtags*) are an easy way for the user to attach a sentiment to the tweet. This category of hashtags is composed of sentiment words, like *"#love*" or *"#bad*". The third kind of hashtags (*sentiment-topic hashtags*) are a mix of the previous two, in which the topical word and the sentiment words appear together without blanks. For example, "*#IHateTrump*" (I hate Trump) directly expresses negative opinion towards President Trump. Hashtags falling in these last two categories are even more informative since they explicitly indicate the overall sentiment the tweet is expressing.

Two additional reasons for choosing Twitter are essentially attributable to the type of information transmission that characterizes the network structure of this social network and to the use that the online community makes of the Twitter platform, compared to other social networks. (Heimann, R., Danneman, 2014)

The first reason refers to Twitter's ability to foster second-order connections, or "weak ties" (Granovetter, 1973). Its network is conceived so that the interpersonal links are not two-way and equal: this means that one can be a follower of a user without being followed by that user. According to this approach, it is possible for any user to create a connection (however weak) with the President of the United States, while it would be much more difficult to become friends with him on a platform like Facebook. For the same principle, Twitter is able to convey a greater exchange of information between users who share content frequently and less active users: the use of the platform is for many is actually aimed just at listening, instead of sharing content, so all these people do is keep themselves exposed and updated on activities, news and ideas spread by others. The

second reason relates in fact to the use of Twitter to collect information and comments or to contact companies for requests or complaints about their products and services.

The only drawback of Twitter when used to collect data for sentiment analysis is something that characterizes all social networks: the noise. Many of the extrapolated texts can be irrelevant or useless for the analysis at hand or may not contain a sentiment (content like news or press releases is an example of this).

Twitter has an *Application Programming Interface* (*API*)[3] that allows programs to access its tweets by query term. In the next paragraph, there will be a first description of the programming language and software that was chosen to apply sentiment analysis to the collected data.

### 2.4.2. **Python**

Using Python as an environment to conduct a social media mining analysis brings several advantages. In addition to performing the functions of a simple calculator, the flexibility of Python allows to assist the user in the manipulation of even large datasets, both for the calculation of basic functions and for the application of algorithms and complex mathematical operations, as well as statistical processing and production of graphs. It has many functions dedicated to advanced mathematical calculation and statistical analysis and allows to realize new functions easily recallable, if necessary, by the user. Python also provides access to user-available feature and program libraries.

New libraries are continuously created by developers and made accessible instantly to the user. In particular, many collections for text mining are made available, making the range of tools wide for the benefit of the researcher who faces an analysis like the one that is the object of this thesis.

The large number of research jobs and data mining manuals that use Python as a standard clearly makes it easier for anyone who approaches this type of analysis to be guided in applying a model on a real case.

[3] More information about the Twitter API can be found at http://apiwiki.twitter.com/ .

Python can be defined as an *object-oriented* language: “each variable it uses is an entity that has certain defined attributes and methods” (Bird, Klein, & Loper, 2009).

It can be analyzed in three ways:

1. As an interpreted language, Python facilitates user exploration.
2. As an object-oriented language, Python permits data and methods to be encapsulated and re-used easily whenever the user needs them.
3. As a dynamic language, Python allows attributes to be added to objects very easily and allows variables to be typed dynamically, facilitating rapid code development.

Python is already packed with an extensive standard library, including components for graphical programming, numerical processing, and web connectivity, but in addition to those for the purpose of this thesis *NLTK* and *Textblob* are going to be used, respectively for Natural Language Processing and Sentiment Analysis.

*NLTK*, the Natural Language Toolkit, is a library of open source program modules, tutorials and problem sets, considered to be the leading suite when building Python programs to work with human language data. It provides user-friendly interfaces to over 50 corpora and lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, tagging and parsing. (Loper & Bird, 2002)

Textblob, on the other hand, is a Python library integrating a wide range of machine learning algorithms for all kinds of sentiment analysis applications. This package’s emphasis is instead put on ease of use, performance, documentation, and API consistency. (Pedregosa et al., 2011)

# 3. Social media campaigns

## 3.1. Social media marketing

Social media and social networks are platforms that use the Web to share and foster user-generated content (Kaplan & Haenlein, 2010).

Social media in particular can be defined as web-based services that allow users to create their own identities, engage in public or private conversations, share content, find other people, create and manage relationships, build reputations, and join social groups (Kietzmann et al., 2011).

Social media platforms can be divided into more specific categories by defining their different characteristics: collaborative projects, blogs, content communities, social networking sites, virtual game worlds, and virtual social worlds are all platforms with different types of users and usage. To create such a classification scheme, Kaplan and Haenlein (2010), in their research, relied on a set of theories in the field of media research (*social presence*, *media richness*) and social processes (*self-presentation*, *self-disclosure*), the two key elements of social media. When considering social presence and media richness, applications to which lower scores were assigned are collaborative projects (e.g., *Wikipedia*) and blogs, as they are often text-based and hence only allow for a relatively simple exchange. Slightly above those the authors put content communities (e.g., *YouTube*) and social networking sites (e.g., *Twitter*) which, along with text-based communication, enable the sharing of pictures, videos, and other forms of media. The platforms with the highest levels of social presence and media richness for the authors are virtual games and social worlds (e.g., *World of Warcraft*), which try to replicate direct interactions and all of their dimensions in a completely virtual environment. Regarding self-presentation and self-disclosure, blogs are positioned higher than collaborative projects, as the latter tend to be focused on more specific content domains. Social networking sites are in a similar position since they allow for more self-disclosure than content communities. Finally, for the research, virtual social worlds require a higher level of self-disclosure than virtual game worlds, as the latter are guided by stricter rules that force users to behave in a certain way. (Kaplan & Haenlein, 2010)

Brands can use social media to communicate with audiences as they do with traditional media, with the difference that consumers can also use these platforms to communicate with each another. Social media platforms have thus transformed the traditional role of the audiences, making them both recipients and initiators of content (Hanna, Rohm, and Crittenden, 2011).

The introduction of social media as an interaction form has radically disrupted the communication process itself. Traditionally, the communication model has consisted of the four elements *source–message–channel–receiver* (*SMCR*); the traditional processes within this model involved encoding, decoding, response, feedback, and noise (Berlo, 1960). The emergence of social media, however, has made the nature of communications more complex and inclusive. Scholars like Mangold and Faulds (2009) have recreated a structure for the communications framework to incorporate the newly added elements of social media, conceptualizing a *new communication paradigm*, that emphasizes how much the brands' control over their content, together with its timing and frequency, is being severely eroded.

Traditionally, companies had a significant control over how their brand was perceived through the management of their promotion mix, including their advertising, public relations, and promotions. Now, however, consumers have the possibility to interact with one another to create conversations about the brand.

The new communications paradigm model (Mangold & Faulds, 2009) shows that social media can be considered as a hybrid between the traditional promotional tools and an avenue for customers to interact and create word-of-mouth. Therefore, even if social media can help a company communicate more efficiently, it can be difficult to deal with the uncertainty always present in this kind of marketplace.

In 2010, Hennig-Thurau developed the "pinball" framework, which explains the effects of new media on customer relationships through a similarity with the game: companies release a "marketing ball" into the environment, which new media bumps and diverts in chaotic and unpredictable ways. Marketing managers can still use marketing tactics or "flippers" to guide the ball; however, the ball does not always move where intended.

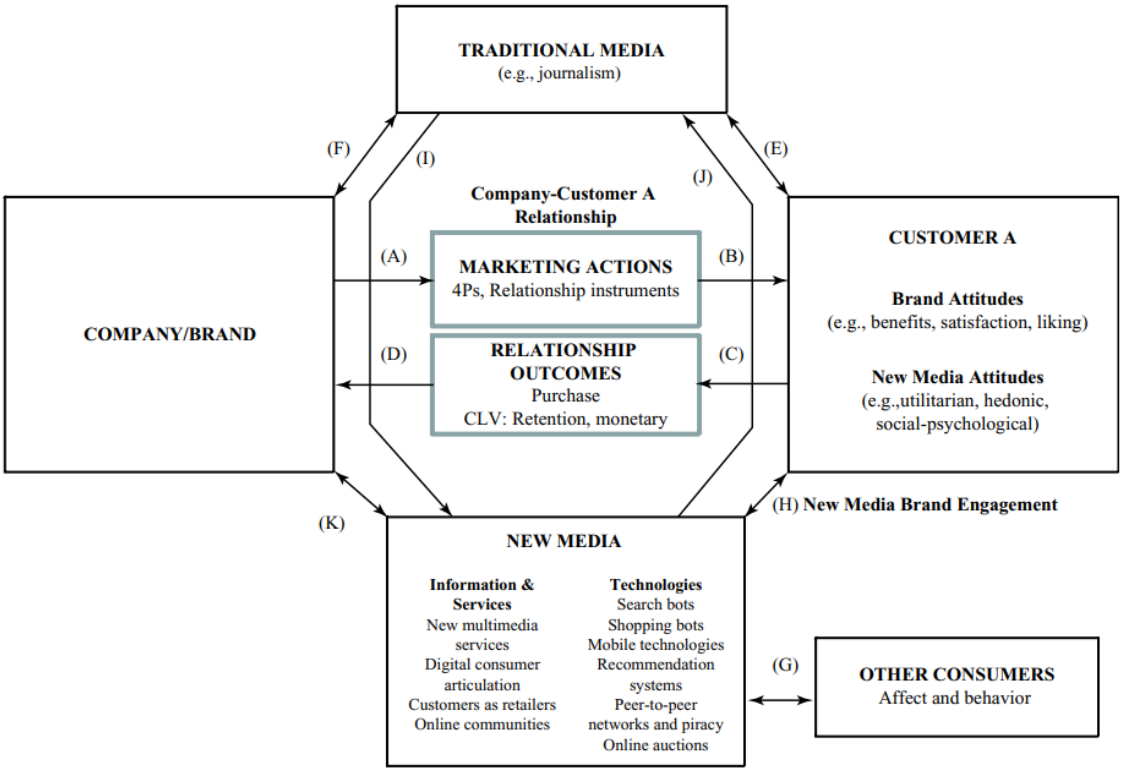


*Figure 2: The "pinball" framework (Hennig-Thurau, 2010)*

Figure 2 shows the conceptual framework created by the authors that clarifies the role of new media in customer relationships.

Traditionally, companies and their brands (shown in the left square) actively influenced customer relationships through their marketing actions (unidirectional arrow A) and, both actively and reactively, through public relation (bidirectional arrow F). Up until the advent of new media, customers (shown in the right square) were predominantly passive receivers of marketing and products information (arrows B and E), with companies who were able to mostly avoid negative mass media coverage by having quasi-complete control over the messages that ended up shaping their brand and thus the final outcome of customer-brand relationships, such as customer retention (arrow C) through their own actions and decisions.

The bottom portion of the scheme, on the other hand, illustrates how the rise of new media creates a disruptive change in the marketing environment. Today, in fact, the information about a brand has turned into a multidirectional, interconnected flow and this change has made it difficult to predict. Marketers have lost the complete control they

previously had over their brands and now can only participate in a "conversation" about the brand. This explains why in the era of new media, managing customer relationships can be compared to the game of pinball: companies launch a marketing "ball" that consists of their brands and brand-building messages into an unpredictive environment, which can be deviated or often accelerated by new media that act as "bumpers", changing its course in chaotic ways. After the marketing ball is in play, marketing managers can continue to guide it with agile use of the "flippers", but the ball does not always go where they intended it to, and their slightest miscalculation can be amplified into a crisis. (Hennig-Thurau et al., 2010)

Castronovo and Huang (2012) proposed an alternative take on the effect of social media platforms in marketing. The authors pointed out the connections between marketing activities on social media and other marketing activities like brand community, customer relationship management (*CRM*), and search engine optimization (*SEO*), illustrating that the effects of marketing activities on social media are pervasive and can have an impact on companies' entire marketing communications strategies. (Castronovo & Huang, 2012)

The rising importance of social media in the marketing field has stimulated researchers to thoroughly analyze the phenomenon, in order to provide guidelines for businesses to use social media in an effective way. Research has mainly focused on the types of content that brands share on social media and their effects on their customers.

The content of the ideal post is still questioned among scholars, with some suggesting that the most effective type of content is pictures (Hansson, Wrangmo and Søilen, 2013), others arguing about the best type of messages to share between entertaining and informative (Cvijikj and Michahelles, 2013), interactive and responsive (Burton and Soboleva, 2011), amusing and philanthropic (Zhang, Jansen and Chowdhury, 2011), or appropriate and informal (Kwok and Yu, 2013). Jansen et al. (2009), along with marketers[4] think that brands should use Twitter as their preferred social medium because of its characteristics as a feedback mechanism that also learns from customers' posts and engages them in content.

[4] https://www.convinceandconvert.com/social-media-strategy/twitter-engagement/

## 3.2. Twitter marketing

Twitter is recognized by researchers as the ideal social medium for brands that seek to build relationships with their key stakeholders (Hennig-Thurau et al. 2010). Twitter is mainly devoted to information dissemination, that can be accomplished through word-of-mouth, spreading via many small cascades triggered by ordinary individuals (Bakshy et al., 2011). However, firms can also derive benefits from using Twitter to interact with their customers. In fact, many companies typically use it to communicate with a large number of followers in a *one-to-many* form. In addition to that, they can use the *one-to-one* mechanism to interact with single individuals by replying to them or retweeting their content. (Burton and Soboleva, 2011)

What makes Twitter so invaluable to companies is its nature as an online *listening tool* (Crawford, 2009), that allows its users to create online engagement with them even without having interactions and just following their content.

Jansen et al. (2009) was able to create a way to sort companies' tweets based on their content in four hierarchical categories:

1. *Sentiment*:

   Tweets that contain the expression of an opinion concerning a brand. The sentiment can be either positive or negative.
2. *Information Seeking*:

   Tweets that contain the expression of a need to address some gap in data, information, or knowledge concerning a brand.
3. *Information Providing*:

   Tweets that provide data, information, or knowledge concerning a brand.
4. *Comment*:

   Tweets than just name a brand in a text where the brand is not the primary focus. (Jansen et al., 2009)

This distinction, however, ended up being troublesome in their research, in fact many of the tweets were classified in two or more categories, thus not generating a clear distinction between their contents.

A study by Taecharungroj (2017) on Starbucks' communications strategy on Twitter evolved and improved Jansen's classification by reducing the content categories to three:

1. *Information-sharing content:*

   Tweets in this category are used to communicate valuable information to followers. The emphasis is on content that benefits its receiver like practical tips about the product or service, store promotions and official announcements.

2. *Emotion-evoking content:*

   Tweets in this category have their main purpose in evoking positive emotions in followers, such as happiness, excitement, serenity and delight. They mostly include pictures, closely followed by content like storytelling, inspirational quotations, or humorous and witty messages.

3. *Action-inducing content:*

   Tweets in this category attempt to persuade the community to take a desired action, such as purchasing a product or service, participating in an event, or registering to something. They are often presented in the form of imperative sentences and their most common subtype is sales promotion, followed by questions and event participation.

There are two primary and very different reasons that explain why every company is compelled to have a Twitter presence (Thomases, 2010).

The first reason has less to do with bigger brands and more to do with niche or young companies, the kinds of businesses that need more effort in their marketing in order to get on par with their competitors. Through a correct use of Twitter, they can boost their website traffic and search engine rankings. Search engines tend in fact to privilege new content just like what Twitter and other social media platforms provide. Search engine crawlers specifically look for this type of content for their indexes, which in turn help with search visibility. In addition to this, the Twitter *#Explore* function itself plays a key role in helping people to find new information and trends, taking the role of a full-fledged real time search.

The second reason does not pertain small-scale marketing; it has to do with the initial resistance to Twitter from bigger brands who feel they should act only when their audience, or competition, has already established their presence. By getting to use the platform as early-movers, they can be allowed initial mistakes and learn from them in front of their audience.

The ways in which Twitter helps brands are:

- *Direct-to-Consumer (DTC) Opt-In Marketing:*

  *DTC* consists in what companies create to give their users the possibility to keep directly in touch with them. Normally it is operated through mailing lists, but Twitter has the unique ability to create a seemingly one-to-one relationship between the consumer and the brand. It can be *active* (e.g. a consumer tweets to the brand and brand replies back) or *passive* (e.g. a consumer follows the brand and has the potential to be exposed to everything the brand publishes on Twitter).

- *Provide Direct Customer Service:*

  Twitter allows consumers to get in contact with companies in a less time-consuming way. Instead of telephone lines with predefined hours, contact forms on websites or mails, if the user is able to convey his problem in 280 characters or less, he can get an answer from an expert right away.

- *Build Customer Loyalty and Retention:*

  As mentioned before, Twitter can boost loyalty and retention through the customer service a brand is able to deliver. This kind of loyalty, however, can often be passive. Twitter can also be a channel to cultivate loyalty and retention in an active way, for example creating exclusive promotions or deals for the brands' followers.

- *Promote*:

  One of the clearest uses of the platform is to share promotions, coupons or in-store deals through the business' social account.

- *Provide Instant Updates and Alerts:*

  When offering services or products that are time-sensitive, it can be important to communicate updates and news in a faster way and Twitter gives that possibility to firms.

- *Thought leadership:*

  Brands that want to be perceived as leaders in their field can find in Twitter an excellent vehicle for their ideas and messages: through tweets they can release research findings, pose evocative questions, and take stances on different issues.

- *Manage crises:*

Unhappy or slighted customers, through the power social media, can at best post a legitimate complaint to all their friends and at worse launch a true brand warfare when their grievances go unaddressed. A brand can respond and contain such problems by keeping an eye to its Twitter messages.

- *Create spokespeople:*

  After an aforementioned crisis, a well-treated customer can not only be retained but also become a brand advocate or even an evangelist. All of this without comporting relevant costs for the company.

- *Entertain*:

  Entertaining tweets, particularly the ones that link to entertaining content like funny videos or bizarre photos, can have a viral effect on the platform. Entertainment-oriented tweets are some of the most commonly retweeted and receive some of the highest volume of views and interaction between all the content tweeted.

- *Get instant feedback:*

  After a quick view of the reactions and comments under a tweet, a brand can gauge the pulse of its audience in a simple way. This way firms can also test new product or logo ideas directly with their public

- *Branding and awareness building:*

  Finally, and most importantly, Twitter can spread Word-Of-Mouth about the brand and let the market know about it or reinforce its existing perception.

Expanding on the last segment, business can build their brands and credibility on Twitter in different ways. Firstly, they have the possibility to listen to their customers and look for common threads in their comments that can be used to learn things they can then apply to the rest of their marketing efforts.

Secondly, they maintain consistency in tones and messages between their offline communications and their tweets so that the consumers do not perceive any differences in the overall brand messages.

Thirdly, the companies build on their network to deliver a consistent and stronger message: they communicate with their clients about their products and services while at the same time they engage with them about their needs, so they turn spokespeople into brand ambassadors. (Thomases, 2010)

Companies do all this through the community manager and his team, someone to oversee all of their social media initiatives.

### 3.3. The community managers and their teams

The community manager, often erroneously referred to as the social media manager, has the responsibility of managing how a brand is portrayed and perceived with the online public, but he is also responsible for developing and overseeing the execution of strategies that are in constant flux due to the environment in which they're being executed. Community managers need to be individuals with 360-degree vision and need to have empathy with the community in order to involve and engage the brand's stakeholders. (Moretti & Tuan, 2015)

The primary role of the community manager is to represent the customer to the shareholders. This includes listening, monitoring, and interpreting what the brand's community is talking about, as well as engaging with the customers by responding to their requests and needs both in a private and in public form

Community managers must also act as brand evangelists; in this capacity, they must serve as corporate promoters of events, products, and upgrades to customers by using techniques like conversational discussions.

One of their most important roles is the one of communicator. This task mandates them to be fluent with all forms of jargon within social media communication: from forums, to other social media, to podcasts, and to Twitter itself.

This individual is also responsible for mediating disputes within the brand's community, eventually turning to consumer advocates for assistance, and trying to work through challenges presented by potential detractors.

In an editorial strategy and planning capacity, the community manager works with multiple internal stakeholders to conceive, plan for, produce, and publish the necessary content to keep the brand's public community fresh and current.

This figure does not operate on its own, it is in fact generally supported by three subordinates (Owyang, 2011):

- *Social Analyst*:

  This figure is responsible for measurement and reporting across the entire program and for individual business units, he uses brand monitoring, social analytics, web analytics, and traditional marketing tools and is responsible for measurement and reporting of time, costs, quality, risks and results across the entire program or for individual business units. He can also act as a *SEO* (Search Engine Optimizer), providing the brand with the highest online visibility, especially on search engines.

- *Social Media Manager*:

  This figure is responsible of coordinating all the brand's business units to launch social media initiatives. The Social Media manager may straddle internal and external communications, direct resources, and formulate program plans. He can be the sole figure in charge of web contents, so that it retains a unique voice and he curates the filtering and aggregation of online information. This figure is commonly the one tasked with the complete management of the brand's Twitter account.

- *Corporate Social Strategist*:

  This figure is responsible for the overall vision and accountability towards investments. The strategist is primarily internally facing and supports the choices of the top management. He is also in charge of online promotion tools and of integrating online and offline in order to gain and maintain relationships.

The degree of separation between these roles tends to depend on the size of the business: smaller ones incorporate all of the functions under a single community manager, while bigger ones expand it to a full-blown team, also often dividing the social media manager role between different people (Moretti & Tuan, 2015).

The social media manager role is in fact the most delicate, because it can be easily undermined by a low consistency in the posts and it is mostly based on the sense of continuity of message and sentiment that the brand can create with its messages.

## 3.4. Hashtags on Twitter

The sense of consistency and common message between different posts can be achieved, among other ways, with a correct use of hashtags.

The use of Twitter to coordinate discussions and focused communications has been a key to its legitimization, or "*debanalisation*" (Rogers, 2013), and the increased use and recognition has brought an increased academic and journalistic attention on it. In every case, the hashtag is what has been recognized as the "killer app" for Twitter's role as a platform to foster the emergence of new publics, where *publics* are meant as new entities that are being generated, renovated, and coordinated via dynamic networks and social connectivity, organized primarily around new issues or events rather than pre-existing social groups (Warner, 2005).

Semantically, tagging in Twitter is done with a focus on filtering and directing texts so that it appears in certain content streams rather than to index messages for a later retrieval (Huang et al, 2010).

Before Twitter gained its popularity and ended up influencing most of the other social networks, the selection of tags in social media sites was often done after the creation of the content itself, the key concepts that the user wanted to underline were distilled into short strings of text added to the document, image, or resource posted. In contrast, tagging in Twitter has established itself as a method for filtering and promoting content, rather than as a tool for recalling it (Huang et al, 2010).

In his study, Huang (2010) found out that the concept of a priori tagging in Twitter, while seemingly unintuitive at the time, was created with different goals in mind from those of other social platforms. For example, a user that observes the rise of a compelling trending topic that sports a particular hashtag, may be inclined to take the tag associated with it and compose his own personal tweet on the subject. Thus, the presence of the hashtag itself makes it more likely that a user ends up writing the tweet and participates in the phenomenon than if they had not been inspired.

In addition to that, following and posting tweets to an existing hashtag conversation allows Twitter users to communicate with a community of interest around the particular topic without the need of going through the process of establishing a following relationship with any of the other participants; as a matter of fact, it is even possible to

follow the stream of messages linked by a hashtag without being a Twitter user (the aforementioned *#Explore* function works as a normal webpage accessible to anyone), and it's becoming more common for a screened or even complete version of the hashtag stream to even be broadcasted alongside television news coverage or other particular televised events (Bruns & Burgess, 2015).

Hashtags practices are far from static: they are born by a process of *adoption*, happening when a newly coined hashtag is embraced by a critical mass of users and subsequently disseminated through Twitter, and they die by a process of *abandonment*, happening when said mass of users stop attaching the specific hashtag to their tweets over a period of time, until the tag's appearances become scarce (Huang et al, 2010).

The communities that rise from hashtags are both overlapping with the creator's own and completely new, in fact the social layers overlap: all public tweets marked with a certain hashtag are visible both to the user's existing followers and to anyone else following the hashtag stream. At the same time, it is possible for a user to post to a hashtag conversation and not to follow said conversation: they might want to include the topical hashtag in order to make their tweets visible to others, thus increasing its potential exposure, but they may still continue to focus only on tweets coming in from their established network.

The power of hashtags can also go beyond the inclusion of one: a user can respond to a *hashtagged* tweet and so that may be seen as carrying out the conversation in front of a wider audience, by comparison with the more limited visibility which a non-hashtagged response would have.

The possibility to reply to tweets containing hashtags allows for a way to easily measure community participation. It is possible, in fact, to measure the extent to which the contributors are actively interacting with one another by sending publicly visible *@replies*[5] or retweeting each other's messages. A high volume of such responses would indicate that users are not only tweeting into the hashtag stream, but also following what the others are currently posting. The higher the number of messages like these are contained in the hashtag stream and the greater the total number of engaged participants,

[5]*@replies* are tweets which contain the username of the original message recipient, prefixed by the *'@'* symbol.

the more the hashtag community can be said to act as an actual community (Bruns & Burgess, 2015).

Research has also shown a secondary use for hashtags that surpasses the sorting one: *metacommunication* (Daer et al., 2012). Users employ these tags to offer their own thought over the post itself, creating a "metacommentary" instead of a topic mark.

The metacommunicative function can work in 5 ways:

1. *Emphasizing*: it is used to add emphasis or call attention to something in the post or something that the post refers to; usually expressed without judgment as a comment or reflection (e.g. *#LateNight*).
2. *Critiquing*: it is used when the purpose of the post is to express a judgment or verdict regarding the object of discussion (e.g. *#WhatIsHeThinking*).
3. *Identifying*: it is used to refer to the author of the post and expresses an identifying characteristic or mood (e.g. *#IHateMyself*).
4. *Iterating*: it is used to express humor by referring to a well-known internet meme or happening in internet culture (e.g. *#WokeUpLikeThis*)
5. *Rallying*: it functions to bring awareness or support to a cause; it can either be used in marketing campaigns to gain publicity (e.g. *#TasteTheFeeling*). (Daer et al., 2012)

### 3.4.1. An history of the hashtag on Twitter

In order to understand hashtags, their analysis must start from their conception.

In the early years of Twitter, following its launch in 2006, the social network had almost none of the many functionalities that it does today. Its users were invited to answer the question "*What are you doing?*" in less than 140 characters, to follow their friends' accounts, and little else (Weller et al., 2014).

Many of the capabilities, applications and cultural functions of Twitter, that make its role in the public communication so important, were not developed by its creators, instead they were user-led innovations, only later integrated into the architecture of the main system by Twitter, Inc. Such innovations include but do not limit to the *@reply* format

for addressing or mentioning other users, the integration of multimedia file uploads into tweets and, most significantly for this work, the idea of the *hashtag* as a way to coordinate conversations.

As a concept the hashtag was generated on *Web 2.0* platforms and came to popularity thanks to its use in *Flickr*[6]. Its emergence on Twitter was originally proposed by Chris Messina with a post on his tech blog entitled *"Groups for Twitter, or a Proposal for Twitter Tag Channels* and with a tweet on Twitter itself (shown in Figure 3). Messina referred to his idea as a proposal for "improving contextualization, content filtering and exploratory serendipity within Twitter" by creating a system of "channel tags" using the hash or pound (#) symbol, allowing people to follow and contribute to conversations on particular topics of interest. (Messina, August 2007)

While it would seem that the original idea was centered around the formation of Twitter user groups based on common interests, Messina argued that he was instead interested in creating a way to easily follow conversations and interest groups in order to "have a better eavesdropping experience on Twitter"; therefore, rather than "groups", hashtags would result in building *ad hoc channels* to which pools of users could pay selective attention.

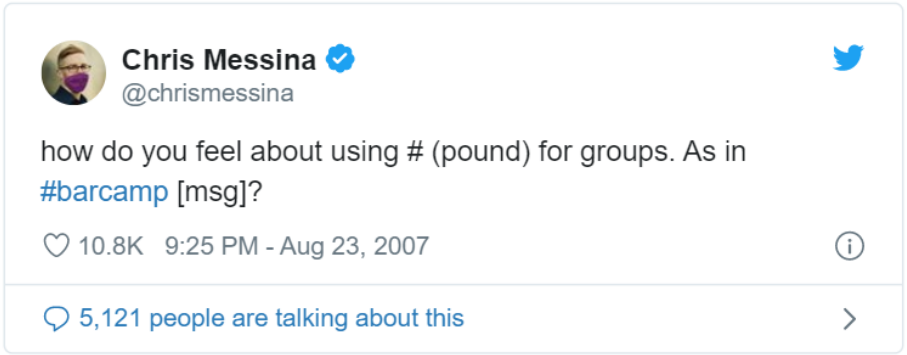


*Figure 3: Messina's original tweet about the use of hashtags*

[6] *Flickr* is an image and video hosting service. It was created by Ludicorp in 2004.

Messina's idea was not employed at first, at least until the 2007 San Diego bushfires[7] demonstrated a clear use-case for it while he continued highlighting the strength of hashtags to coordinate information (Messina, October 2007).

Over time the custom became a habit in the social and communicative languages of the platform's user community and finally in the system itself, with the inclusion of hashtags into search results and trending topics.

The simplicity and ease-of-use of hashtags allowed them to exceed the platform they were initially intended for and nowadays it is possible to see examples of them in most of the social networks available. This is largely attributable to their basic simplicity and to the absence of any top-down usage regulation, meaning that all a user has to do to create one is typing the hash symbol followed by any string of alphanumeric characters.

### 3.4.2. Hashtag communities as *ad hoc publics*

What emerges from the observations about the nature of hashtag conversation and the story behind their popularity is a picture of hashtag communities not as separate and enclosed entities, but as both macro-level spaces which overlap with the flow of messages across longer-term networks, and the micro-level spaces, intended as communicative exchanges conducted in the form of *@replies* between users that find each other through the hashtag (Bruns & Moe, 2014).

Twitter's user generated system of hashtags allows it to stand out from other social spaces thanks to the platform's unique capability to respond quickly to incoming issues and events. Thanks to Twitter's hashtags the traditional part where the news story must be written, edited and published, where the interested parts must be invited, and the event set up is completely skipped: the platform condenses such processes to a single action, and the focused publics it forms are virtually ad hoc, since they are brought into existence for that particular need.

[7] The October 2007 California wildfires, also known as the San Diego bushfires were a series of about thirty wildfires that began igniting across Southern California on October 20.

The animate nature of the conversations happening within hashtag networks can also provide interesting insights into the deeper functioning of such ad hoc publics: it enables researchers to differentiate the various roles played by different individuals (e.g. information sources, leaders, commenters, conversationalists, or *lurkers*[8]), and to study how the public reacts to new stimuli. Such insights can also offer new outlooks on the interconnection of the community with the other social spaces beyond Twitter itself, and on the relative importance of such spaces. (Bruns & Burgess, 2015)

As to Twitter Inc., it's safe to say that the company both acknowledges and fosters the creations of ad hoc publics: the gradual changes to the platform's implementation and use of hashtags have transformed the ways they are experienced from users, to the extent that the publics emerging through hashtags might be defined as ad hoc.

For example, in 2012 Twitter introduced official "hashtag pages" for certain events, like the European Football Championships of 2012[9]. These pages are similar to a normal hashtag search result, with the difference that they represent carefully curated depictions of the public communication and comments over a specific event, often favoring particular sources deemed as more reliable, as part of Twitter's efforts to be perceived more as a media company than a social networking service (Dijck, 2011).

Twitter's efforts can be also perceived in their algorithm: in fact, for a normal user, a Twitter search will by default return a list of *"Top Tweets"* - instead of a complete view of all the content available, which is what the algorithm thinks are the most trustworthy and relevant results. This of course does not end up canceling the work on as hoc publics, that constitutes one of the platform's strengths, but it results in a more curated and consciously assembled list of results.

The flexibility and ease of creating such publics and communities, when they are needed and without restrictions, is what gave Twitter recognition as the preferred platform for events discussion. This recognition is evident in the common use of the platform by media organizations, politicians, and, most importantly, industries and firms, that choose to carry out part of their marketing strategies or their public interactions on it.

[8] In Internet culture, a *lurker* is typically a member of an online community who observes but does not participate in the discussion at hand.
[9] Twitter hashtag page for the EURO 2012 Cup: https://blog.twitter.com/2012/euro-2012-follow-all-the-action-on-the-pitchand-in-the-stands

## 3.5. Branded Hashtags

A brand interested in laying out a marketing campaign, after establishing its goals and targets, can decide how to harness the potential of hashtags on social media.

In fact, a study by Dan Zarrella has shown[10] that a tweet that contains at least one hashtag was 55% more likely to be *ReTweeted* than tweets that did not. However, to tap into this trend, competition must be taken care of first. An overused hashtag, in fact, could involve a large quantity of content around it, so it would be difficult to bring out a particular post in an ever-updated repository. In contrast, a less used hashtag used would not allow the post to gain the desired visibility due to the low visitor rate.

After determining the competition rate of the various hashtags on that particular topic, the company needs to understand if the popularity of the brand is high enough to compete on higher grounds (e.g. a company like *Coca-Cola* might want to ride the popularity of hashtags like *#ValentinesDay* because it has less risk of being overwhelmed by the stream of other tweets). If the brand is not recognized enough, the company should instead aim to insert less used and less clicked hashtags that may allow greater visibility. The fewer accounts they use that hashtag, the greater the chance that a niche audience will stumble on the post. (Shin et al., 2018)

The number of hashtags to include in a single post is also important, in fact empirical research (Richter, 2014) shows that posts which contain 1 to 2 hashtags have on average 50% more interactions than post with up to 5 hashtags and double the interactions than post with up to 10.

Once the company has decided the hashtag to use, it can move its effort to creating a second one that takes up the name or pay-off of the company to pair with the chosen one. Originality at this stage is essential in order to create a unique hashtag that differs from the competition. In addition, a successful hashtag could be a way to get free advertising from its audience chatter and through sponsorship from satisfied customers.

[10] Data retrieved from http://danzarrella.com/new-data-use-quotes-and-hashtags-to-get-more-retweets/

When creating company hashtags, it is important to follow certain principles in order for them not to be useless or even negative for the brand. Firstly, a hashtag should be created after having assessed how the competition has operated in the field and what can work. Then the language used must come from an assessment of the target audience: common-use words or more sophisticated languages must be used, depending on the firm's customers. Acronyms can be used only if the relevant public already knows them, while for longer phrases an uppercase letter for each new word eases the comprehension of the whole. (Yang et al., 2012)

After understanding the many facets of how hashtags work, a brand can think of undertaking a hashtag campaign. It consists of creating an absolutely unique and unambiguous hashtag, never used before, posting new content together with this hashtag and inviting the public to do the same. Companies' experiences show that a hashtag that triggers emotions while being still connected to the source brand has a stronger appeal on customers: Ben & Jerry decided to use *#CaptureEuphoria* and Nike went with *#MakeItCount* not only because they spread more emotions than their respective brands, but because the emotions they convey are deeply linked to the brands' messages. The hashtag created must also be consistent within different social networks in order not to create confusion in the mind of consumers.

Hashtag campaigns' use has risen in recent years among companies active on social media, in fact in 2015 70% of the most used hashtags on Twitter were brand related (Simply Measured, 2015). This happens because these campaigns have very low costs and, in return, allow brands to create awareness and an image for the brand by developing a free sponsorship that feeds itself with the posts of social users. Companies this way are able to adopt a *consumer-centric* approach, where they can build and consolidate interactive relationships with their audience with the final objective of creating engaging and loving consumers (Stathopoulou et al., 2017).

Hashtag campaigns are key in building a buzz around brands and driving engagement among audiences. They can be used to introduce a new product, spread content from a blog or a video, to get more interaction with customers, to create awareness for a cause and so on.

Despite their helpfulness, it has been studied (Lin et al., 2013) that some of these hashtags fail to achieve the popularity of others. Lin, when studying their use in

presidential campaigns, divided them in two categories: “winners” and “also-rans”. In his study he found that the number of times a hashtag was retweeted as well as the popularity of the users mentioning the hashtag lead to a faster growth for “winner” and “also-ran” hashtags alike; additional replies and the number of unique retweet sources also supported their persistence. However, the findings unexpectedly suggested that the number of retweets can tend to inhibit the growth of hashtags; this leads to the conclusion that there can be higher-order processes that lead to practical limits or tipping points for these posts’ total reach. The way that Lin explains these controversies is by drawing from organizational ecology theories used to describe the birth, growth, and death of organizations and communities.

Ecologists, like Carroll and Hannan (2000), discuss in fact organizations as entities which, just like hashtags, coordinate their consumers’ behavior through actions that suggest the kinds of attitudes and messages that are appropriate for them. The comparison continues moving to the environment, that in both cases provides a limited supply of resources - laborers and customers for organizations and attention, users and ideas to express for Twitter hashtags. Companies are able to survive and prosper when their identity is specific enough that individuals know exactly what to expect from them, but vast enough that they can appeal to a range of people and address different needs.

In Carroll and Hannan’s studies, similar organizations are able to thrive as their population grows, while at the same time gaining more attention and legitimacy. Nonetheless, at some point a limit is reached and the environment can no longer support most of these companies, so in the end only a few remain.

On the other hand, an additional explanation for the growth and limits of hashtags comes from the research on growth and sustainability involving the comparison of cultural forms, such as words, names or networks. Both exposure and “fitness” play key roles in these models, suggesting motivations for both self-reinforcing growth and saturation (Steyvers & Tenenbaum, 2005). Lin’s (2013) model adds an additional factor to consider: the native and developing properties of the communication and interactions between those that use the hashtags, serving to weaken the determinism suggested by fitness-based explanations.

The author concludes with the creation of a method, called *conversational vibrancy*, which is what influences the growth and persistence of distinct classes of hashtags, and

is based on how features such as topicality, interactivity, diversity, and prominence interact with the communities producing and following tweets containing these hashtags. (Lin et al., 2013)

### 3.5.1. Branded Hashtags in television

The increased usage and popularity of branded hashtags has also brought TV broadcasters' attention on it. They have in fact quickly adapted to this shift in consumer behavior by integrating these words in their TV programs as a way to increase viewers' participation (Page, 2012), *audiencing* (Highfield et al., 2013), and to promote information sharing (Gleason, 2013).

It is important to remember that consumer interact with these hashtags in addition to the ones they are subject to when they are watching the commercial breaks between those programs; these act as *cross-channel connectors* and allow brands to link different social media discussions to their campaigns and to further facilitate their audience's engagement and participation. (Stathopoulou et al., 2017)

The integration of hashtags in and between TV programs shows a growing interest of the broadcasters' category in interacting with, and, most importantly, tracking and analyzing audiences about their programming.

After studying branded hashtag users, Stathopoulou (2017) found that the more original and novel the whole advertisement with a branded hashtag is, the more likely it will be for consumers to actually engage with the brand advertised. His team's findings did not change between humorous or warmth advertisements that include hashtags.

The same reasoning applied to relevance and appropriateness of hashtags, which were perceived as positive for the consumers and raised their likeliness to engage in branded content creation. In addition to this, however, their results shown that consumers that already were more familiar with the brand advertised were more likely to engage with that brand though the offered hashtags. Brand familiarity in fact is able to act as a moderator in the relationship and it decreases the overall effort and time required by consumers to interact through hashtags with the brand when they are shown on TV

advertisements. However, brand familiarity was also found to negatively moderate the relationship between resolution and hashtag engagement, meaning that if a consumer is already familiar with the brand, relevance and appropriateness will not affect him the same way. (Stathopoulou et al., 2017)

The overall result of Stathopoulou's research and the proliferation of the use of branded hashtags show that there can be a novel way to make advertising, not always focusing on hard sell and traditional techniques, but instead on creating smart and warm branded content that consumers can interact with because they want to.

# 4. Sentiment analysis on the campaigns

## 4.1. Method

In order to capture the widest range of sentiments that Twitter users express in response to campaign hashtags, this work will analyze campaigns with different kinds of public receptions.

The method utilized to export and analyze the tweets, as already introduced, uses the Twitter API to gain access to the tweet database. Access to the Twitter API must be requested justifying the purpose and type of work that the user will employ it for, and it is granted in one of four different types of access:

- *Standard APIs*: this type of admission to the data provides limited access to tweets, in particular it allows the user to publish and engage with content and to analyze the past 7 days of tweets by filtering and sampling them.
- *Premium APIs*: this type of admission to the data provides a broader access to tweets by increasing the day limit to 30 and allowing all the operations included in the Standard APIs.
- *Enterprise APIs*: this type of admission to the data provides the broadest access, allowing the user to access the full archive of tweets and all the possible operations with them, including tweets batching.
- *Ads APIs*: this type of admission to the data provides the tools to create advertising in the Twitter Ads platform. (Twitter, 2020)

For the purpose of this work the *Standard APIs* access was requested so that most of the analysis functions were allowed and the only downside to the access was the day limit.

In addition to that downside, however, another was found, empirically, regarding the maximum number of accesses to the Twitter database, in fact the APIs limit the number of available tweets into blocks of 15 minutes, stopping any other access when a certain limit has been crossed until the 15 minutes pass and the timer resets.

The first downside was solved in two ways: firstly, the chosen hashtags were popular enough that they generated aver 500 tweets per week and up to 9000 tweets for the most popular, making their samples representative enough; in addition to this, to practically

overcome this API's shortcoming, different weeks of tweets were collected during the drafting of this work, so that the final analysis could be made on a bigger sample.

The second downside is easily avoided by only accessing one campaign at the time, thus limiting the total number of tweets asked to the platform in a 15-minutes window.

### 4.1.1. Python code

The Python code was created on *Jupyter Notebook*[11], an open-source web application that allows to create and share documents that contain both code, visualizations and narrative text. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, data visualization and machine learning, among others.

This particular program was chosen because of its clarity in showing both code and comments in separate "batches" that allow for a separate and distinct analysis.

The packages used were *os*, *pandas*, *tweepy*, *re*, *string*, *textblob*, *preprocessor* and *nltk*. Both *os* and *pandas* were used for the handling of the csv files where the tweets and their data were stored. While *tweepy* was used to access the Twitter APIs, *re* and *string* were used to manage the resulting data and finally, *textblob*, *preprocessor* and *nltk* were used to clean and analyze the text resulting from the tweets.

Moving to the actual code, a step-by step explanation, together with the actual batch of code discussed, will be provided in the next pages.

The first step when creating a code (Figure 4) that deals with the Twitter APIs is to set the personal credentials, thanks to which the single user can access the data and the dates to access (if the API category allows to access more days).

```
# Pass Twitter credentials to tweepy
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

# Set two ideal date variables for the date range
start_date = '2020-02-01'
end_date = '2020-05-01'
```

*Figure 4: First part of the Python code*

[11] https://jupyter.org/

The second part of the code (Figure 5) moves to define and describe the *csv* file, where the tweets will be subsequently stored.

Firstly it defines the columns in the files where the following data will be stored: tweet's id, tweet's timestamp (*created_at*), tweet's unique link (*source*), original text, processed text, sentiment, polarity, subjectivity, language, favorite count, retweet count, original author, sensitive (yes/no type variable), hashtags, user mentions, tweet or user's generic geolocation (*place*) and its precise location if available.

```
# Set columns of the csv file
COLS = ['id', 'created_at', 'source', 'original_text','clean_text', 'sentiment','polarity','subjectivity', 'lang',
        'favorite_count', 'retweet_count', 'original_author', 'possibly_sensitive', 'hashtags',
        'user_mentions', 'place', 'place_coord_boundaries']

# Set Happy Emoticons
emoticons_happy = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]', '8)', '=)', ':}',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD', '=-D', '=D',
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', ':-P', ':P', 'X-P',
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)', '>;)', '>:-)',
    '<3'
    ])

# Set Sad Emoticons
emoticons_sad = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||', '=L', ':<',
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":'(", ':\\', ':-c',
    ':c', ':{', '>:\\', ';('
    ])

# Set Emoji patterns
emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                           u"\U0001F680-\U0001F6FF"  # transport & map symbols
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)

# Combine sad and happy emoticons

emoticons = emoticons_happy.union(emoticons_sad)

# Set Stop words:

stop_words = set(stopwords.words('english'))
```

*Figure 5: Second part of the Python code*

After defining the columns, the code sets the different sentiments for two types of emoticons. This method has been used by itself in Go, Bhayani and Huang's (2009) work and has proven successful, thus integrating it in a machine learning-led sentiment analysis can improve and enrich its results.

Afterward the code defines emoji[12] patterns in order to exclude them in a later cleaning, merges the two types of emoticons and sets the stopwords, which are words that are filtered out before the processing of the data starts.

[12] Emojis are digital images that can be added to messages in electronic communication in order to express a particular idea or feeling. Retrieved from https://dictionary.cambridge.org/dictionary/english/emoji

After these preliminary passages, the code moves to define the two main functions that will be used to save and clean the tweets.

```
def clean_tweets(tweet):

    # After tweepy preprocessing the colon left remain after removing mentions
    # or RT sign in the beginning of the tweet
    tweet = re.sub(r':', ' ', tweet)
    tweet = re.sub(r',Ä¶', ' ', tweet)
    tweet = re.sub(r'[^\x00-\x7F]+',' ', tweet) # drop non-ASCII characters
    tweet = emoji_pattern.sub(r'', tweet) # remove emojis from tweet

    # Check tokens against stop words, emoticons and punctuations

    word_tokens = word_tokenize(tweet) # It returns a list of words and puntuation symbols
    filtered_tweet = []

    for w in word_tokens:
        if w not in stop_words and w not in emoticons and w not in string.punctuation:
            filtered_tweet.append(w)

    return ' '.join(filtered_tweet)
```

*Figure 6: Third part of the Python code*

The first function that the code defines is *"clean_tweets"* (Figure 6), it essentially removes all the previously defines stopwords and emojis from the texts, in addition to Twitter's own symbols and language, like mentions and the retweet signs. After this process is over, the function returns the clean tweets, ready to be inserted in their relative column in the final csv file.

The second function defined is the most important and much longer than the first, thus it had to be divided in two figures (Figures 7 and 8) to fit in a page.

The function is called *"write_tweets"* (Figure 7) and firstly, it looks for the destination csv file to see if there is an existing file with the same name that already contains some older tweets, in order to possibly add the new data to it; in case it does not exist, the function creates a new csv file.

After this first passage, the functions gives instruction to the *tweepy* cursor on how to iterate the tweets search: it instructs it not to save non-English tweets in order not to create problems during the analysis and on how to save the different data coming from every tweet on the right variable.

```
def write_tweets(keyword, file):

    # If the file exists, then read the existing data from the CSV file.
    # Otherwise create a new one.
    if os.path.exists(file):
        df = pd.read_csv(file, header = 0)
    else:
        df = pd.DataFrame(columns = COLS)

    # Page attribute in tweepy.cursor and iteration
    for page in tweepy.Cursor(api.search, q = keyword, count = 200, include_rts = False, since = start_date).pages(100):
        for status in page:
            new_entry = []
            status = status._json # Convert the status in json

            ## Check whether the tweet is in english or skip to the next tweet
            if status['lang'] != 'en':
                continue

            # When running the code, the below code replaces the retweet amount and
            # Number of favorires that are changed since last download.
            if status['created_at'] in df['created_at'].values:
                i = df.loc[df['created_at'] == status['created_at']].index[0]
                if status['favorite_count'] != df.at[i, 'favorite_count'] or \
                   status['retweet_count'] != df.at[i, 'retweet_count']:
                    df.at[i, 'favorite_count'] = status['favorite_count']
                    df.at[i, 'retweet_count'] = status['retweet_count']
                continue


        # Tweepy preprocessing is called for basic preprocessing
        # It removes:  URLs, Hashtags, Mentions, Reserved words (RT, FAV), Emojis, Smileys

            clean_text = p.clean(status['text'])

            # Call clean_tweet method for extra preprocessing

            filtered_tweet = clean_tweets(clean_text)

            # Pass textBlob method for sentiment calculations:

            blob = TextBlob(filtered_tweet)
            Sentiment = blob.sentiment

            # Seperate polarity and subjectivity in to two variables

            polarity = Sentiment.polarity
            subjectivity = Sentiment.subjectivity

            # New entry append

            new_entry += [status['id'], status['created_at'], status['source'], status['text'],
                          filtered_tweet, Sentiment, polarity, subjectivity, status['lang'],
                          status['favorite_count'], status['retweet_count']]
```

*Figure 7: Fourth part of the Python code*

In addition to the cleaning capabilities of the *preprocessor* package, *tweepy* has some of its own. After scraping the tweets, the function utilizes both packages to clean the resulting data, removing the URLs, Hashtags, Mentions, Reserved words (*RT*, *FAV*), Emojis and Smileys it contains.

When the cleaning is completed, the function passes the processed text to *textblob* for the sentiment calculation, which it computes through the use of a Naïve-Bayes classifier. *Textblob* determines the sentiment in two different variables that are finally added in their respective rows on the csv file: polarity and subjectivity. The first variable is a float within the range [-1.0 (very negative), 1.0 (very positive)] and it measures the overall sentiment of the tweet, while the second is a float within the range [0.0, 1.0] where

0.0 is very objective and 1.0 is very subjective, meaning that it tries to understand if the tweet does not contain an opinion and if it does it measures how objective it is.

```
            # To append original author of the tweet
            new_entry.append(status['user']['screen_name'])

            # To append if there is sensitive data:
            try:
                is_sensitive = status['possibly_sensitive']
            except KeyError:
                is_sensitive = None
            new_entry.append(is_sensitive)

            # Hashtagas and mentiones are saved using comma separted
            hashtags = ", ".join([hashtag_item['text'] for hashtag_item in status['entities']['hashtags']])
            new_entry.append(hashtags)
            mentions = ", ".join([mention['screen_name'] for mention in status['entities']['user_mentions']])
            new_entry.append(mentions)

            # Get the location of the tweet if possible:
            try:
                location = status['user']['location']
            except TypeError:
                location = None
            new_entry.append(location)

            try:
                coordinates = [coord for loc in status['place']['bounding_box']['coordinates'] for coord in loc]
            except TypeError:
                coordinates = None
            new_entry.append(coordinates)

            single_tweet_df = pd.DataFrame([new_entry], columns=COLS)
            df = df.append(single_tweet_df, ignore_index = True)
            csvFile = open(file, 'a', encoding = 'utf-8')
        df.to_csv(csvFile, mode = 'a', columns = COLS, index=False, encoding="utf-8")
        csvFile.close()

# Declare keywords as a query for three categories

nintendodirect_keywords = '#NintendoDirect OR #nintendodirect'
mycalvins_keywords = '#MyCalvins OR #mycalvins OR #Mycalvins OR #myCalvins'
priceless_keywords = '#PricelessSurprises OR #pricelesssurprise OR #PricelessSurprise OR #Pricelesssurprise'

# File location path
nintendodirect_tweets = "nintendodirect_data.csv"
mycalvins_tweets = "mycalvins_data.csv"
priceless_tweets = "priceless_tweets_data.csv"


# Call main method passing keywords and file path
write_tweets(nintendodirect_keywords,  nintendodirect_tweets)
write_tweets(mycalvins_keywords, mycalvins_tweets)
write_tweets(priceless_keywords, priceless_tweets)
```

*Figure 8: Fifth part of the Python code*

After these measures, the function moves on (Figure 8) to write all the possible data it can find about the single tweets on the csv file.

It starts by finding the tweet author, then it defines if the tweet is sensitive and contains unsafe content for analysis, it then saves hashtags and mentions in the tweet in a separate column, then lastly, it tries to obtain a location for the tweet. In case the tweet is geotagged, the function saves the geotag in a csv column, if not, it goes on the profile of the user that posted the tweet, looks for his source and it saves it as representative of the tweet.

The final part of the code in Figure 8 also defines three different searches on hashtag campaigns in a three parts process:

1. The keywords are defined and assigned to an object. In the case of this research hashtags are being analyzed, but, since Python makes differences between uppercase and lowercase characters, but Twitter does not, all possible combinations are saved.
2. The file location paths for the final csv files are defined and assigned to their respective objects.
3. The *write_tweets* function is called using the two objects just defined for keywords and paths.

The following part of the code (Figure 9 to 12) is shown applied to the first case in this thesis. Since the other cases are the same in their processes and vary only in their keywords and results, the first is taken as representative.

```
# Import DF:

nintendo_direct_df = pd.read_csv(nintendodirect_tweets)

# Not interested in subjectivity 0 (drop those rows):

nintendo_direct_df = nintendo_direct_df[nintendo_direct_df['subjectivity'] != 0]
nintendo_direct_df.reset_index(drop = True, inplace = True)

# Fill Nan in hastags clumn:

nintendo_direct_df['hashtags'] = nintendo_direct_df['hashtags'].fillna('')
```

*Figure 9: Sixth part of the Python code*

This final part of the code (Figure 9) moves to analyze and interpret the result of the sentiment analysis. The first steps it takes are to import the database that will be studied and to entirely drop the columns that do not contain subjective tweets (the ones where the variable *subjectivity* is equal to 0), since they will not be useful for the successive measurements.

The next operation that the code performs is the filling of the empty columns in the "*hashtags*" column. This is done so that, when analyzing the single words in a later stage, it will not give problems.

```
# Get tot. polarity and subjectivity:

direct_polarity_avg = round(nintendo_direct_df["polarity"].mean(), 3)
direct_subjectivity_avg = round(nintendo_direct_df["subjectivity"].mean(), 3)

print('Polarity: {}\nSubjectivity: {}'.format(direct_polarity_avg, direct_subjectivity_avg))
```

```
Polarity: 0.152
Subjectivity: 0.41
```

*Figure 10: Seventh part of the Python code*

After the aforementioned processes are completed, the code measures an average of the remaining rows' polarity and subjectivity columns (Figure 10). This general value (rounded to the third number after the zero) is useful for a quick approach to the data; in fact, it allows to grasp the overall measures for the entire dataset.

For a more in-depth approach, however, the code continues by evaluating the single hashtags used in conjunction with the ones the research sets out to measure (Figure 11). To do so, the code creates a list of all the hashtags used in every tweet and finds the most commonly used, after dropping the rows previously set as empty.

Finally, the code picks the 10 most common hashtags from the list and pairs them with the mean of the sentiment associated with their original tweets.

```
# Get all hashtags:

all_hash = []
for i in range(len(nintendo_direct_df)):

    h = nintendo_direct_df["hashtags"][i] # Get hashtags of row i
    h_lst = h.split(", ") # get each tag as string
    h_lst = [h.lower() for h in h_lst]
    all_hash.extend(h_lst) # Add all tags to the list

# Find top 10 hashtags:

count_hash = Counter(all_hash)
del count_hash[''] # drop nan count
best_hash = [h for h, c in count_hash.most_common(10)]

# Find avg Polarity of top hashtags:

dic_avg_pol = {}
for h in best_hash:
    polarity_arr = nintendo_direct_df[nintendo_direct_df["hashtags"].str.contains(h, case = False)]["polarity"]
    mean = polarity_arr.mean(skipna = True)
    dic_avg_pol[h] = round(mean, 3)

dic_avg_pol
```

```
{'nintendodirect': 0.141,
 'nintendoswitch': 0.136,
 'nintendo': 0.139,
 'legendofzelda': 0.429,
 'minidirect': 0.056,
 'majorasmask': 0.402,
 'mariobros': 0.27,
 'gameandwatch': 0.223,
 'e32020': 0.017,
 'gaming': 0.085}
```

*Figure 11: Eight part of the Python code*

The last part of the code (Figure 12) analyzes the locations of the processed tweets. It does so in a similar way to what it previously does with hashtags, finding the 10 locations where the keywords that the research wants to measure are used the most.

```
# Get all places:

all_place = []
for i in range(len(nintendo_direct_df)):

    p = nintendo_direct_df["place"][i] # Get places of row i
    all_place.append(p) # Add all tags to the list

# Find top 10 places:

count_place = Counter(all_place)
del count_place[''] # drop nan count
best_place = [p for p, c in count_place.most_common(10)]

best_place
```

```
[nan,
 'somewhere north of Toronto',
 'Ohio',
 'Louisiana, USA',
 'O Fallon',
 'United States',
 'Polska',
 'Existence',
 'England, United Kingdom',
 'onlycheapdeals@gmail.com']
```

*Figure 12: Ninth part of the Python code*

## 4.2. The hashtag campaigns

### 4.2.1. *#NintendoDirect*

The first campaign that this thesis will analyze was launched by Nintendo for the first time in 2011 and has been used until today to group tweets and posts that relate to the firm's online presentations and live shows[13], where information regarding the company's content or franchises is presented, such as news about upcoming games and consoles.

The format of the *Nintendo Direct* has been mostly an internationally available live streaming lasting 30 to 50 minutes, usually presented simultaneously in English by the president of Nintendo of America and in Japanese by the president of Nintendo of Japan.

[13] https://www.youtube.com/playlist?list=PLPh3p_yYrx0CjUKA7c0f8K20wMD-crPMx

It has been aired in various styles, including "*Nintendo Direct Mini*" and "*Indie Highlights*", respectively a shorter version of the traditional *Direct*, capping at 15 minutes and a presentation focusing on titles created by independent developers coming to Nintendo consoles.


*Figure 13: Nintendo of America's tweet to announce the latest Nintendo Direct*

Since these kinds of presentations have been the most important way for the company to announce important news and updates, their online following is massive, with peeks of millions of users per video and discussions about them that continue for weeks or even months after a single event has ended. Because of this, other companies in the same business have tried to mimic the strategy with various degrees of success[14].

Due to the peculiar nature of this public live announcements, the hashtag allows for the creation of *ad-hoc publics*, that enable users to comment the single announcements immediately as they are presented and to express their feelings about them. The homonymous hashtag is of course not only used during these presentations, that usually happen around 10 times per year, but is used to express feelings resulting from them even

[14] https://www.playstation.com/it-it/explore/state-of-play/

after their airing, or the will to get more news and announcements before a *Direct* is even announced. Because of this, a sentiment measurement during a month when there has been no *Nintendo Direct* can lean towards the negative side of the spectrum, while calculating the sentiment of a month that has had one or more presentations will yield a very different result.

Looking at the data resulting from the sentiment analysis of a total of over 7000 tweets that contain the hashtag *#NintendoDirect*, we have a Polarity score of 0,152 and a Subjectivity score of 0,410.

The average Subjectivity score, together with a rapid qualitative assessment of the tweets, show that most of the comments relative to *Nintendo Directs* are actually coming from news sources, that just post about the latest announcements without expressing their opinion.

The low but positive Polarity score, on the other hand, shows that the sentiment linked with these kinds of announcements is mainly positive, suggesting the company to continue operating this way.

Looking at the other most common hashtags, together with their sentiments, allows a company to see what topics are commonly associated with their campaign. In this case we can see that #NintendoSwitch is the second hashtag by popularity and has a Polarity score of 0,108; being the only Nintendo console that users talk about and having a positive sentiment show its strength on the gaming market.

Looking at the other results, *#LegendOfZelda* is the fourth hashtag by popularity with a Polarity of 0,429, while *#MajorasMask* is sixth with a polarity of 0,402, meaning that discussions that involve topics surrounding this popular videogame series[15] have an overall higher sentiment and users commonly talk about it. *#MarioBros* is also in the top 10 hashtags by use with a lower but still positive Polarity of 0,270.

The presence of both game series allows to understand that they are a positive driver for sentiment and users and fans are eager to talk and to share their personal feelings about them.

[15] The Legend of Zelda is an action-adventure video game series that started in 1986 and was developed and published by Nintendo and originally designed by Shigeru Miyamoto and Takashi Tezuka. The Legend of Zelda: Majora's Mask is the sixth episode in the series, and it debuted in 2000.

At last, in the tenth position we also have #E32020, the hashtag of the now canceled gaming expo of June 2020[16]. The discontent with the unfortunate annulment of the event and the news about the lack of Nintendo's presence during it, bring its Polarity score to a lower 0,017.

When analyzing the top locations by usage of the hashtag, locations inside the United States are prominent, with a first position occupied by the Country, followed by different American cities and states like Ohio, Louisiana and New York, in addition to the United Kingdom. This shows the success of the campaign in the US, where conversations about this topic are more common, but also more in general in English-speaking countries, suggesting a focus on these for future campaigns.

### 4.2.2. *#MyCalvins*

The second campaign this work will analyze was originally released by Calvin Klein to promote its new underwear collection with a video on Youtube on January 6, 2015 and is still active by this day[17].

The video[18], starring singer Justin Bieber and model Lara Stone, was uploaded firstly on the brand's channel and then on the singer's. Since its release, it has generated a total over 9,7 million views and 27.000 likes. A print campaign with Bieber was also released following the premiere of the video. The peculiarity of the video is that it is entirely shot in black and white and has no text or voice until the end, when *#MyCalvins* appears onscreen, followed by the Calvin Klein Jeans logo and a link to *calvinklein.com*.

The hashtag, *#MyCalvins*, is incorporated directly into the campaign in order to foster audience engagement. The campaign video contains narratives of self-expression, sexuality, and pop culture and is mostly focused on the target of young Millennials between the ages of 18 and 25, with Justin Bieber acting as the initial social influencer.

[16] E3, also known as the Electronic Entertainment Expo, is a trade event for the video game industry held in June in Los Angeles. Its latest edition has been fully canceled on account of the 2019–20 coronavirus pandemic.
[17] Campaign page: https://www.calvinklein.com.au/mycalvins
[18] Original video: https://www.youtube.com/watch?v=K0t-aBAYym8

The description of the video that started the campaign contains, at the very end, a message that encourages audience participation to the campaign by showing their own *#MyCalvins* look on social media.

After the initial launch of the *#MyCalvins* campaign with Bieber, Calvin Klein continued to build on it by partnering with models, musicians, and social influencers, using and encouraging to use the phrase, "I ____ in *#MyCalvins*", where the blank space was left to be filled by the users, letting them express their ideas and experiences with the brand and its products.

The most remarkable part of this campaign is the flux of User Generated Content and audience engagement it is able to create. In fact, the *#MyCalvins* hashtag incorporated into the ads prompted high audience engagement and generated more than 1.6 million interactions in just the first 48 hours post-launch (Fumo, 2015).

While Calvin Klein has obviously used influencers and celebrities to take part in the campaign, the flow-on effect has been so heavy that users who were not even brand ambassadors started posting pictures of themselves in their CK underwear to take part in the global campaign. Calvin Klein fueled that positive reaction with another successful idea, starting a dedicated micro-website where the best user-generated content was recognized by the brand and shown on a curated social wall.

After the effective start of the campaign in 2015, Calvin Klein continued to build on its success by incorporating all of its other brands, alongside underwear, in the ads and partnering with many additional influencers, actors, musicians and models including Kendal Jenner, Kate Moss, Bella Hadid and Zoe Kravitz. The message remained around the phrase "I ____ in *#MyCalvins*", with different fillings depending on the endorsers and with an open invitation to the world to fill the blank[19].

In 2017 Calvin Klein further expanded the campaign with the launch of the *"Our Family. #MyCalvins"* series of advertising and videos, revolving around interpretations of the term family together with the older message and with the inclusion of celebrities siblings[20].

[19] The campaign ads can be found at https://campaignsoftheworld.com/outdoor/calvin-klein-spring-2016-ad-campaign-mycalvins/

[20] *"Our Family. #MyCalvins"* video: https://www.youtube.com/watch?v=90CCjgX0n20&list=UUuf6cvFcYdpHZ_EjSIcBZlg&index=156

Finally, in 2019 Calvin Klein launched its latest campaign, "I Speak My Truth *#MyCalvins*." featuring singers and celebrities like Billie Eilish in a series of videos sharing their personal truths in an emotional way. In the brand's words, the new message shows "today's most influential voices telling their own stories, in their own words – and invites others around the world to do the same."[21]

Looking at the data resulting from the sentiment analysis of a total of over 5000 tweets that contain the hashtag *#MyCalvins*, we have a Polarity score of 0,058 and a Subjectivity score of 0,672.

The low but positive Polarity score is difficult to explain without looking at the tweets, that in fact show a positive response relative to the brand and a negative one relative to some of the personalities involved with the ads.

The high Subjectivity score, on the other side, shows that the almost the 70% of the tweets come from people that express emotions with their messages.

Analyzing the most common hashtags with their Polarity values, we can see that the most successful is *#briefs* with a Polarity of 0.75, demonstrating how the original object of the campaign still receives the strongest push from it, and suggesting CK to continue focusing on this segment. *#MyThruth* is the third hashtag for popularity and the second for sentiment, with a value of 0.6, already showing the strength of the newest campaign launched, as well with its adoption rate.

Another positive reception can be found in the hashtag *#Maluma*, showing that the cooperation with the Colombian singer has been fruitful. Other hashtags like *#CalvinKleinPerformance* (0.261) and *#ModernCotton* (0.267) have a positive sentiment, demonstrating how the campaign has a positive effect on the materials and collections advertised through it.

A negative sentiment, on the other hand, can be seen when analyzing #CKCoachella (-0.3), showing that the reception for the collaboration between the brand and the event[22] has been negative and suggesting Calvin Klein to try to pin down what has caused this bad perception and to try to understand it in order not to repeat the same mistakes in the future.

[21] https://www.calvinklein.com.au/mycalvins
[22] https://www.calvinklein.com.au/coachella

The top locations paint a very broad picture about the geolocation of the people that engaged with the campaign. Canada, United States, Philippines, Brazil and France all appear to be in the most common locations, showing a global outreach for CK.

### 4.2.3. *#TasteTheFeeling* and *#ShareaCoke*

The third campaign this thesis will analyze was launched in 2016 by The Coca-Cola Company to unite its different brands *Coca-Cola, Diet Coca-Cola, Coca-Cola Zero, and Coca-Cola Life* under one slogan: "Taste the Feeling".

The campaign was launched with television and web ads and it featured an "anthem" written for it by the late musician Avicii, together with singer Conrad Sewell[23]. The anthem was also used in the campaign of the UEFA Euro 2016 cup and Rio 2016 Olympic Games, of which The Coca-Cola company was a sponsor.

The videos of the campaign show a series of different emotionally charged moments in which the different types of Coke are enjoyed, finishing with all the Coca-Cola products coming together under the red Coca-Cola logo with the *#TasteTheFeeling* hashtag on the bottom of the picture. The overall core message of these ads is: "The simple pleasure of drinking any Coca-Cola makes the moment more special" (Heilpern, 2016). The company has linked the campaign to such a positive sentiment with the hope of changing the conversation around its drinks, which has been plagued by critics' concerns around obesity and other health issues.

The Company's Chief Marketing Officer, Marcos De Quinto, stated[24] that "This is a powerful investment behind all Coca-Cola products, showing how everyone can enjoy the specialness of an ice-cold Coca-Cola, with or without calories, with or without caffeine. […] The bigness of Coca-Cola resides in the fact that it's a simple pleasure - so the humbler we are, the bigger we are. We want to help remind people why they love the product as much as they love the brand."

[23] Campaign video available at: https://www.youtube.com/watch?v=F82W3tKtr8c

[24] https://www.coca-colacompany.com/au/media-centre/media-releases/coca-cola-announces-one-brand-global-marketing-approach

Coca-Cola's ads aired in more than 200 countries and were mainly targeted at millennials, the characters in the ad are in fact seemingly selected from the audience the company is targeting. It is young people – boys and girls having fun and partying. The ads do nothing to describe the product itself, but instead assume that its audience is already familiar with both the brand and its drinks.

The purpose of the ad is thus of simple brand recall. The scenes in the ad are picked from the common life and attempt to represent the lifestyle of the target generation.

The ad tries to connect with millennials' feelings and tries to reflect the same passion that is associated with the young generation. The message that the ads end up sharing is that Coca-Cola is made for the youth and that people can add joy into their lives simply by adding these signature Coca-Cola moments.

Moving to the sentiment analysis of the *#TasteTheFeeling* hashtag, run on a total of almost 12000 tweets, it has a Polarity of 0.345, and a Subjectivity of 0.543.

The positive Polarity shows that the objective of improving the perceived sentiment with the brand has been a success, in fact the indicator shows that the average opinion on the matter is positive. The Subjectivity value instead shows that the majority of the tweets written are from actual people instead of descriptions of news and facts.

Analyzing the additional hashtags shows how much a campaign can change its course when a particular event occurs. In fact, the two keywords with the higher sentiment are *#StayHome* (0.645) and *#StaySaveStayHome* (1.0) and are connected to The Coca-Cola Company because of the important stand that it took on the Covid-19 matter.

The company, in fact, decided to put social distancing messages with its logo on ads in Times Square in New York (Figure 14) and creating an ad dedicated to spreading the message to share on social networks[25].

[25] Coca-Cola's video on Covid-19: https://www.youtube.com/watch?v=oRfV2xHrvss

The very high sentiments connected to the two hashtags show how much good an important stand like this can make on a company's reputation. The hashtag *#StaySaveStayHome*, for example, has the higher possible sentiment value at 1.0, showing that the level of admiration and positivity generated this way around the brand is difficult to obtain with a standard marketing campaign.


*Figure 14: Coca-Cola advertising in Times Square*

Looking at the geotags connected to the tweets gives another interesting information about this campaign: its success in African countries. In fact, most of the tweets analyzed come from Uganda, Nigeria, Kenya and South Africa.

The explanation of this phenomenon is possible by analyzing the tweets qualitatively; a process that shows that African food-delivery businesses, like *Jumia Food* and *SafeBoda*, started a partnership with the Coca-Cola company to create promotions in their menus when combined with their beverages with any order in order to both promote their services and convince people to stay at home during the pandemic.

The fourth campaign this thesis analyzes was also launched by the Coca Cola Company, but in 2016. It initially involved swapping parts of the logo on their products

with common names in the regions where it was running[26]. This encouraged customers to find bottles and cans with names that had a personal meaning to them, share them with their friends and family and then tweet about their experiences, using the hashtag *#ShareaCoke*.

After its first year, the campaign evolved to introduce the possibility for customers to order online personalized bottles with any chosen name on it, in addition to introducing terms like "*Star*" or "*BFF*" for those with more unusual names, that were not represented.

By analyzing the sentiment results of two campaigns run by the same brand, it is possible to grasp the difference in consumer's reactions to the emotions that they create.

The sentiment analysis of the *#ShareaCoke* hashtag, was run on a total of almost 10000 tweets, and resulted in a Polarity of 0.325, and a Subjectivity of 0.665.

The positive Polarity value is on par with the analysis of previous campaign, while the higher Subjectivity value (0.665 vs 0.543) shows that this hashtag is used more by people expressing their feelings and thoughts than by brands that use it to promote news and for advertising.

A key result from the sentiment analysis is that, by studying the sentiment values of the additional hashtags it is possible to evaluate the different brand perception of Coca Cola when connected to either of the two campaigns. In fact, the polarity of the hashtag #CocaCola, when connected to #TasteTheFeeling is 0.52 while the polarity when connected to #ShareaCoke is 0.8.

This represents a big difference in the reaction that customers have to the two different campaigns: the Taste the Feeling campaign, probably because of the weaker social participation, makes consumers perceive the brand as positive, but the sharing and commonality embedded in the Share a Coke campaign pushes the evaluation of the brand almost to the maximum value. The embedded sociality can also be derived from another hashtag in the most common 10 connected to the campaign: *#StrongerTogether* has again an evaluation of 0.8 and represents the core message of the campaign.

It is also interesting to notice how the brand has tried to keep the campaign active over time by creating the *#NationalHaveaCokeDay* hashtag (with a positive polarity of

[26] https://www.coca-colacompany.com/au/news/share-a-coke-how-the-groundbreaking-campaign-got-its-start-down-under

0.302 and second on the most used list), that works as a revival for the campaign over years and is successfully able to bring it back on the spotlight.

In this hashtag's case, the geotag analysis shows that its popularity is strong almost only in the United States, followed by Brazil. In this case the results are not influenced by smaller brands' campaigns since this hashtag (as also shown by the higher subjectivity value) is mostly oriented to Coca-Cola's customers to express their emotions.

### 4.2.4. *#ElonMusk*

The last hashtag this work will analyze is different from the other campaigns in the sense that it was not created or developed by a brand. In fact, Tesla, the electric car company, has distinguished itself for not adhering to traditional marketing and instead relying on different and novel techniques[27].

Tesla was able to understand how advertising in its segment had become less effective than before and realized that instead having fans sell their product was more effective than anything else they could do. Thus, their advertising budget is almost null. Their main focus is on turning customers into fans. (Andersen, 2017)

They accomplish this task through their authenticity and controversy[28]: their CEO, Elon Musk, is a social-savvy and is open about his life and thoughts on Twitter and has publicly smoked weed, sold flamethrowers, sent his first Tesla car into orbit and smashed two shatter-proof windows of their new model on its presentation. All of this, however, resulted in the sale of more than 250 thousand units of that car upfront on the same day[29].

This happened because his characteristics end up creating a wealth of user-generated content around every new announcement and communication, that in turn can yield much bigger revenues than standard commercials.

[27] https://www.ninjamarketing.it/2019/11/25/cybertruck-tesla-elon-musk-pickup/

[28] https://www.businessinsider.com/tesla-problems-2019-autopilot-elon-musk-tweets-2019-6?IR=T#musks-ambien-use-reportedly-worries-board-members-11

[29] https://www.businessinsider.com/elon-musk-reports-250000-pre-orders-for-tesla-cybertruck-2019-11?IR=T

Tesla's main driver that brings together the user-generated content around the brand and, simultaneously, their constant campaign can be found in the CEO itself, connected to the hashtag #ElonMusk.

Analyzing over 56000 tweets that contain the hashtag #ElonMusk paints an uncertain picture about the Tesla brand. The polarity is positive but low at 0.123 while the subjectivity is average at 0.505. This last value shows that half of the tweets shared on the social network are actually just news reports and announcements regarding the brand, instead of opinions.

Among the most popular hashtags, Tesla is present at the second position with a negative polarity of -0.036. This value could be explained from the continuous controversies surrounding the brand, one of which was reported in the news at the time of the writing of this thesis and regards Elon Musks son and, more precisely, the name he and his partner have chosen for him[30]. Both #ElonMuskBaby, #Grimes and #XAEA12 are in fact among the top ten hashtag and all of them have a positive value of polarity that range from 0.082 to 0.213, meaning that even personal events like this can bring people's perception about the brand forward, thanks to the close association between the CEO and its company.

SpaceX, Musk's aerospace company is also in the most used hashtag with a positive polarity of 0.275, showing that, due to the higher degree of separation between the second company and its leader, the former end up being less affected by the latter's personal businesses.

Considering the top locations, 50% of them are in the United States, showing a clear predominance of this market in the company's influence.

In this particular case, the hashtag (*#ElonMusk*) is difficult to frame in a particular role. Its use started with the Tesla brand and was employed to link its CEO to the brand. Over time, however, the individual became more popular than the firm and their roles have been turned, so that now Elon Musk's actions can have a negative impact on Tesla's perception.

All of this results in the impossibility, for the algorithm, to understand whether it is measuring Musk's or Tesla's sentiment. This distinction, however, is possible if a human

[30] https://edition.cnn.com/2020/05/06/entertainment/grimes-elon-musk-baby-name-intl-scli/index.html

researcher analyzes the data at hand, which, in this case, shows a clear focus on the person instead of the brand. This can be determined from the fact that only 1 of the 10 most used hashtags are connected to Tesla, while 4 are personally related to its CEO.

# Conclusion

Sentiment analysis has been presented as a merge between qualitative and quantitative analysis, with the objective of using one to improve the other. Kelle, in his numerous studies (1998, 2002, 2003, 2006), showed in fact one could aid the other, while raising data quality, by supplementing quantitative data with qualitative insights.

The mix of the two methods that is in play in sentiment analysis can be considered a sequential quantitative-qualitative design (*quan* > *qual*). A quantitative research is carried in order to better narrow the problem areas and research questions, that are further investigated with the help of qualitative methods and data.

In this case the qualitative method analyzes an amount of tweets that no human could and yields numerical results about the sentiment and the top keywords, afterwards a quantitative exploration and study of the results allows the researcher to fully grasp the outcome of the campaign analyzed.

This design helps to cope with two general problems of quantitative research: the difficulty to understand the quantitative data without the proper sociocultural knowledge and the doubt that the research is focusing only on remote or marginal cases.

The method employed to quantitatively calculate the sentiment used a Naïve Bayes classifier with a corpus created with dictionary-based approach. This approach has been adopted since it represents the most resource-efficient and fastest way to create a sentiment set. In fact, by starting from a sentiment dictionary that contains human-tagged words and expanding it using machine learning, it is possible to analyze through Natural Language Processing almost any kind of text. In the case of this research, tweets lend themselves particularly well to the analysis since they are notably short (no more than 280 characters) and tend to contain their author's full opinion.

The merits of the human encoding, used to create the original corpus, are that, based on the researcher's ability, each text can be encoded with a very low margin of error, regardless of the language used, the context of discussion, the use of metaphors or rhetorical figures.

In addition to the corpus-based approach, this thesis also employs a second way to better classify tweets. An emoticon dictionary, with each symbol and its sentiment value, is used to assign their values to tweets, instead of deleting them, as a corpus-based

approach normally would do. This method has been used by itself in Go, Bhayani and Huang's (2009) work and has proven successful, thus integrating it in a machine learning-led sentiment analysis can improve and enrich its results, by considering the entirety of the tweet instead of a part of it.

Other key differentiators from past research are the introduction of a subjectivity measurement and the analysis of the most used hashtags and most common locations. The first has made it possible to exclude texts that did not contain their authors' opinions in them from the sentiment count, giving an overall better result in the final sentiment analysis. The second gives marketers and researchers the possibility to not only analyze the sentiment behind a hashtag, but also understand what the main topics discussed from that community are and where those discussions are taking place.

Moving to the matter analyzed, hashtag campaigns are of particular interest in recent literature since they are the embodiment of a new type of marketing environment. The rise of these new communication channels has caused a disruptive change in the marketing environment. In fact, the information about a brand has turned into a multidirectional, interconnected flow and this makes it difficult to predict.

Marketers have lost the complete control they previously had over their brands and campaigns and now can only participate in a "conversation" about the brand.

In the era of new media, managing customer relationships can be compared to the game of pinball: companies launch a marketing "ball" that consists of their brands and brand-building messages into an unpredictive environment, which can be deviated or accelerated by new media and interactions that act as "bumpers", changing its course in chaotic ways. After the marketing ball is in play, marketing managers can continue to guide it with agile use of the "flippers", but the ball does not always go where they intended it to, and their slightest miscalculation can be amplified into a crisis. (Hennig-Thurau et al., 2010)

Analyzing different hashtag campaigns has shown the potential of the sentiment analysis to aid marketers in their work in different situations and with different numbers and end results.

With all the new kinds of marketing communications, always keeping the situation under control becomes of utmost importance: it can be easy for a campaign to derail or for it not to produce its intended results, but all it needs to regain momentum is a push in

the right direction. This kind of action can only be performed by merging an in-depth quantitative analysis of the entire campaign made by a machine with a qualitative assessment made by the marketer.

The main limitation of this work lies into the type of permission used to analyze the tweets, that only gave access to a limited number of them at the time. Future studies can improve on this by designing a workaround to access Twitter's data without using its APIs, thus limiting however the type of data available to be accessed, or by purchasing an *Enterprise API* access from Twitter.

# Appendix

## Values for the hashtag campaigns:

1. #NintendoDirect:

```
N=7268
Polarity: 0.137
Subjectivity: 0.415

{'nintendodirect': 0.133,
 'nintendoswitch': 0.108,
 'nintendo': 0.125,
 'legendofzelda': 0.429,
 'minidirect': 0.056,
 'majorasmask': 0.402,
 'mariobros': 0.27,
 'gameandwatch': 0.223,
 'switch': 0.108,
 'e32020': 0.017}

[nan,
 'United States',
 'Ohio',
 'Louisiana, USA',
 'O Fallon',
 'somewhere north of Toronto',
 'New York, USA',
 'Existence',
 'Polska',
 'England, United Kingdom']
```

2. #MyCalvins

```
N=5132
Polarity: 0.058
Subjectivity: 0.672

{'mycalvins': 0.021,
 'calvinklein': 0.137,
 'mytruth': 0.6,
 'ckcoachella': -0.3,
 'briefs': 0.75,
 'maluma': 0.35,
 'calvinkleinunderwear': -0.04,
 'calvinkleinjeans': 0.167,
 'calvinkleinperformance': 0.261,
 'moderncotton': 0.267}

[nan,
 'United States',
 'Canada',
 'Los Angeles, CA',
```

```
 'Republic of the Philippines',
 'San Diego, CA',
 '-in my own skin-',
 'Curitiba, Brasil',
 'Paris, France',
 'Chicago, IL']
```

3. #TasteTheFeeling

```
N=11764
Polarity: 0.345
Subjectivity: 0.543
```

```
{'tastethefeeling': 0.2,
 'stayhome': 0.645,
 'staysafestayhome': 1.0,
 'cocacola': 0.52,
 'actors': 0.075,
 'artists': 0.075,
 'comics': 0.075,
 'filmmakers': 0.05}
```

```
['Kampala, Uganda',
 'Uganda',
 'Mbarara and Namanve',
 'United States',
 'Kampala',
 'Entebbe | Kigezi | Kisaasi',
 'Entebbe, Uganda ',
 'Uganda, Kenya & South Africa',
 nan,
 'La République de Libertalia']
```

4. #ShareaCoke

```
N=9658
Polarity: 0.325
Subjectivity: 0.665
```

```
{'shareacoke': 0.305,
 'nationalhaveacokeday': 0.306,
 'cocacola': 0.8,
 'coke': 0.329,
 'strongertogether': 0.8,
 'teamjl': 0.037,
 'yummy': 0.5,
 'support': 0.5,
 'restaurants': 0.5,
 'shrimppoboy': 0.5}
```

```
[nan,
 'Kingsport, Tennessee, USA',
 'Amapá, Brasil',
 'Bristol Baby, TN',
 'Encinitas, CA US',
 'Harrisburg, NC',
 'North Carolina,USA',
 'The Swamp',
 'Charlotte, NC',
 'Orlando, Florida']
```

5. #ElonMusk

```
N=56782
Polarity: 0.123
Subjectivity: 0.505
```

```
{'elonmusk': 0.107,
 'tesla': -0.036,
 'billgates': 0.381,
 'pandemic': -0.296,
 'elonmuskbaby': 0.213,
 'spacex': 0.275,
 'grimes': 0.082,
 'coronavirus': 0.104,
 'california': 0.01,
 'xaea12': 0.124}
```

```
[nan,
 'Worldwide',
 'United States',
 'California, USA',
 'Los Angeles, CA',
 'USA',
 'India',
 'London, England',
 'California',
 'Mumbai, India']
```

**Complete Python code:**

```
import os
import pandas as pd
from statistics import mode
from collections import Counter
import tweepy
import re
import string
from textblob import TextBlob
import preprocessor as p
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Twitter credentials for the app
consumer_key = 'REDACTED'
consumer_secret = 'REDACTED'
access_key= 'REDACTED'
access_secret = 'REDACTED'

# Pass Twitter credentials to tweepy
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

# Set two ideal date variables for the date range
start_date = '2020-02-01'
end_date = '2020-05-01'

# Set columns of the csv file
COLS = ['id', 'created_at', 'source', 'original_text','clean_text',
'sentiment','polarity','subjectivity', 'lang',
        'favorite_count',    'retweet_count',    'original_author',
'possibly_sensitive', 'hashtags',
        'user_mentions', 'place', 'place_coord_boundaries']

# Set Happy Emoticons
emoticons_happy = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]', '8)',
'=)', ':}',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD', '=-
D', '=D',
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', ':-P',
':P', 'X-P',
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)', '>;)',
'>:-)',
    '<3'
    ])
```

```
# Set Sad Emoticons
emoticons_sad = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||', '=L',
':<',
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":'(",
':\\', ':-c',
    ':c', ':{', '>:\\', ';('
    ])

# Set Emoji patterns
emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"    #  symbols  &
pictographs
                           u"\U0001F680-\U0001F6FF"   # transport &
map symbols
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)

# Combine sad and happy emoticons

emoticons = emoticons_happy.union(emoticons_sad)

# Set Stop words:

stop_words = set(stopwords.words('english'))

def clean_tweets(tweet):

    # After tweepy preprocessing the colon left remain after removing
mentions
    # or RT sign in the beginning of the tweet
    tweet = re.sub(r':', ' ', tweet)
    tweet = re.sub(r',Ä¶', ' ', tweet)
    tweet = re.sub(r'[^\x00-\x7F]+',' ', tweet) # drop non-ASCII
characters
    tweet = emoji_pattern.sub(r'', tweet) # remove emojis from tweet

    # Check tokens against stop words, emoticons and punctuations

    word_tokens = word_tokenize(tweet) # It returns a list of words
and puntuation symbols
    filtered_tweet = []

    for w in word_tokens:
```

```
        if w not in stop_words and w not in emoticons and w not in
string.punctuation:
            filtered_tweet.append(w)

    return ' '.join(filtered_tweet)

def write_tweets(keyword, file):

    # If the file exists, then read the existing data from the CSV
file.
    # Otherwise create a new one.
    if os.path.exists(file):
        df = pd.read_csv(file, header = 0)
    else:
        df = pd.DataFrame(columns = COLS)

    # Page attribute in tweepy.cursor and iteration
    for page in tweepy.Cursor(api.search, q = keyword, count = 200,
include_rts = False, since = start_date).pages(100):
        for status in page:
            new_entry = []
            status = status._json # Convert the status in json

            ## Check whether the tweet is in english or skip to the
next tweet
            if status['lang'] != 'en':
                continue

            # When running the code, the below code replaces the
retweet amount and
            #  Number  of  favorires  that  are  changed  since  last
download.
            if status['created_at'] in df['created_at'].values:
                i        =        df.loc[df['created_at']        ==
status['created_at']].index[0]
                if     status['favorite_count']     !=     df.at[i,
'favorite_count'] or \
                   status['retweet_count']       !=       df.at[i,
'retweet_count']:
                    df.at[i,           'favorite_count']           =
status['favorite_count']
                    df.at[i,           'retweet_count']           =
status['retweet_count']
                continue


        # Tweepy preprocessing is called for basic preprocessing
        # It removes:  URLs, Hashtags, Mentions, Reserved words (RT,
FAV), Emojis, Smileys
```

```
            clean_text = p.clean(status['text'])

            # Call clean_tweet method for extra preprocessing

            filtered_tweet = clean_tweets(clean_text)

            # Pass textBlob method for sentiment calculations:

            blob = TextBlob(filtered_tweet)
            Sentiment = blob.sentiment

            # Seperate polarity and subjectivity in to two variables

            polarity = Sentiment.polarity
            subjectivity = Sentiment.subjectivity

            # New entry append

            new_entry   +=   [status['id'],   status['created_at'],
status['source'], status['text'],
                             filtered_tweet,    Sentiment,    polarity,
subjectivity, status['lang'],
                             status['favorite_count'],
status['retweet_count']]

            # To append original author of the tweet
            new_entry.append(status['user']['screen_name'])

            # To append if there is sensitive data:
            try:
                is_sensitive = status['possibly_sensitive']
            except KeyError:
                is_sensitive = None
            new_entry.append(is_sensitive)

            # Hashtagas and mentiones are saved using comma separted
            hashtags    =    ",    ".join([hashtag_item['text']    for
hashtag_item in status['entities']['hashtags']])
            new_entry.append(hashtags)
            mentions = ", ".join([mention['screen_name'] for mention
in status['entities']['user_mentions']])
            new_entry.append(mentions)

            # Get the location of the tweet if possible:
            try:
                location = status['user']['location']
            except TypeError:
                location = None
```

```
            new_entry.append(location)

            try:
                coordinates     =     [coord     for     loc     in
status['place']['bounding_box']['coordinates'] for coord in loc]
            except TypeError:
                coordinates = None
            new_entry.append(coordinates)

            single_tweet_df       =       pd.DataFrame([new_entry],
columns=COLS)
            df = df.append(single_tweet_df, ignore_index = True)
            csvFile = open(file, 'a', encoding = 'utf-8')
    df.to_csv(csvFile,  mode  =  'a',  columns  =  COLS,  index=False,
encoding="utf-8")
    csvFile.close()

# Declare keywords as a query for three categories

nintendodirect_keywords = '#NintendoDirect OR #NintendoDirect!'

# File location path
nintendodirect_tweets = "nintendodirect_data.csv"


# Call main method passing keywords and file path
write_tweets(nintendodirect_keywords,  nintendodirect_tweets)

# Import DF:

nintendo_direct_df = pd.read_csv(nintendodirect_tweets)

# Not interested in subjectivity 0 (drop those rows):

nintendo_direct_df                                                  =
nintendo_direct_df[nintendo_direct_df['subjectivity'] != 0]
nintendo_direct_df.reset_index(drop = True, inplace = True)

# Fill Nan in hastags clumn:

nintendo_direct_df['hashtags']                                      =
nintendo_direct_df['hashtags'].fillna('')

# Get tot. polarity and subjectivity:

direct_polarity_avg  =  round(nintendo_direct_df["polarity"].mean(),
3)
direct_subjectivity_avg                                             =
round(nintendo_direct_df["subjectivity"].mean(), 3)
```

```
print('Polarity:  {}\nSubjectivity:  {}'.format(direct_polarity_avg,
direct_subjectivity_avg))

# Get all hashtags:

all_hash = []
for i in range(len(nintendo_direct_df)):

    h = nintendo_direct_df["hashtags"][i] # Get hashtags of row i
    h_lst = h.split(", ") # get each tag as string
    h_lst = [h.lower() for h in h_lst]
    all_hash.extend(h_lst) # Add all tags to the list

# Find top 10 hashtags:

count_hash = Counter(all_hash)
del count_hash[''] # drop nan count
best_hash = [h for h, c in count_hash.most_common(10)]

# Find avg Polarity of top hashtags:

dic_avg_pol = {}
for h in best_hash:
    polarity_arr                                                    =
nintendo_direct_df[nintendo_direct_df["hashtags"].str.contains(h,
case = False)]["polarity"]
    mean = polarity_arr.mean(skipna = True)
    dic_avg_pol[h] = round(mean, 3)

dic_avg_pol

# Get all places:

all_place = []
for i in range(len(nintendo_direct_df)):

    p = nintendo_direct_df["place"][i] # Get places of row i
    all_place.append(p) # Add all tags to the list

# Find top 10 places:

count_place = Counter(all_place)
del count_place[''] # drop nan count
best_place = [p for p, c in count_place.most_common(10)]

best_place
```

## Summary

In the marketing research field, a distinction can be made between the quantitative and the qualitative method (Neumann, 2011). Quantitative research focuses on the collection of data from a large sample of respondents from a defined population and uses statistical, mathematical and computational techniques for data analysis in order to generalize the result to the entire population. Qualitative research instead focuses on the exploration of a phenomenon in a more unstructured way, with the objective of obtaining valuable insights about attitudes, beliefs and emotions of the narrow group of subjects on which the research is carried. (Newman, 1998)

The type of information that the researchers obtain using one or the other greatly differs: the output of quantitative analysis is data in a raw form, the rigid and objective research approaches that this method uses yield a formalized and mathematic solution to the problem, basing on a standardized sequence of instructions. On the contrary, the output of qualitative research are the subjective opinions that the researcher is able to create through a personal reasoning after interpreting the data gathered from the subjects.

The most significant difference between these two types of analysis relies in their measurement process: what links the data to the concepts.

The number of differences between qualitative and quantitative research have created, starting from the 1980s, an antipathy between the two methods, that created full-fledged “paradigm wars”.

Even with their differences, both of the methods of measurements intimately connect how we perceive and think about the social world with what we find in it (Neumann, 2011). Nevertheless either of the two received their share of criticism: quantitative research has been criticized as a rigid approach that ignores the inherent subjectivity of human social interactions (Holstein and Gubrium, 1995) while qualitative research is described as a subjective and non-scientific method that lacks structural coherence (Poggenpoel and Myburgh, 2005).

Despite the contradictory methodologies used, recent development in research methodologies (Kelle, 2006; Olsen, 2004; Srnka, 2007) suggest that a new approach can be conceived by integrating the two, in order to improve both the rigor and the connection to the data at hand.

The problem in the past literature has been a chauvinism on the researchers' own doctrines. This mindset has led to a lack of answers regarding arguments stressing methodological limitations of both qualitative and quantitative research. A researcher siding for a particular doctrine tended to answer a problem of his tradition by emphasizing problems of the other tradition. In this way problems that could have been solved using a dialectic approach were simply neglected. This has led to a situation where the potential of quantitative and qualitative methods to cope with problems of the competing method has not been utilized. (Kelle, 2006)

Sentiment analysis has been presented as a merge between qualitative and quantitative analysis, with the objective of using one to improve the other. Kelle, in his numerous studies (1998, 2002, 2003, 2006), and after considering the developments brought by the aforementioned "paradigm wars", showed in fact that one method could aid the other, while raising data quality, by supplementing quantitative data with qualitative insights.

The mix of the two methods that is in play in sentiment analysis can be considered a sequential quantitative-qualitative design (*quan* > *qual*). A quantitative research is carried in order to better narrow the problem areas and research questions, that are further investigated with the help of qualitative methods and data. Quantitative methods are in fact suited to give an overview about the matter under study and can describe its subjects on a macro level, whereas qualitative methods can be utilized to tap into the local knowledge in order to develop grounded hypotheses that cover relevant phenomena.

In this case, the qualitative method analyzes an amount of tweets that no human could and yields numerical results about the sentiment and the top keywords, afterwards a quantitative exploration and study of the results allows the researcher to fully grasp the outcome of the campaign analyzed.

This design helps to cope with two general problems of quantitative research: the difficulty to understand the quantitative data without the proper sociocultural knowledge and the doubt that the research is focusing only on remote or marginal cases.

With the advent of the Internet and the World Wide Web, the collection of data analyzable with traditional methods of research, both quantitative and qualitative, has changed. Access to user-generated content is allowed in an immediate and spontaneous

way: this results in the possibility to gather information about the opinion of a population and the expression of subjective and personal ideas of many subjects. (Miller, 2006)

The web and especially the social networking platforms offer to researchers large amounts of data. Even for qualitative research, traditional tools such as focus groups and questionnaires, when carried out online, allow to collect considerably larger volumes of data in a considerably shorter time. (Kaden, Linda & Prince, 2011)

However this evolution of research is not without its drawbacks: as Kotler (2010) explains: the people that have access to the internet and actually use it to express their opinions and thoughts about a relevant subject for research are not certainly representative on the entire population the researcher should refer to. A marketing research carried out online is not suitable for every kind of product: its representation is closely linked to variables such as the degree of computerization of the population, the type of consumers who use social media to communicate and the type of digital platform that is taken into consideration in the survey.

Another aspect to consider when approaching a web analysis is that, unlike offline survey, messages posted online (referred to as *User Generated Content – UGC*) are written spontaneously by users and received to the researcher in a dirty and unordered manner (referred to as *online chatter*). It is nonetheless important to highlight the value and at the same time the limitations of this type of data: on the one hand, their unstructured nature requires a greater effort than the typical offline survey that follows a standardized script; on the other hand, the spontaneity of the data received by the user and not addressed by the researcher allows to collect free and unguided opinions, possibly revealing links and information not initially considered. (Tirunillai & Tellis, 2014)

In this context, sentiment analysis is used as a synthesis between traditional research methods in an online search context, that is, the extraction of information from the Internet, with the awareness of the limits, but also of the advantages that a survey of an online-only population obviously presents.

When talking about sentiment analysis we enter the field of *Natural Language Processing* (*NLP*), the sector of Artificial Intelligence and computer sciences that deals with the relationship of computers with human language, or "natural language". Natural language understanding is, in fact, the main challenge of AI and deep learning techniques,

through which it is possible to teach machines to recognize words, understand texts and communicate with humans (Bates, 1995).

The core of any NLP work is the topic of natural language understanding, accomplished by humans through seven different levels (phonetic, morphological, lexical, syntactic, semantic, discursive and pragmatic), that has to be carried over to machines.

The literary definition of sentiment analysis used in this thesis was originally developed by Liu in 2010 and describes the operation of classifying a sentence as expressing a positive or negative opinion as *sentence level sentiment classification*, which can be further defined as following:

"Given an opinionated document *d* which comments on an object *o*, determine the orientation *oo* of the opinion expressed on *o*, i.e., discover the opinion orientation *oo* on feature *f* in the quintuple (*o, f, so, h, t*), where *f = o* and *h, t, o* are assumed to be known or irrelevant." (Liu, 2010)

Since the most functional techniques in existence for sentiment classification are based mostly on supervised learning, this approach was chosen for the thesis' algorithm.

Supervised learning is a type of machine learning, which can be defined as the set of techniques that allows a machine to learn and perfect itself in a certain skill. The *machine* is defined as a computer or software that uses the algorithms to whom the skill is taught. Machine learning is also called *automated learning* as algorithms become able to perform the task automatically and independently of the instructions of a human researcher, learning instead from the data itself. (Bishop, 2006)

The field of machine learning is characterized by two main types of tasks: the aforementioned supervised, and unsupervised. The key difference between the two types is that supervised learning is accomplished using a *ground truth*, i.e. having prior human-made knowledge of what the output values for the samples should be. Consequently, its goal is to learn a function that, given a sample of data and intended outputs, best approximates the relationship between input and output observable in the data. Unsupervised learning, in contrast, does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points. It then looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision. (Hinton, 1999)

In particular, a supervised predictive classification model is used to categorize words into values based on their sentiment and subjectivity.

Predictive classification modeling can be defined as the task of approximating a mapping function (*f*) from input variables (*x*) to discrete output variables (*y*) (Kotsiantis et al, 2007).

There is a wide variety of machine learning techniques that are commonly used in supervised classification tasks. Using a supervised leaning approach in sentiment analysis in fact requires a data corpus, which serves as a preparation document for classification learning. The classification, in turn, can be executed in different ways based on the theorems applied.

The basic functions available for classification include: *Naïve Bayes*, *Support Vector Machines*, *Decision Trees* and *Maximum Entropy*.

A *Naïve Bayes* classifier was used in this work, which is a probabilistic classifier based on applying Bayes' theorem that assumes that attributes are conditionally independent. In fact, its key difference from other classifiers is that Naïve Bayes assumes that the features are independent of each other and have no type of correlation between them. However, as easily imagined, this is not the case in real life. This naïve assumption of features being uncorrelated is thus the reason why this algorithm is named "naïve".

We can describe how the classifier operates starting by defining *P(x)* as the probability that an event *x* occurs: it is calculated as the number of the desired outcome divided by the total number of outcomes. Conditional probability, on the other hand, is the likelihood that an event *x* occurs given that another event *(y)* that has a relation with event *x* has already occurred. The probability of event *x* given that event *y* has occurred is denoted as *P(x|y)*. Finally, a joint probability is the probability of two events occurring together and is denoted as *P(x) × P(y)*.

Bayes' Theorem can be thus defined as:

$$P(x|\mathrm{y}) = \frac{\mathrm{P(x)} \times \mathrm{P(y)}}{\mathrm{P(y)}}$$

Or the probability of event *x*, given that event *y* occurs equals to the probability that *x* and *y* occur together divided by the probability of *y*.

This classifier is constructed based on the frequency of occurrence of each feature per class in the training data set. And under the assumption of features being independent, *P(x1, x2 , … , xn | yi)* it can be written as:

$$P(x1, x2, \ldots, xn|y_i) = P(x1|y_i) \times P(x2|y_i) \times \ldots \times P(xn|y_i)$$

The classification is conducted by deriving the maximum posterior which is the maximal *P(x1, x2,…xn|X),* applying the Bayes theorem assumption. This assumption greatly reduces the required calculations by only counting the class distribution and not their interdependence. Even though the assumption is not valid in most real cases since the attributes are dependent, Naïve Bayes is able to perform impressively in a number of different contexts (Lewis, 1998).

This classifier has some advantages and disadvantages over its substitutes: firstly, the assumption that all features are independent makes its algorithm very fast compared to others, therefore it is prone to works with high-dimensional data such as text classification or spam detection. On the downside, because of the aforementioned assumption it is less accurate than other algorithms (Rish, 2001).

The method employed to quantitatively calculate the sentiment used a Naïve Bayes classifier with a corpus created with dictionary-based approach. This approach has been adopted since it represents the most resource-efficient and fastest way to create a sentiment set. In fact, by starting from a sentiment dictionary that contains human-tagged words and expanding it using machine learning, it is possible to analyze through Natural Language Processing almost any kind of text. In the case of this research, tweets lend themselves particularly well to the analysis since they are notably short (no more than 280 characters) and tend to contain their author's full opinion.

The merits of the human encoding, used to create the original corpus, are that, based on the researcher's ability, each text can be encoded with a very low margin of error, regardless of the language used, the context of discussion, the use of metaphors or rhetorical figures.

In addition to the corpus-based approach, this thesis also employs a second way to better classify tweets. An emoticon dictionary, with each symbol and its sentiment value,

is used to assign their values to tweets, instead of deleting them, as a corpus-based approach normally would do. This method has been used by itself in Go, Bhayani and Huang's (2009) work and has proven successful, thus integrating it in a machine learning-led sentiment analysis can improve and enrich its results, by considering the entirety of the tweet instead of a part of it.

Other key differentiators from past research are the introduction of a subjectivity measurement and the analysis of the most used hashtags and most common locations. The first has made it possible to exclude texts that did not contain their authors' opinions in them from the sentiment count, giving an overall better result in the final sentiment analysis. The second gives marketers and researchers the possibility to not only analyze the sentiment behind a hashtag, but also understand what the main topics discussed from that community are and where those discussions are taking place.

Moving to the matter analyzed, hashtag campaigns are of particular interest in recent literature since they are the embodiment of a new type of marketing environment. The rise of these new communication channels has caused a disruptive change in the marketing environment. In fact, the information about a brand has turned into a multidirectional, interconnected flow and this makes it difficult to predict.

Marketers have lost the complete control they previously had over their brands and campaigns and now can only participate in a "conversation" about the brand.

In the era of new media, managing customer relationships can be compared to the game of pinball: companies launch a marketing "ball" that consists of their brands and brand-building messages into an unpredictive environment, which can be deviated or accelerated by new media and interactions that act as "bumpers", changing its course in chaotic ways. After the marketing ball is in play, marketing managers can continue to guide it with agile use of the "flippers", but the ball does not always go where they intended it to, and their slightest miscalculation can be amplified into a crisis. (Hennig-Thurau et al., 2010)

Twitter was used to gather the data to be analyzed since it is recognized by researchers as the ideal social medium for brands that seek to build relationships with their key stakeholders (Hennig-Thurau et al. 2010). The platform is mainly devoted to

information dissemination, that can be accomplished through word-of-mouth, spreading via many small cascades triggered by ordinary individuals (Bakshy et al., 2011). However, firms can also derive benefits from using Twitter to interact with their customers. In fact, many companies typically use it to communicate with a large number of followers in a *one-to-many* form. In addition to that, they can use the *one-to-one* mechanism to interact with single individuals by replying to them or retweeting their content. (Burton and Soboleva, 2011)

What makes Twitter so invaluable to companies is its nature as an online *listening tool* (Crawford, 2009), that allows its users to create online engagement with them even without having interactions and just following their content.

In particular, the hashtag function in Twitter was under the lens of this work since its use transcends the content aggregation that other platforms make it to be, to also encompass functions such as metacommunication, defined as the inclusion of a personal thought after a comment in the form of a hashtag, and creation of ad-hoc publics, defined as the natural formation of publics and communities around a specific topic, when they are needed and without restrictions.

A brand interested in laying out a marketing campaign, after establishing its goals and targets, can decide how to harness the potential of hashtags on social media.

In fact, a study by Dan Zarrella has shown[31] that a tweet that contains at least one hashtag was 55% more likely to be *ReTweeted* than tweets that did not.

Hashtag campaigns' use has risen in recent years among companies active on social media, in fact in 2015 70% of the most used hashtags on Twitter were brand related (Simply Measured, 2015). This happens because these campaigns have very low costs and, in return, allow brands to create awareness and an image for the brand by developing a free sponsorship that feeds itself with the posts of social users. Companies this way are able to adopt a *consumer-centric* approach, where they can build and consolidate interactive relationships with their audience with the final objective of creating engaging and loving consumers (Stathopoulou et al., 2017).

[31] Data retrieved from http://danzarrella.com/new-data-use-quotes-and-hashtags-to-get-more-retweets/

After a thorough explanation of the method used to write the Python code, a sentiment analysis has been carried on five different hashtags connected to five different campaign across four brands.

The first case, *#NintendoDirect*, has been an example of the creation of ad-hoc publics. In this case centered around the discussion of the latest news in videogames.

The second case, *#MyCalvins*, has shown the potentiality of letting the consumers express their own emotions and connect them to a brand in creating an enormous response from the public.

The third and fourth cases, *#TasteTheFeeling* and *#ShareaCoke*, were analyzed to show the differences in the usage and reception of branded hashtags from the same company (Coca-Cola in this case), which highlighted how positive emotions and sharing are more successful in improving the overall brand perception.

The fifth and final case, *#ElonMusk*, has shown how the technology behind the method still necessitates help from a human researcher to better discern between results, when the investigated topic is uncertain.

Analyzing different hashtag campaigns has thus shown the potential of the sentiment analysis to aid marketers in their work in different situations and with different numbers and end results.

With all the new kinds of marketing communications, always keeping the situation under control becomes of utmost importance: it can be easy for a campaign to derail or for it not to produce its intended results, but all it needs to regain momentum is a push in the right direction. This kind of action can only be performed by merging an in-depth quantitative analysis of the entire campaign made by a machine with a qualitative assessment made by the marketer.

The main limitation of this work lies into the type of permission used to analyze the tweets, that only gave access to a limited number of them at the time. Future studies can improve on this by designing a workaround to access Twitter's data without using its APIs, thus limiting however the type of data available to be accessed, or by purchasing an *Enterprise API* access from Twitter.

## Bibliography

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

- Andersen, M., Dauner, T., Lang, N., & Palme, T. (2017). What Automakers Can Learn from the Tesla Phenomenon. BCG Perspectives. URL: https://www.bcgperspectives.com/content/articles/automotive-what-automakers-can-learnfrom-tesla-phenomenon/Accessed, 17(11).

- Andreevskaia, A., & Bergler, S. (2006, April). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In 11th conference of the European chapter of the Association for Computational Linguistics.

- Andrews, T. (2012). What is social constructionism?. Grounded theory review, 11(1).

- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 65-74).

- Barton, A. H., & Lazarsfeld, P. F. (1955). Some functions of qualitative analysis in social research (No. 181). Bobbs Merrill.

- Bates, M. (1995). Models of natural language understanding. In Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, pp. 9977–9982

- Berlo D. (1960). The process of communication. New York: Holt, Reinhart and Winston.

- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

- Blumer, H. (1940). The problem of the concept in social psychology. American Journal of Sociology, 45(5), 707-719.

- Bramer, M. (2007). Avoiding overfitting of decision trees. Principles of data mining, 119-134.

- Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. Cognition, 118(1), 2-16.

- Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 804-812). Association for Computational Linguistics.

- Bruns, A., & Burgess, J. (2015). Twitter hashtags from ad hoc to calculated publics. Hashtag publics: The power and politics of discursive networks, 13-28.

- Bruns, A., & Moe, H. (2014). Structural layers of communication on Twitter. In Twitter and society (Vol. 89, pp. 15-28). Peter Lang.

- Bryman, A., & Bell, E. (2001). The nature of qualitative research. Social research methods, 365-399.

- Carroll, G. R., & Hannan, M. T. (2000). Why corporate demography matters: Policy implications of organizational diversity. California Management Review, 42(3), 148-163.

- Castronovo, C., & Huang, L. (2012). Social media in an alternative marketing communication model. Journal of marketing development and competitiveness, 6(1), 117-134.

- Ceron, A., Curini, L., and Iacus, S. M. (2014). Social media and sentiment analysis. L'evoluzione dei fenomeni sociali attraverso la rete. Springer, Milano

- Chomsky, N. (1965). 1965: Aspects of the theory of syntax. Cambridge, MA: MIT Press.

- Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

- Court, D. and Elzinga, D. (2009). The consumer decision journey. [online] McKinsey & Company.

- Crawford, K. (2009). Following you: Disciplines of listening in social media. Continuum, 23(4), 525-535.

- Cruz, L., Ochoa, J., Roche, M., & Poncelet, P. (2016, September). Dictionary-Based Sentiment Analysis applied to specific domain using a Web Mining approach. In 3rd. Annual Internacional Symposium on Information Management and Big Data (p. 80).

- Cui, H., Mittal, V., & Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. In AAAI (Vol. 6, No. 1265-1270, p. 30).

- Curtis, C. W. (1968) Linear Algebra, page 62, Allyn & Bacon, Boston

- Cvijikj, I. P., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. Social network analysis and mining, 3(4), 843-861.

- Daer, A. R., Hoffman, R., & Goodman, S. (2014, September). Rhetorical functions of hashtag forms across social media applications. In Proceedings of the 32nd ACM International Conference on The Design of Communication CD-ROM (pp. 1-3).

- Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

- Dijck, J. V. (2011). Tracing Twitter: The rise of a microblogging platform. International Journal of Media & Cultural Politics, 7(3), 333-348.

- Ding, X., & Liu, B. (2007, July). The utility of linguistic rules in opinion mining. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 811-812).

- Everitt B.S., Skrondal A. (2010), Cambridge Dictionary of Statistics

- Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

- Filstead, W. J. (1970). Qualitative methodology: Firsthand involvement with the social world. Markham Pub. Co..

- Føllesdal, D. (1979). Hermeneutics and the hypothetico-deductive method. Dialectica, 33(3-4), 319-336.

- Fortin, D., Uncles, M., Burton, S., & Soboleva, A. (2011). Interactive or reactive? Marketing with Twitter. Journal of Consumer Marketing.

- Fumo, N. (2015) Unpacking Calvin Klein's Wildly Successful #MyCalvins Campaign, Racked.com. retrieved from https://www.racked.com/2015/10/15/9534325/calvin-klein-mycalvins-justin-bieber-kendall-jenner

- Gage, N. L. (1989). The paradigm wars and their aftermath a "historical" sketch of research on teaching since 1989. Educational researcher, 18(7), 4-10.

- Gama, J., and de Carvalho, A.C. (2009). "Machine Learning". In M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Technology, Second Edition pp. 2462-2468. Information Science: Hershey, PA.

- Glaser, B., & Strauss, A. (1967). The discovery of grounded theory. 1967. Weidenfield & Nicolson, London, 1-19.

- Gleason, B. (2013). # Occupy Wall Street: Exploring informal learning about a social movement on Twitter. American Behavioral Scientist, 57(7), 966-982.

- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

- Granovetter, M.S. (1973). The Strength of Weak Ties. In American Journal of Sociology, Vol. 78, pp. 1360–1380

- Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. Business horizons, 54(3), 265-273.

- Hansson, L., Wrangmo, A., & Søilen, K. S. (2013). Optimal ways for companies to use Facebook as a marketing channel. Journal of Information, Communication and Ethics in Society.

- Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics (pp. 174-181). Association for Computational Linguistics.

- Heilpern, W. (2016). Coca-Cola just launched a massive new ad campaign to change the conversation around sugary drinks. Business Insider. Retrieved from https://www.businessinsider.com/coca-colas-taste-the-feeling-campaign-2016-1?IR=T

- Heimann, R., Danneman, N. (2014). Social Media Mining with R. Packt Publishing, Birmingham

- Heinz, W. R., Kelle, U., Witzel, A., & Zinn, J. (1998). Vocational training and career development in Germany: Results from a longitudinal study. International Journal of Behavioral Development, 22(1), 77-101.

- Hennig-Thurau, T., Malthouse, E. C., Friege, C., Gensler, S., Lobschat, L., Rangaswamy, A., & Skiera, B. (2010). The impact of new media on customer relationships. Journal of service research, 13(3), 311-330.

- Highfield, T. (2013). Following the yellow jersey: Tweeting the Tour de France. In Twitter and society (pp. 249-261). Peter Lang.

- Hinton, G. E., Sejnowski, T. J., & Poggio, T. A. (Eds.). (1999). Unsupervised learning: foundations of neural computation. MIT press.

- Holstein, J. A., & Gubrium, J. F. (1995). The active interview (Vol. 37). Sage.

- Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010, June). Conversational tagging in twitter. In Proceedings of the 21st ACM conference on Hypertext and hypermedia (pp. 173-178).

- Jansen, B. J., M. Zhang, K. Sobel, and A. Chowdury. 2009. "Twitter Power: Tweets as Electronic Word of Mouth." Journal of the American Society for Information Science and Technology 60 (11): 2169–2188

- Jurafsky, D., & Martin, J. H. (2009). Information extraction. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 725-743.

- Kaden, R. J., Linda, G., & Prince, M. (Eds.). (2011). Leading edge marketing research: 21st-century tools and practices. Sage.

- Kalekin-Fishman, D. (2001, September). David Silverman (2001). Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 2, No. 3).

- Kanayama, H., & Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 355-363).

- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

- Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & javad Rajabi, M. (2014, September). Advantage and drawback of support vector machine functionality. In 2014 international conference on computer, communications, and control technology (I4CT) (pp. 63-65). IEEE.

- Kelle, U. (2006). Combining qualitative and quantitative methods in research practice: purposes and advantages. Qualitative research in psychology, 3(4), 293-311.

- Kelle, U., & Lüdemann, C. (1998). Bridge assumptions in rational choice theory: methodological problems and possible solutions. Rational choice theory and large-scale data analysis. Westview Press, Boulder, 112-125.

- Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 54(3), 241-251.

- Kotler, P., Armstrong, G. (2010). Principles of Marketing. Prentice Hall, Boston

- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

- Kwok, L., and B. Yu. 2013. “Spreading Social Media Messages on Facebook an Analysis of Restaurant Business-to-Consumer Communications.” Cornell Hospitality Quarterly 54 (1): 84–94.

- Lewis, D. D. (1998, April). Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

- Liddy, E. D. (1998). Enhanced text retrieval using natural language processing. Bulletin of the American Society for Information Science and Technology, 24(4), 14-16.

- Liddy, E. D. (2001). Natural language processing.

- Lillis, A. M. (1999). A framework for the analysis of interview data from multiple field research sites. Accounting & Finance, 39(1), 79-105.

- Lin, Y. R., Margolin, D., Keegan, B., Baronchelli, A., & Lazer, D. (2013, June). # Bigbirds never die: Understanding social dynamics of emergent hashtags. In Seventh International AAAI Conference on Weblogs and Social Media.

- Liu B., (2006). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer

- Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627-666.

- Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

- Lundberg, G. A. (1941). Case-studies vs. statistical methods-an issue based on misunderstanding. Sociometry, 4(4), 379-383.

- Maldonado, J. L. (1993). On ambiguity, confusion and the ego ideal. International journal of psycho-analysis, 74, 93-100.

- Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. Business horizons, 52(4), 357-365.

- Marouli, G. (2014). Comparison between Maximum Entropy and Naïve Bayes classifiers: Case study; Appliance of Machine Learning Algorithms to an Odesk's Corporation Dataset.

- Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. [online] Forbes.com. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#b970bde60ba9

- Messina, C. (2007, August 25). Groups for Twitter; or a proposal for Twitter tag channels. FactoryCity [Blog]. Retrieved from http://factoryjoe.com/blog/2007/08/25/groups-for-twitteror-a-proposal-for-twitter-tag-channels

- Messina, C. (2007, October 22). Twitter hashtags for emergency coordination and disaster relief. FactoryCity [Blog]. Retrieved from https://factoryjoe.com/2007/10/22/twitter-hashtags-for-emergency-coordination-and-disaster-relief/

- Miller, J. (2006). Online marketing research. The handbook of marketing research, 110-131.

- Mitkov, R. (2014). Anaphora resolution. Routledge.

- Mollenkopf, H., Wahl, H., Baas, S., Rooij, R., Tacken, M., Marcellini, F., ... & Scharf, T. (2002, October). Social, structural, and psychological perspectives on outdoor mobility of older people: Findings from the European Project" MOBILATE". In Gerontologist (Vol. 42, pp. 248-249)

- Moretti, A., & Tuan, A. (2015). The social media manager as a reputation's gatekeeper: an analysis from the new institutional theory perspective. *ISSN 2045-810X*, *153*.

- Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002, July). Mining product reputations on the web. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 341-349).

- Neuman, Lawrence, W. (2011). Social Research Methods: Qualitative and Quantitative Approaches, 7th Ed. Persons: Boston

- Newman, I., Benz, C. R., & Ridenour, C. S. (1998). Qualitative-quantitative research methodology: Exploring the interactive continuum. SIU Press.

- Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering (Vol. 1, No. 1, pp. 61-67).

- Olsen, W. (2004). Triangulation in social research: qualitative and quantitative methods can really be mixed. Developments in sociology, 20, 103-118.

- Owyang, J., Jones, A., Tran, C., & Nguyen, A. (2011). Social business readiness: how advanced companies prepare internally. San Mateo.

- Page, R. (2012). The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. Discourse & communication, 6(2), 181-201.

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

- Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-

- Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

- Perez, S. (2018). TechCrunch. [online] Techcrunch.com. Available at: https://techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/

- Poggenpoel, M., & Myburgh, C. P. H. (2005). Obstacles in qualitative research: Possible solutions. Education, 126(2), 304-312.

- Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. International journal of nursing studies, 47(11), 1451-1458.

- Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. Qualitative research in accounting & management.

- Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

- Rambocas, M., & Gama, J. (2013). Marketing research: The role of sentiment analysis (No. 489). Universidade do Porto, Faculdade de Economia do Porto.

- Richter, F. (March 20, 2014). Brands Shouldn't Overuse Hashtags. Retrieved from https://www.statista.com/chart/2032/hashtags-affect-interaction/

- Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

- Rittman, R., Wacholder, N., Kantor, P., Ng, K. B., Strzalkowski, T., & Sun, Y. (2004). Adjectives as indicators of subjectivity in documents. Proceedings of the American Society for Information Science and Technology, 41(1), 349-359.

- Rogers, R. (2013, May). Debanalizing Twitter: The transformation of an object of study. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 356-365).

- Sayer, R. A. (1992). Method in social science: A realist approach. Psychology Press.

- Sayer, R. A. (2000). 2000: Realism and social science. London: Sage.

- Seale, C. (1999). Quality in qualitative research. Qualitative inquiry, 5(4), 465-478.

- Shin, J., Chae, H., & Ko, E. (2018). The power of e-WOM using the hashtag: Focusing on SNS advertising of SPA brands. International Journal of Advertising, 37(1), 71-85.

- Simply Measured. (2015). 2015 Instagram Industry Report. Retrieved from https://www.slideshare.net/SRG726/simply-measureds-instagram-influencer-report

- Srnka, K. J., & Koeszegi, S. T. (2007). From words to numbers: how to transform qualitative data into meaningful quantitative results. Schmalenbach Business Review, 59(1), 29-57.

- Stathopoulou, A., Borel, L., Christodoulides, G., & West, D. (2017). Consumer branded# hashtag engagement: can creativity in TV advertising influence hashtag engagement?. Psychology & Marketing, 34(4), 448-462.

- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science, 29(1), 41-78.

- Taecharungroj, V. (2017). Starbucks' marketing communications strategy on Twitter. Journal of Marketing Communications, 23(6), 552-571.

- Thomases, H. (2010). Twitter marketing: An hour a day. John Wiley & Sons.

- Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. Journal of Marketing Research, 51(4), 463-479.

- Tsang, E. W. (2014). Case studies and generalization in information systems research: A critical realist perspective. The Journal of Strategic Information Systems, 23(2), 174-186.

- Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

- Twitter (2020) . data retrieved from https://developer.twitter.com/en/apply-for-access

Volume 10 (pp. 79-86). Association for Computational Linguistics.

- Walby, S. (2001). Against epistemological chasms: The science question in feminism revisited. *Signs: Journal of Women in Culture and Society*, *26*(2), 485-509.

- Wang, H., Liu, L., Song, W., & Lu, J. (2014). Feature-based sentiment analysis approach for product reviews. Journal of Software, 9(2), 274-279.

- Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1031-1040).

- Warner, B. R., McGowen, S. T., & Hawthorne, J. (2012). Limbaugh's social media nightmare: Facebook and Twitter as spaces for political action. Journal of Radio & Audio Media, 19(2), 257-275.

- Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). Twitter and society (Vol. 89, p. 447). P. Lang.

- Wiebe, J., Bruce, R., & O'Hara, T. P. (1999, June). Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics (pp. 246-253).

- Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012, April). We know what@ you# tag: does the dual role affect hashtag adoption?. In Proceedings of the 21st international conference on World Wide Web (pp. 261-270).

- Yu, H., & Hatzivassiloglou, V. (2003, July). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

- Zhang, M., & Ye, X. (2008, July). A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 411-418).

- Zhang, M., Jansen, B. J., & Chowdhury, A. (2011). Business engagement on Twitter: a path analysis. Electronic Markets, 21(3), 161.

- Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1(1-4), 43-52.

- Znaniecki, F. (1934). The method of sociology. Octagon Books.