Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Miguel Vilela Carvalho

A Data Analysis approach to study events' influence in Social Networks

November 2018

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Miguel Vilela Carvalho

A Data Analysis approach to study events' influence in Social Networks

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Paulo Jorge Freitas de Oliveira Novais

November 2018

ACKNOWLEDGEMENTS

This thesis is the end result of almost one year of work, for which many people contributed. Some had a direct impact, while many others influenced me in a more indirect way.

First and foremost, I would like to thank my supervisor Professor Paulo Novais for motivating me into this work and accompanying me till its end. In the beginning several were available and, through a healthy discussion, we figured what the most interesting work was possible. When developing a thesis, it is of the uttermost importance being interested in it.

Also, a major thanks to Marco Gomes for the continuous support he provided me during the work on this research project. I thank him for the understanding showed for the various problems that I have encountered at certain times and for his help in solving them.

Both meant a great deal and showed a great availability to discuss many of the doubts that I had throughout this thesis. For that, they have my utmost respect and gratitude.

A special thanks to my entire family, specially to my parents Armindo and Paula, and my brother José, who rode with me in this journey that now culminates with this thesis. They were right there, with me, the whole time.

Thank you for a lot of good advice and guidance, you are the reason I'm here today.

ABSTRACT

Nowadays, the assimilation of web content, by each individual, has a considerable impact on our' everyday life.

With the undeniable success of online social networks and microblogs, such as Facebook, Instagram and Twitter, the phenomenon of influence exerted by users of such platforms on other users, and how it propagates in the network, has been attracting, for some years computer scientists, information technicians, and marketing specialists.

Increased connectivity, multi-model access and the rise of social media shortened the distance between almost every person in the world, more and more content is generated. Extracting and analyzing a significant amount of data is not a trivial task, Big Data techniques are essential.

Through the analysis of this interaction, an exchange of information and feelings, it is entirely imaginable its usefulness in understanding complex human behaviours and so, help diverse organization's decision-making. Influence maximization and viral marketing are among the possibilities.

This work is intended to study what is the impact and role that an event's social influence has and how does it propagate, particularly on its surrounding territory. This influence is inferred by analysis of the online platform's data, by applying intelligent techniques, right after its extraction. The final step is to validate the results with data from different sources. Helping businesses through actionable and valuable knowledge is the ultimate goal.

This document contemplates an introductory section where the study subject and its State of the Art are addressed. Next, the problem and what direction to take to solve it are discussed.

Keywords: Social networks, Intelligent Techniques, Social Influence.

RESUMO

Atualmente, a assimilação de conteúdo Web, por cada indivíduo, tem um impacto considerável no nosso quotidiano.

Com o inegável sucesso de redes sociais e microblogs, como por exemplo Facebook, Instagram e Twitter, o fenômeno de influência exercida, por utilizadores de tais plataformas, em outros utilizadores e como se propaga na rede tem atraído, por alguns anos, informáticos, tecnicos de informação e especialistas em marketing.

O aumento da conectividade, o acesso multi-modal e a proliferação dos meios de comunicação social reduziram a distância entre quase todas as pessoas do mundo, mais e mais conteúdo é gerado. Extrair e analisar uma grande quantidade de dados não é uma tarefa trivial, são essenciais técnicas de Big Data.

Através da análise desta interação, troca de informações e emoções, é perfeitamente imaginável a sua utilidade na compreensão de complexos comportamentos humanos e, portanto, ajudar na tomada de decisão de diversas organizações. A maximização da influência e o marketing viral estão entre as possibilidades.

Este trabalho destina-se a estudar qual é o impacto e o papel que a influência social de um evento tem e como se propaga, particularmente no território envolvente. Esta influência é inferida pela análise dos dados de plataformas online, aplicando técnicas inteligentes, logo após a sua extração . O passo final é validar os resultados com dados de diferentes fontes. Ajudar empresas através do conhecimento valioso e atuável é o objetivo final.

Este documento contempla uma seção introdutória, onde o assunto de estudo e o seu estado da arte são abordados. De seguida, é discutido o problema e a direção a seguir para o solucionar.

Palavras-chave: Redes sociais, Técnicas Inteligentes, Influência social.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

In an increasingly complex and connected world, the demand by business organizations for processes that help them make accurate and rapid decision-making, based on real data, has never been higher. At the same time, the growth of data available for study does not slow down, and we live in the Big Data era.

Currently, daily access to the web is facilitated, accompanied by the evolution of devices, and is already a formality. There is a deep and frequent interaction between individuals, to the point that there is no filter on what it is decided to be shared, resulting in a digital footprint that people leave in their day-to-day communication, actively or passively. This is the perfect stage for the study of social behaviours.

Through the knowledge extraction of web content, both social media, where there is a network of users sharing all kinds of information, as web activity, traces left by people when using search engines or when visiting websites, many purposes are possible. This dissertation's focus is the study of social influence on the Web, applied to an event (intrinsic to a region).

Earlier works related to the diffusion of information or influence state that it is possible to trace and measure the evolution of this interference, on how to act or think, with success. Comparing to these, three of the points that are gone after, seeking improvement, are the use of few and similar data sources, the appropriateness of simple metrics and the closed world concept. The idea is to enrich the collected datasets with external resources or results, mixed-sources, so that the implemented metrics are precise, representative of the real world. The result´s validation by analyzing patterns of influence around the event location is also of concern.

These referred sources represent a relevant and reliable source of information, and there are objects of analysis at different spatial and temporal points, but also some difficulties, such as computational problems, heterogeneity, and data privacy. Of the set of several utilities, we must highlight the discovery of new trends and patterns, using explanatory models, characterization of groups that are of interest and the know-how and to whom redirect investment through ads, for example. The prediction of inflows and other phenomena for subsequent years and the study of the role that influence holds in specific highlight events or commemorative dates are also relevant points of interest.

In the absence of any data, it is be necessary to elaborate information capture mechanisms, to pre-process and clean (if the data does not fit into the analysis universe), followed by a transformation process, so that later on, using tools and techniques of Data Science and learning systems, Machine Learning, useful and actionable knowledge can be achieved.

1.2 OBJECTIVES AND HYPOTHESIS

The primary purpose of this dissertation is to generate useful knowledge, associated with a physical event, through the analysis of information extracted from web content. This analysis is performed after the extraction of data from multiple online platforms, of varied typologies and content.

The objectives of this dissertation are therefore:

1. Implementation of all necessary infrastructure for information extraction;
2. Validation of information capture methods and models;
3. Specification of a computational system for the analysis of the objects of study of this work;
4. Validation of the computational system's results.

The ultimate goal is to answer the question - **Is it feasible, through a social networks' data extraction and analysis methodology, to quantify and validate an event's influence, on a specific region?**

From the refereed goal, several questions can be formulated: How does the diffusion of an event's influence evolve? In this diffusion, is there a correlation between different social media platforms? Is it possible to correlate estimated influence with other sources' patterns of data surrounding the event? Or even with real-world phenomenon, such as an impact on the regional economy?

To answer this questions, specific methods and steps should be taken in order to investigate and retrieve the intended knowledge.

1.3 ORGANIZATION OF THE THESIS

After presenting and contextualizing the problem in question and the objectives to resolve it (chapter one), the definitions of this work's related terms are given. Descriptions, which are essential to fully understand the next ones, along with some literature review and associated works that certainly helped to find what the best direction is to achieve the desired results. In these studies are specified what problems other authors excelled.

In chapter three, is suggested an approach to the problem, how to tackle it. This includes giving an explanation on the chosen methodology and the issues that may arise trying to reach the solution.

Chapter four and five present how the data is collected, processed and what insights can be inferred with which techniques.

Chapter six demonstrates the solution previously studied through a case study, where the procedure followed, and the analysis's results are detailed and discussed.

Finally, in chapter seven, conclusions are formed and discussed, what is already done, some limitations and possible directions that can be followed in the future.

Some references that helped to substantiate the problem are given on the last pages of the document.

2

LITERATURE REVIEW

2.1 INTRODUCTION

Over the last decade, the online world has been expanding exponentially, generating more and more content. As Moniz (2017) states, such development may, mainly, be justified by:

1. **Increased connectivity** - The age of information in which we currently live has been defined by the appearance, and exponential growth of Internet users
2. **Multi-modal access** - The number of Internet users has risen at an accelerated pace, accompanied by the evolution of devices. Today, these provide near full-time access to the Internet. The main contributor to this new reality is the rise of mobile technology, both concerning the devices themselves and mobile data connectivity
3. **The rise of Social Media** - Social media platforms have greatly influenced the Internet due to their ability to not only connect users but mainly for allowing users to generate content of various types. These platforms continue to expand at a great pace, increasing their number of users, and the amount of content generated by them. Additionally, with the growing connectivity between users, these platforms form powerful mechanisms that are highly important for both information dissemination and information search.

The combination of these three factors caused a severe shortening of the distance between every other person in the world. This phenomenon plays an essential role in shaping users' behaviors.

In this chapter, the work described is framed, by presenting a thorough literature review on social media, web content, social influence, and information propagation.

Firstly it is clarified the definition of frequent and relevant terms in this work, and secondly exposed the current state of the art, where tasks have been extensively studied and analyzed, revealing distinct approaches. A discussion on their strengths and shortcomings is also introduced.

2.2 BACKGROUND

In a field of study, such as computer science, that never stops evolving, a flood of new terms and definitions arise with new technologies or techniques, and consequentially causing conflict with existing ones. To resolve this doubt sometimes is better to relax some denominations and explain the processes and purposes.

2.2.1 *Big Data*

The definition of **Big Data** has changed and evolved several times. The easiest way to understand what it means is by explaining the famous reference to the three core V's: volume, velocity and variety, along with the ones who entered the debate in recent years: veracity, validity, volatility and value (Demunter, 2017).

Briefly, **volume** refers to the exploding quantity of data concerning observations — in orders of magnitude of gigabytes, terabytes, petabytes (and soon exabytes or zettabytes?) — and to the variables observed; **velocity** refers to how quickly this data is generated and their resolution in time.

Variety refers to the many different types of data, such as natural language textual data (e.g., social media posts), photos (posted on, say, Instagram or Facebook), website logs, videos (e.g., camera surveillance), recordings or geo-coded data.

Veracity and **validity** touch upon the quality, reliability and usefulness of big data. The sheer volume of data and observations does not guarantee quality. On the contrary, the unwanted bias and noise in most big data sources are without a doubt some of the more complicated challenges for statisticians.

Volatility refers to how long the data remains relevant and how long it should be kept, bearing in mind the billions of impulses registered every second or the legal framework on retaining personal data

The **value** of big data is twofold: firstly, for statisticians as a potentially richer or timelier data source; and secondly, for businesses and policy-makers in an era of data-driven decisions. Those businesses or organizations holding the data are a particular case. Data is a valuable, marketable asset and can make these stakeholders reluctant to grant access to the data they contain.

The same referred author organizes big data sources into a taxonomy. The diagram below outlines the most common ones.

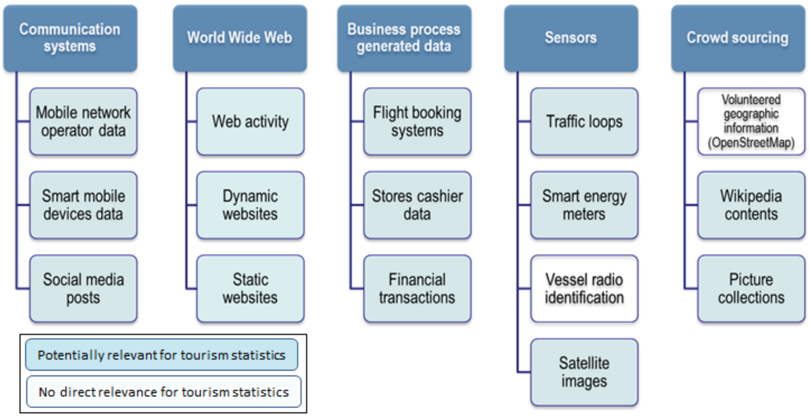

Figure 1: Taxonomy for big data sources from Demunter (2017).

Some of these sources are of particular interest in this work. The next paragraphs describe them.

Whether or not they intend to do so, people leave their digital footprint when using social media. **Social media posts** can be an information source on people's movements and behaviour. Gomes et al. (2014) uses knowledge about social interactions as a basis for informed decision support in situations of conflict.

Traces left by people while using search engines (e.g., Google Trends data) and visiting websites (e.g., Wikipedia page views), **web activity** can give an indication of which topics interest people at each moment in time.

A derived source such as page views of Wikipedia articles can be a proxy for visits to a destination, measured through the traveller's web activity.

Traveler's web activity, usually, **dynamic websites** feature structured data and an interface to access and consult a (dynamic) database. Typical examples include tripadvisor.com, booking.com or airbnb.com. In general, the data is obtained via web scraping, where pieces of relevant data are extracted from the web pages returned by dynamic website.

Static websites are composed of a limited amount of web pages which do not change frequently. Data is obtained by extracting the contents from the HTML source code and transforming them, clustering them into meaningful information for further analysis.

Taking tourist accommodation as an example, websites can give information on the activity status of establishments and their location, on the number of rooms and bed places available, and on standard prices.

Wikipedia contents are also relevant. Both web activity and its underlying web contents can be relevant.

For a decade now, people have been sharing pictures online, **picture collections** rather than in printed photo albums. The smart devices used to take the images typically log the location and time stamp.

Solutions to big data problems will not come from incremental improvements to business. Instead, they require rethinking how data analysis is managed Agrawal et al. (2016). From data, the ambition is to get insights.

Therefore Big data analysis involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Fortunately, existing computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problems.

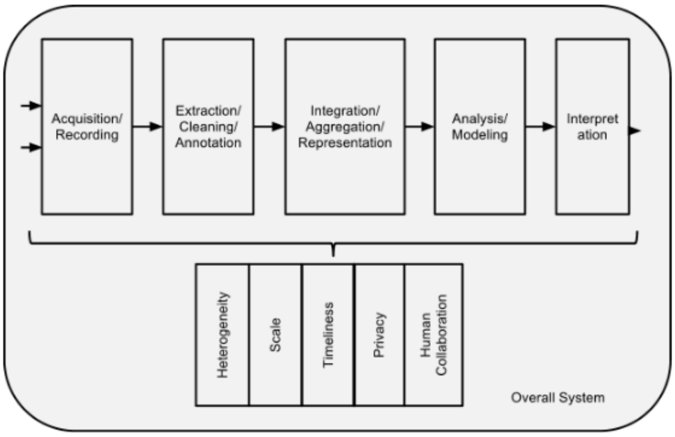

Figure 2: Big Data pipeline and its main challenges Agrawal et al. (2016).

Data Acquisition and Recording is the first step, big data does not arise out of a vacuum: it is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

Frequently, the information collected will not be in a format ready for analysis, **Information Extraction and Cleaning** is essential. We cannot leave the data in this form and still effectively analyze it. Instead we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis.

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository, **Data Integration, Aggregation, and Representation** is needed. Data analysis is considerably more challenging than only locating, identifying, understanding, and citing data. For useful large-scale analysis all of this has to happen in a completely automated manner.

Methods for querying and mining Big Data, **Query Processing, Data Modeling, and Analysis**, are fundamentally different from the traditional statistical analysis of small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query, and mining interfaces, scalable mining algorithms, and big-data computing environments.

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. **Interpretation** cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. In short, it is rarely enough to provide just the results.

The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result in interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone.

2.2.2 *Data Extraction*

As it is normal in such an evolving field, there are terms constantly emerging with similar definitions. In the specific case of data extraction, **web scraping** and **web crawling** are the most dominant ones.

Web crawling is used to index the information on the page using different kinds of bots. Basically it is used by major search engines, such as Google. In other terms they are the major web crawlers. Here we get generic information, where as scraping we get specific information.

In simple terms, Web scraping is the process of automatically requesting a web document and collecting information from it.

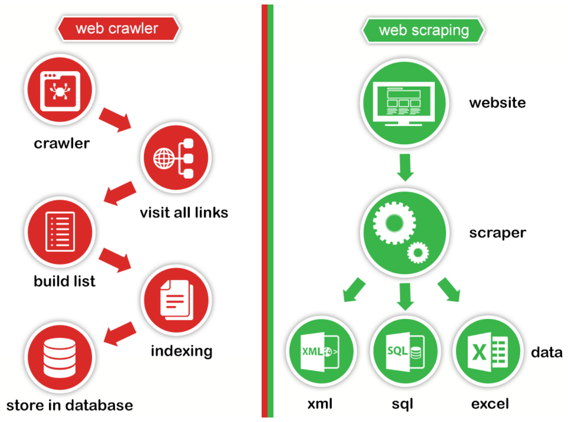

Figure 3: Distinction between web scraping and crawling Anonymous (c).

Information is one of the most valuable item in the whole world. This means that those who have the information take all possible precautions to protect them against copying, which in their eys is like stealing. Web scrapers are simply the algorithms that can automate the information extraction from publicly available sources.

It is easy to reach an abusive state to the websites' servers. For example, web scrapers can send far more requests per second than a human would do. Having this in mind when scraping, not one legal issue may rise.

Some rules in order to respect who is on the other side of the problem, are common sense in this field. It is always a good practise having a little common sense and respect the entity from which the data is extracted. Having regard to robots.txt protocol and also to a minimal submission of forms and javascript execution is a good start.

2.2.3 *Intelligent Techniques*

As data is growing at a faster pace, new terms associated with processing and handling data are coming up. Data science, data analysis, data analytics, data mining and machine learning, for example, are considered intelligent techniques.

Commercial organizations have realized that there is enormous value hiding in the data and are employing the techniques related to these to achieve that value. Ultimately what they all intend to produce is insights, things that may not have been known otherwise, such as Carneiro et al. (2015b), in which a stress estimation is performed.

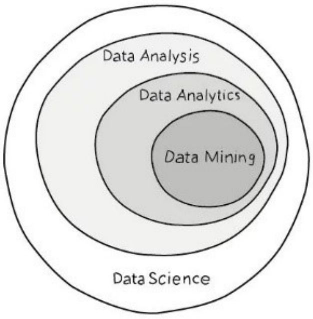

Figure 4: Data science terms from Anonymous (b).

Insight on these terms and how they differentiate is presented by Heiler (2017).

Data scientists are responsible for coming up with data-centric products and applications that handle data in a way which conventional systems cannot. The process of **data science** is much more focused on the technical abilities to manage any type of data. Unlike others, it is responsible for assessing the impact of data in a specific product or organization.

While data science focuses on the science of data, **data mining** is concerned with the process. It deals with the process of discovering newer patterns in big data sets. It might be apparently similar to **machine learning** because it categorizes algorithms. However, unlike machine learning, algorithms are just a part of data mining. In machine learning algorithms are used for gaining knowledge from data sets. However, in data mining algorithms are only combined as the part of a process. Unlike machine learning, it does not completely focus on algorithms.

Data science is the denomination that includes all the other ones, comprising everything that has to do with preparation, cleansing, and analysis, dealing with both structured and unstructured data. There are also two more specializations between the broadest and most confined terms someone can have, **data analysis** and **data analytics**.The analysis is a heuristic activity, where scanning through all the data the analyst gains some insight. Analytics is about applying a mechanical or algorithmic process to derive the insights, for example, running through various data sets looking for meaningful correlations between them.

One of the difficulty lies on how to get all the information that matters from data. Each technique has one purpose: discovering knowledge, distilling knowledge in form of rules or discovering optimal solutions.

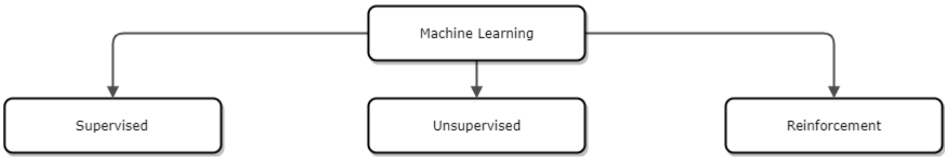

Figure 5: Types of machine learning algorithms.

The three types of machine learning algorithms are now elucidated Jain (2015).

Supervised Learning / Predictive model, as the name suggests, is used to predict the future outcome based on the historical data. Predictive models are normally given clear instructions right from the beginning as in what and how it needs to be learned.

For example: Supervised Learning is used when a marketing company is trying to find out which customers are likely to churn. We can also use it to predict the likelihood of occurrence of perils like earthquakes, tornadoes, etc. with an aim to determine the Total Insurance Value.

Some examples of algorithms used are: Nearest neighbour, Naive Bayes, Decision Trees, Regression, etc.

Unsupervised learning / Descriptive model is used to train models where no target is set and no single feature is important than the other.

The case of unsupervised learning can be: When a retailer wishes to find out what is the combination of products, customers tends to buy more frequently. Furthermore, in the pharmaceutical industry, unsupervised learning may be used to predict which diseases are likely to occur along with diabetes.

Example of algorithm used here is: K- means Clustering Algorithm.

Reinforcement learning is an example of machine learning where the machine is trained to take specific decisions based on the business requirement with the sole motto to maximize efficiency (performance). The idea involved in reinforcement learning is: The machine/ software agent trains itself on a constant basis based on the environment it is exposed to, and applies it's enriched knowledge to solve business problems. This continual learning process ensures less involvement of human expertise which in turn saves a lot of time!

An example of algorithm used is the Markov Decision Process.

There is a subtle difference between Supervised Learning and Reinforcement Learning. The last one essentially involves learning by interacting with an environment. Learns from its experience, preferably from its continual trial and error learning process whereas in supervised learning an external supervisor provides examples.

2.2.4 *Social Media and Web Content*

A Portuguese study, realized between 19 of July and 7 of August (2017), seemed to know the habits of users of social networks, the sites they know and use the most, the features they value most, the personalities and brands they follow, how often they access the sites and with which they publish information on those sites, as well as the time they devote to them or the equipment they use to access them. 818 interviews were done, a sample stratified by Marktest Consulting (2017), being proportional to the population that constitutes the universe, according to the variables gender and age.

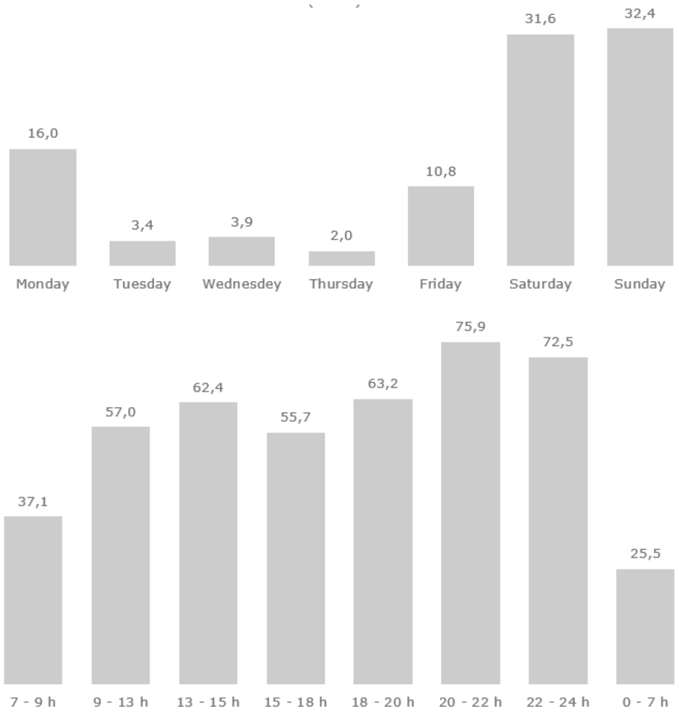

Figure 6: Periods of access to social networking sites, by Marktest Consulting (2017)

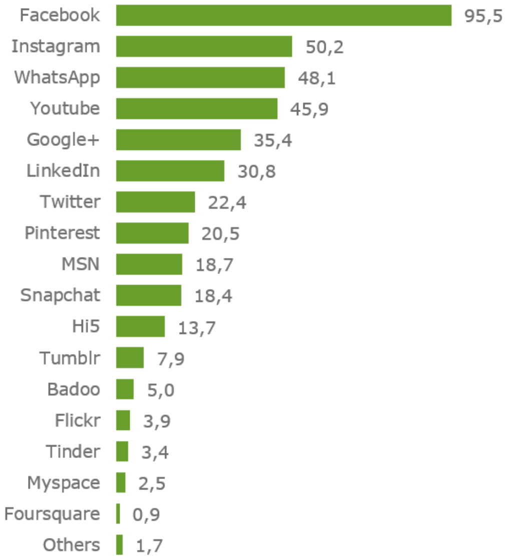

Figure 7: Social network websites the Portuguese have an account, from Marktest Consulting (2017)

From the figures above is quite perceptible that, along with Facebook, Instagram, WhatsApp and YouTube are the social network websites with more accounts and consequentially more impact in Portugal. This study also states that Monday, Saturday and Sunday, between 20 and 24 o'clock, are the periods with the highest peak of activity.

These stats are of great interest on understanding which online platforms are more relevant and so, the ones that should be selected to study.

One mission of this work is to extract and mix several sources, content from different types of online platforms, resulting in a expanded spectrum and more richness to the data.

From professional or image online platforms to options for bloggers and social video networks, it is easy to find something that matches everyone needs. There are all kinds of online platforms, it is possible to have it all, and can be divided by the user's intention on using or the network content. **Social network websites**, such as Facebook and Google+, are social media channels that don't tend to focus on any one specific topic and instead aim to attract a broad spectrum of users. Users communicate, click, and share different content types that reflect their interests, making the themes on such social networks just as diverse as the users themselves.

Image networks websites, like Pinterest and Instagram, are image sharing platforms, visual aspects are put at the forefront: the focus here lies on publishing photos and videos; comments play a smaller role. What counts the most is entertaining posts that leave a strong and lasting visual impression on target groups.

The origins of **Blogging networks websites**, Twitter and Tumblr, can be traced back to the blogging scenes. Users of both networks share different types of content (news, links, images, videos). Those who chose to follow a particular timeline are then presented with news regarding this profile in their timeline. Users who aren't registered with the account also can view all the content that is posted.

Professional networks websites, like LinkedIn, have created themselves a niche market by catering explicitly to a more professional environment. This platform model facilitates an exchange between business partners, employees as well as applicants and companies to occur.

YouTube is by far the most widely used channel of its kind, **video networks websites**, and innovative ideas are awarded on this channel, allowing some individuals to turn their hobby into a career.

At this point there might be some confusion related to social networks and online social networking site. The former is a mobile or web-based software platform that allows users to interact with their social connections.

While a network, like a group, is a collection of people, a **social network** includes something more: a specific set of connections between people in the group. These ties and the particular pattern of these, are often more important than the individual people themselves Christakis and Fowler (2009).

Other authors, such as Chen et al. (2013), classify a social network as a possibly directed graph. It may be homogeneous, where all nodes are of the same type, or heterogeneous, in which case the nodes fall into more than one kind.

Examples of homogeneous networks include the underlying graphs representing friendships in basically all of the social networking platforms (e.g., the list of "friends" in Facebook), as well as the graphs representing co-authorship or co-worker relationships in collaboration networks.

Examples of heterogeneous networks include rating networks consisting of people and objects such as songs, movies, books, etc. Here, people may be connected to one another via friendship or acquaintance, whereas objects (songs, movies, etc.) may be linked with one another using of similarity of their metadata.

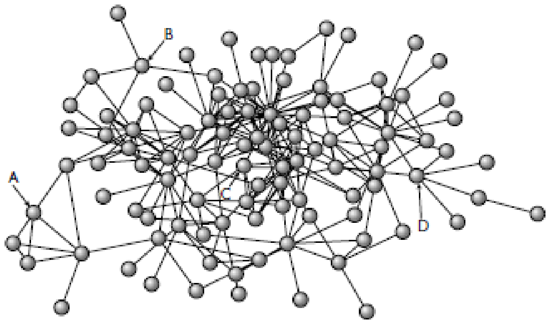

Figure 8: Natural network of close friendships among 105 college students, each circle represents a student, and each line a mutual friendship Christakis and Fowler (2009).

A social network is formally represented by a graph, where nodes are users and edges are relationships that can be either directed or not depending on how relationships are managed by the system. More precisely, it depends on whether it allows connecting in a unilateral (e.g., Twitter social model of following) or bilateral (e.g., Facebook social model of friendship) manner.

An influence graph is often represented as a weighted, directed graph with edge weights indicating how much influence a particular source node has on its destination.

2.2.5 *Social influence analysis*

Milgram's experiment, conducted in the 1960s, involved giving a few hundred people who lived in Nebraska a letter addressed to a businessman in Boston, more than a thousand miles away. They were asked to send the letter to somebody they knew personally. The goal was to get it to someone they thought would be more likely than they to have a personal relationship with the Boston businessman. And the number of hops from person to person that the letter took to reach the target was tracked. On average, six hops were required (a friend is one degree away, a friend's friend is two degrees, and so on). This amazing fact initiated a whole set of investigations into the small-world effect.

However, just because we are connected to everyone else by **six degrees of separation** does not mean that we hold sway over all of these people at any social distance away from us. Christakis and Fowler (2009) research has shown that the spread of influence in social networks obeys the **Three Degrees of Influence Rule**. Everything we do or say tends to ripple through our network.

The term social influence analysis or just influence analysis has the same meaning as the analysis of the diffusion of information or influence through a social media platform. In this area there are several terms subject to confusion and the next paragraphs try to enlighten them.

But firstly, when and how can we say that there is influence between users? There are at least two different phenomena surrounding users' behavior that are different from influence, but may appear to be as such. Causes of **correlation** in social networks can be categorized into roughly three types Anagnostopoulos et al. (2008).

The first is **influence** (also known as induction), where the action of a user is triggered by one of his/her friend's recent actions. An example of this scenario is when a user buys a product because one of his/her friends has recently bought the same product.

The second is **homophily**, which means that individuals often befriend others who are similar to them and hence perform similar actions. For example, two individuals who own Xboxes are more likely to become friends due to the common interest.

The third is **environment** (also known as confounding factors or external influence), where external factors are correlated both with the event that two individuals become friends and also with their actions. For example, two friends are likely to live in the same city, and therefore to post pictures of the same landmarks in an online photo sharing system.

Developing accurate propagation models, predictive or explanatory, is crucial in effectively taking business actions in the social networking space, marketing.

Although there has been some interesting work in this direction, this is by far the area of which we know the least: it is mostly unclear why specific information propagates while other information does not, measuring influence remains a difficult task (in large part because all social network data is partial), and successful application of models depends on a number of external factors that are difficult to quantify.

Having influence is not just about the numbers of followers or friends, it is about the reaction. It is possible to follow the traces to see what happens to that content and where it ends up.

Eventually, the ingredients of an information diffusion process taking place in an social network can be summarized as follows Bonchi et al. (2011): (i) a piece of information carried by messages, (ii) spreads along the edges of the network according to particular mechanics, (iii) depending on specific properties of the edges and nodes.

The stream, produced by the members of the network, can be viewed as a sequence of decisions (i.e., whether to adopt a certain topic or not), with later people watching the actions of earlier people Guille et al. (2013). Therefore, individuals are influenced by the actions taken by others. This effect is known as social influence, and is defined as follows:

Social Influence, a social phenomenon that individuals can undergo or exert, also called imitation, translating the fact that actions of a user can induce his connections to behave similarly. Influence appears explicitly when someone “retweets” someone else for example.

Based on the social influence effect, information can spread across the network through the principles of herd behavior and informational cascade which Guille et al. (2013) also defines.

Herd behavior, a social behavior occurring when a sequence of individuals make an identical action, not necessarily ignoring their private information signals.

Information Cascade, a behavior of information adoption by people in a social network resulting from the fact that people ignore their information signals and make decisions from inferences based on earlier people’s actions.

Bonchi et al. (2011) as a similar idea of influence in social networks: when users see their social contacts performing an action, they may decide to perform the action themselves (e.g., people buy items their friends buy). Influence for performing an action, may come

(i) from outside the social network, (ii) because the action is popular, or (iii) by the social contacts in the network.

Some authors also refer this the diffusion as **contagion** and for Christakis and Fowler (2009) there are two fundamental aspects of social networks. First, there is a connection, which has to do with who is connected to whom. Second, there is contagion, which pertains to what, if anything, flows across the ties. It could be buckets of water, of course, but it also could be germs, money, violence, fashions, kidneys, happiness, or obesity.

The usual way contagion is thought is if one person has something and comes into contact with another person, that contact is enough for the second person to get it. The spread of health behaviors such as vaccination is often modeled as a simple contagion process, similar to biological contagion, where each exposure event contributes equally to the probability of adoption of the behavior. However, there is increasing evidence that the process of social transmission of behaviors is governed by the process of complex contagion, where social reinforcement - i.e., multiple exposures from different peers - are necessary for adoption Campbella and Salathé (2013).

2.3 STATE OF THE ART

Seeing ourselves as part of a superorganism allows us to understand our actions, choices, and experiences in a new light. If we are affected by our embeddedness in social networks and influenced by others who are closely or distantly tied to us, we necessarily lose some power over our own decisions. Christakis and Fowler (2009) believe that our connections to other people matter the most, and that by linking the study of individuals to the study of groups, the science of social networks can explain a lot about human experience.

And, as referred to in the last section, our connections do not end with known people. People we may not even recognize can start chain reactions that eventually reach us.

On Christakis and Fowler (2013) a review of previous work, with several years, is done and the authors explicitly state they stand behind it. Referring to an offline study, in which the medical records of about 12,000 patients were analyzed, they show stimulating evidence regarding social contagion. It is suggested that obesity may spread in social networks in a quantifiable and discernable pattern that depends on the nature of social ties. Social distance appears to be more important than geographic distance within these networks. There was not focus on causation but instead on correlation, but frequent exposure to local environmental factors were ruled out as an explanation for their observations. There was also account for sources of confounding.

Separating correlation and causation is a difficult task. Some researchers have tackled the problem of homophily vs influence has been tackled by some researchers. Anagnostopoulos et al. (2008) used two techniques in order to do this, shuffle test and edge reversal test. The conclusion was that while it is true that influence does not play an essential role in users' tagging behavior in Flickr, there is some limited effect by looking at the difference between similar tags, some propagation of the misspelled versions was identified.

Aral et al. (2009) describe a statistical method for distinguishing these two phenomena. By analyzing the day-by-day mobile service adoption behavior of over 27 million Yahoo! users in Yahoo! instant messaging network, they show that homophily explains >50% of the perceived behavioral contagion in mobile service adoption. These findings demonstrate that homophily can account for a great deal of what appears at first to be a contagious process. It was found that different subsets of the population, characterized by distributions of individual and relational characteristics, such as the strength of ties and local clustering, display various susceptibilities to potential influence.

An experimental approach on Facebook to measure the spread of information sharing behaviors was made Bakshy et al. (2012). A random sample of all Facebook users who visited the site was acquired, comparing situations where both influences via the feed and external correlations exist (the feed condition), to circumstances in which only external correlations exist (the no feed condition). By balancing the behavior of individuals within these two conditions, is possible to determine the causal effect of the medium on information sharing. Weak ties are argued to have access to more diverse information because they are expected to have fewer mutual contacts. Although stronger ties are individually more influential, it is the more abundant weak ties who are responsible for the propagation of novel information. The majority of influence results from exposure to individual weak ties, which indicates that most information diffusion on Facebook is driven by pure contagion.

As for the speed and types of diffusion, the study of Zhao et al. (2012) on Facebook networks, whose ties are within five American universities, find that information pushing speeds up the information exchange within an online social network. For the networks with very high clustering coefficients, selecting weak ties preferentially can speed up the information propagation.

Lerman and Ghosh (2010) tries to understand how information spreads through the social network by measuring how the number of in-network votes a story receives. Posting a link on Twitter is analogous to submitting a new story on Digg, and retweeting the post is analogous to voting for it. Digg networks are dense and highly interconnected. A story posted on Digg initially spreads quickly through the network slowing down after the story is promoted to Digg's front page. The Twitter social network is less dense than Digg's, and stories spread through the network slower, but they continue spreading at this rate as the story ages and generally penetrate the network farther.

Another aproach is made by Stieglitz and Dang-Xuan (2013). In this paper, it is explicitly examined whether the affective dimensions of Twitter messages (positive and negative sentiment) occurring in social media content are associated with a user's information sharing behavior. They found that emotionally charged Twitter messages tend to be retweeted more often and more quickly compared to neutral ones. Regarding the tweeting behavior of influential users in the Twitter network, they tend to post more emotionally charged tweets. In doing so, their influence may increase even more because their emotionally charged content would be more likely to be disseminated.

Furthermore, do topics play a key role? Analyzing how tokens known as hashtags spread on a network defined by the interactions among Twitter users, such as Romero et al. (2011) helps to answer this question. They found that tags of different types and topics spread in different mechanics. Along with it, some of the most significant differences in hashtag adoption provide intriguing confirmation of sociological theories developed in the off-line world. In particular, the adoption of politically controversial hashtags is primarily affected by multiple repeated exposures, while such repeated exposures have a much less critical marginal effect on the adoption of conversational idioms. They provide a large-scale validation of the complex contagion.

While the existence of influence can be difficult to detect, some investigators do not altogether dismiss the role played by influence.

On the other hand, there have been other studies revealing the genuine and certain existence of social contagion and influence. Huang et al. (2012) show that even after removing the effects of homophily, there is clear evidence of influence. For instance, they find that people rate items recommended by their friends higher than they otherwise would.

With the rise of autonomous bots all around social networks, should they be taken in account? In Brachten et al. (2017) the strategies and influence of social bots were analyzed based on relevant features and network visualization on a German state election in 2017, via Twitter. Possibly due to the concentration on the German language as well as the elections regionality, identified bots showed no signs of collective political strategies and low to none influence. This away, it is expected they do not interfere on the influence estimation.

Regarding the usefulness of the sources discussed in chapter 2, Demunter (2017) approaches this subject having in mind their application to tourism. Thus it can be related to this work, validating the impact of an event influence with data from a regional economy, for example. He starts by explaining a methodological barrier on social media posts, in particular related to the selectivity bias: the inclusion probability or likelihood that an individual or event will be observed is highly correlated with the intensity of activity (namely the frequency of posting on social media). This limits its usefulness to detecting short-term trends rather than volume information or longitudinal trends.

He continues by saying Web activity can indicate of which topics interest people at each moment in time. Searching for information on tourism destinations or page views of Wikipedia articles related to destinations can do much to help predict tourism flows. Interest via search queries or visiting websites does not always lead to a physical visit or a purchase, but a correlation has been found several times Sharpe et al. (2016). A refined analysis (e.g., destination names in combination with search terms such as 'hotel' or 'metro') could increase the correlation with tourism visits.

A decade ago, Girardin et al. (2008) examined the potential of digital pictures to uncover the presence and movement of people in a city. In their study they highlighted the relevance of this data for tourism (and urban) planning: 'information about who populates different parts of the city at different times can lead to the provision of customized services (or advertising), accurate timing of service provision based on demand, and, in general, more synchronous management of service infrastructures'.

2.4 SUMMARY

Most of the works concerning information and influence propagation in online or offline social networks regard one of next three areas: i) Detecting interesting, bursty, topics; ii) modeling diffussion processes, explanatory and predictive models; iii) identifying influential spreaders.

Some limitations were found. A few authors use techniques with poor quantitative indication of the existence of influence, and others do not consider topics. Furthermore, it is not easy to provide formal verification of results. Can we pinpoint social networks and behaviors, where influence is indeed prevalent, and verify our tests? Also, what happens when different sources of social correlation are present, as is usually the case? Only some works inspected had in consideration several social networks sources. What if at least one source from every type of social network, by content, is taken into account? An event estimated influence could be correlated with an impact on the regional economy?

As made clear by the previous descriptions, the potential of social media and web activity data is immense. Also, its broad definition concerns a very diverse set of data types besides text, where each type contains particular properties.

All these important questions might be tricky to answer and probably require the design of controlled user experiments.

Relaxing the closed world assumption and scalability are additional adversities when researching influence propagation on social networks.

To summarize, the existence of influence and its effectiveness for applications such as enumerated above depend on the datasets and what the goal is to do with them. There is both evidence supporting and challenging it, found from different datasets by researchers. For a given situation, careful analysis of evidence should first be undertaken before deciding whether to adopt a it. There is still much to be done in this area of investigation.

3

THE PROBLEM AND ITS CHALLENGES

3.1 INTRODUCTION

The abundance of big data sources capable of capturing facets of the social influence phenomenon makes it abundantly clear that this is not an outdated area, it still is on the frontline.

In this chapter is suggested an approach to the problem, how to tackle it. This includes explaining on the chosen research and development methodologies, along with the problems that may arise trying to reach the solution. Considering possible courses of action and selecting the best one are crucial tasks.

3.2 RESEARCH METHODOLOGY

Similar research goals can be sought in completely different ways depending on the accessibility and proximity of experts, synergies with ongoing research projects and so forth. Because of this research context, the chosen research strategy, represented in Figure 9, is based on the following activities:

1. Update the acquired knowledge by reviewing recent and state-of-the-art publications;
2. Design and develop the different parts of the proposed models enlarging the scope gradually in an iterative process;
3. Experiment on and evaluate the system;
4. Attend conferences and workshops to present partial results and to learn of existing state-of-the-art advancements;
5. Redesign the system with the feedback obtained from all the above means;
6. Develop and deploy the final system for context and behavioural analysis in real-world-like scenarios to gather results;

7. Disseminate the obtained knowledge and experiences to the research community.

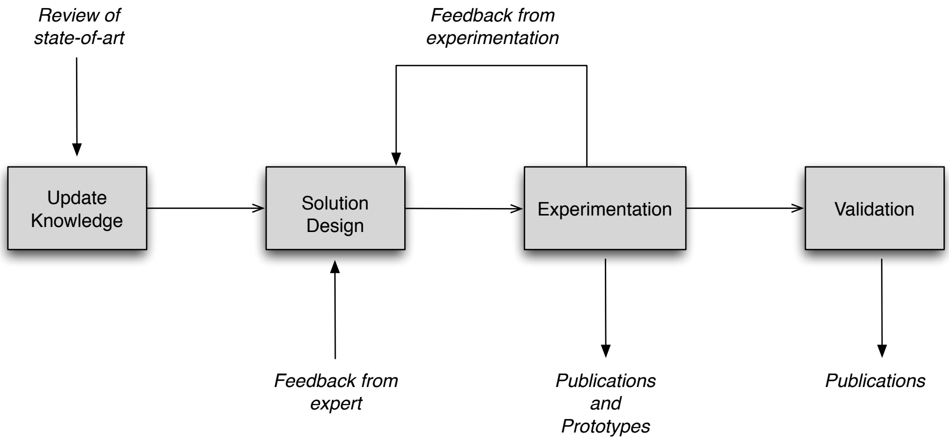

Figure 9: Schematic view of the research process.

This research process is the action-research methodology composed of five different phases:

- Diagnosing: identifying the problem;
- Action planning: considering possible courses of action;
- Taking action: selecting a course of action;
- Evaluating: analyzing the consequences of the course of action;
- Specifying results: identifying general findings.

These phases were applied to all the outlined research activities with the aim of providing rigour, reflexive critiques, and continuous challenges.

3.3 Proposed Approach - Solution

An event has sereval stages subjectable to analysis, before, during and after. To fully understand the role of social influence, it would be ideal to study the three moments. To perform this task, the system data acquisition component would have to be running when the event is not yet influential, weeks or months before its start. Given the potential amount of data and consequent difficult analysis, the pipeline methodology in figure 2 will be followed. This pipeline emerged in 2012, almost a decade after CRISP–DM (Cross Industry Standard Process for Data Mining) Process.

Usually, models developed in the context of online social networks assume that people are only influenced by actions taken by their connections, such as liking or sharing. To put it differently, it is considered a closed world and assumed that information spreads because of informational cascades. That is why the path followed by a piece of information in the network, the diffusion or influence graph, is often referred to as the spreading cascade.

This provides knowledge about where and when a piece of information propagated but not how and why did it spread. Therefore, there is a need for models that can capture and predict the hidden mechanisms. In this scope, there are explanatory models and predictive models.

Explanatory models, static or dynamic, as the name suggests, make it possible to retrace the path taken by a piece of information and are very useful to understand how information propagated.

Predictive models aim at predicting how a specific diffusion process would unfold in a given network, from temporal and/or spatial points of view by learning from past diffusion traces. We classify existing models into two development axes, graph (Independent Cascades and Linear Threshold) and non-graph based approaches. While the first focus on network structure, the former do not assume the existence of a specific graph structure and have been mainly developed to model epidemiological processes. Nodes are classified into several classes or states.

Both models typologies are of interest, generating different analysis on the data previously extracted from social networks and social media, but in this work there is a focus on explanatory models. From one point of view, it is useful to an organization to prepare or manage the event while it still is in progress, on the other hand, understanding how to tackle the next edition with more valuable knowledge is also essential.

What good does this work intend to bring to society? Essentially, from data analysis the main idea is to provide actionable knowledge to companies, helping in their decision making. It is intended to study the following applications: viral marketing, exploiting the word-of-mouth effect in a social network to achieve marketing objectives through self-replicating viral processes, and influence maximization, an NP-hard optimization problem.

Having a more significant impact through maximizing and improving social influence an event is a obvious matter of preoccupation. More influence usually results in more profits a business has.

In a more objective application's description, the planned computational system optimally results in the know-how to:

- Identify communities and infer its type and specific roles played by different users in the community;
- Customize services advertising;
- Provide accurate timing of service, based on demand;
- Implement a more synchronous management of service infrastructures;

- Identify the impact on the collaborative economy.

Of course that before being taken as ready to act on the real world it would have to be validated against data extracted from web activity and content, particularly from tourism related websites and distinct online networks. This way it is perceived if the estimated influence and, consequentially, its conclusions are precise.

3.3.1 *System Architecture*

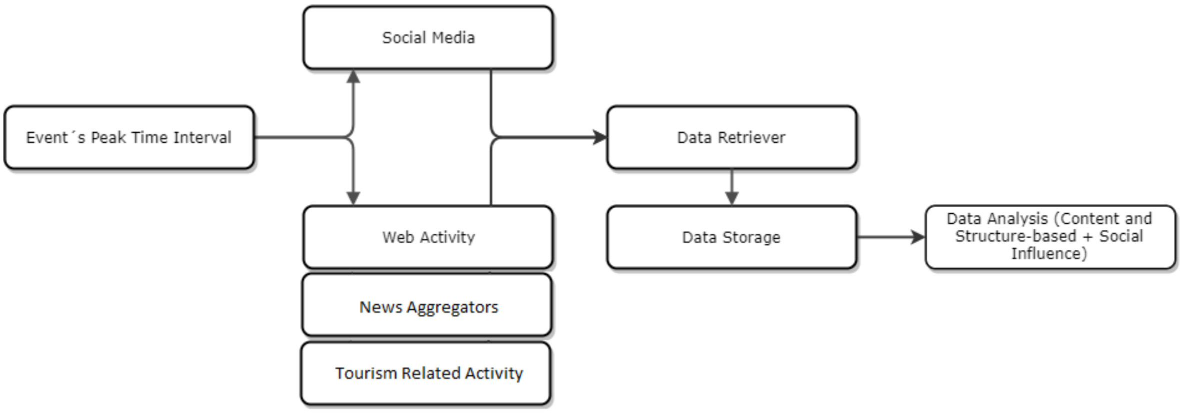

Figure 10: Proposed solution with its integrating components.

The above solution components and interaction were idealized having the following steps in consideration:

1. Determine the event's peak time interval, searching on Google Trends by its designation, in which data from online platforms will be extracted;

2. Visits to specific Wikipedia pages visits also measured, to increase the accuracy;

3. Collect information from top news aggregators (Google news and Sapo) and other possible popular news websites: title, headline, publication date, the news outlet and its position in the ranking;

4. Analyze pattern of visits and reviews on a sample of Portuguese tourism websites, such as TripAdvisor, Expedia, Airbnb, and Booking, regarding accommodations, activities, and restaurants. Possible search by event's location;

5. Collect patterns and numbers of searches by the event on video networks, like YouTube;

6. Extract posts geo-data, temporal data, user details, content, visualizations, reactions, shares and posts' comments from online social networks (Facebook, Twitter, Flickr,

and Instagram) using specific event #'s to model and measure the evolution of influence in the network. This data is exported to an influence graph representing user's details and relations, as well as the diffusion of influence;

7. Analyze correlations between the evolution of influence on online social networks the analysis on data extracted from the other sources. It is also studied the impact on the collaborative economy;

8. From the extracted data and through its analysis discover relevant insights, conforming with the applications discussed in the previous section.

3.4 CHALLENGES

The main challenges with the development of this work regard the data, its extractio, nprocessing, and analysis.

Privacy preserving methods, data quality, scalability, overfitting, models' accuracy, and assertive results are the difficulties expected to be faced and overcome.

Because of technical API limitations, there is a data acquisition bottleneck potentially responsible for missing data. To overcome this issue, if possible, one approach is to scrape data as efficiently as possible. Sampling methods that consider both network topology and users' attributes, such as activity and localization, allows capturing information diffusion with lower error in comparison to random or activity-only based sampling.

3.5 SUMMARY

To summarize, the most critical factor to consider is the data, every process is related to it, from extracting to analysis. Not being trivial tasks, it is mandatory to follow one or a handful of proven methodologies, as the ones presented. Thereby, some steps are always essential to adopt: acquisition, cleaning, integration, analysis and interpretation.

The objectives and hypothesis proposed in section 1.2 are feasible, quantifiable and possible to conquer. If it is thoroughly followed a rethought management of data and studied works, in which their authors excelled, plus the concept this work tries to prove.

4

SOCIAL MEDIA AND WEB CONTENT

4.1 INTRODUCTION

Analysis refers to breaking a whole into its separate components for individual examination, therefore an interdisciplinary field. Data analysis is a process for obtaining raw data and converting it into information in an understandable manner, a structured one.

To get there, data retrieval is essential, and data is collected from a variety of distinct sources.

There has been increasing concern about privacy of individuals when it comes to social networks and general web activity. Disclosing information on the web is a voluntary act of an individual and in most cases users are unaware of who can access their data and how potentially their data will be used. Now, more than ever, personal information is requested in many daily life activities on the web. While some of this sharing is not considered dangerous, risk and many interested entities are searching to gain on its exploitation.

Because of these and in light of recent scandals regarding privacy rights, there has been a growing sensibility to the matter. Data privacy means freedom from unauthorized intrusion, which implies techniques such as data anonymization and security against privacy breaches, from who owns the data.

That was not the case earlier this year when it was revealed Cambridge Analytica, a company that had worked on Donald Trump's US presidential election campaign, had harvested the personal data of millions of people's Facebook profiles without their consent and used it for political purposes. It has been described as a watershed moment in the public understanding of personal data and precipitated a massive fall in Facebook's stock price and calls for tighter regulation of tech companies' use of data. The scandal was significant for inciting public discussion on ethical standards for social media companies. Consumer advocates called for greater consumer protection in online media and right to privacy.

As for what scraping is concerned, from my excursions into this subject, it is a slightly grey area. Though realistically speaking, as long as ethical scraping is considered (such as not overloading the web servers with constant rapid scrape requests) or otherwise causing undue harm or business loss, that might be a low risk in general. It's not like the information is privileged and not already out in public.

Many scrapers are written in Python language to facilitate the step of further processing the collected data. Scrapers can be written using frameworks and libraries for web crawling, such as Scrapy or Selenium. BeautifulSoup and lxml, libraries for parsing HTML(and XML) style documents and extract certain bits wanted, were also considered. All are free, but the choice fell mainly on Scrapy and occasionally the last ones. Selenium has the advantage of handling JavaScript, but it lacks speed and multifunctionality.

Scrapy is a rather big framework, it's not necessarily hard or complicated, but sometimes a whole framework is not necessary to scrape some data. That's where the alternative comes in, sticking with requests library to retrieve the pages and BeautifulSoup to parse those pages for bits of data needed.

Scrapy is an open-source web-crawling framework written in Python. Initially designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. Scrapy project architecture is built around ”spiders”, which are self-contained crawlers that are given a set of instructions. Following the spirit of other don't repeat yourself frameworks, it makes it easier to build and scale large crawling projects by allowing developers to reuse their code.

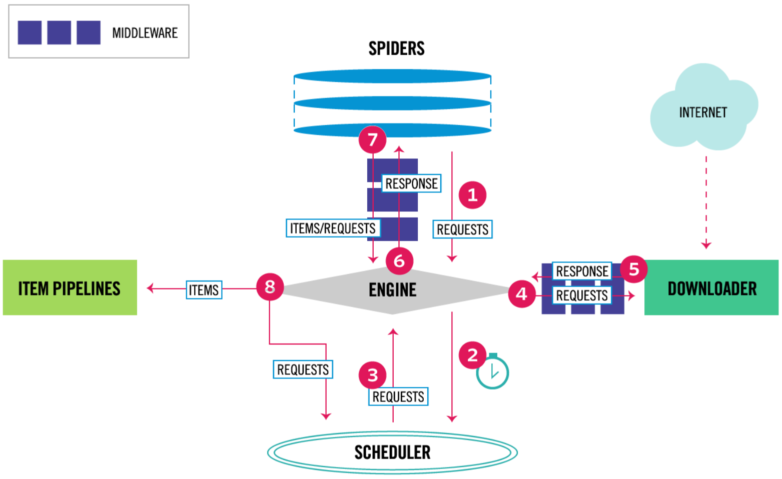

Figure 11: Scrapy framework architecture, Scrapy.

It has the tools to manage every stage of a web crawl, to name a few:

- Requests manager - which is in charge of downloading pages and the great bit about it is that it does it all concurrently behind the scenes, so the user gets the speeds of concurrency without needing to invest a lot of time in concurrent architecture.

- Selectors - which is used to parse the html document to find the specific bits wanted. BeautifulSoup does the same thing, and its usage is possible instead of scrapy Selectors if preferred.
- Pipelines - after retrieving the data it may be passed through various pipelines which are basically bunch of functions to modify the data.

In a project created with this framework, it is also easy to define, in a settings file, user-agents, identification is a good practise, obedience to robots.txt and requests download delays (usually 5 seconds and above), to esteem the connection established.

An example of what a simple Scrapy spider looks like is presented right next:

```
import scrapy
from myproject.items import MyItem

class MySpider(scrapy.Spider):
    name = ’example.com’
    allowed_domains = [’example.com’]

    def start_requests(self):
        yield scrapy.Request(’http://www.example.com/1.html’, self.parse)
        yield scrapy.Request(’http://www.example.com/2.html’, self.parse)
        yield scrapy.Request(’http://www.example.com/3.html’, self.parse)

    def parse(self, response):
        for h3 in response.xpath(’//h3’).extract():
            yield MyItem(title=h3)

        for url in response.xpath(’//a/@href’).extract():
            yield scrapy.Request(url, callback=self.parse)
```

Right away comes to attention its ability to scrape specific pages and parse the response as conveniently as catching the content needed by XPath or CSS selectors and follow links to the next page, for example.

On the next sections each project coded or reused to extract raw data will be summarized, as well as underlying strategies adopted since there is a methodology defined for every type of data (subsection 3.3.1). There was the effort of using only free tools and trying to escape highly restricted APIs, which in few cases was not trivial.

According to the solution proposed in the previous chapters, we defined a couple of data extraction occasions, before and after the studied physical event. This away is possible a comparison of the same searches separated by distance in time. Needless to say that these processes have to be identical.

In social media, a group of hashtags' variations is taken into account in the scrapers' search. A diversity of keywords related to the event are used in the web activity, as well as the cities surrounding involving it. The first is also useful to news websites' searches and the former to tourism-related activity. This imposes uniformity by using the same sets of keywords in different scrapers.

After collecting data and before getting too intelligent techniques, some steps are always fundamental, data cleaning, integration, and exploratory analysis.

4.2 SOCIAL MEDIA

Social media was the main focus in this work, its epicentre. From Marktest Consulting (2017) study it is inferable straight away that Facebook dominates amongst social network websites, with almost 96% of Portuguese users having an account. The difference to Instagram, second one in the ranking, is enormous. Being elected as a top social network, carries some unique characteristics, the range of subject matters it covers is more significant and more importance is given to insights extracted from its data.

Unfortunately and despite the effort applied to contradict what happened consequently to Cambridge Analytica scandal, Facebook's data had to be dismissed of this study. Public feed posts were already inaccessible on start, leaving access only to posts on pages, groups, and events, via Graph API, but it got worse. On February 2018, user information stopped being included in responses unless the request was made with a Page access token. User and app access tokens became even more restricted, it was implemented a rigorous review on all apps registered. Consequently, the decision of dropping Facebook as a data source was made.

Ideally, all the major online networks websites would be scrapped, from each topology, following an unified strategy: extract public feed posts geo-data, temporal data, user details, users relations, content, visualizations, number of likes/favorites, shares/retweets and posts' comments, using specific event #'s, saving the results on CSV(comma-separated values) format in order to model and measure the evolution of influence in the network. Included in these referred topologies are Social networks, Image networks, Blogging networks, and video networks websites.

Regarding the first one, Google+ was not deemed worthy of replacing Facebook, a lot of personal data from random users is hidden and also lacks user connections data. Additionally, in a sense, the percentage, shown in 7, indicating 50% of Portuguese users have an account in Google+ is misleading, since everyone registered to a Google service has, automatically, a Google+ account.

As for image networks Instagram and Flickr were the selected ones, leaving Pinterest outside of the equation, there is currently no way to search for pins with specific tags or keywords through the API, and demographic data is limited.

Blogging and video networks are well represented, with Twitter and YouTube having the lead.

Twitter

After some research on projects publicly made available, it was possible to reuse a solution called TweetScraper, from jonbakerfish, able to get tweets from Twitter Search. It is built on Scrapy without using Twitter's API. The standard one only provides a 7-day search endpoint that does not produce full-fidelity data and supports a smaller set of query operators. The crawled data is not as clean as the one obtained by the APIs, but the benefits are getting rid of the API's rate limits and restrictions.

Tweets attributes						
usernameTweet	description	statuses_count	nbr_favorite	is_reply	datetime	followers_count
ID	in_reply_to	is_retweet	user_id	favourites_count	medias	text
friends_count	nbr_reply	query	has_media	nbr_retweet	url	user_location
Retweets attributes						
status_id	user_id	user_name	screen_name	user_description	followers_count	favorites_count
Favorites attributes						
status_id	user_id					

Table 1: Data extracted from Twitter.

User information and retweets had to be extracted alternatively, based on the tweet id and through tweepy, a python library for Twitter API. Favorites were only available by sending a request, simulating opening the details pop-up and so it is limited. These extensions were coded in scrapy's pipelines.py.

Instagram

Posts attributes						
user_id	username	full_name	profile_pic_url	media_count	follower_count	following_count
date	media_type	code	view_count	video_duration	location	pic_url
like_count	comment_count	caption	caption_media_id	tags		
Comments attributes						
post_id	username	is_verified	full_name	user_pk	created_at	created_at_utc
comment_like_count	text	comment_pk				
Favorites attributes						
post_id	username	is_verified	full_name	pk		

Table 2: Data extracted from Instagram.

Once again the option to reuse was on the table and the crawler presented, simonseo, was chosen because most of its competitors seem to either require a browser or a developer account. This one utilizes a Python wrapper for the Instagram private API with no 3rd party dependencies, and thus no developer account is required. The only requirements are a Instagram username and password, and either a Hashtag or a target file.

Flickr

Flickr does not fit in the same level as the other selected sources of data (a bit lower importance), and because of that, it is not expected that much high flow of requests. After checking the official API, Flickr, there were no hindrances in moving in that direction.

Its photos search method returns a list of photos matching given tags and range of upload dates. Only photos visible to the calling user are returned. Being authenticated returns public images.

A scrapy project was built from scratch to accommodate and process these requests.

Photos attributes							
id	user_id	nmr_comments	date_taken	alias	views	title	real_name
username	url	location	nmr_favorites	user_location	date_posted	desc	
like_count	comment_count	caption	caption_media_id	tags			
Comments attributes							
comm_id	post_id	authorname	realname	datecreate	author	_content	
comment_like_count	text	comment_pk					
Favorites attributes							
post_id	username	favedate	realname	nsid			

Table 3: Data extracted from Flickr.

Youtube

To scrape YouTube, another satisfactory unofficial method was discovered HermanFassett, implemented in Heroku, a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.

Similarly to the previous scraper, a simple scrapy project was built, having a YouTube search query and the search results' page as options. In order to maintain results related to the event, only 25 pages for each keyword are scraped.

Videos attributes								
views	duration	url	uploader	title	rank	snippet	upload_date	term

Table 4: Data extracted from Youtube.

4.3 WEB ACTIVITY

In addition to digital footprint left when using social media, traces left by people while using search engines and visiting websites, indicate which topics interest people at each moment in time, the factor that unifies all data sources.

In this range of scrapers, the search for viable methods of data extraction went much more smoothly. Since it is usual for someone to undergo a simple search online on the destination or event that plans to go, Wikipedia page views and Google Trends data illustrate well a general interest on a specific subject. Also, there is a widespread common knowledge that people resort to Google's search engine, and usually, just after, to Wikipedia articles.

In this work is especially meaningful sources which can be a proxy for visits to a destination or interest in events.

Google Trends

May be defined as a website by Google that analyzes the popularity of top search queries in Google Search across various regions and languages. The website uses graphs to compare the search volume of different queries over time.

An unofficial API for Google Trends was available, GeneralMills, once more explored through a scrapy project. It allows simple interface for automating downloading of reports from Google Trends. Main feature is to enable the script to login to Google on behalf of the user to enable a higher rate limit.

The more relevant API method is Interest Over Time: returns historical, indexed data for when the keyword was searched the most, as shown on Google Trends' Interest Over Time section.

Data attributes	
date	query_value

Table 5: Data extracted from Google Trends.

Wikipedia

The Wikimedia REST API, team offers access to Wikimedia's content and metadata in machine-readable formats. Focused on high-volume use cases, it tightly integrates with Wikimedia's globally distributed caching infrastructure. As a result, API users benefit from reduced latencies and support for high request volumes.

The REST API along with its documentation is available for all major Wikimedia projects. The spider developed with scrapy includes the following projects: Portuguese, English, Spanish, German, Italian and French, without discriminating access method, a daily granularity and between given dates.

Data attributes			
timestamp	article	views	project

Table 6: Data extracted from Wikipedia.

4.4 NEWS AGGREGATORS

Traditional media, or as some refer to as old media, has been used in the advertising world for years. These forms of communication are the steadfast ways that businesses have reached both consumers and other companies for decades. The time when news was primarily consumed on traditional media platforms is over. Now, more people consider online news sites and social media – instead of TV, radio and print newspapers – to be their primary sources of news.

Even tough the news is now primarily consumed online, it is worth noting that it is still the traditional media companies originating from TV and newspapers who are behind the most preferred online news sites.

Even though online news sites is the main source of news in all countries – with some exceptions – they face a challenge regarding trust from their users. Trying to surpass it, was followed a strategy of scraping the leading websites of news in Portugal, Jn, Dn, Público, Expresso and Correio da Manhã, in addition to the best aggregator nationally, Sapo (limited to 1500 news for each keyword), and internationally, Google News. The formers are considered to have more weight.

Beacause of these data sources having similar attributes, there is a basic structure, presented next.

News attributes			
Pubdate	Title	Url	term

Table 7: Data extracted from news websites.

In extent, both aggregators' script also capture rank, publisher (only Google News), body and score (only Sapo).

The Scrapy framework was suitable to develop all these scrapers, but because of the nature of a few of these websites, using a lot of javascript calls and custom search by Google, the technique of searching network packages moving around had to be refined.

4.5 TOURISM RELATED ACTIVITY

As it has already been stated, searching for information on tourism destinations can do much to help analyze and predict tourism flows. Not being the owner of a website, it is no trivial to access visits counters, would legitimize influence through correlation with other sources.

So validating the impact of an event's influence with data from a regional economy was one of the solutions encountered. In particular listings' characteristics, as well as their reviews' details, frequency and and how they vary in time, subject to external factors. Hotels, rental houses, activities, and restaurants are of interest.

Tripadvisor, Expedia, Booking, and Airbnb are some of the most recognized tourism listings websites, therefore were the preferred ones. Wanting to scrape tourism websites led to the necessity of adapting and fixing previous projects for them to work. These websites are constantly trying to block data leakage, for example changing their layouts, many javascript calls in the HTML page and remodeling classes names.

Tripadvisor

This website was the one that gave more fight, so that it was possible to extract the desired data. The scraper developed to that end had to undergo several changes through time, resulting in more than one version coded.

In the end prevailed two different implementations, one for hotels, requests plus lxml, and another duplicated to restaurants and activities, scrapy project.

To collect reviews from each listing BeautifulSoup and requests module were experimented. In this case, requests with forms are included to choose the option "All languages" and to expand reviews' text, there was no way of avoid it.

Hotels attributes									
best_provider	country	highlights	hotel_amenities	hotel_url	id	lat	lng	locality	max_price
min_price	name	nmr_rooms	official_description	rank	rating	ratings_average	ratings_excellent	ratings_poor	ratings_terrible
ratings_very_good	region	review_count	room_amenities	room_types	stars	street_address	things_to_do	top_amenities	type
Restaurants attributes									
rating_overall	url	id	lng	region	features	rating_value	rank	cuisine	rating_food
good_for	name	lat	rating_service	meals	address_street	average_prices	rating_atmosphere	address_city	
Activities attributes									
rank	region	rating_overall	id	url	type	nmr_reviews	lat	lng	address
name	address_city								
Reviews attributes									
bubble_rating	id	id_restaurant	num_reviews_reviewer	restaurant_name	review_body	review_date	review_title	reviewer_country	reviewer_name

Table 8: Data extracted from Tripadvisor.

Tripadvisor is a tourism aggregator website. Adding restaurants and to-do activities to hotels is handy and, as expected, its scraper had a lot of data to contribute. It was no revelation the fact it was the most consumer of time. Plus the website implements the time-out technique frequently.

Booking

Towards Booking scraping there were no adversities, a simple Scrapy project was enough. Hotels listings details and reviews are also separated.

Only Portuguese, English, German, French and Spanish reviews were scrapped.

Listings attributes						
lat	price	lng	header	nmr_reviews	stars	review
score	distance	url	score_loc	nome	id	
Reviews attributes						
hotel_name	score	date	title	negative_content	hotel_url	reviewer_name
hotel_id	positive_content	reviewer_location	reviewer_nmr_reviews			

Table 9: Data extracted from Booking.

Airbnb

After being forced to rewrite the Airbnb houses scraper, because of null results, a better alternative was found. Inspecting the network packages moving around in the website, it was trivial to detect requests to a URL suggesting it belonged to an API. It was a surprise to have acess to an API, since it is no usual in these type of websites.

stevesie is competent on getting Airbnb listings by location keyword (e.g., neighborhood, city, state, town). Listings Details method enables to gather highly detailed information about a specific listings. Reviews task is also used, retrieving reviews in any language.

Listings attributes									
id	is_superhost	localized_city	person_capacity	room_and_property_type	star_rating	query	price	url	lat
reviews_count	name	lng	city	beds	bathrooms	bedrooms	is_host_highly_rated	picture_count	
Reviews attributes									
comments	reviewer_id	created_at	listing_id	rating	language	reviewer_name	id		

Table 10: Data extracted from Airbnb.

Expedia

Expedia.com is a travel booking website that can be used to book airline tickets, hotel reservations, car rentals, cruises and vacation packages. But following the previous scrapers, exception to Tripadvisor, only Hotels were chosen to be scrapped.

SamHO666, a data crawler based on the Scrapy framework, was found also found in the github community and later tested. After some transformations and touch-ups, not having to search for region ID and results' filtering by search city, it was presumed ready. All reviews are covered.

Hotels attributes							
price	hasAvailableOffer	hotelName	itemType	neighborhoodOrCityName	superlativeMessage	hotelStarRating	infositeUrl
totalReviews	guestRating	structureType	reviews	hotelId			
Reviews attributes							
reviewId	userLocation	ratingRoomComfort	positiveRemarks	userName	ratingOverall	locationRemarks	ratingHotelCondition
reviewSubmissionTime	hotelId	reviewText	ratingRoomCleanliness	ratingService	title	negativeRemarks	

Table 11: Data extracted from Expedia.

4.6 DATA PREPARATION

After data acquisition through scrapers and recording to CSV files, to comply with figure 2, the immediate next tasks are information extraction and cleaning , as well as data integration, aggreagtion and representation.

No matter how someone views it, data preparation and preprocessing tasks constitute a high percentage of any data-centric operation, be it of a descriptive or predictive nature. It can also be a collection of the most frustrating tasks for a data practitioner.

The previous section details which data the analysis starts with. It is the raw data. Moving forward, comes into the horizon everything from data sourcing right up to, but not including, model building. That is the vague, yet oddly precise definition we may also assume.

Generally, data preparation is quite a broad concept that encompasses things such as:

- collection methodology;
- formatting (strings, numbers, delimiters, etc);
- normalization (acronyms, typos, formatting, encoding,null/empty, etc);
- filling in / removing missing values;
- de-duplication;
- merging multiple datasets/sources;
- sampling;
- specific preparation methods depending on analytical technique.

Python is more of a general-purpose language with a rich set of libraries for a wide range of purposes. It's as good for mathematics, engineering, and deep learning issues as for data manipulation and visualizations. Guido van Rossum introduced Python back in 1991. It has since become an extremely popular general purpose language, and is widely used within the data science community. The significant versions are currently 3.6 and 2.7, and everyone in the community recognizes logo. The first was selected, it is the currently feature-improved version of Python.

Figure 12: Python logo, Vinícius Damaceno (2016).

Since the scrapers adopted and developed were coded in Python, having already in mind data manipulation and analysis, we decided to start this data preparations steps in the same language, just right after trying Rapidminer.

It provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analytics. Having in mind the idea of speeding up the data preparation step, we ended up burning some unnecessary time switching to Python after realizing it was not worth it. Without academic license this program is useless, reading ability limited to 10000 rows, and with one, analysis time shrinks to only one month.

Packages such as pandas, NumPy and scikit-learn make Python a solid option for advanced machine learning applications and together make up a python scientific ecosystem.

The **NumPy** library provides the means for performing n-dimensional array manipulation, which is critical for data science work. N-dimensional arrays couldn't be easily accessed without NumPy functions that include support for linear algebra, Fourier transform, and random-number generation.

Pandas can be described as one of the most well known data manipulation and analysis library. It is an unconditional library of choice for analytics, data processing and data science.

The **Matplotlib** library provides you with a MATLAB-like interface for creating data presentations of the analysis performed. Without this library, generating an output that people outside of the data science community could easily understand would not be so simplified. Seaborn library, in the same category, was also deemed to try out.

The **Scikit-learn** library is one of some Scikit libraries that build on the capabilities provided by NumPy and SciPy to allow Python developers to perform domain-specific tasks. In this case, the library focuses on data mining and data analysis, tasks explored in the next chapter.

Some of these initial major tasks are now explained, being this an iterative and sometimes cyclic process, which means that some of them might be revisited several times. It is redundant to say that, to fulfill these, some data exploration and understanding is treasured, like assuring data quality and precise measurements.

Data Integration

In many real-world situations, the data that we want to use come in multiple files. We often need to combine these files into a single one to properly analyze the data.

In this case eliminating duplicate copies of repeating data is often unavoidable, even it is applied by unique attributes or by the entire row.

Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved. It is a process in which heterogeneous data is retrieved and combined as an incorporated form and structure. It involves combining data residing in different sources and providing users with a unified view of them. The new set enhances the information.

Dealing with Missing Values

Different types of data and processes suggest different best practices for dealing with missing values. Understanding the situation and relying on Pandas powerful methods is crucial.

The former handles missing values, NaN, as the user pleases. Because of them some dropping might be chosen to happen, on instances or on attributes. Usually the best aprocah if the proportion of NaN data is much bigger than the other. Filling them with the attribute mean value or mode is also an option.

There are all sorts of strategies for dealing with missing data, and none of them are applicable universally.

Dealing with Outliers

There are times when including outliers in modeling is appropriate, and there are times when they are not (regardless of what anyone tries to imply). This is situation-dependent.

It is not recommended excluding any outlier in the principal analysis (unless positive certainty they are mistaken). In science, often you discover new conclusions precisely when focusing on such outliers.

So outliers can be the result of poor data collection, or they can be genuinely good, anomalous data. These are two different scenarios, and must be approached differently.

In this work real data is extracted, still subjected to errors. Therefore it is expected outliers adding value to the analysis.

Dealing with Imbalanced Data

Starting to look at real, uncleaned data one of the first things noticed is that it's a lot noisier and imbalanced. The classical data imbalance problem is recognized as one of the major ones in the field of data mining, as most algorithms assume that data is equally distributed. In the case of imbalanced data, majority classes dominate over minority classes, bias towards majority classes.

A commonly used strategy is called resampling, which includes undersampling and oversampling techniques. If one balances the dataset by removing the instance from the overrepresented class, then it is called undersampling. Oversampling can be achieved by adding similar instances of an underrepresented class to balance the skewed class ratio. Synthesizing new minority classes is also viable. Resampling could be done with or without replacement.

Another strategy is doing nothing. Sometimes luck comes into play and nothing needs to be done, it works without need for modification.

When considering evaluation metrics, one must take precautions and do not use accuracy evaluation methods with imbalanced data. There can only be progress if measuring the right thing.

Data Transformations

Transforming data is one of the most important aspects of data preparation and one which requires more finesse than most others. When missing values manifest themselves in data, they are generally easy to find, contrasting with data transformations, it is not trivial to know if they are essential.

Transforms are usually applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied, or to improve the interpretability or appearance of graphs.

In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the dataset, robust scalers or transformers are more appropriate. Some of the most important preprocessing transformations are now listed:

- Standardization, Gaussian with zero mean and unit variance. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

- Non-linear transformation smooths out unusual distributions and is less influenced by outliers than scaling methods. It does, however, distort correlations and distances within and across features.

- Normalization is the process of scaling individual samples to have unit norm. This process can be useful in quantifying the similarity of any pair of samples.
- Encoding categorical features, often features are not given as continuous values but categorical. Such features can be efficiently coded as integers.
- Discretization provides a way to partition continuous features into discrete values. Can transform the dataset of continuous attributes to one with only nominal attributes.
- Generating polynomial features add complexity to the model by considering nonlinear features of the input data.

Feature Selection

Feature selection is also called variable selection or attribute selection in the data (such as columns in tabular data) that are most relevant to the modeling problem interested.

Having too many irrelevant features can decrease the accuracy of the models. Three benefits of performing feature selection before running into modeling are:

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

4.7 SUMMARY

It has already been repeated several times that getting good raw data is precious, specifically in the branch of social networks websites (recent scandals did not contribute smoothing this fact).

Any data scientist has to be capable of collecting data, scraping or simple requests to APIs, because it is not merely handed by organizations we want data from. The thing is, only with luck, a dataset or repository is found respecting all desired requirements. Usually there are imperfections and limitations, predominantly in projects which consider several data sources.

This happens due to websites' owner's strategies against data access grant. The bigger the organization, more attention to security is done. Consequently, to get desired data more effort and time a researcher has to expend. Having no experience in scraping or crawling, it was an harsh and, at the same time, enjoyable experience learning on this specific area.

On the other hand, simply scientist dependant, it comes data preparation. A task that can be tiresome, takes up a lot of time and has many variants and possible techniques for a single problem. It requires a lot of attention to detail. Despite this, it is crucial to a work's succcess and many argue that it takes almost 80% of their time in data preparation. Fortunately, Python's scientific environment eased this journey.

5

DATA ANALYSIS

5.1 INTRODUCTION

In the previous chapter, we approached and detailed how data acquisition and preparation tasks apply in this thesis's study. After dealing with unstandardized, unstructured or inconsistent data and combining data from different sources, with different formats, we assume having in our possession sources of data in a format, quality and structure suitable for further analysis.

Being so we are ready for the next assignment, Data analysis is the process of searching, inspecting, transforming and modeling data with the end goal of surfacing or coming up with useful information conclusions, and supporting decision-making. It has different facets and approaches, reaching a myriad of techniques under a variety of names, while being used in different business, science, and social science domains.

Figure 13: Data Analysis, Anonymous (a).

Data analysis is a proven way for organizations and enterprises to gain the information they need to make better decisions, serve their customers, and increase productivity and revenue. The benefits of data analysis are almost too numerous to count, but in this work it is intended to answer the hypothesis raised in chapter 1.

Some derivative questions are also aimed to be resolved, how does the diffusion of an event's influence evolve? In this diffusion, is there a correlation between different social media platforms? Is it possible to correlate estimated influence with other sources' patterns of data surrounding the event? Or even with real-world phenomenon, such as an impact on regional economy?

There are a variety of specific data analysis methods, some of which will be covered in the next sections, like exploratory data analysis, text analysis, and online social networks analysis.

5.2 EXPLORATORY DATA ANALYSIS

Understanding data before working with it isn't just a pretty good idea. It is a priority. In Mawer its definition is clarified: At a high level, EDA is the practice of using visual and quantitative methods to understand and summarize a dataset without making any assumptions about its contents.

This type of task that can be revisited several times, this lack of presumptions is handier and can be used to understand the data in the data preparation step, often called initial data analysis. It is also a crucial step to take before diving into machine learning or statistical modeling because it provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results. Descriptive statistics are included.

Chloe also states that EDA usually involves a combination of the following methods:

- Univariate visualization of and summary statistics for each field in the raw dataset, involves describing the distribution of a single variable;
- Bivariate visualization and summary statistics for assessing the relationship between each variable in the dataset and the target variable;
- Multivariate visualizations to understand interactions between different fields in the data;
- Dimensionality reduction to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data;
- Clustering of similar observations in the dataset into differentiated groupings, which by collapsing the data into a few small data points, patterns of behavior can be more easily identified.

The main reason for differentiating univariate and bivariate analysis is that bivariate analysis is not only simple descriptive analysis, but also it describes the relationship between two different variables. Quantitative measures of dependence include correlation, such as Pearson's correlation coefficient, a measure of the linear correlation between two variables X and Y, when both variables are continuous.

5.3 TEXT ANALYSIS

Text analysis applies statistical, linguistic, and structural techniques to extract and classify information from textual sources. Tasks can therefore range from text categorization, text clustering, or simply sentiment analysis.

Within the social chatter being generated every second, there are vast amounts of hugely valuable insights waiting to be extracted. With sentiment analysis, we can create insights about consumers' reactions to announcements, opinions on products or brands and even track opinion about events as they unfold. News and tourism listing's reviews are also of interest in searching insights into text attributes.

TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks, unfortunately, after testing, we realized it is seriously limited by Google translation API (free plan), not only in detecting languages, but also in translating them. Its great advantage was the capability to merge everything in one tool, but we had to seek another one.

Polyglot is similar to TexBlob, but in addition to not having rate limit, it supports language detection in 196 languages and translation is not needed. Polyglot has polarity lexicons for 136 languages, returning a range of values for each words' polarity consisting in three degrees: +1 for positive words, -1 for negatives and 0 for neutral. The score of a tweet text, for example, is the sum of all its words.

Word clouds are also thought to be helpful in visualizing what words are more used and aggregate more importance by its frequency.

5.4 TIME SERIES ANALYSIS AND FORECASTING

Time is one of the most important varibles in this work, it connects data from entirely different sources. A time series is a series of data points indexed in time order. It is a well-known fact that visualizing time series information is always better than viewing the same information in a tabular format.

After loading the data in pandas, we would like to convert this data into time series, that can be manipulated naturally and easily.

The first step, after loading the data and converting it to a time series object, is to plot the data and obtain simple descriptive measures, looking for trends and seasonal fluctuations. Observations taken on two or more variables, making possible to use the variation in one-time series to explain the variation in another series. This may lead to deeper understanding.

Comparing different trends becomes difficult when faced with a significant difference in the scale of at least one trend. Two techniques that are viable to rescale time series data are normalization and standardization. Like normalization, standardization can be useful, and even required in some machine learning algorithms. Standardization requires the ability to estimate the mean and standard deviation of observable values. These values, as well as the maximum and minimum for normalization, can be estimated from the data.

In order to evaluate if there is agreement between sources we have to compare them, a time series similarity between them can be performed. Correlation is affected by the existence of trends in the data, which is expected to encounter. With Dynamic Time Warping, we want to assess how similar two time series, which can be of different lengths and frequencies, are. In other words, we try to find the smallest difference between the two time series.

The algorithm implemented is described as follows:

```
def DTWDistance(s1, s2):
    DTW={}

    for i in range(len(s1)):
        DTW[(i, -1)] = float(’inf’)
    for i in range(len(s2)):
        DTW[(-1, i)] = float(’inf’)
    DTW[(-1, -1)] = 0

    for i in range(len(s1)):
        for j in range(len(s2)):
            dist= (s1[i]-s2[j])**2
            DTW[(i, j)] = dist + min(DTW[(i-1, j)],DTW[(i, j-1)], DTW[(i-1, j
                -1)])

    return math.sqrt(DTW[len(s1)-1, len(s2)-1])
```

The applications of this technique indeed go beyond speech recognition. Dynamic time warping can essentially be used to compare any data which can be represented as one-dimensional sequences. This includes video, graphics, financial data, and plenty of others.

Euclidean distance, a very common metric is disposed of consideration, it often produced pessimistic similarity measures when it encounters distortion in the time axis.

Forecasting these time series, which represent trends, we have to take in account several factors, characteristic of the data extracted. From the literature and searching for possible candidates, ARIMA (Autoregressive Integrated Moving average) and Holt's winter method are some of the Time series models that are very popular among the data scientists. Both are able to handle seasonality and trend if observed.

On results section, depending on the type of data, they will be subjected to further clarification, since there the case study is already detailed.

5.5 SOCIAL NETWORKS ANALYSIS

Social Network Analysis element is essential for extracting knowledge from social networks, it investigates social structures through the use of networks and graph theory. For network analysis to be applicable, theory from sociology or other social and behavioural sciences should give reasons to believe that the structure of ties is linked to behaviour, opinions, or social position of the members of the network.

Traditionally, from this type of analysis we are able to map and measure the relationships and interactions that are developed between people, groups and other network nodes. The network nodes can be the users and their profiles, while the connections can be the relations or, again, interactions between them.

Measuring influence is a difficult task, in large part because all social network data is partial and successful application of models depends on a number of external factors that are difficult to quantify. Earlier, when focusing on the strategy to follow, it was expected the possibility of incomplete data, but through the course of this work we realized that relations data among social networks were lacking. Therefore, instead of knowing a user's friends, there will be a spotlight in his interactions with other users in the same network.

From many definitions of social influence presented in section 2.2.5, and taking into account the previous limitations,the most elucidative is: a social phenomenon that individuals can undergo or exert, also called imitation, translating the fact that actions of a user can induce his connections to behave in a similar way Guille et al. (2013). Influence appears explicitly when someone "retweets" someone else, "likes" or "shares" a post for example.

- User A mentions another user, B: A influences B;
- User A replies to B's post: A influences B;
- User A retweets B's post: A influences B;
- User A likes B's post: A influences B.

This logic can be expanded an applied to influence estimation along the time in the social networks studied.

This analysis has a strong component related to visual results. Exploration of the data is done through displaying nodes and ties in various layouts, and attributing colors, size and other advanced properties to nodes.

Gephi, an open-source network analysis and visualization software package, is probably the most popular network visualization package out there. It's strength is that it is able to produce very high-quality visualizations. It can also handle relatively large graphs - the actual size will depend on the computer used (particularly RAM). It does have the ability to calculate a few of the more common metrics such as degree, centrality, etc, in a simple, fast and modular away.

Social networks can be analyzed in various ways depending on a user's needs. In this work, social networks analysis is categorized into three main broad applications, explored hereafter.

5.5.1 *Statistical Analysis*

Focus on analyzing the structural properties of a network by measuring values of actors and relations. With statistical analysis we can examine the connectivity behavior of nodes in the network and identify patterns.

This includes understanding a range metrics defined in Anonymous (d) such as:

- Centrality - refers to a group of metrics that aim to quantify the "importance" or "influence" (in a variety of senses) of a particular node (or group) within a network. Examples of common methods of measuring "centrality" include betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality, and degree centrality.
- Distance - the minimum number of ties required to connect two particular actors, as popularized by Stanley Milgram's small world experiment and the idea of 'six degrees of separation'.

Two more concepts should be introduced, bridge and structural holes. The first may be described as an individual whose weak ties fill a structural gap, providing the only link between two individuals or clusters. It also includes the shortest route when a longer one is unfeasible due to a high risk of message distortion or delivery failure. The second is the absence of ties between two parts of a network. Finding and exploiting a structural hole can give an entrepreneur a competitive advantage.

5.5.2 *Community Detection*

Community detection is an essential challenge in the area of social network analysis. This challenge is highly related with clustering problem and tries to identify parts of the network that have similarities on their linkage behavior.

Community structure in the context of networks, refers to the occurrence of groups of nodes in a network that is more densely connected internally than with the rest of the network. Among many different community detection approaches, there are two main ones: the graph structure of the network which is called the topology-based community detection approach, and the textual information of the network nodes under consideration which is named the topic-based community detection approach.

Two more definitions are essential from Anonymous (d):

- Clustering coefficient - a measure of the likelihood that two associates of a node are associates. A higher clustering coefficient indicates a greater tendency to associate with only a select group;
- Cohesion - the degree to which actors are connected directly to each other by cohesive bonds. Structural cohesion refers to the minimum number of members who, if removed from a group, would disconnect the group.

5.5.3 *Evolution in Social Networks*

Because of the dynamic nature of online social networks, where new members join or leave the social graph, or one aggregates more importance in a specific range of time than another, many changes take place regarding the structure of the network and the communities.

Earlier in this thesis, it was stated the intention to extract data from different time periods, at least before and after the event takes place, to compare them. In this area we can analyze which communities change most and identify their patterns. Additionally, we can provide information about the nodes that are involved in this dynamic behavior.

5.6 SUMMARY

Having already some understanding of the data, originating through the previous data pipeline step, data preparation, and ready to analyze and work data we may proceed to try to get valuable insights.

Once again Python's scientific environment was appropriated to the problem of data analysis. This task comprises several more specific ones, exploratory data analysis, text analysis, time series analysis and, the final one, social networks analysis. Other tools were tried and evaluated to fulfill these objectives, Polyglot to analyze, understand, and derive meaning from human language; Gephi in visualizing the global connectivity of social networks interactions and examining network traffic during major events, along with more traditional network analysis topics.

Efficiently following the procedure for extracting actionable knowledge from several social networks, web activity, news aggregators and tourism-related activity, we presented the different elements that our methodology has and the steps that each one follows in order to achieve its design goal.

It is easy to get excited and want to pursue more and more different analysis techniques. Trying to solve a problem already at hands by looking for discussions in the community, another interesting analysis comes to our attention, but we have to restrain ourselves and realize that the time for study is limited. Focus on the present task is crucial.

6

CASE STUDY

6.1 INTRODUCTION

Until now, all this work presented aimed at explaining the methodology proposed in order to carry out our initial aspirations and answer the defined hypothesis.

Is it feasible, through a social networks'data extraction and analysis methodology, to quantify and validate an event's influence, on a specific region?

At least one case study has to be selected for the methodology study and validation. Almost a year ago this was discussed, and it was concluded that for a proper study and a satisfactory sample of raw data, a major physical event had to be chosen. The range of options included only events that would be realized in Portugal, so that social influence analysis could be validated against web activity, news aggregators and tourism-related activity data.

Having already some in mind, the option fell on WRC Vodafone Rally de Portugal. It was not a random choice. After ten years in the Algarve and Baixo Alentejo, the long-standing dream had become true: Vodafone Rally de Portugal returned to the North of the Country with a resounding success and with an impressive level of competitiveness, as it happened in 2017, in what turned out to be a memorable edition of the race, with records that should prevail for a long time.

Justifying this choice comes a few of this event's distinct properties:

- Rally of Portugal, along with the Rally of Argentina, was the World Rally Championship's stage with the highest number of spectators following the race 'in loco';
- According to official estimates, last years' event reached almost one million spectators, more precisely 950,000, over the three days of the event;
- It has international coverage and attracts attention from the most distinct locales;
- There is an estimated rise in rally enthusiasts following this rally for several years;

- It is a historical rally, it already captivates rally supporters for 50 years, which means it has a solid adept foundation;
- It is a distributed event, divided through a region that comprises several cities;
- For several years, the event maintains similar dates, during May, for its realization;
- The Rally of Portugal will have 17 WRC cars, a record in this world championship;
- There is a claim that Northern region registered a rise on tourism profits around the event's duration, reflecting the "extremely positive" impact of the Rally of Portugal.
- There is also an assertion it has a direct economic return in the local economies. Restaurants, hotels and supermarkets will have been the business areas with the most benefits in the various municipalities where it passed.

As a justification for the increase of the audience throughout this time, it is considered that fans are enjoying the appeal of more powerful and spectacular WRC cars, as well as the most exciting pilots' title fight of the decade, with six riders already having won evidence on behalf of the four World Constructors.

Is is interesting to validate this rise, especially taking note of the hundreds of thousands of tourists and visitors, both national and foreign, who have traveled to the North Region and their primary motivation to attend Rally Portugal, generating an unequaled tourist flow and whose visits reveal high satisfaction rates.

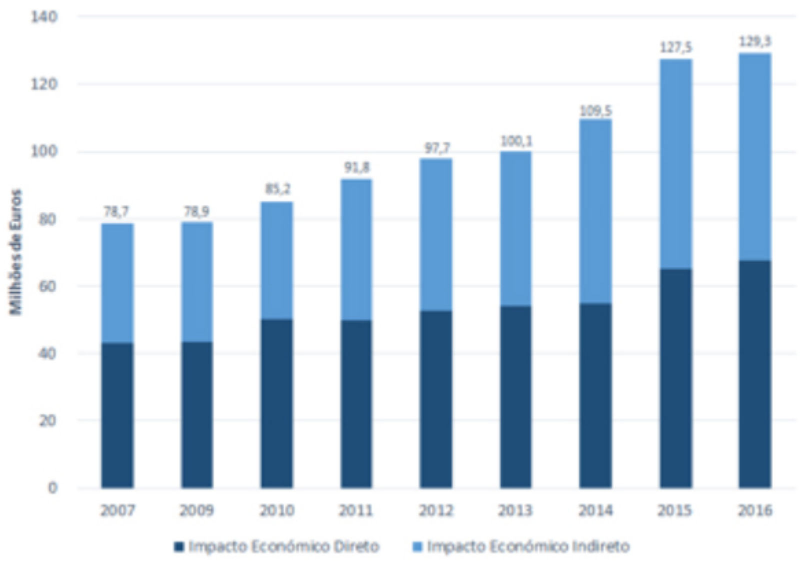

Figure 14: Obvious rise of WRC Rally de Portugal's impact over the years (light blue - indirect economic impact; dark blue - direct economic impact), de Portugal.

The president of Porto and Northern Portugal tourism, Melchior Moreira, in Lusa (a), reveals ″On a number of fronts, Porto and the North surpassed all expectations in May, with 25% more than in the same period of last year, 2016. And it is also worth pointing out a trend of increasing tourist revenue, which places us increasingly as a destination of quality″.

Another news outlet, Ambitur, explains more in detail that, for three consecutive years, this event has produced extremely positive results, whether in public capture, viewing in the media, or in creating income for the territory of the region. The numbers coming in from the 2017 edition show that there was an impact of 136 million euros in the economy of the territory, of which 71 million euros of direct impact (seven million more global impact than in the 2016 edition); almost one million fans, indicating that among the spectators, 39.1% made their first visit to the Region with the purpose of Rally de Portugal; 92.9% of Rally de Portugal fans intend to return to the Region, including in the winter (69.2% - decrease in seasonality), assuming that Porto and Northern Portugal excel by hospitality, landscape, gastronomy, heritage and culture.

Carlos Barbosa, president of the Automobile Club of Portugal (ACP), noted that in 2017, the Rally of Portugal originated ″60 million euros in food and overnight stays in three days″ and gave about ″30 million in IVA to the State″, Lusa (b).

6.2 EXPERIMENT SETUP

In this section we describe the methodology that we follow for this case study and the approach developed.

The first step, in properly analyzing this case study, is to infer this even's peak time interval. Therefore searching Google Trends data is a priority, allowing to recognize around which time intervals, data from other sources would have to be collected, maximizing the focus in the month which aggregates more interest.

On the last years WRC Rally de Portugal has remained on the calendar around the third and fourth week of May. As explained previously, this competition moved from south to north of Portugal in 2015, so it makes sense this work contemplates 2015, 2016, 2017 and 2018 editions.

From consulting the data, it was possible to check an intermittent and shy rise of interest from the beginning of the year until April. In May there is an exponential boost as expected, but as a surprise this enthusiasm drops quickly after the event. July and following months register almost minimum values, and it is registered an abrupt drop.

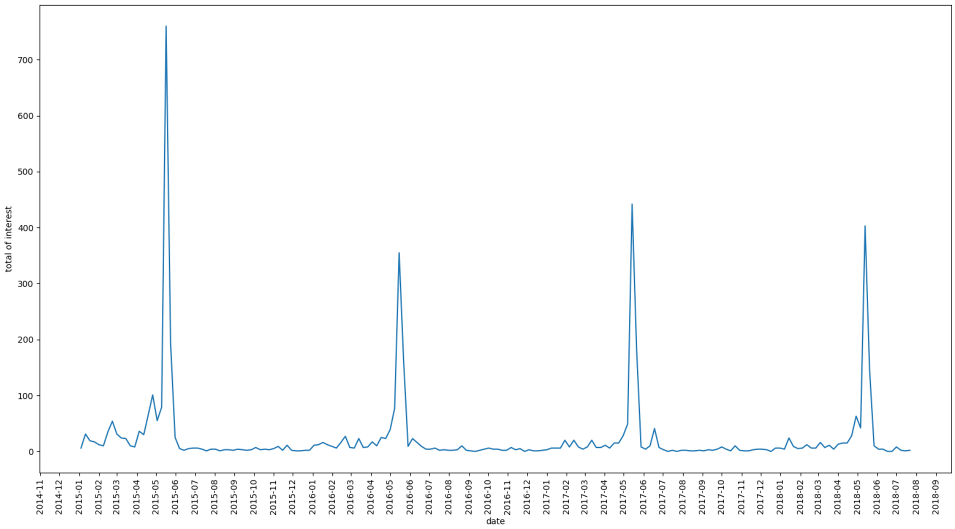

Figure 15: Search volume over time from Google Trends scraper.

With this, two data extractions steps were defined, one anticipating the rally, in the middle of May, and the other one around 25 of July. Additionally to an unexpected delay in developing the tools to scrape the desired data, it was not possible to automate all the extraction process. Because of these, there are no precise extraction dates, but all the scrapers, in which data range specification was implemented, respect a range. This includes almost 3 years and 7 months, from 01-01-2015 to 25-07-2018.

With the need to analyze several editions, comes difficulties related to uniformity. Despite the similarity of this event's characteristics throughout the years, not always the same stages take place in the same cities. For example, Braga was elected to be included in 2017, but not in the other years.

After reviewing the routes selected in all editions under analysis, 12 cities were contemplated: Paredes, Guimarães, Lousada, Viana do Castelo, Caminha, Ponte de Lima, Cabeceiras de Basto, Amarante, Porto, Vieira do Minho, Fafe, and Braga. The power stage in Fafe and street stages, Braga in 2017 and Porto, are the ones with the most media attention.

All of these are used as keywords searching for web activity, as well as in tourism-related activity. Wikipedia also regards Rally de Portugal's articles in the projects stated before. Google Trends also has additional targets: 'vodafone rally de portugal','vodafone rali de portugal','wrc portugal','rally de portugal' and 'rali de portugal', with and without edition's year complementing.

As for Youtube and News aggregators, the same set of targets previously listed are used, but without the edition's year and involved by quotation marks, trying to restrain odd results.

Social networks' scrapers are implemented to search by hashtag, or tag in the Flickr case, except YouTube. Several variations have to be convention, more precisely seven, #rallydeportugal, #ralideportugal, #wrcportugal, #rallyportugal, #raliportugal, #vodafonerallydeportugal and #vodafoneralideportugal.

With understanding of the case study, joined with a clear-cut extraction methodology and the analysis methods listed in chapter 5, the stars are aligned to reach valuable insights and trustworthy results.

6.3 RESULTS

Not diverting from the main goals is essential, and some rethinking on which data to use and how to present it was needed. It is tempting to dive into the big pile of data collected, run statistics on every dataset and try to find insights in the 4 main groups of data, social media, web activity, news aggregators and tourism related activity.

So that we may separate the problem into parts, clarifying it, two main analysis are granted in this section. We focus on this study's core, social networks, trying to study and estimate the social influence and reach derivative meaningful conclusions, and its validation against other sources.

Searching for patterns between 17 different data sources can be exhausting and confusing, so in order to simplify the analysis, a good strategy planning, data preparation and exploratory data analysis had to be performed.

Exploratory Data Analysis

The scrapers developed already had some degree of data preparation, but the data, at this point can be considered raw and divided by each source. The first step was integrating the datasets, some of them extract the data by search, resulting in many different CSV files. In almost all cases, the event's posterior extraction ranges all the data, from 2015 to mid 2018, but others, because of their topology, have different instances between the extraction's interval. That is the case of Youtube and news aggregators. To expand and enrich the load of data to work with, these suffered an append among the first and second retrievals.

All had to be subjected to data preparation, which ranges from treatment against duplication, by url or id's, some more parsing, type conversion, replacing, or not, NaN values and dropping instances and columns. Pandas handles all of these very well.

In overall, we ended up with the following:

- Twitter - 27569 tweets, 104574 favorites, 79457 retweets, 335 replies and 15616 mentions;
- Instagram - 17936 posts, 541096 favorites, 8917 replies and 13322 mentions;
- Flickr - 733 posts,218 favorites and 250 replies;
- Youtube - 1130 videos;
- News aggregators - 4653 news;
- Google trends - 186 entries for rally searches and 186 for the cities search;
- Wikipedia - 6696 entries for rally searches and 73882 for the cities search;
- Tripadvisor - 71364 reviews from 835 hotels, 189628 reviews from 2818 restaurants and 139873 reviews from 869 to-do activities;
- Booking - 265211 reviews from 1723 accommodations;
- Expedia - 21296 reviews from 797 hotels;
- Airbnb - 59378 reviews from 1420 accommodations.

Is is clear that Flickr and Youtube are the weak links in the social networks representation, the first because of data quantity and the former because of interaction lacking. Next we center on the online social networks.

To give an idea of the distribution of values and relations between relevant attributes, for each of the four social networks are now given some understandings of the data. Granting this kind of attention to the other sources would be overwhelming, therefore they will be approached later on.

Twitter has been rising in the Portuguese's usage over the years, being an excellent sample of the spread of influence through its users.

Users who twitted in the interval of time extracted, are predominantly from Portugal, Spain, Paraguay, Argentina, Germany and France (in order). It makes sense that this event attracts people from countries known to have enthusiastic rally fans, also that closer someone is to the event's location, there is more exposition and interest in it.

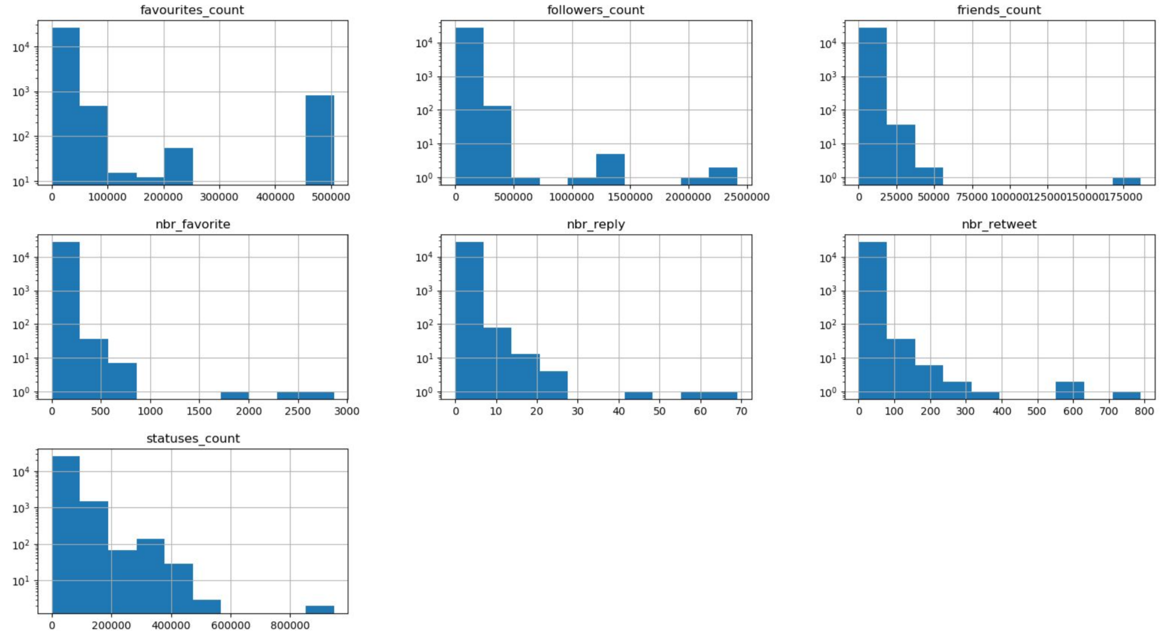

Figure 16: Distribution of numeric attributes in tweets.

Additionally, Argentina and France were in the WRC calendar right before Portugal, there might be a residual interest representing this fact.

To evaluate the common notion that having more followers or statuses implicates having, for example, more likes, it is usual to plot them against each other.

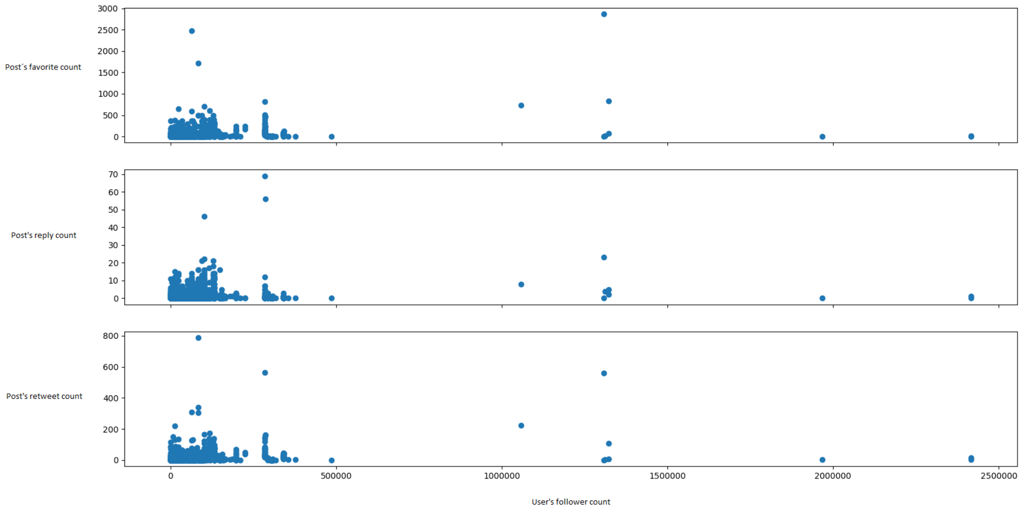

Figure 17: User followers count against tweets favorites,reply and retweet count (Twitter).

From the image above can be inferred a couple of facts. Having more followers does not ensure popularity, some of the most followed users have tweets with a low rate of success, which, in some cases, might be related to the tweet recent date. Only on the follower count plotted against the tweet favourite count can be seen a tweet that surpassed the first two clusters.

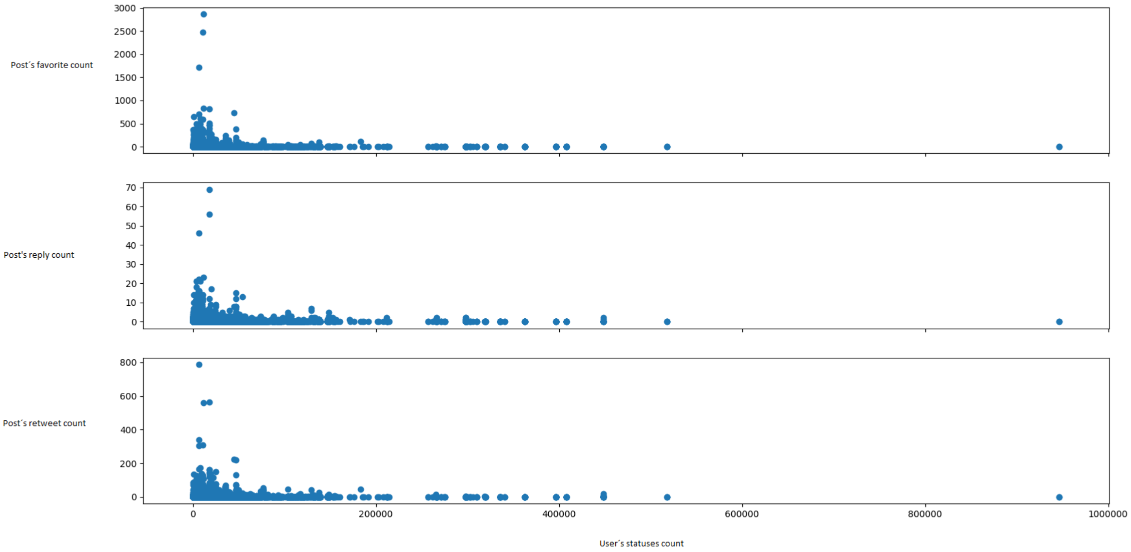

Figure 18: User statuses count against tweets favorites, reply and retweet count (Twitter).

All of a tweet's metrics that indicate influence are inversely related to a a high number of statuses. Being persistent does not confirm being influential.

Turning to Instagram, the social network with more weight and better represented among the ones studied, comes to light better distributed values.

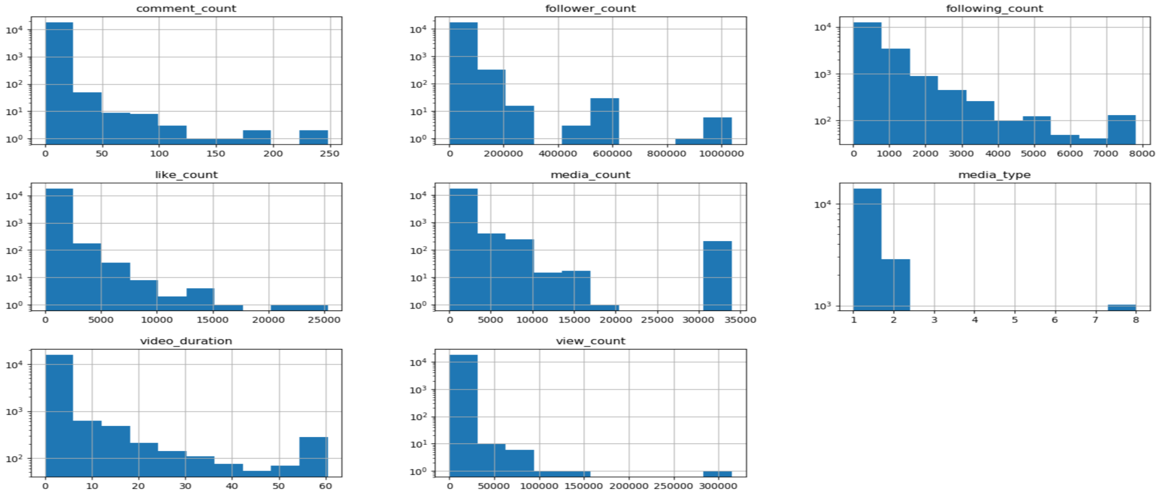

Figure 19: Distribution of numeric attributes in Instagram posts.

In the developed Instagram scraper, instead of where the user is from, the location of the post is revealead. Only 36% of them chose to include this information, but it is a good sample. Porto, Fafe, Exponor, Amarante and Vieira do Minho made the top 5 of locations with more posts.

The first three from the previous top also include the top locations by number of likes. Here Matosinhos and, a more general term, Portugal rise to the elite.

The relationship among an user followers count and its posts metrics, view, like and comment count is similar to the Twitter one. Excepting the number of comments, users with mid range follower count are, without a doubt, the ones with best ranked posts.

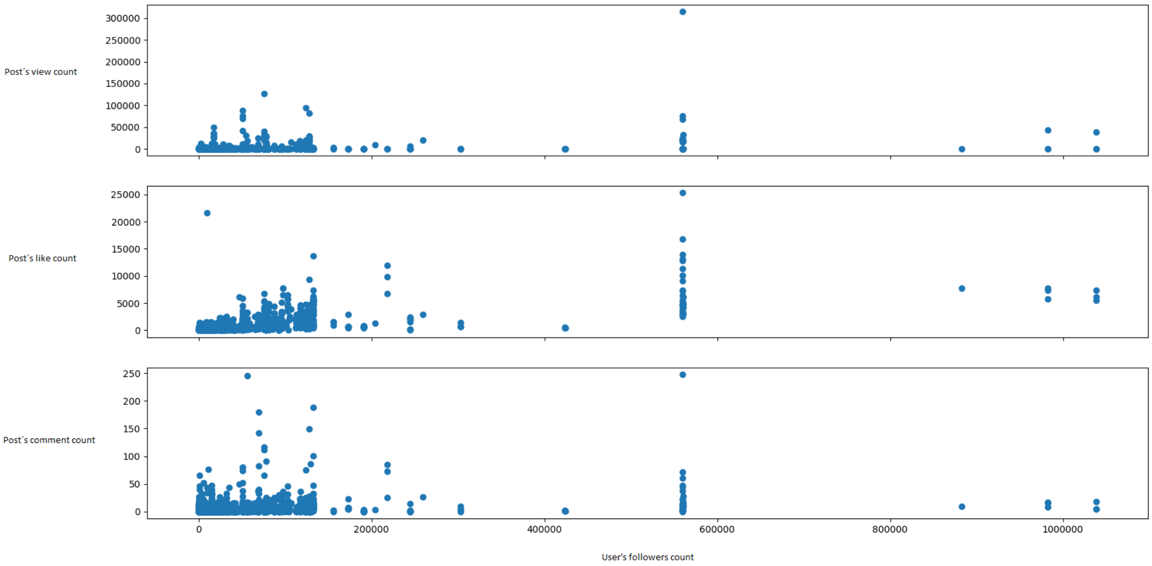

Figure 20: User followers count against post view, favorite and reply count (Instagram).

Very distinctive from the previous, Youtube data did not able us to generate as many intuitions, but an interesting plot follows right away. It was expected ranks well distributed, but the attention goes to the clearly separated parallel lines.

With most recent videos having more scattered upload dates, the dataset analyzed reveals that videos with several years have more defined dates for appearing in the search, middle of May and start of August. This indicates that the most popular videos are the ones uploaded right before and a couple months after the event, which looking at the data is meaningful, it includes preparation for the current rally, as well as reliving other editions and post event productions.

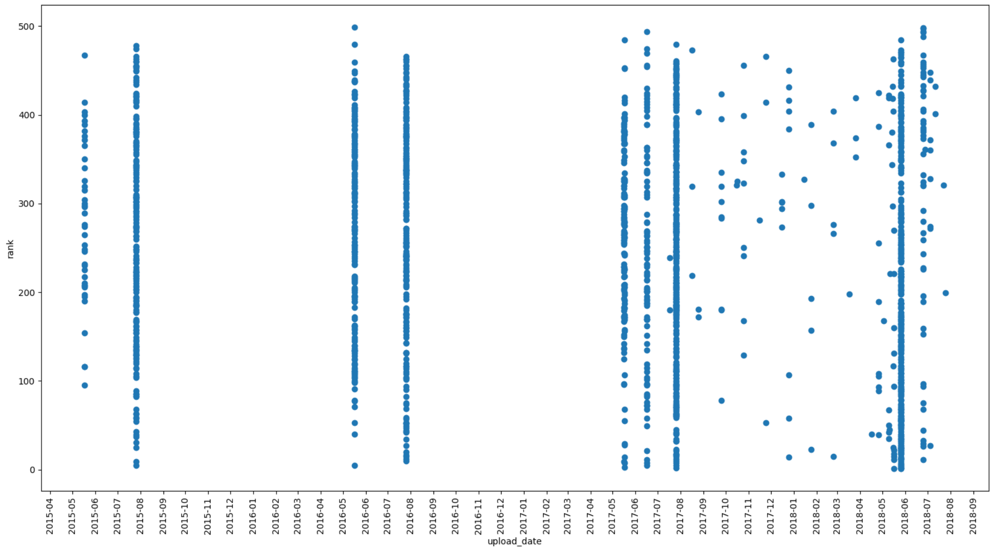

Figure 21: Video upload date versus rank on the searches (Youtube).

FIA World Rally Championship, the official WRC Championship account, WRCantabria, a motorsports Spanish account, and J-Records (lacks description), dominate in terms of sum of views.

Data extracted from FLickr was not very rich, it was envisioned a more substantial sample, but photos containing tags related to this event are controlled almost fully by a subset users. Since it is a social network focused on photography, its users range from amateurs to professional ones, sharing details on how the photo was taken.

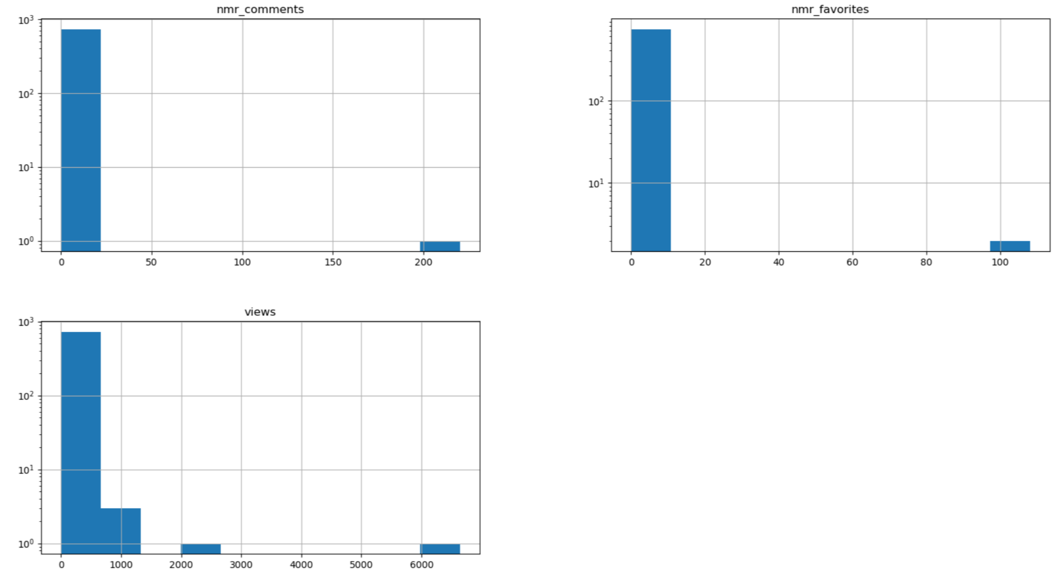

Figure 22: Distribution of values in Flickr.

The extracted data is uneven, composed by some highly popular photos and others which seem to be ignored. It is not an interactive network, only the attribute view count has a mean above 1, well above, 109. The user @_Rjc9666_ made the post with the most views, but in sum, @SérgioFotografia is the one which aggregates more.

Any extra value expected to arise, may come from the relation between an image upload date and its interactions from other users through time.

Text Analysis

This type of analysis could be implemented for all data sources that have a text attribute, but, as it was already been stated, focus on social networks has to be maintained.

Polyglot was the selected natural language pipeline, being therefore applied to text included in tweets, Instagram captions, Youtube and Flickr titles.

Gaining insights from text analysis is not a trivial task, what may sound natural to people, might be very difficult to computers. Drawing relations between words written is dependent on what is perceived as the context. So, in order to optimize the results, the text has to be clean and without mistakes, not usually the case in social networks text.

Even after lowering the letters, extracting some not recognized unicode characters, mentions (@'s), hashtags (#'s), multiple white spaces and urls, the text polarity evaluation can be erroneous. Two examples of some mistaken cases on cleaned text from tweets are:

- "world champion sebastien ogier crashes" - Neutral (0.0);
- "impresionante un año más el público en el tramo. se habla de 300.000 almas repartidas en 11 kilómetros. algo que sólo se puede ver en este deporte. ¡vivan los rallys!" - Negative (-0.5).

It is intuitive that the first has a negative connotation, but because of "world champion" it is classified as neutral. The Spanish one emphasizes and gives credit to this rally's public's unique adhesion, something only capable of being seen in that sport. Strangely Polyglot gives a negative value.

Apart from some exceptions, the overall picture was satisfactory, as expected text classification as neutral monopolizes the distribution. A comparison with only English text evaluations is also presented.

Instagram poses as the source with the most percentages of positive relatively to the negative ones, almost double, while Youtube data demonstrate more negative titles than positive, possibly because of titles aimed to attract more users.

	All languages			Only English		
	Negative (%)	Neutral (%)	Positive (%)	Negative (%)	Neutral (%)	Positive (%)
Instagram	11.1	67.9	21.0	5.5	72.7	21.8
Twitter	20.2	60.1	19.7	8.1	65.4	26.5
Flickr	2.2	96.3	1.5	2.4	96.0	1.6
Youtube	12.3	77.2	10.5	6.9	80.9	12.2

Table 12: Polarity classification on text from social networks text.

Filtering by language shows a decreasing percentage in negative polarity, spreading, mainly, to the neutral class. We suspect it is related to a better accuracy of text polarity evaluation. FLickr data is predominantly neutral.

For the wordclouds generated, the text was filtered against stop words, NLTK(Natural Language Toolkit) in Python has a list of stopwords stored in 16 different languages and emojis, specific ranges of unicodes.

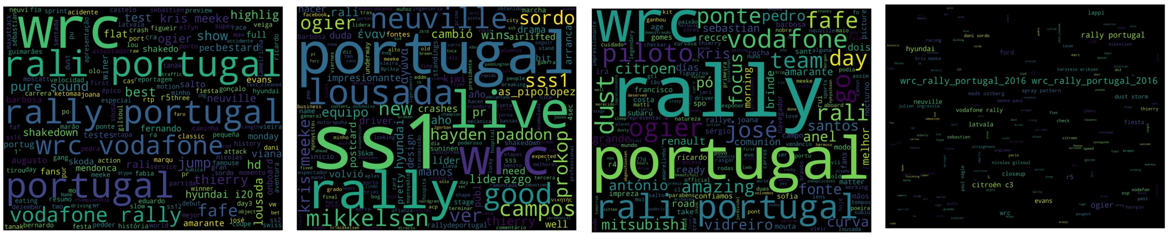

Figure 23: Wordcloud from Youtube, Twitter, Instagram and Flickr.

Apart from words predicted to have a higher frequency, such as the ones used in searches and the most notorious pilots, comes to our attention adjectives denoting satisfaction towards the event. Examples of these comprise ”good”, ”best”, ”amazing”.

Lousada, the first special stage (SS1), Amarante, and Fafe, where the rally ends, are well represented. In Fafe there is one of the most famous attractions of all the stages, the ”jump”. It almost appears to exist a pilgrimage to the sites surrounding it, therefore it makes sense it is referred many times by users in the web.

Social Influence Analysis and Forecasting

Following the logic we defined as what is considered influence in social networks interactions between users, we are able to track its evolution and dispersion through time.

Time data associated with these interactions is of the essence. Some extra time was devoted in the initial steps of developing the scrapers to ensure this data was on the datasets. Unfortunately it was not possible to all the social networks.

For Instagram, posts and comments, for Twitter, posts and replies, for Flickr, posts, likes and comments, and, finally, for Youtube's videos (includes number of views) date is associated. Mentions, @someone, were extracted from Instagram posts, tweets posts and replies, being this away possible to track them over time.

For each of these, it is now presented their time series. The values were summed and grouped by day. Comparing different ranges of values and from different sources, normalization or standardization is advised, so we applied the well known formula:

```
(Value - Value.mean()) /Value.std()
```

Standardized values are useful for tracking data that is otherwise incomparable. Using dataframes in Pandas facilitates this calculation, easily calculating the mean and standard deviation from a given value.

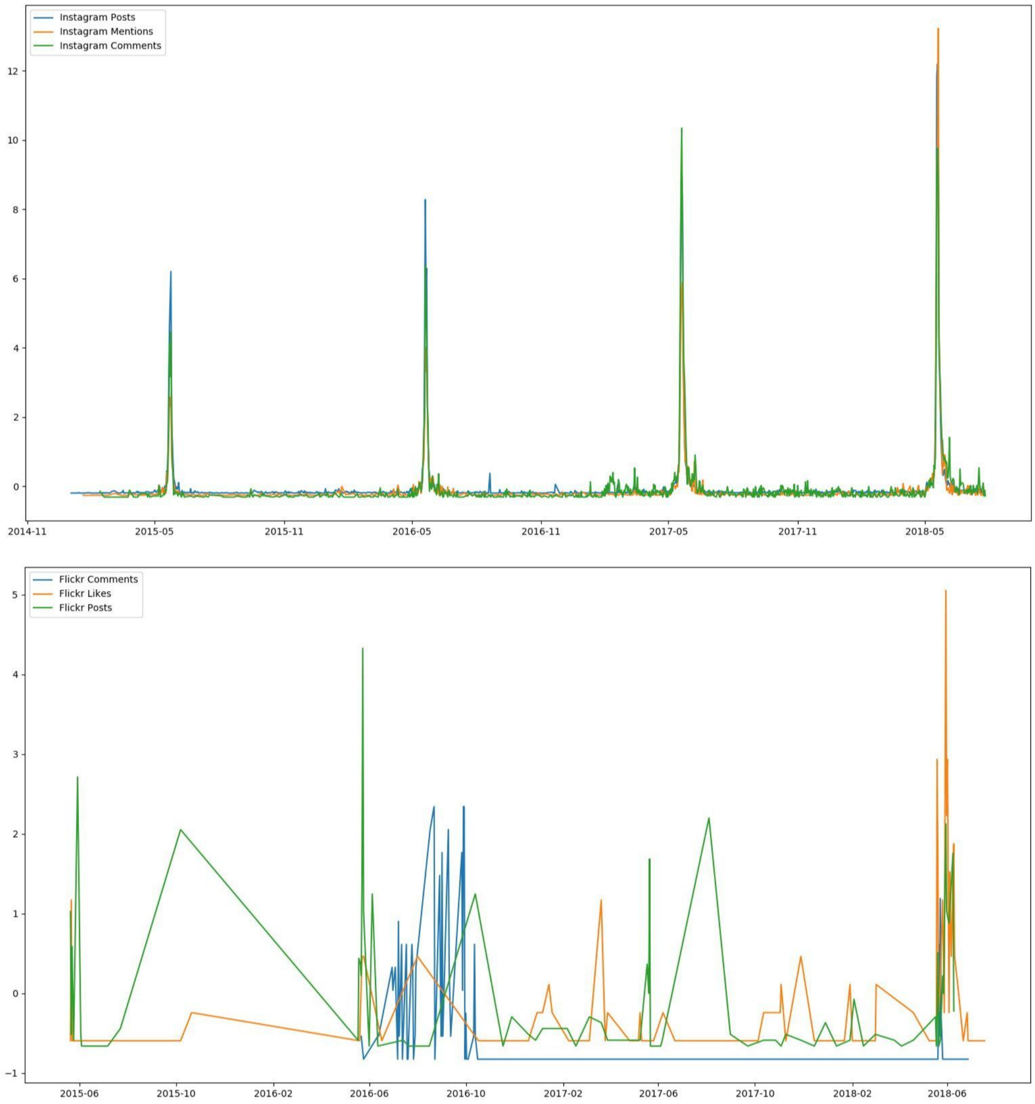

Figure 24: Trends from two of the social networks analyzed, Instagram and Flickr.

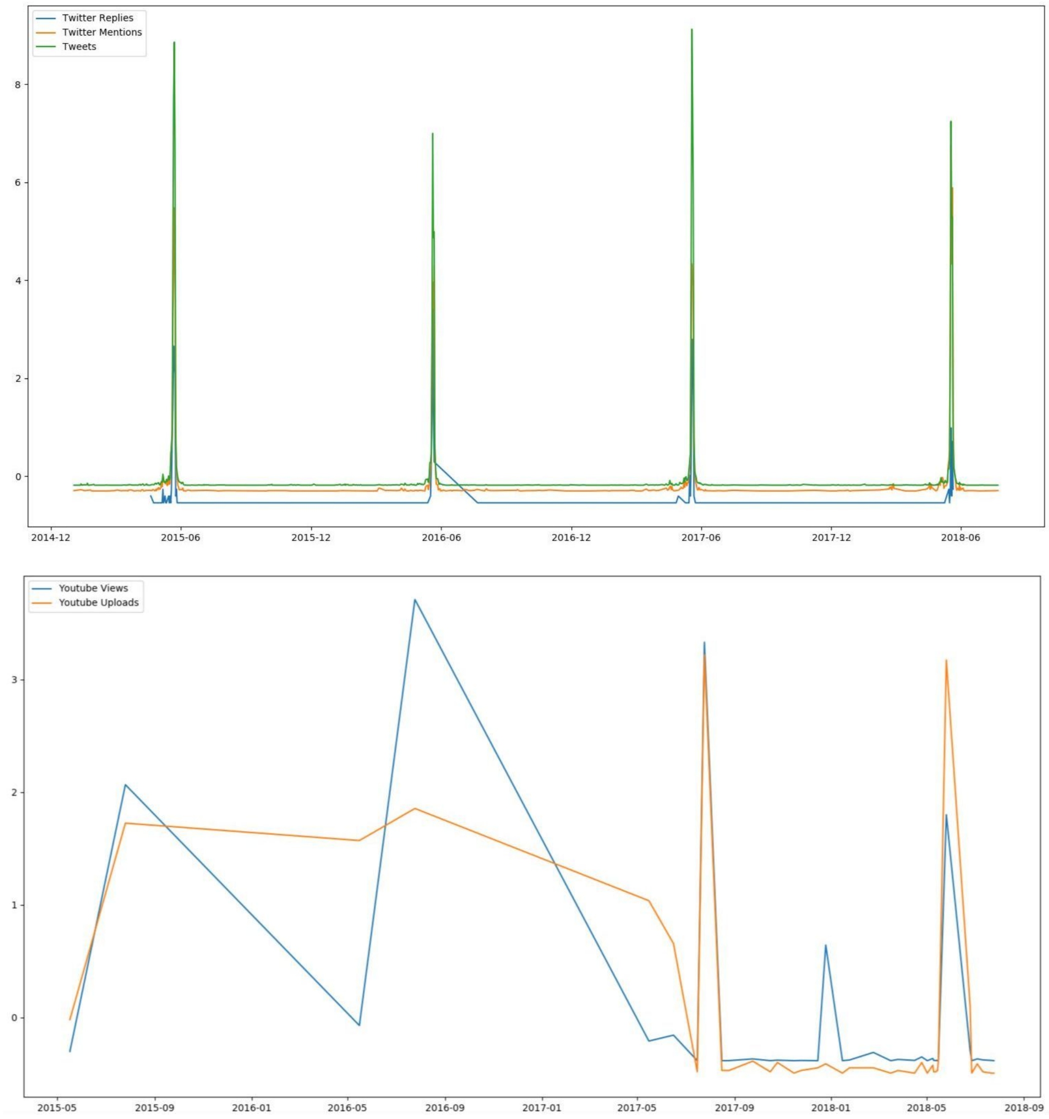

Figure 25: Trends from another two of the social networks analyzed, Twitter and Youtube.

Twitter and Instagram have very well defined peaks around the event (mid May to the end of the month), whereas Youtube and Flickr, as we discussed before, show the lack of data with a good distribution over the years. The wider peaks demonstrate that those lines are linking points from further than desired dates. In Flickr, the metrics related to interactions do not seem to have a order, nor correlation. In the Youtube case it may be observed much more views than videos uploaded, in past years, and the inverse in recent past.

Having access to the estimated percentage of utilization by users from these companies, a metric that represent the relationship can be formulated. Reminding these percentages: Instagram - 50.2%, Youtube - 45.9%, Twiter - 22.4% and Flickr - 3.9%. To enrich this opportunity of quantifying the influence through time, posts count will be also taken into consideration.

Building a metric dumping values into it is trivial, but tuning it and being able to access its quality, no. Joining the trends within each source gives a good idea of the overall influence it has, specially if the values are correlated with each other, which is the case of Instagram and Twitter. This way we built two metrics foreseeing that sources with lack of data are not well suited. Some experimentation was done, such as dropping or not Flickr and Youtube or just downgrading their percentage, ultimately reaching the following:

```
estimated_influence1=(youtube_sum*0.375).add(instagram_sum*0.41,fill_value=0).
    add(flickr_sum*0.032,fill_value=0).add(twitter_sum*0.183,fill_value=0)

estimated_influence2=(twitter_sum*0.33).add(instagram_sum*0.66,fill_value=0)
```

In both metrics, the percentages used are transformed from the simple rule of three, enclosing 100%.

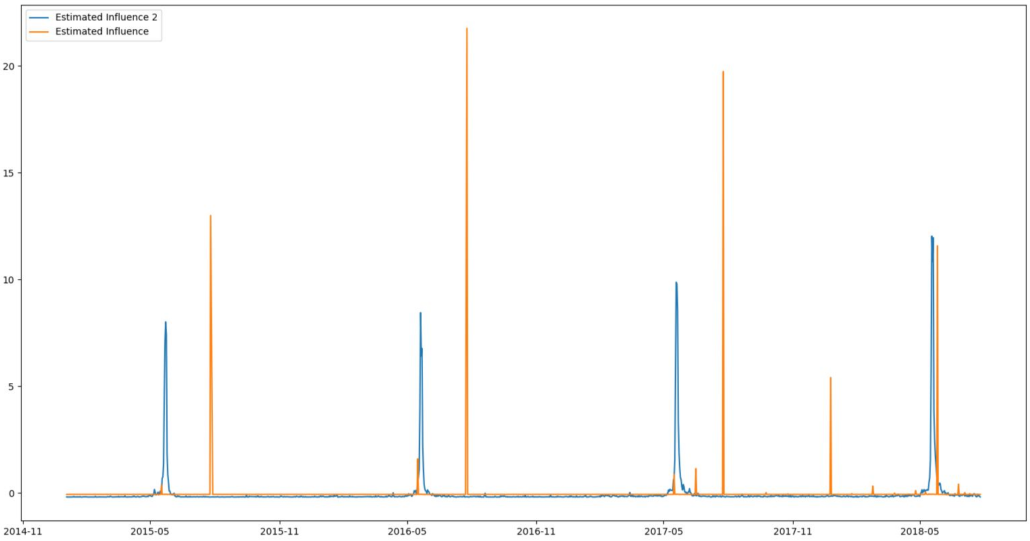

Figure 26: Comparison between metric of influence 1 and 2.

As expected dropping Flickr and Youtube as sources for the metric gives a better pattern. Giving the former 37% has a great impact on its quality. In 2018, one agrees with the other as a result of a Youtube's well fitted data for this period.

Having this in mind, estimated influence number two is the selected one to continue our investigation.

The step forward is validation, which can be performed comparing this estimate with trends from other sources. These range from news aggregators and web activity (Google

Trends and Wikipedia), to tourism related activity (Tripadvisor, Expedia, Booking and Airbnb).

As previously stated, Dynamic Time Warping algorithm helps in doing it. It allows many-to-one point comparisons,differentiating from the Euclidean point-to-point distance. Illustrating the situation always elucidates what is discussed and transports our conclusions to the image.

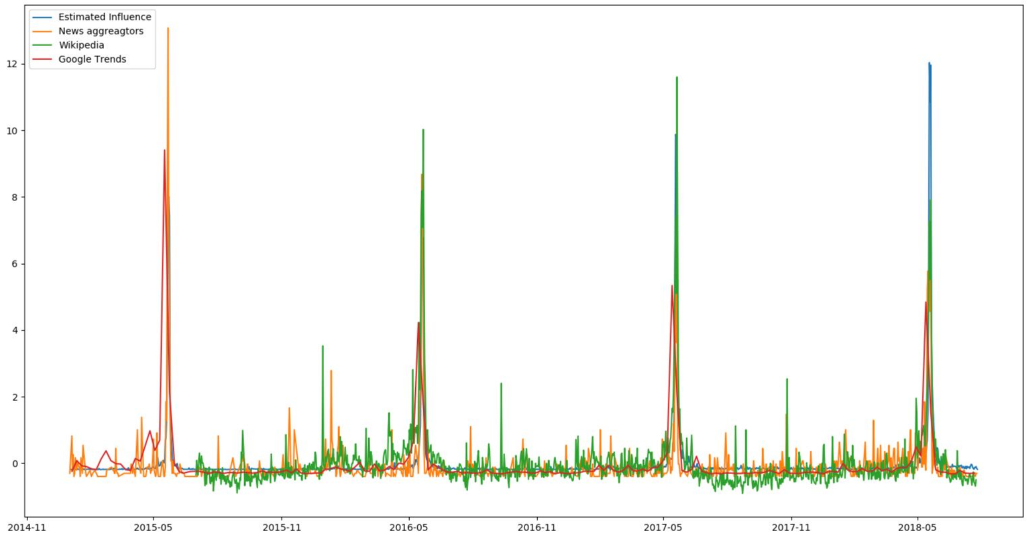

Figure 27: Comparison between estimated influence and distinct sources, news aggregators and web activity.

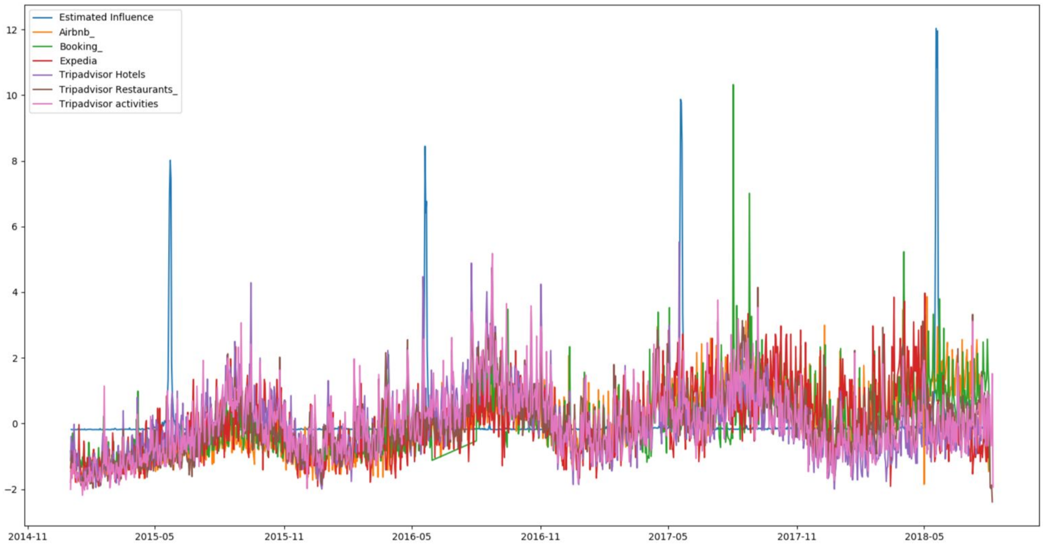

Figure 28: Comparison between estimated influence and tourism related activity.

The same technique of building the trends for other sources is followed. In parts, the number of news from publishers, with interest over time from Google Trends, Wikipedia page's number of views and quantity of reviews from the tourism world are collected and grouped by day. Several time series are created in the interval of dates we were incited to study.

With just looking at the images, and comparing them, we perceive a much better fit between estimated influence and web activity and news trends.

The most similar two time series that can be, result in a DTW distance of 0. In this experimentation, and due to its derivation from Instagram and Twitter, the first has a better similarity distance, of 6, to the influence resulting from the metric.

Estimated Influence (2015-2018)			
Instagram	6.124	Airbnb	32.040
Twitter	9.508	Tripadvisor Hotels	30.413
Youtube	24.797	Tripadvisor Restaurants	32.705
Flickr	21.137	Tripadvisor Activities	33.556
Google Trends	16.271	Booking	29.821
Wikipedia	15.283	Expedia	34.665
News aggregators	16.889		

Table 13: Dynamic Time Warping calculation among estimated influence and other pertaining trends.

The fit previously stated presents interesting outcomes, encompassing values as 15,16 and 17, which ables us to infer a close relation. Even better than with other social networks, Youtube and FLickr (did not enter the metric). Contrasting with this tight fit comes the relation with tourism related activity, in which we denote a different pattern, because of its seasonality peaking in the late summer, not May-June. Here we are considering the range of dates from 2015 to July of 2018.

Having a closer look at figure 27, it is possible to observe a finer connection among estimated influence and Airbnb, Booking, Expedia and Tripadvisor, just in the begginig of May 2018. Therefore we were lead to repeat the DTW calculations, filtering all the time series, so that we can, or not, infere similiarity in these sources just in 2018.

Estimated Influence (2018)			
Instagram	1.257	Airbnb	13.739
Twitter	3.622	Tripadvisor Hotels	13.899
Youtube	6.577	Tripadvisor Restaurants	14.529
Flickr	7.849	Tripadvisor Activities	14.261
Google Trends	5.395	Booking	13.112
Wikipedia	5.049	Expedia	13.432
News aggregators	6.098		

Table 14: Dynamic Time Warping calculation among estimated influence and other pertaining trends, focusing just in 2018.

Having just into account 2018 values, ables us to conclude an even better similarity distance in the group of sources present. All of them seem to straighten their relationship with the pass of the years, but this is more evident for the tourism related activity, as its peak in 2018 comes sooner, right in the beginning of May, anticipating the the Vodafone Rally de Portugal realization. The updated values in the anterior table confirms this conclusion.

With this successful validation we can try to infer future values, forecasting what will happen several years forward. Reaching this goal requires analyzing if there is trend or seasonality in the data (Figure 25 is the clearer one with estimated influence plotted through time).

A pattern that repeats itself every year is called seasonality an that figure shows exactly that, with these repetitions happening in May-June (seasonality of one year). Having a trend means that a time series mean values do not stay the equal along the time.

To elucidate these definitions and prove the existence of the referred patterns, one more plot is presented below.

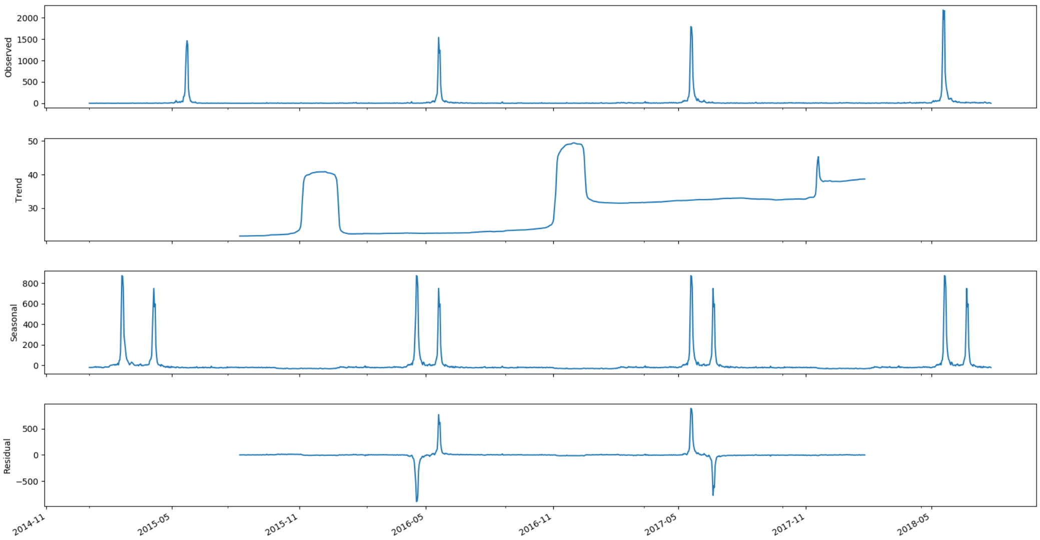

Figure 29: Decomposition of the time series into trend, seasonal and residual data with frequency parameter of 365 days.

Using Holt's winter method will be the best option among the rest of the models beacuse of the seasonality factor. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations, using Python and Statsmodels we can implement it, regarding also the existence of an upward trend.

Exponential smoothing methods assign exponentially decreasing weights for past observations. The more recent the observation is, the higher the weight would be assigned, which is intuitive.

We have to begin splitting the time series into train and test sets, leaving 20% of the data to test, starting in 2018. The data was also resampled by month.

```
model = ExponentialSmoothing(train, seasonal=’mul’,trend=’add’,
    seasonal_periods=12).fit()
pred = model.predict(start=test.index[0], end=test.index[-1])
future=model.forecast(48)
```

Exploring the combination of the additive method, where the seasonal variations are roughly constant through the series and the multiplicative method,in which seasonal variations are modified proportionally to the level of the series is essential. We came to the best Root mean squared error (RMSE) of 5.808 with the above combination. The most common metric used to measure accuracy for continuous variables was chosen.

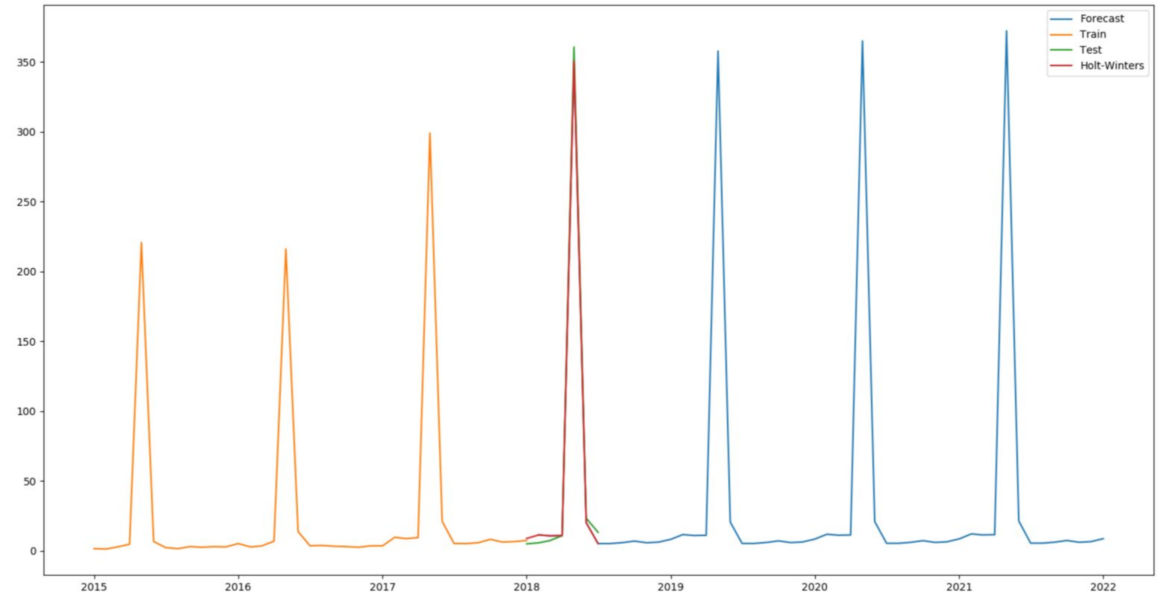

Figure 30: Train, test and forecast series plotted together.

The forecast was defined to be 48 steps, or months in this case, ahead. The model forecasted a similar pattern from the past, from May 2015 to one year later the peak is slightly smaller, increasing the following two years. The predicted version beahves also this away, but in a not so expressing manner.

The mean from each year was calculated in pandas and does not share the anterior described pattern, not decreasing in any year.

Estimated and Forecasted Influence		
Year	Count	Mean
2015	12	21.045
2016	12	22.317
2017	12	32.392
2018	12	37.157
2019	12	37.983
2020	12	38.752
2021	12	39.521

Table 15: Mean values for the original time series plus the forecasted one.

Concluding our analysis on social influence, we will now dive into an analysis more focused on the structure of the networks formed, combining all the possible interactions recorded.

Gephi will be our crutch, facilitating a more appealing and interactive away of showing the results from the analysis. With this goal all the data we have had to be reevaluated and rearranged. A scprit was developed to create two csv files, nodes and edges, avoiding being forced to write this information in the gexf format.

So that we are able to manipulate the time bar in Gephi, the data has to be associated with an interval of dates, the end of the this interval will be considered infinite. Having this in mind we chose to import into this tool, import spreadsheet function, the data previously approached in the time series analysis, from Instagram and Twitter. Tracking the dates is of the essence.

The nodes files are, at minimum, composed by id, label and date. This is the first date in which a user appears for the first time in the network. Some extra attributtes of that user are also mapped into each node if available.

The edges files have, at least, source, target, kind of interaction, id and date. As the influence goes from one person to another, the edges arechosen to be directed.

	Instagram	Twitter
Number of nodes	11510	4403
Number of edges	14919	6660

Table 16: Distribution of nodes and edges.

Customizing networks' nodes and edges is one of the main advantages of using Gephi and the same strategy was implemented for both.

Coloring the nodes by their attributed modularity ables us to perceive if the network is in fact a collection of smaller sub-communities or clusters which engage among them.

Rezising them by PageRank comes to light who is more influential. It is a measure of how likely a user is to reach a specific node from others in a network. PageRank was chosen because it was referred many times in the literature read, while preparing for this work.

Application of the ForceAtlas 2 layout algorithm, tunned for some extra gravity to mantain the nodes closer, alongside just a slight use of Expansion algorithm helps in making a sense of the network.

Firstly, we will look at Instagram. From the image we distinguash four main communities, being one of them the center of the network, in pink (37%). In the former the users which aggragate more influence are: @rallyportugal, @thierryneuville, @amikkelsenrally and @officialwrc, two official rally accounts and two pilots. In green, representing 13%, arises @fia_wtcr, but not being considered having an influential role. In blue, 6.67% , @ricardocosta.jr is highlighted, a Portuguese rally driver. Not so expressing, in black with 2.5%, appears @iloverally_, a more general rally page.

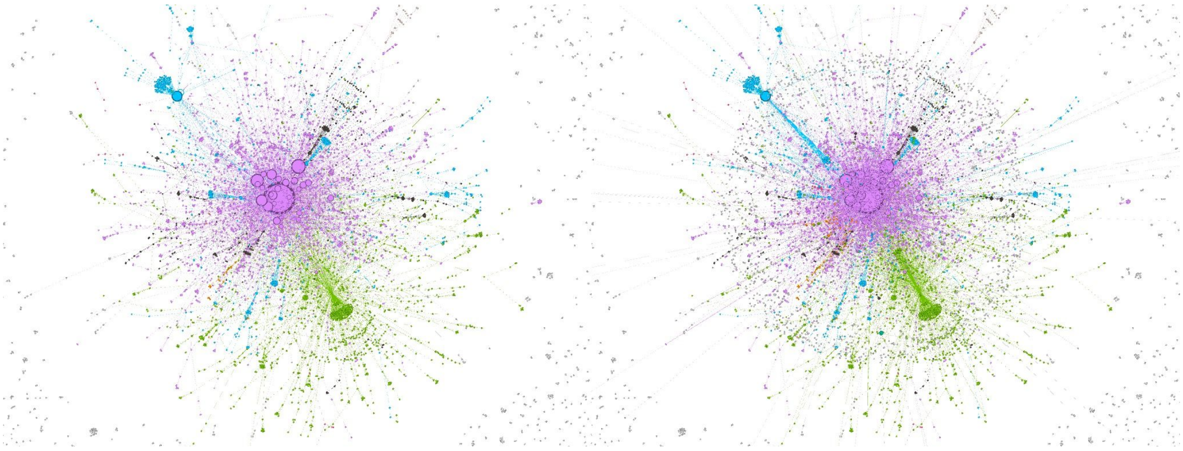

Figure 31: Evolution of the Instagram network, from before to after the event.

Using a dynamic model in Gephi ables its user to control the time and analyse the differences between several points in time.

In our case, relating to the event, a time step was defined in April and the other one at the end of the collected data. It is clear an intensification of interaction, paying special attention to the blue cluster. @ricardocosta.jr plays a very important role in propagating influence from the periphery to the the center.

Some other measures of how a user is perceived as influential is presented in the table below.

Betweenness centrality measures how often a node appears on shortest paths between nodes in the network and the, most common, degree, which quantifies the volume of interactions a user has. These offer distinct information, valuable to marketing strategies, depending of the specific goal is.

Top users	Instagram		
	Degree	Betweenness centrality	PageRank
1	@fia_wtcr	@colin_clark_rally	@rallyportugal
2	@rallyportugal	@mariomonteirophoto	@amikkelsenrally
3	@officialwrc	@fia_wtcr	@thierryneuville
4	@ricardocosta.jr	@henningsolberg1	officialwrc
5	@thierryneuville	@ msportld	@ricardocosta.jr
6	@hmsgofficial	@hsmgoofficial	@hmsgofficial
7	@wrcmotorsport	@officialwrc	@sebogier
8	@krismeeke	@amikkelsenrally	@otttanak
9	@msportld	@rallyportugal	@krismeeke
10	@danisordorallye	@skodamotorsport	@danisordorallye

Table 17: Top 10 users comparison between 3 different aproaches in Instagram.

In second comes Twitter, with a different distribution of the network and a not so obvious rise and spread of influence between sub-communities. It is a more centered network, the difference between cluster's size is here quite evident.

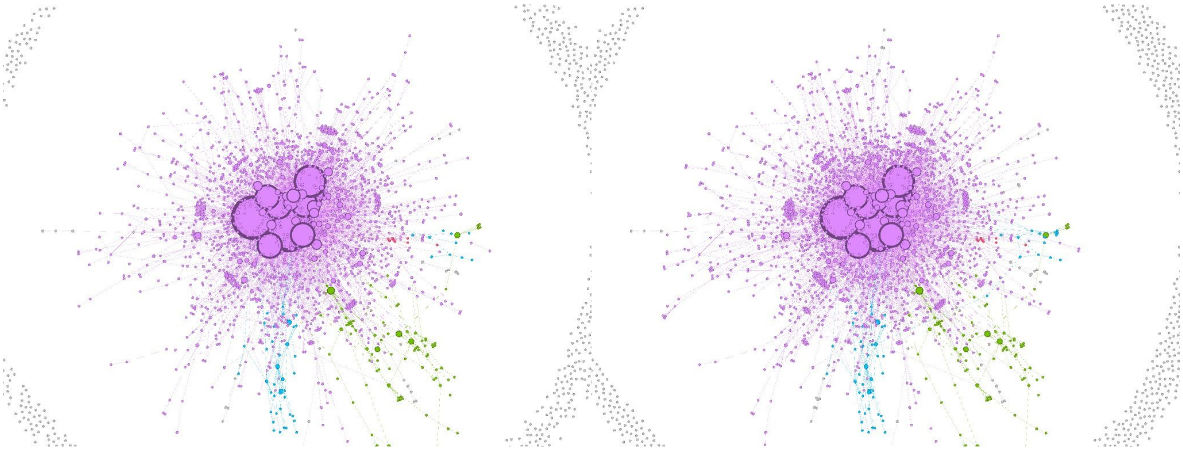

Figure 32: Evolution of the Twitter network from before to after the event.

In pink, 48,6% of the nodes, is main the cluster, which is the center of the network. Here we identified, similarly to Instagram, official Wrc and brands accounts, in addition to the most famous pilots who participate in the world championship. Only two more communities are possible of identification, the green one with 2.2% and the blue cluster, 1.1%, being these the minor ones from the bunch.

The same influence metrics are now applied to the case of Twitter.

Top users	Twitter		
	Degree	Betweenness centrality	PageRank
1	@OficialWRC	@MSportLtd	@rallyportugal
2	@rallydeportugal	@vwrallytheworld	@OficialWRC
3	@krismeeke	@WRCrtp	@krismeeke
4	@SebOgier	@Motorseries	@SebOgier
5	@JariMattiWRC	@MikkaAnttila	@JariMattiWRC
6	@DaniSordo	@frank_wrc	@DaniSordo
7	@rallyparadise	@OpensTightens	@HyundaiWRC
8	@thierryneuville	@OficialWRC	@thierryneuville
9	@OpensTightens	@JariMattiWRC	@CitroenRacing
10	@CitroenRacing	@TorsteinEriksen	@MSportLtd

Table 18: Top 10 users comparison between 3 different aproaches in Twitter.

6.4 DISCUSSION

In this section we try to relate the beginning of this thesis to its end, and what best way to do it than by answering the hypothesis and derivated raised in the initial chapters.

Is it feasible, through a social networks'data extraction and analysis methodology, to quantify and validate an event's influence, on a specific region?

From the scrapers to the final insights in the last section, we describe a methodology that has proven to, by applying it to the study case of the Vodafone Rally de Portugal, being able of measuring with a satisfactory rate of success social influence in social networks and also validating it against sources from different areas. This implies impact on the event's region.

Is it possible to correlate estimated influence with other sources' patterns of data surrounding the event? It was observed a tight fit (2015-2018) between estimated influence and Google Trends, News aggregators and Wikipedia data, with the help of the Dynamic Time Warping technique. As for tourism websites activity this kind of similarity was just found filtering the time series, including only 2018 data. In overall all the sources seem to improve their affinity with our estimation.

Or even with real world phenomenons, such as an impact on regional economy? It was possible to relate estimated influence with data from tourism related activity, the best indicator of direct influence in the locale where the event takes place, but just in 2018, which ables us to correlate this variation with some statements from interested parties. More precisely, with the affirmations concerning the Rally's influence on the rise of the economic impact on the North region of Portugal. In addition, it was observed a good relation with other indicators of influence in the web. Web activity can give an indication of which topics interest people at each moment in time. So, searching for information on tourism destinations or page views of Wikipedia articles related to destinations can do much to help predict tourism flows.

How does the diffusion of an event's influence evolve? As shown before with images captured in Gephi, the spread of influence can be traced between different sub communities and individuals, having regard for the time component. In the more structural analysis on Instagram, this is evident. There is also an interesting insight on how it evolves over time. The mean from the influence estimation from 2015 to 2017 has a very similar trend to the data on the economic impact in figure 14, both show an upward trend, accelerating and slowing down in the same periods.

In this diffusion, is there a correlation between different social media platforms? A clear correlation is seen among Instagram and Twitter over all studied years. Youtube's and FLickr's data also show a close relation to these, but focusing in instances from recent past.

6.5 SUMMARY

This chapter aggregates the more technical and insightful part of the work, application of the suggested methodology to an event with physical connotation.

We covered an in depth description of the case study, as well as all the stages that comprise the data analysis approach. From experiment setup to results and discussion on exploratory data analysis, text analysis and the most significant part, social influence analysis and forecasting.

In the mid range of this works steps, many doubts raised beacuse of lacking of data, in quantity and quality for some sources. For some it was the opposite, questioning what to do with all this data. Social networks are applying more and more restrictions, getting specially closed. In the beginning there was the mistaken dream of getting the data in an ideally manner, which is rarely the case.

All went well and the questions formulated in the beginning were answered in a confident away.

7

CONCLUSION

This chapter aims to present a summary of the work carried out in this project, as well as some conclusions and presentation of the scientific contribution. There are also some aspects in which the project can be improved and some ideas for future work.

7.1 WORK SUMMARY

The demand by business organizations for processes that help them make accurate and rapid decision-making, based on data extracted from real-world sources, has never been greater. Also rising, is the growth of data available for study.

Generation of useful knowledge, associated to a physical event, through the analysis of information extracted from web content is our main purpose. This analysis is performed after the extraction of data from multiple online platforms, of varied typologies and content.

Of the set of diverse utilities from this data extraction, we must highlight the prediction of inflows and other phenomena for subsequent years and the study of the role that influence holds in specific highlight events.

Throughout this thesis we raise several questions, which, further from the beginning, are all answered. The main hypothesis formulated is the following: is it feasible, through a social networks' data extraction and analysis methodology, to quantify and validate an event's influence, on a specific region? To reach meaningful conclusions we had to fulfill four basic objectives.

Implementation of all necessary infrastructure for information extraction and consequent validation of information capture methods complement each other and are the first two ones.

Starting with data extraction, it required a lot of work, due to strategies against data access grant and the fact that all the data extraction, for every source considered, had to be done uniformly and regarding the time factor. For each of this sources, a scraper was developed when official ways of acquiring data (APIs) were not available or were unsatisfactory. All the data was saved in a highly compatible format to posterior treatment and analysis, CSV.

In order to validate this data and ensure it was authentic, manual tests were done by comparing the results of this step with what websites showed to an average user, using a browser. These tests were performed before the initial phase of analysis, and so, here we worked with samples, it would be unreal to check all the instances from the final extractions.

Data preparation came next, it required a lot of attention to detail and patience to explore all the data in almost a raw form. We end up with data ready to be analyzed and start to infer some conclusions.

Specification of a computational system for the analysis of the objects of study of this work is the third objective. This analysis comprises exploratory analysis, text analysis, time series analysis and online social network analysis. For every type of analysis a Python script was developed and with the combination of them all we have a system, or methodology, to, mainly, infer the social influence associated with an event. The metric proposed, and obviously, its study are given.

Having a studied and well thought strategy, prior to analysis, gave us the tools to conclude that, indeed, we may quantify and validate the influence associated to a major physical event. We can confirm this with decent confidence, showing all the processes applied for influence estimation and consequent validation.

Similarity was found between this and trends from other extracted sources, giving a certain richness to our conclusions. The case study that was chosen, Vodafone Rally de Portugal, ended up to be perfect. We can say that our work, based in the results, was validated against studies, and affirmations, from referred interested parties to the event. The last objective, validation of the computational system's results, on the grounds of these insights, is also accomplished.

All the main objectives, from section 1.2, were, therefore, fulfilled and were crucial in ending with a proper work.

7.2 LIMITATIONS AND FUTURE WORK

Although the system developed has satisfied our raised objectives, there is space for improvements.

A set of challenges were raised while developing the system, we were forced to adopt a different strategy just in this thesis' initial phase. More concretely to data from social networks, that, as we are well aware, pay now more attention do data privacy and security problems.

Facebook is the social network with greater expression in Portugal and many other countries. Not including it in our work, not really by choice, was a step back and can be considered a limitation.

A more concise, planned and automated data extraction methodology would also make a great impact in his work. Working with data from 17 distinct sources can be overwhelming and this would certainly make the difference.

The development of a Graphical User Interface, which would provide a researcher, with no programming experience, the ability to use its functionalities to other case studies is, definitely, also an interesting step for the future.

More than the two last points, writing one paper in conferences indexed in the ISI Web of Knowledge has priority and will, for sure, be completed in the near future.

Concluding, and having in mind all the data extracted and what was achieved until now, there is a certain will to continue this work and improve it. Much more can be done with all this data and more time. This is a area with some tradition but that continues to amaze researches and attract a lot of attention. Our study may have the capability to add value to other areas, such as ambient assist living, Costa et al. (2007) and Costa et al. (2014), behaviour analysis, Carneiro et al. (2015a) and Rodrigues M. (2005), or online conflict resolution, Carneiro et al. (2014).

BIBLIOGRAPHY

Divyakant Agrawal, Elisa Bertino, Susan Davidson, Michael Franklin, Alon Halevy, H. V. Jagadish, Sam Madden, Yannis Papakonstantinou, Kenneth Ross, Cyrus Shahabi, and Shiv Vaithyanathan. Challenges and opportunities with big data, 2016. URL http://valsoftservices.com/big-data-implementation/.

Ambitur. Rali de portugal volta ao porto e norte. URL https://www.ambitur.pt/rali-de-portugal-volta-ao-porto-e-norte/.

Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM, 2008.

Anonymous. Data analysis, a. URL https://www.springboard.com/learning-paths/data-analysis/. Accessed: 2018-10-10.

Anonymous. What is the difference between data science, data analysis, big data, data analytics, data mining and machine learning?, b. URL https://onthe.io/learn/en/category/analytic/What-is-the-difference-between-Data-Science,-Data-Analysis,-Big-Data,-Data-Analytics,--Data-Mining-and-Machine-Learning%3F.

Anonymous. Web scraping vs web crawling, c. URL http://prowebscraping.com/web-scraping-vs-web-crawling. Accessed: 2018-10-10.

Anonymous. Social network analysis, d. URL https://en.wikipedia.org/wiki/Social_network_analysis. Accessed: 2018-10-10.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):22, 2011.

Florian Brachten, Stefan Stieglitz, Lennart Hofeditz, Katharina Kloppenborg, and Annette Reimann. Strategies and influence of social bots in a 2017 german state election-a case study on twitter. *arXiv preprint arXiv:1710.07562*, 2017.

Ellsworth Campbella and Marcel Salathé. Complex social contagion makes networks more vulnerable to disease outbreaks. 2013. Scientific Reports. 2013;3:1905. doi:10.1038/srep01905.

Davide Carneiro, Paulo Novais, Francisco Andrade, John Zeleznikow, and José Neves. Online dispute resolution: an artificial intelligence perspective. *Artificial Intelligence Review*, 41(2):211–240, Feb 2014. ISSN 1573-7462. doi: 10.1007/s10462-011-9305-z. URL https://doi.org/10.1007/s10462-011-9305-z.

Davide Carneiro, Paulo Novais, José Miguel Pêgo, Nuno Sousa, and José Neves. Using mouse dynamics to assess stress during online exams. In Enrique Onieva, Igor Santos, Eneko Osaba, Héctor Quintián, and Emilio Corchado, editors, *Hybrid Artificial Intelligent Systems*, pages 345–356, Cham, 2015a. Springer International Publishing. ISBN 978-3-319-19644-2.

Davide Rua Carneiro, Paulo Novais, José M. Pêgo, Nuno Sousa, and José Neves. Using mouse dynamics to assess stress during online exams. 2015b. URL http://hdl.handle.net/1822/40893. Lecture notes in computer science series", ISSN 0302-9743, vol. 9121.

Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown, 2009.

Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

Angelo Costa, Paulo Novais, and Ricardo Simoes. A caregiver support platform within the scope of an ambient assisted living ecosystem. *Sensors*, 14(3):5654–5676, 2014. ISSN 1424-8220. doi: 10.3390/s140305654. URL http://www.mdpi.com/1424-8220/14/3/5654.

Ricardo Costa, Paulo Novais, José Machado, Carlos Alberto, and José Neves. Inter-organization cooperation for care of the elderly. In Weijun Wang, Yanhui Li, Zhao Duan, Li Yan, Hongxiu Li, and Xiaoxi Yang, editors, *Integration and Innovation Orient to E-Society Volume 2*, pages 200–208, Boston, MA, 2007. Springer US. ISBN 978-0-387-75494-9.

Rally de Portugal. Mundial de ralis em portugal vale mais de 898 milhÕes. URL http://www.rallydeportugal.pt/content.aspx?menuid=2&eid=2399.

Christophe Demunter. Tourism statistics: Early adopters of big data? 2017. EUROSTAT Commission.

Flickr. Documentação api. URL https://www.flickr.com/services/api/. Accessed: 2018-10-10.

GeneralMills. pytrends. URL https://github.com/GeneralMills/pytrends. Accessed: 2018-10-10.

Fabien Girardin, Francesco Calabrese, Filippo Dal Fiorre, Assaf Biderman, Carlo Ratti, and Josep Blat. Uncovering the presence and movements of tourists from user-generated content. In *Intn'l Forum on Tourism Statistics*, 2008.

Marco Gomes, Javier Alfonso-Cendón, Pilar Marqués-Sánchez, Davide Carneiro, and Paulo Novais. Improving conflict support environments with information regarding social relationships. In Ana L.C. Bazzan and Karim Pichara, editors, *Advances in Artificial Intelligence – IBERAMIA 2014*, pages 779–790, Cham, 2014. Springer International Publishing. ISBN 978-3-319-12027-0.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

Leonard Heiler. Difference of data science, machine learning and data mining, 2017. URL https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining.

HermanFassett. youtube-scrape. URL https://github.com/HermanFassett/youtube-scrape. Accessed: 2018-10-10.

Junming Huang, Xue-Qi Cheng, Hua-Wei Shen, Tao Zhou, and Xiaolong Jin. Exploring social influence via posterior effect of word-of-mouth recommendations. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 573–582. ACM, 2012.

Kunal Jain. Machine learning basics for a newbie, 2015. URL https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/.

jonbakerfish. Tweetscraper. URL https://github.com/jonbakerfish/TweetScraper. Accessed: 2018-10-10.

Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010.

Lusa. Turismo do porto e norte congratula-se com crescimento registado em maio, a. URL https://www.dn.pt/lusa/interior/turismo-do-porto-e-norte-congratula-se-com-crescimento-registado-em-maio-8645053.html.

Lusa. Carlos barbosa destaca retorno recorde do rally, b. URL http://www.rallydeportugal.pt/content.aspx?menuid=2&eid=3630.

Marktest Consulting. Os portugueses e as redes sociais 2017. Folheto, 2017.

Chloe Mawer. The value of exploratory data analysis. URL https://www.kdnuggets.com/2017/04/value-exploratory-data-analysis.html. Accessed: 2018-10-10.

Nuno Miguel Pereira Moniz. *Prediction and Ranking of Highly Popular Web Content*. PhD thesis, Faculdade de Ciências da Universidade do Porto, 2017.

Santos M. Rodrigues M., Novais P. Future challenges in intelligent tutoring systems – a famework, recent research developments in learning technologies. 2005. Proceedings of the 3rd International Conference on multimedia and Information Communication Technologies in Education (m-ICTE2005), A. Méndez Villas, B. Gonzalez Pereira, J. Mesa González, J.A. Mesa González (Eds), Publishers Formatex, ISBN 609-5994-5, pp 929-934.

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

SamHO666. Expedia-crawler. URL https://github.com/SamHO666/Expedia-Crawler. Accessed: 2018-10-10.

Scrapy. Architecture overview. URL https://doc.scrapy.org/en/latest/topics/architecture.html.

J Danielle Sharpe, Richard S Hopkins, Robert L Cook, and Catherine W Striley. Evaluating google, twitter, and wikipedia as tools for influenza surveillance using bayesian change point analysis: A comparative analysis. 2016. Published online 2016 Oct 20. doi:10.2196/publichealth.5901.

simonseo. instagram-hashtag-crawler. URL https://github.com/simonseo/instagram-hashtag-crawler. Accessed: 2018-10-10.

stevesie. Unofficial airbnb api. URL https://stevesie.com/apps/airbnb-api. Accessed: 2018-10-10.

Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248, 2013.

Wikimedia Services team. Wikimedia rest api. URL `https://wikimedia.org/api/rest_v1/#!/Pageviews_data/get_metrics_pageviews_per_article_projectaccess_agent_article_granularity_start_end`. Accessed: 2018-10-10.

Vinícius Damaceno. Conhecendo a linguagem python – parte 1, 2016. URL `http://viladosilicio.com.br/conhecendo-a-linguagem-python-parte-1/`.

Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, and Ke Xu. Information propagation in online social networks: a tie-strength perspective. *Knowledge and Information Systems*, 32 (3):589–608, 2012.