IN SUPREMÆ DIGNITATIS

1343

UNIVERSITÀ DI PISA

Department of Philology, Literature and Linguistics

Master Degree in Digital Humanities

MASTER THESIS

Filter Bubbles in the Italian Twittersphere: a data-driven analysis

Author:
Concetta Chiara Spampinato

Supervisors:
Prof. Giulio Rossetti
Dr. Letizia Milli

April
Academic Year 2021/2022

Filter Bubbles in the Italian Twittersphere: a data-driven analysis

Author:
Concetta Chiara Spampinato

Supervisors:
Prof. Giulio Rossetti
Dr. Letizia Milli

Institute:
Università di Pisa, April, Academic Year 2020/2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Nowadays social media have acquired an important role in daily news consumption, but they may also become a venue of selective exposure. The personalization of new content and Internet applications has represented an important digital progress within the past years. In relation to online news consumption the concepts of *Filter Bubble* and *Echo Chambers* emerged in this scenario, gaining a focal point both in science and in popular press. Various factor such as homophily, information overload, congeniality bias, and filter bubbles may lead people to expose themselves to congenial information, consuming only information that align with their beliefs and excluding them from the contradicting one.

This framework tries to figure out whether or not if in the Italian Twittersphere there is the presence of Filter Bubbles, focusing not on the conventional way of conceiving the concept of filter bubble, but revisiting it, analyzing both the content of tweets, but seeing in particular if the behavior and the way of using the platform are more or less uniform between a user and his circle of friends. This is done through a data-driven analysis of Italian users: with a focus on hashtags, topics, sentiment analysis and community discovery, it was possible to find evidence that to an extent, Twitter is a place where Filter Bubbles are present, supporting findings of some earlier academic studies.

An other main feature of this thesis is the realization of a Dashboard in order to display the statistical, content and sentiment analysis made on the downloaded data.

ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
BA	Barabási–Albert
CC	Clustering Coefficient
CM	Configuration Model
CSS	Cascading Style Sheets
ER	Erdős–Rényi
HTML	HyperText Markup Language
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NRCL	National Research Council Canada Lexicon
SNA	Social Network Analysis
SNS	Social Network Service
URL	Uniform Resource Locator
WS	Watts-Strogatz
WWW	World Wide Web

1 INTRODUCTION

In this digital era, the rapid development of the Internet and social networking sites (SNS) have made the world more interconnected than ever before, making them essential in our daily life. By increasing online communication, Internet offers a continuous-expanding abundance of opinions and information.

The new digital tools we use every day could create knowledge and new possibilities, but they also might become a risk and distort the reality. In fact, the web is a tool that can offer great opportunities, but can also isolate people, creating polarization of public opinion.

If users are already more likely to select and share content that reinforces their pre-existing opinions, the algorithms of social networks enhance this tendency, causing the so called *confirmation bias*. By applying filters to each user's navigation, they are deprived of getting in touch with specific contents, making them view always the same ones. The search for ideological similarity is reflected also connections and interactions among users. By remaining within their own bubble and getting only ideas that confirm their own believes, it becomes difficult to change their mind or get closer to others way of thinking. This is what we called the *Filter Bubble* effect.

When it comes to filter bubble, Twitter has gained a focal spot in the social network environment, in particular when talking about political polarization [1–5].

This thesis aims to figure out whether or not if in the Italian Twittersphere there is the presence of filter bubbles, not only by analysing the content of tweets, - whose main characteristics was the topic/themes recognition of tweets and sentiment analysis classification -, but in particular, we try to

discover if users have similar attitudes when using Twitter, like hours of usage, or seeing which is the most used device.
The reasons which lead to the choice of this platform were mainly two: the possibility that Twitter API gives, which allow developers to do complex queries, and the easy recognition of topics and emotions since tweets are necessarily short texts.
The second goal of this project is the creation of a dashboard, for displaying the results of the analyses, and giving the possibility of a user to see how he is similar or different in respect to his friends.

The chapters of the thesis each address a different aspect of what is summarized until now.
In **Chapter 2**, a brief introduction of the social media Twitter, one of the most popular social networks used by Internet users, is given. Created in 2006, it has found a lot success thanks to the speed with which it disseminates information and the brevity that characterizes his texts. After the Twitter overview, we will focus on the traditional explanation of filter bubbles, that can lead to the creation of echo chambers. Explanation of how and when they were found out are given, as well as the reference to several work, which have discovered how they can affect the thinking of people, bringing them to selective exposure, that can be extremely dangerous when talking about politics.

In **Chapter 3** all the techniques used for the project realization will be examined in depth. Starting with the explanation of what content analysis is, we will also make an overview of Natural Language Processing techniques, focusing in particular on Sentiment Analysis, and on its possible methods of application. A paragraph will be devoted also to Social Network Analysis, which investigates the structure and peculiarities of social relations through the use of notions of network, underlying the concepts of graph theory.

The **Chapter 4** will contain the main analyzes on which the project is focused on. Both the methodology used for data extraction and data storage, as well as all the processes and results of the performed analyzes are explained.

The **Chapter 5** focuses on the dashboard. Initially explaining the reasons that prompted its creation, we try to show the functioning of Dash - the original low-code framework for rapidly building data apps in Python -, seeing in detail the used components and the structure of the Twitter Dashboard.

In the **6** and final Chapter, conclusions are drawn, briefly summarizing the obtained results, pointing out possible improvements and future developments.

2 AN INTRODUCTION TO TWITTER

2.1 TWITTER OVERVIEW

> *«Twitter is the best way to connect with people, express yourself and discover what's happening»*[1].

With approximately 330 million active users worldwide, and defined as a micro-blogging tool[2], Twitter[3] is a big gold mine of data.
With over half a million tweets shared per day every day[4], the open system of Twitter creates a perfect place for users to respond to other users: a vibrant forum for public discourse, where communication becomes easier between individuals by allowing them to exchange personal message.

The social media platform was founded in 2006 by the Obvious Corporation of San Francisco: Jack Dorsey, a computer scientist, had the idea of create a service that allows a person to communicate with a small number of individuals through the use of SMS.
Initially to the project was given the name of "twttr", taking inspiration from the already known *Flickr*, a Canadian-born photosharing website. But only on March 21, the day when Dorsey published the first tweet, "just setting up my twttr", Twitter begins to be officially developed.
The first prototype came tested only among the employees of Odeo, a company of which Dorsey was part of, while the final version was released to the public on July of the same year. In 2007 Twitter became an independent company.

[1] This is how Twitter defines itself.

[2] Microblogs are concise blog posts (under 300 words) that can have images, GIFs, links, infographics, videos, and audio clips. These small messages are sometimes called micro posts.

[3] The name "Twitter" comes from the English verb to tweet which means "chirp."

[4] Twitter Statics at `https://www.internetlivestats.com/twitter-statistics/`

From 2007 many changes have been done to the social media platform, but a meaningful one was the one made in 2017. At the first, the micro-blogging network allows people to write one sentence of up to 140 characters, but in November 2017 Twitter doubled the available character space from 140 to 280 characters, bringing people to use abbreviations much less than before [6]. These tweets are shared with the so-called *Followers*, users who have decided to follow what is written and published by a user. Twitter allows the relationship between two users even in one way only, that is, a user can follow another even if the latter does not follow him in turn.

In Twitter, the use of *Retweets* is frequent: the practice to read a comment and replicate it to our followers to give it more visibility or simply because we appreciate the comment made by that particular user. When a user retweets, the new tweet copies the original one in it. Furthermore, the retweet attaches an **RT** and the **@username** of the user who posted the original tweet at the beginning of the retweet. For instance: if the user @username posted the tweet text of the original tweet, a retweet on that tweet would look this way: "RT @username: text of the original tweet".

Hashtags and *Mentions* are other important features in this social platform. Preceded by the # symbol, the hashtags are used to point out the topic, the context, the main meaning of the tweet that has just been published.

Since June 2009, Twitter introduced hyperlinks on all hashtags, going forward then in the following year to create a list of "Trending topics", i.e a list of the most used hashtags. Today any user can create or use a hashtag by writing the hashtag character # in front of a word or sentence, as long as it is written without spaces between words, of the tweet.

A mention is instead when someone uses the @ sign immediately followed by the Twitter Handle, where the intention is indeed to 'mention' that particular user in the tweet: a simple method of directing the text to a user, or have him read it more quickly.

2.2 WHAT IS A FILTER BUBBLE

The personalization of new content and Internet applications has represented an important digital progress within the past years. In relation to online news consumption the concepts of *Filter Bubble* and *Echo Chambers* emerged in this scenario, gaining a focal point both in science and in popular press.
These concepts postulates that homophily[5] and content personalization lead to an increased exposure to conforming opinions, along with the hiding of contrasting positions.

The concept of filter bubble is often misunderstood with the one of echo chambers. Although they are sometimes used interchangeably, the terms indicate different phenomena:

- Echo chamber is environment where a person only encounters information or opinions that reflect and reinforce their own;
- Filter bubbles are spaces created by our previous online behaviors that determine what content we will see on our feed and with what hierarchy of importance it will appear.

Eli Pariser was the first one to ever talk of filter bubble. According to Pariser, December 2009 is the time the individualization on the internet begun, when Google launched a new feature that could individualize their users' search result.

> Starting that morning, Google would use fifty-seven *signals* —everything from where you were logging in from to what browser you were using to what you had searched for before— to make guesses about who you were and what kinds of sites you'd like. Even if you were logged out, it would customize its results, showing you the pages it predicted you were most likely to click on [7].

[5] Homophily is a well-established phenomenon in sociology, that has been seen to occur frequently in social networks: users with similar contexts have a nature of connecting with one another constantly, creating personal networks that tend to be more homogeneous than heterogeneous. In this way there is a higher chance of bonding with like-minded people and not with dissimilar ones.

This means that two user searching the same things end up with different results based on several different things, such as location, earlier searchers, interests. As a user, to receive an individualized feed of information which becomes bias, is what we now called *filter bubble*, often caused by algorithms that decide what the user should see. Whether it's search engines, SNS, or social media, users are less exposed to different points of view, shutting them up into their own bubble of information.
Furthermore, in a filter bubble, there are less probabilities to have chance encounters that bring insight and learning, and so less creativity, which is sparked by the collision of ideas from different disciplines and cultures.

Pariser highlights three main elements when talking about pre-selected personalization and filter bubbles:

- isolation, people are alone in their personalized information bubble;
- invisibility, the bubble is not visible, hence most people do not what kind of information about themselves is collected and analyzed; this potentially leads to the misbelief that the presented set of information is unbiased;
- lack of choice, people do not choose to enter the "filter bubble" actively, but are put into it passively.

These elements brings other several outcomes: narrower self-interest, dramatically increased confirmation bias[6], overconfidence, decreased motivation to learn, lack of curiosity, decreased creativity and ability to innovate, decreased diversity of ideas and people; basically a skewed picture of the world.
Eventually, users won't see the things that aren't in their interests, since, how Pariser says, the filter bubble "will often block out the things in our society that are important but complex or unpleasant. It renders them invisible" [7].

[6] Phrase coined by English psychologist Peter Wason. It is a mental process that consists in seeking, selecting and interpreting information in order to pay more attention, and therefore to attribute greater credibility, to those that confirm one's beliefs or hypotheses and, vice versa, ignore or belittle information that contradict them.

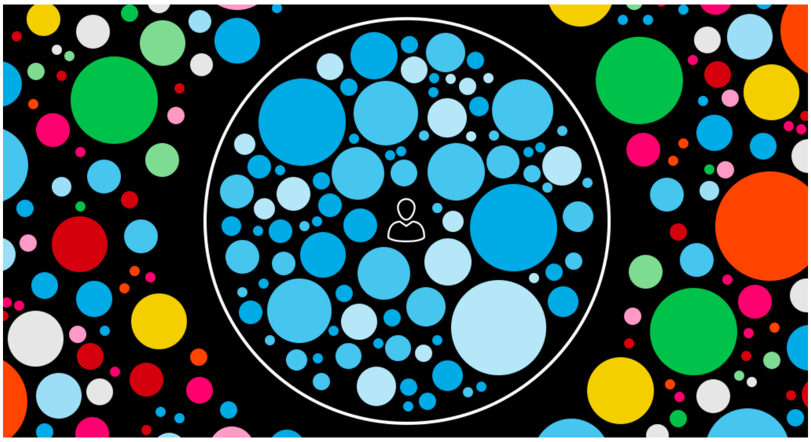


Figure 2.1: Visualization of Filter Bubbles

In his review, Dahlgren amplifies Pariser's view of filter bubbles, stating that the latter one can be seen at two different levels: technical level and societal level[7].

The first one refers to the filter bubble as an "immediate technological situation in which any single choice affects the content recommended by personalization algorithms, thereby narrowing the type of content available over time" [8].

The second one, bring us to "see the causes and consequences of these choices and technologies for humans and society, and, more importantly, for the political process and democracy over time"[8].

There are two main research streams on the filter bubble. The first one mainly focuses on the impact of recommendation systems which consider different feature of the user such as his search behaviour and his demographic information for suggesting new content, creating a filter bubble for the information the user receives [9].

The second wave of studies focuses on the role of users rather than recommendation system technologies [10, 11]. Through a study on Facebook content conducted by Bakshy, Messing and Adamic, it is found out that only 5%–8% of the content provided to people with various political viewpoints is based on their profile [12]. On the other side, different recent studies, done by Cadwalladr and Graham-Harrison and Lazer et al.,

[7] Pariser considers both levels as a chained argument, where one thing leads to another, without making any distinction.

[8] Ibidem.

—who discussed about the widespread usage of bots on social networks to influence political campaigns — show how the role of these platforms in creating filter bubbles could not be totally ignored [13, 14].
After the publication of Pariser's book, the filter bubbles topic received wide attention in academia, in the media and in industry.
Different studies have been done, showing most of the times the negative consequences associated with them like [15]:

- a decline in user's trust: in the long term, the lack of transparency provided by the filtering mechanisms of the social media platforms, can result in the changes to the user's usage experience and a decrease in their trust;
- limiting people's access to information: users rely on a limited number of sources for news that are not subject to professional editorial policies and are often ideologically biased;
- social fragmentation: filter bubbles result in a self-confirming feedback loop for users who are subject to like-minded information. In the long term, this phenomenon will create communities that become increasingly polarised and fragmented;
- the proliferation of fake news: lack of access to factual news (from outside the bubble) results in the spread of emotionally charged and biased news within the bubble, the credibility of which will never be checked or questioned;
- extremism: ideological polarisation through the filter bubbles in social media will help the spread of extremist viewpoints.

One of the most discussed negative consequence of filter bubbles and echo chambers, which has been cited numerous times and more clearly in the literature, is the polarisation of political discussions in social media.
Many scholars claim also that echo chambers threaten democracies [16].
Examples of the impact of the consequences of filter bubbles are the results of The United Kingdom European Union membership referendum, known as Brexit, and the unforeseen results of the USA election in 2016 [17], which will be deepen in section 2.3.
Positive effects of filter bubbles are not deepen, and the only one positive effect of filter bubbles on individuals found in the literature is the one

found in the paper written by Bozdag and Timmermans [18], which is the reduction of information overload. As a matter of fact, according to them, the need of reducing information overload is the main reason for introducing individualizing algorithms in the first place.

2.3 twitter, filter bubbles, echo chambers and polarization

Pariser, in his book, spent some words even for micro blogging platform. He considered the social network less "filtered" compared to Facebook, or Google, but it has still a sort of filtering too. As Pariser said:

> [...] even Twitter, which has a reputation for putting filtering in the hands of its users, has this tendency. Twitter users see most of the tweets of the folks they follow, but if my friend is having an exchange with someone I don't follow, it doesn't show up. The intent is entirely innocuous: Twitter is trying not to inundate me with conversations I'm not interested in. But the result is that conversations between my friends (who will tend to be like me) are overrepresented, while conversations that could introduce me to new ideas are obscured [7].

Additionally, although users often follow their family, friends, and colleagues, they also often choose to follow many additional accounts that they find agreeable to their point of view, reflecting their interests and preferences.

This biased process of information selection creates the filter bubbles phenomenon, recognized also by the Twitter's CEO, Jack Dorsey. Through their selection of certain accounts, Twitter users are often limited in the information provided from these accounts. In his interview, the founder gave an example of how during the social media firestorm in the months before the Brexit vote, many users only saw tweets from people advocating for or against the United Kingdom remaining within the European Union and not both [5].

In regard to this, an analysis made by City, University of London, explores the geographic dependencies of echo-chamber communication on Twitter by analyzing 33,889 tweets from the Brexit referendum campaign period, between 15th April and 23rd June 2016. The average distance between users who sent pro-leave messages was just 22km and the average for remain supporters was 40km. It was found that 69% of pro-leave messages were interactions with other pro-leave accounts, and 68 % of pro-remain messages were with other pro-remain accounts. Just 9% of tweets by leave supporters were sent to remain supporters, who similarly only sent 10% of messages to pro-leave users [19].

From the many available research papers on-line, can say that Twitter has been one of the main platform from where academics took data for analyzing and doing studies focusing on filter bubbles, echo chambers and polarization.

An et al. for example observe extreme polarization among media sources in Twitter in their study in 2014 between Conservatives and Liberals [20].

Another study done by Matteo Cinelli et al. analyzes three different datasets collected on Twitter related to controversial topics: gun control, Obamacare, and abortion, where the analysis focuses on finding homophily in the interaction networks and bias in the information diffusion toward like-minded peers [21].

In 2016, Hong and Kim, considering the social media activities of members of the 111th U.S. House of Representatives, found out that politicians with extreme political ideologies had more Twitter followers than their more moderate peers. This polarization remained valid even after the control for the number of times the sample politicians were mentioned in print newspapers [1].

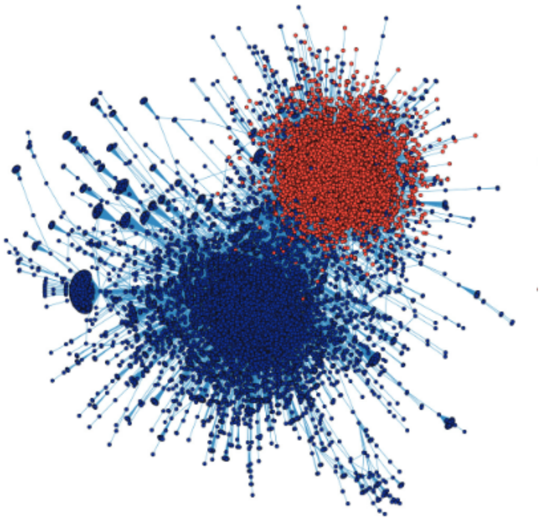


Figure 2.2: The political retweet networks of 2010 U.S. congressional midterm elections, where the red cluster is made of right-leaning users, while the blue cluster is made of left-leaning users. Picture taken from[2].

Being inside a filter bubble is comfortable, because it's more simple for us to digest viewpoints we already harmonize. But when it comes to politics and news, it's good to actively seek out the other side and not completely rely on social media to do it. If not, we may feel self-assured about our views, when there are perspectives and information out there that would better to know, that could even change our views entirely.

3 DATA ANALYSIS TECHNIQUES

In this chapter we will introduce and explain all the techniques used for the analysis of the project, recalling the definition of Content Analysis. We will also discuss about Sentiment Analysis, its related terms and concepts, and a short review of commons approaches in this field, based on both Machine Learning and Lexicon methods.

3.1 CONTENT ANALYSIS

Content analysis has been around for decades and has been implemented in several different fields of study. It is a general term for a number of various strategies used to analyze text and to examine content be it written, audiovisual, or verbal [22]. Krippendorff defines content analysis more specifically as “a research technique for making replicable and valid inferences from texts to the contexts of their use" [23].

Content analysis can be either quantitative or qualitative. Quantitative content analysis helps answer ‘what’ questions, while qualitative content analysis helps answer ‘why’ questions [24], transforming communication materials, like news article or tweets in our case, into a manageable way that helps make inferences and drawing conclusions from the data [25]. Considered as systematic coding and categorizing approach, it used for exploring large amounts of textual information in order to determine patterns and trends of words used, their frequency, their relationships, and the structures and discourses of communication, making possible to analyze data qualitatively and at the same time quantify the data [26]. Two examples of quantitative and qualitative approaches can be:

- quantitative approach: an analysis of campaign speeches for the frequency of terms such as *unemployment*, *jobs*, and *work* and use statistical analysis to find differences over time or between candidates, in order to research the importance of employment issues in political campaigns;
- qualitative approach: locating the word *unemployment* in speeches, identifying what other words or phrases appear next to it, which can be *economy*, *inequality*, and analyze the meanings of these relationships to better understand the intentions and targets of different campaigns [27].

Furthermore, Prasad highlights that content analysis relies on three basic principle of scientific method, which are objectivity, systematic, and generalizability:

- objectivity: content analysis is conducted according to explicit rules that ensure different researchers that can generate the same results when analyzing the same messages or documents;
- systematic: how the analysis of the content is done systematically according to its rule system, where by the possibility of including only materials which support the researcher's ideas is eliminated;
- generalizability: the last principle ensures that the results of the analysis can be transferred or applied to other contexts [28].

To summarize, we can say that content analysis aims to describe the characteristics of the document's content by examining what people say, to whom, and with what effect [29].

3.2 NATURAL LANGUAGE PROCESSING

Natural language processing is an interdisciplinary field that combines computer science, computational linguistics, methods of AI as well as cognitive science. It deals with using computers to derive meaning from human languages in order to execute tasks [30].

Examined from the scientific and engineering perspective NLP has differing goals — former dealing with modeling the cognitive structures involved in

making sense and producing human languages, the latter dealing with developing practical applications that enable interaction between computers and natural languages.
There are plenty applications of Natural Language Processing: systems like chatbots or Siri are able to understand written or spoken text and provide a service to users; self-completion and self-correction systems are now implemented in the user interfaces of search engines; NLP techniques can be used to analyze consumer opinions for advertising or marketing purposes, or to understand whether a document is relevant or not for the purpose of a research.

3.2.1 *Sentiment Analysis*

The area of NLP concerned with tracking what people think about some topic or product is sentiment analysis, also known as opinion mining. To be more specific, sentiment analysis indicates the set of techniques and procedures for the study and analysis of textual information, in order to detect evaluations, opinions, attitudes and emotions related to a certain entity, which can be a product, a person, a topic, etc.

This type of analysis has evident and important applications in the political, social and economic fields. For example, a company might be interested in hearing consumer opinions about her products. But potential buyers of a particular product or service will also be interested in knowing the opinion and experience of someone who has already purchased or used the product. On the other hand, also a public figure, (politics, entertainment, sport) might be interested in what people think of him. Let's imagine a political figure, who wants to know what people think of him, in order to monitor and control the consent for his next possible re-election, or how his election campaign is going on.
An opinion can be formally defined as a quintuple, consisting of:

- an entity;
- an aspect of such entity;

- a sentiment about the aspect, which can be positive, negative or neutral, or it can be expressed with different intensity levels (for example from 1 to 5 stars);
- an opinion holder;
- the time when the opinion was expressed [31].

3.2.2 *Sentiment Analysis Approaches*

There are two main techniques for sentiment analysis: machine learning based and lexicon based.

- Machine learning approach, can be roughly divided into two groups: the so-called supervised methods and the unsupervised methods.
  Two sets of documents are needed: a training and a test set. A training set is used by an automatic classifier to the differentiating characteristics of documents, and a test set is used to check how well the classifier performs. Several machine learning techniques have been adopted to classify the reviews, like Naive Bayes (NB), support vector machines (SVM) which have achieved great success in sentiment analysis.
- Lexicon based approach: involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. Classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. The basic steps of the lexicon based techniques are:
  1. Preprocess each text, like URL and noisy characters removal, etc.
  2. Initialize the total text sentiment score: s ← 0.
  3. Tokenize text. For each token, check if it is present in a sentiment dictionary
     If token is present in dictionary,
     i. If token is positive, then s ← s + w.
     ii. If token is negative, then s ← s - w.

4. Look at total text sentiment score s,
   (a) If s > threshold, then classify the text as positive.
   (b) If s < threshold, then classify the text as negative [32].

The lexicon based approach can be divided in dictionary-based and corpus-based approaches.

The first one is based on the activity of building a small set of word which usually are organized in a hierarchy structure, with an associated polarity value. Unfortunately, the context in which words are used is not taken into account, and it is also possible to find words without polarity.

The second method can solve previous problems related to find words that could express opinions inside a specific context. This approach is based on syntactic recurrent patterns and models, which are used inside corpus of big dimensions in order to list words expressing a certain polarity.

When talking about sentiment analysis, another classification is the one based on the structure of the text. As already discussed, the most well-studied problem in the field of sentiment analysis is the sentiment polarity classification. Typically, this task is considered as a multi-class classification problem, meaning that given a subjective text, the goal is to determine whether the general tone of the text is positive, negative or neutral.
This task can be conducted at various levels of granularity: from the sentiment polarity associations of words and phrases, to the sentiment of sentences, chat messages, and tweets, to the analysis of sentiment in product reviews, blog posts, and entire documents.
Analysing in detail the structure of the text:

- Document Level: this type of analysis is based on the overall sentiment expressed by opinion holder, where it can be a blog or reviews for example. Studying the document as a single unite and classifying it according to the type of opinion that the entire text reveals, whether it is negative or positive is the main task in this level. This level of analysis is based on the assumption that the whole document discusses only one topic and thus cannot be applied to documents that contain opinions on more than one entity.

- Sentence Level: sentiment analysis at the sentence level aim to assign labels such as positive, negative, or neutral to whole sentences. This type of analysis is closely related to the concept of subjectivity classification, which distinguishes the so-called objective sentences, those sentences that express an objective data or fact, from the subjective sentences that instead express a point of view and an opinion.

- Feature Level: the idea of this level of analysis is to assign a sentiment value to a word. Both the document and the sentence level analyses do not discover what exactly people like or not. This level instead performs a finer-grained analysis, looking at the opinion itself instead of looking at language constructs like documents, clauses or paragraph. It is based on the idea that an opinion consists of a sentiment, positive or negative, and a target of opinion.

3.3 SOCIAL NETWORK ANALYSIS

The SNA is is an interdisciplinary research field. Emerging as a key technique in modern sociology, during time it has also gained significant popularity in many fields, such as anthropology, biology, communication studies, economics, geography, information science, political science, public health, as well as social psychology, development studies, sociolinguistics, and computer science.

In general, social networks can be viewed as a set of connected entities. This is commonly represented by a collection of *vertices*, or nodes, and *edges*, or links. A vertex is an abstraction for an actor in the network whereas edges are relations between these actors. Within mathematics, this is known as a graph [33]. Formally, a graph can be written as G = (V, E), where V is a set of vertices, and E is a set of edges connecting vertices in V.

The links of a network can be directed or undirected. Some systems have directed links, like the WWW, whose URL, point from one web document to the other, or phone calls, where one person calls the other. Other systems have undirected links, like romantic ties: if I date Janet, Janet also dates me, or like transmission lines on the power grid, on which the electric current can flow in both directions [34].

Graphs work in the same way, and they can be *directed*, or *digraph*, if all of its links are directed, while it is called *undirected* if all of its links are undirected.
Furthermore, graphs can be *unweighted*, where links are equally strong, and all weights are 1, or a *weighted* graph, where weighted refers to the fact that relations may be of varying importance. Rapresentations of these graphs is visible in Figure 3.1.

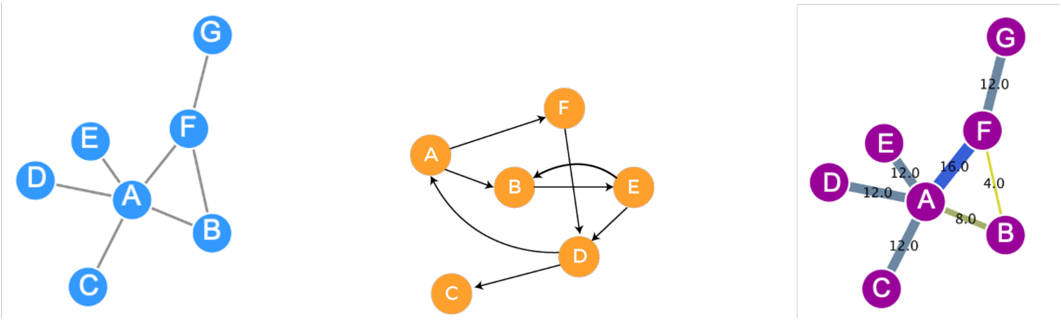


Figure 3.1: Different Types of Graphs - Undirected, Directed, Weighted

Another extension of graph is the *bipartite graph*, or *bigraph*, whose vertices can be divided into two disjoint and independent sets **U** and **V** such that every edge connects a vertex in **U** to one in **V**.

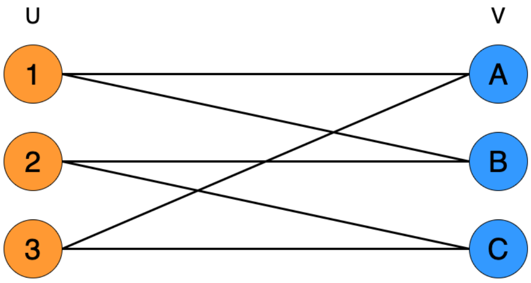


Figure 3.2: Bipartite Graph

3.3.1 *Synthetic Graphs*

There are a number of commonly used methods for generating synthetic graphs. One of them is the *random network*, introduced by Rapoport and Solomonon in 1951 and independently by Erdős and Rényi in 1959. The model embraces the apparent randomness by constructing and characterizing networks that are truly random, where

> *a random network consists of N nodes where each node pair is connected with probability p* [34].

In this model, all graphs on a fixed vertex set with a fixed number of edges are equally likely.
Another model, which interpolate between regular and random networks, is the one proposed by Duncan J. Watts and Steven Strogatz in 1998, called *Small-World*. Formally, the model contains a parameters set including three variables representing group size, number of neighbors, and rewiring probability [35]. This model produces graphs with small-world properties, which include short average path lengths and high clustering.

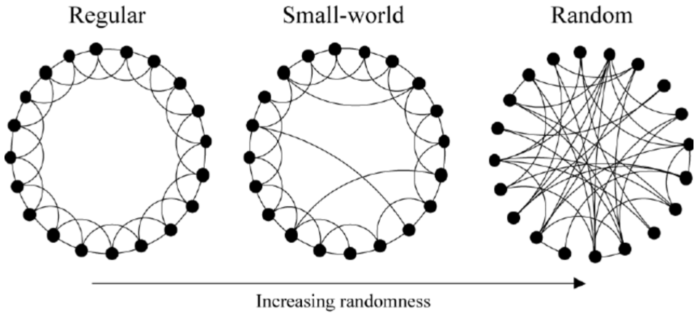


Figure 3.3: Network topology in Small-world model at different p rates

The third presented model is the *configuration model* which is a method for generating random networks whose nodes have pre-defined degree k_i, rather than having a probability distribution from which the given degree is chosen. It is widely used as a reference model for real-life social networks, because it allows the modeler to incorporate arbitrary degree distributions, since such a model is more flexible than the generalized random graph.

The last described model is the *Scale-Free Network*, known also as the Barabási-Albert model (from its inventors Albert-László Barabási and Réka Albert), which use a preferential attachment[1] mechanism for generating

[1] New nodes prefer to link to the more connected nodes, and this what happens in real network, while in random networks, nodes randomly choose their interaction partner.

random scale-free networks and network growth. This model takes into the account the existence of hubs, particular nodes which are more highly connected than others. The main differences between a random and a scale-free network comes in the tail of the degree distribution, representing the high-k region of p_k, and while the random network model assumes that the number of nodes is fixed, real networks are the result of a *growth* process that continuously increases.

3.3.2 *Network Metrics*

Social network analysis provides a set of powerful quantitative graph metrics for understanding networks and the individuals and groups within them. Below we are going to explain the main metrics.

3.3.2.1 *Degree and Degree Distribution*

The main property of a node is the degree. A node's degree is the number of links that are incident to the node. As concerns directed networks, this metric can be expanded into to separate metrics, making a distinction between *in-degree*, the number of incoming links, and *out-degree*, the number of outgoing links.

The degree distribution describes how the links in the graph are distributed among the nodes. The degree distribution of a graph is a function P (*k*) which describes the fraction of the network's nodes which have degree *k*, that is to say the probability distribution of these degrees over the whole network.

3.3.2.2 *Clustering Coefficient*

The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together, defined as the number of directed links that exist between the node's neighbors, divided by the number of possible directed links that could exist between the node's neighbors. The CC of a graph is the average clustering coefficient of all its nodes.

Traditionally, the two versions of the clustering coefficient developed for testing the tendency of nodes to cluster together into tightly knit groups are the *global* clustering coefficient and the *local* clustering coefficient. The global version was designed to give an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.

3.3.2.3 *Path Analysis*

A path is a sequence of interconnected nodes, meaning that each pair of nodes adjacent in the sequence are connected by a link. A shortest path, called also geodesic path, between two nodes in a graph is a path with the minimum number of edges. If the graph is weighted, it is a path with the minimum sum of edge weights. The length of a geodesic path is called geodesic distance or shortest distance. Geodesic paths are not necessarily unique, but the geodesic distance is well-defined since all geodesic paths have the same length [36].

3.3.2.4 *Connected Components*

A connected component of a network graph is a subset of vertices in the graph such that each pair of vertices is connected by a path. Social networks can have several separate connected components, which would indicate that there is no path of connection among the members of one component to the other, even though they belong to the same network. For a directed graph, we distinguish between a strongly connected component and a weakly connected component. A strongly connected component (SCC) is defined as a set of nodes such that there is a path in the network between all pairs of nodes in the set. In contrast, a weakly connected component (WCC) is defined as a set of nodes such that there is a path in the network between all pairs of nodes in set if the all links in the network were viewed as undirected.

For real world graphs and network graphs, where connections are guided by a specific subject, there is usually the presence of a giant component, the largest one, that can contain over 90% of the vertices, and the rest of the network is divided into a large number of small components disconnected from the rest.

3.3.2.5 *Centrality*

The centrality measure is essential when studying network analysis. Through this measure, we can identify which are the most important or central vertices in a network. There are many possible definitions of importance, and correspondingly many centrality measures for nodes in a network. We can identify three major studied indicators when talking about centralities: eigenvector, betweenes and closeness centrality.

Eigenvector centrality measures the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores [37].

Closeness centrality measures the mean distance from a vertex to other vertices. Specifically it is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes.

Betweenness centrality measures the extent to which a vertex lies on paths between other vertices. Vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others [36].

3.3.3 *Communities Detection*

A community is a subset of the users in a social network that is more tightly interconnected than the overall network, or it can be defined as a locally dense connected subgraph in a network [34]. Community discover in a social graph is an interesting task, that can give us a better understanding of a network structure and can give us new insights about how it is organized.

Users in a community tend to interact frequently, sharing interests among groups of users who do not necessarily know each other but who are close together in the social network and trust each other to some extent. There are several types of clustering algorithms that can be applied to different tasks. These algorithms can be divided into two big methods: non-overlapping community detection methods and overlapping community detection methods. In a non-overlapping methods, each node belongs only to one and only one community, in the overlapping methods, instead, a node could be a part of different communities.

3.3.3.1 *Louvain Algorithm*

The Louvain Community Detection method is a widely used approach to identify communities in large network graphs. The Louvain method is a greedy optimization method, introduced specifically for the task of finding communities in networks, and it is divided in two phases: Modularity Optimization and Community Aggregation, where at first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced.

Going deeper in the two phases, Louvain will randomly order all nodes in the network in Modularity Optimization. Then, one by one, it will remove and insert each node in a different community until no significant increase in modularity[2] (input parameter) is verified.

[2] Modularity is a scale value between -1 (non-modular clustering) and 1 (fully modular clustering) that measures the relative density of edges inside communities with respect to edges outside communities.

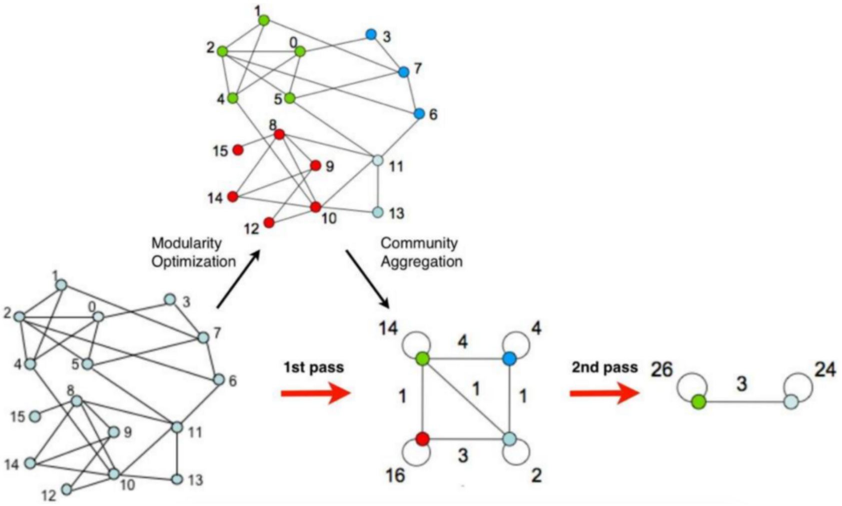


Figure 3.4: Louvain Algorithm. In each iteration modularity is optimized for local community changes first, and newfound communities aggregation second. Iterations stop when it is impossible to increase modularity.

After finishing the first step, all nodes belonging to the same community are merged into a single giant node. Links connecting giant nodes are the sum of the ones previously connecting nodes from the same different communities. This step also generates self-loops which are the sum of all links inside a given community, before being collapsed into one node. Thus, by clustering communities of communities after the first pass, it inherently considers the existence of a hierarchical organization in the network.

4 CASE STUDY

In this chapter we will focus on our case study, exploring step by step the techniques used for analyzing the behaviour of our Twitter users.
The project is divided into different phases:

- data acquisition;
- data cleaning;
- statistical analysis;
- cluster identification;
- data visualization on Dash.

The Python code for the downloading and the analysis part was written mainly on Google Colab Notebooks, while the code for the realization of the dashboard was done in PyCharm.

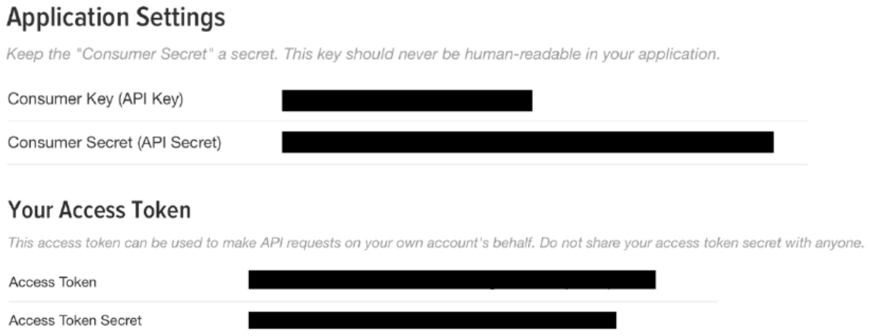
4.1 DATA ACQUISITION

The data acquisition had a main role in the realization of this project. Thanks to Twitter public availability and its API, it was possible to collect more or less 1.500.000 tweets from about 9.000 users. Every time the data were collected, they were stored in different dataframes. Furthermore, since the overall data were also analyzed, it was decided to combine all the dataframes into a single huge dataset.

4.1.1 *Twitter API*

Twitter, as many other social networks, can be accessed via the web or mobile device. Another option provided by Twitter is the programmatic access

to Twitter data through their application programming interface (API). APIs, in general, allow interested parties to access the data and functionality of popular online services, all in a very controlled manner, and Twitter offers one of the most powerful API for most of the currently used programming languages.
Twitter limits free API usage, but, compared to other microblogs, they are quite generous: the API allow developers to do complex queries like pulling every tweet about a certain topic within the last twenty minutes, or target users that specifically live in a certain location, extracting tweets only in a language and so many others.
Before making any API requests to Twitter, it is necessary to create an application, a standard way for developers to gain API access. Once filled a form and got the approval by Twitter, Twitter gives four secret keys, which are essential for creating an app and start to get some data.



Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) [redacted]

Consumer Secret (API Secret) [redacted]

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token [redacted]

Access Token Secret [redacted]

Figure 4.1: Twitter Credential

4.1.2 *Crawling Strategy*

The crawling and the collecting strategy played a major role in the project and took long time to collect the necessary data for a proper analysis. Since it was decided to analyze only Italian tweets, it was useful to stream only the Italian bounding box.

Among all the users given by the latter process, it was selected only one account at random. From the User X, it was decided to download his last 200 tweets, and his friend list. From this list of friends, one of them was chosen and analyzed, collecting also his latest 200 tweets and friends list.

The same process was made multiple times in order to collect as much data as possible. This downloading process was made during the 2021 Spring.

4.1.3 *Tweepy*

As discussed is Section 4.1.1, Twitter's API are quite easy to use, and their popularity resulted in creation of many API wrappers – language-specific kits or packages that wrap sets of API calls into easy-to-use functions. One of these wrappers is *Tweepy*[1], which was used in this implementation for collecting and processing data within Python scripts.

The username and timeline scrapers use Tweepy to carry out the data collection, and Tweepy automatically authenticates with Twitter's API by using access tokens and consumer keys which are provided by Twitter upon creation of a developer account as discussed in Section 4.1.1.

Below is the Tweepy example as shown in the documentation:

```python
import tweepy

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
public_tweets = api.home_timeline()
for tweet in public_tweets:
    print (tweet.text)
```

Listing 1: Python example

4.2 COLLECTING, CLEANING AND PROCESSING DATA

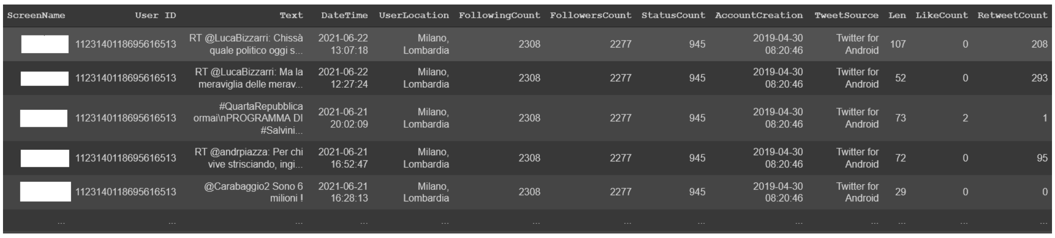
For each tweet, the following attributes are collected:

- *Screen Name*: the screen name of the user who published the tweet;
- *User ID*: the unique ID that every user has;
- *Text*: the text of the tweets, without any changes;
- *DateTime*: when the single tweet was made;

[1] https://docs.tweepy.org/en/stable/index.html

- *Followers Count*: the number of the followers of the user;
- *Following Count*: the number of the user's following;
- *Status Count*: the total number of tweets, since the creation of the account, done by the user;
- *Account Creation*: when the account was created;
- *Tweet Source*: the device used by the user for publishing his status IE Android, Desktop, etc. ;
- *Location*: the location from where the tweet was made;
- *Tweet Length*: the number of characters in every tweet;
- *Like Count*: the number of likes that a tweet received;
- *Retweet Count*: how many times the tweet was retweeted.

Once we obtained the data, they are stored in a DataFrame, where in every row we find a different tweet.



ScreenName	User ID	Text	DateTime	UserLocation	FollowingCount	FollowersCount	StatusCount	AccountCreation	TweetSource	Len	LikeCount	RetweetCount
[redacted]	1123140118695616513	RT @LucaBizzarri: Chissà quale politico oggi s...	2021-06-22 13:07:18	Milano, Lombardia	2308	2277	945	2019-04-30 08:20:46	Twitter for Android	107	0	208
[redacted]	1123140118695616513	RT @LucaBizzarri: Ma la meraviglia delle merav...	2021-06-22 12:27:24	Milano, Lombardia	2308	2277	945	2019-04-30 08:20:46	Twitter for Android	52	0	293
[redacted]	1123140118695616513	#QuartaRepubblica ormai\nPROGRAMMA DI #Salvini...	2021-06-21 20:02:09	Milano, Lombardia	2308	2277	945	2019-04-30 08:20:46	Twitter for Android	73	2	1
[redacted]	1123140118695616513	RT @andrpiazza: Per chi vive strisciando, ingi...	2021-06-21 16:52:47	Milano, Lombardia	2308	2277	945	2019-04-30 08:20:46	Twitter for Android	72	0	95
[redacted]	1123140118695616513	@Carabaggio2 Sono 6 milioni !	2021-06-21 16:28:13	Milano, Lombardia	2308	2277	945	2019-04-30 08:20:46	Twitter for Android	29	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 4.2: Example of one DataSet

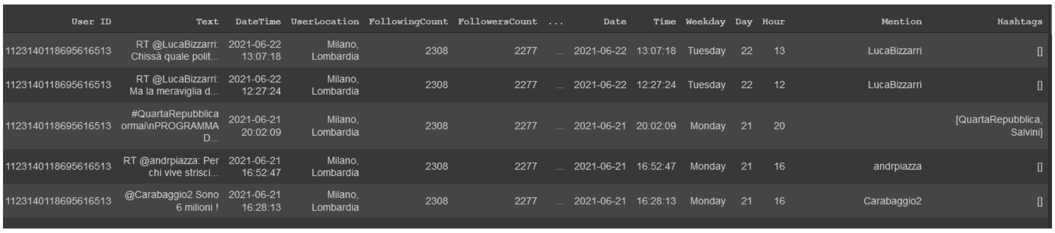
All the attributes were used for doing analysis, except the attribute *location*: after checking its value, it emerged that it contained many missing values, or most of the times, people write fictitious places like: "cittadino del mondo", "Dipende dal tunnel che prendo" or "Somewhere over the rainbow", hence we decided to deleted it.

The DateTime attribute have been spilt into other attributes: *Weekday*, *Day*, *Hour*, in order to facilitate the statistical analysis, and understand in which hours the users use the most the social media, as well as the days of the week.

The dataset contains the "Text" field, which may consist of noise as well as partial and unreliable linguistic data. Hence, in order to analyze linguistic data from Twitter, it is necessary to clean it.
Through this attribute, it is possible to retrieve data on a large set of topics, like find trends related to a specific keyword, or measure brand sentiment. First of all, from the attribute Text, we extracted two other attributes and created two columns:

- hashtag attribute: a column where all the hashtags of the tweet are stored;
- mention attribute: a column where there are all the mentions.

This features extraction was possible thanks to *re module* which gives programmer an embedded functionality inside Python language to operate textual or string dataset.



User ID	Text	DateTime	UserLocation	FollowingCount	FollowersCount	...	Date	Time	Weekday	Day	Hour	Mention	Hashtags
1123140118695616513	RT @LucaBizzarri: Chissà quale polit...	2021-06-22 13:07:18	Milano, Lombardia	2308	2277	...	2021-06-22	13:07:18	Tuesday	22	13	LucaBizzarri	[]
1123140118695616513	RT @LucaBizzarri: Ma la meraviglia d...	2021-06-22 12:27:24	Milano, Lombardia	2308	2277	...	2021-06-22	12:27:24	Tuesday	22	12	LucaBizzarri	[]
1123140118695616513	#QuartaRepubblica ormai\nPROGRAMMA D...	2021-06-21 20:02:09	Milano, Lombardia	2308	2277	...	2021-06-21	20:02:09	Monday	21	20		[QuartaRepubblica, Salvini]
1123140118695616513	RT @andrpiazza: Per chi vive strisci...	2021-06-21 16:52:47	Milano, Lombardia	2308	2277	...	2021-06-21	16:52:47	Monday	21	16	andrpiazza	[]
1123140118695616513	@Carabaggio2 Sono 6 milioni !	2021-06-21 16:28:13	Milano, Lombardia	2308	2277	...	2021-06-21	16:28:13	Monday	21	16	Carabaggio2	[]

Figure 4.3: Dataset with new attributes

After extracting mentions and hashtags, the Text attribute was "cleaned" from noises, in order to do a proper analysis. In particular the phase of cleaning consisted in:

- lower-casing all the letters;
- removal of HTML noise/characters;
- fixing abbreviation: because of character limit in Twitter, people often use abbreviate form of word to fit more characters, so a little dictionary was created, and words like "cmq", "scs" were extended to their original form (comunque, scusa);
- removal of URL, mentions, hashtags, emoji;
- removal of punctuation;

- removal of the "RT" word, the abbreviation for retweet;
- stop words removal;
- tokenization.

Since some hashtags can convey meaning and can have some sentiment it, it was decided to remove only the "#" symbol instead of removing all the words. This phase of text cleaning was performed by using *re module* once again, where the re.sub() function is used to search a patter and replace occurrences of a particular sub-string with another specified sub-string, in this case a white space character.

The set of Stop words was download from NLTK, setting the language in Italian. From this list, it was decided to add other meaningless words for our analysis. Furthermore, since it can happen that some tweets are written in English - it can be that among friends are English user or because an Italian user simply write in English - in a second phase, it was decided to clean also the text from English stopwords. In the Tables below there are some examples of tweets transformation without the removal of stopwords at first, and then with the removal of stopwords.

Orginal Text	**Processed Text**
RT @LucaBizzarri: Chissà quale politico oggi scriverà: a casa nostra si devono rispettare le nostre leggi.	chissà quale politico oggi scriverà”a casa nostra si devono rispettare le nostre leggi”
#QuartaRepubblica ormai PROGRAMMA DI #Salvini Con Porro e Labate OSPITI	quartarepubblica ormai programma di salvini con porro e labate ospiti
Cosa ci facevano #SALVINI e #MELONI al Funerale del Cantante ? #Merlo #19Giugno #sciacalli https://t.co/etTqd8rejr	cosa ci facevano salvini e meloni al funerale del cantante merlo 19giugno sciacalli

Table 4.1: Text Cleaning Example with stop words

Processed Text	**Text Without stop words**
chissà quale politico oggi scriverà a casa nostra si devono rispettare le nostre leggi	chissà politico scriverà casa devono rispettare leggi
quartarepubblica ormai programma di salvini con porro e labate ospiti	programma salvini porro labate ospiti
cosa ci facevano salvini e meloni al funerale del cantante merlo 19giugno sciacalli	salvini meloni funerale cantante merlo 19giugno sciacalli

Table 4.2: Text Cleaning Example without stop words

This text cleaning process was necessary especially for the words frequency task, for trying to understand the main topics of the analyzed user,

and uncover recurring themes, but it is significant to know that the text process done for the sentiment analysis isn't exactly the same, since emoji, for example, are a common way of expressing feelings in Twitter, and so they could convey sentiment.
To understand the effect of pre-processing, a quantitative analysis has been run before taking any action on the corpus.
Key statistics about the length of the Tweets are shown here in summary in Table 4.3 and 4.6.

Average number of words	102
Longest Tweet	152
Shortest Tweet	2

Table 4.3: Statics of Tweets before pre-processing - Friends

Average number of words	14
Longest Tweet	48
Shortest Tweet	0

Table 4.4: Statics of Tweets after pre-processing - Friends

Average number of words	89
Longest Tweet	140
Shortest Tweet	11

Table 4.5: Statics of Tweets before pre-processing - User

Average number of words	12
Longest Tweet	19
Shortest Tweet	1

Table 4.6: Statics of Tweets after pre-processing - User

As we can see, by removing url, mentions and stop words drastically reduced the total number of words, and there's no surprise that some tweets could be empty after the cleaning process. The complete set of stop words used in this project is available in the Appendix A.1.

In Figure 4.4 and 4.5 we find the distribution of words before and after the cleaning process, in red the distribution of friends, in blue the one of the single user.

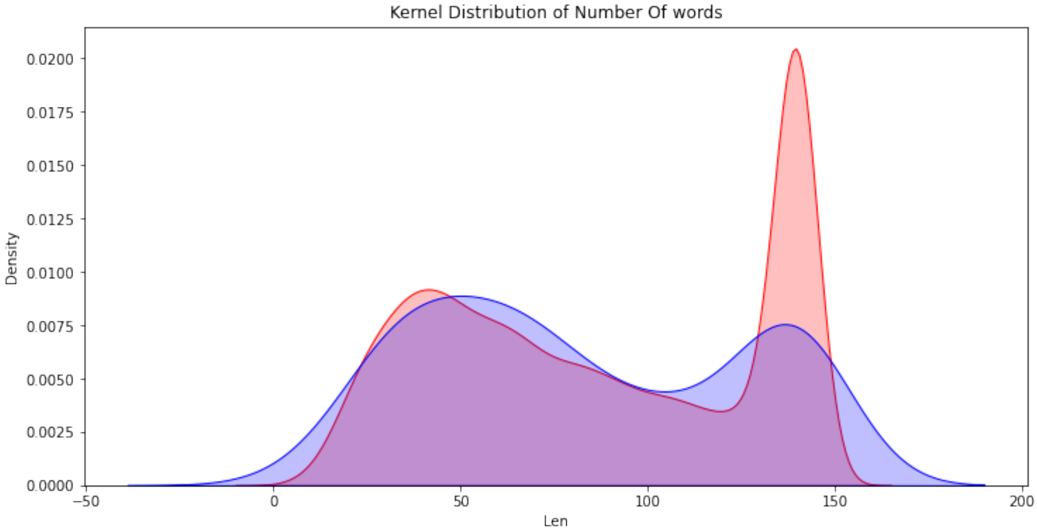


Figure 4.4: Kernel Distribution of words - Original Text

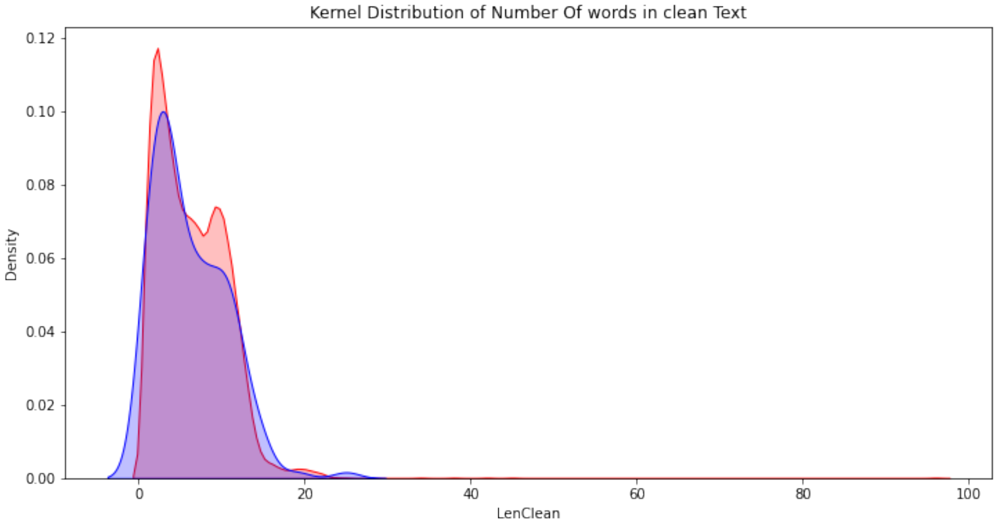


Figure 4.5: Kernel Distribution of words - Clean Text

4.3 DATA ANALYSIS AND TEXT MINING

In this paragraph we will describe the results obtained by the exploration and the analysis of the data. Through the obtained statistics, we will try to describe the behaviours of our users.
Once the dataset is pre-processed, visual data exploration proved a particularly intuitive and fast way for deriving meaning from high volume dataset. Indeed, producing valid visual representations of data aids the cognitive

processes of identifying pattern and trends as well as discover meaningful insight from data. Most of the data visualization was done with the *Plotly* library, and with *wordcloud*, also known as a tag cloud, which is a visual representation of words[2]. An important thing to remember is that all the statics regard the last 200 tweets per every user and not their entire timeline.

4.3.1 *Statistics Overview*

At first it was decided to study the distribution of followers, followings, retweets and likes.

The average number of *followers* of the users in our whole dataset is 340.592. The minimum is 0 followers, while the maximum is 129.636.397, which indicates the most followed and influential person in our dataset, which is Barack Obama. The median is 1466, while the mode is 41.

As regards the static values of the *followings*, the average number is 3062, the minimum of following is 0 and a maximum of 734.844 followings. The mode is 0, while the median is 883.

A high number of following and a low number of followers indicates an average audience of people in their network, miming that those people may use the social network for personal reasons, on the other hand a high number of followers and a low number of followings indicate the popularity of a person.

With regards to *retweet* distribution, the mean is 274 retweets, the most re-tweed post is one with 2.108.723, a post dedicated to the death of the American actor Chadwick Boseman, while the minimum is 0.

The mean of *likes* of tweets is instead of 446, with a max of 2.798.002 (a post of Barack Obama, in which he congrats with the new president of USA, Joe Biden), and a minimum of 0.

Since this framework is focused on the Italian Twittersphere, it was decided to drop all the non Italian user in order to see statics regarding Italy.

By dropping those users, we have an average number of *followers* of

[2] Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide quick and simple visual insights that can lead to more in-depth analyses.

245.567. The minimum is 0 followers, while the maximum is 9.239.755, the account of the soccer team Juventus. As regards the static values of the *followings*, the average number is 1237, the minimum of following is 0 and a maximum of 92.619 followings. As regards the retweet statics, the most re-tweed post written by an Italian user, Fedez, an Italian singer, has 45.948 retweets, where he criticizes the Rai, the national public broadcasting company of Italy, for complaints that he should make during a speech on stage about the Ddl Zan. We can see how numbers dropped, this happens because the English users detected in our dataset were mainly only famous and influential people.

All these statistics were also made for the single user taken as an example, and for his network of friends.
The user X chosen at random in our analysis has 4963 followers, which is far away from the mean value of the whole dataset, and 3109 following, a little bit above the average. The mean of his retweet is 5.2, while the mean of like is 29 per post.
His network of friends has an average of 1354 followings and 1576 followers. The mean of retweet is 165, with a maximum of 563.405 and a mean of 2 likes per post (meaning that there are a lot of post which do not have a like), with a maximum of 2654.

4.3.2 *Time Distribution*

While many social media managers work normal office hours, users on social networks are active around the clock. Given the nature of these networks, users are oftentimes more active on weekends than on weekdays, or during the daylight hours rather that nighttime. In order to extract this information, as discussed is Section 4.2, the datetime of every single tweet was downloaded and splitted in other attributes, making the analysis easier.
As it is possible to see in Figure 4.6 - left side - Twitter's peak times in our dataset are between 12 am (3417 tweets) and 8 pm (3286 tweets), timetables that align with lunch breaks and evening commutes. The busiest hour on Twitter is around 19 pm (3669 tweets). Not surprisingly, activity drops in the early hours, from midnight to 7 am.

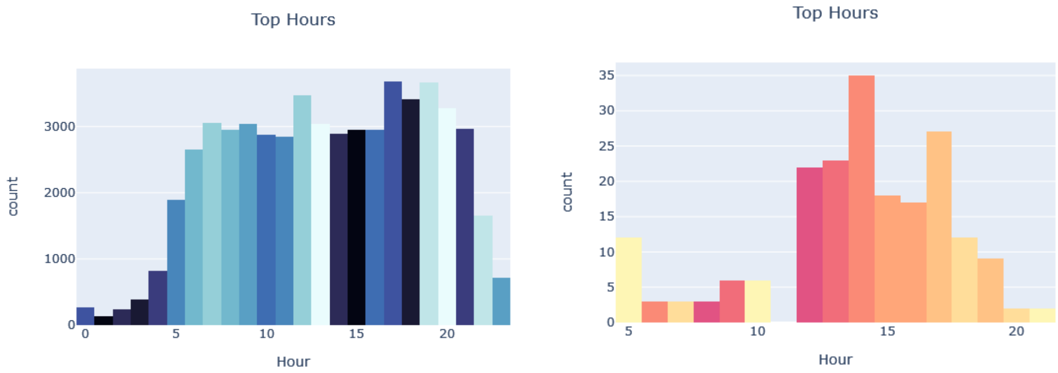


Figure 4.6: Distribution of hours of use - Friends and User

The behaviour of the single user differs a little from the ones of his friends, but not that much. The peaks of his last 200 tweets were done at 2 pm and at 5 pm. Strange enough, there is a peak also at 5 am, so apparently the user wakes up early in the morning, with a drop in the next hours.

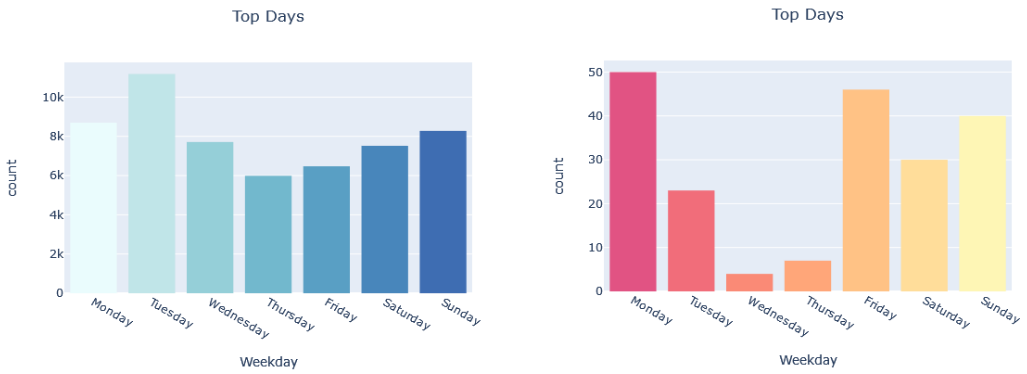


Figure 4.7: Distribution of days of use - Friends and User

As concerns the days of the week of use of the platform, it seems that both the analyzed datasets show an interest in posting at the beginning of the week (Monday) and less interest in posting on Thursday as seen in Figure 4.7.

If all users had recently published 200 tweets, we would have had a timeline that started directly from 2021, but through the graph in Figure 4.8, it is possible to see a timeline of the published tweets that goes up to 2014. Hence we can guess two different scenarios: either some user does not use Twitter assiduously, or some users registered in previous years, have published something in the past but now they no longer use the platform.

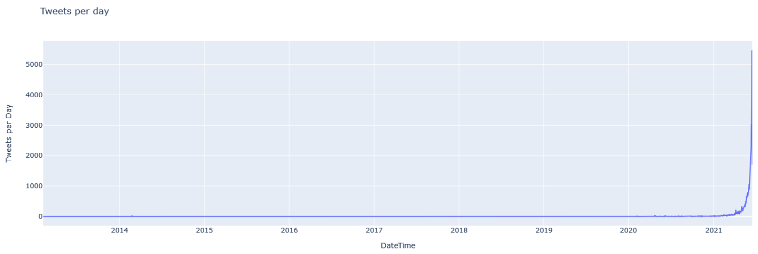


Figure 4.8: Time Analysis - Friends Use

Recalling that only the last 200 tweets of each user have been extracted, it is normal to see an increase in 2021, and in particular, in Figure 4.9, we see how there is an increase in June, the period in which the data extraction was made.

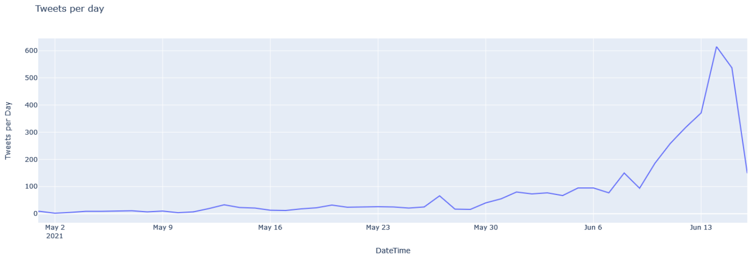


Figure 4.9: Time Analysis - Friends Use Detail

The single user analysis proves that he is a quite active as a user since he made 200 tweets in less than a week, with an average of 33 tweets at day.

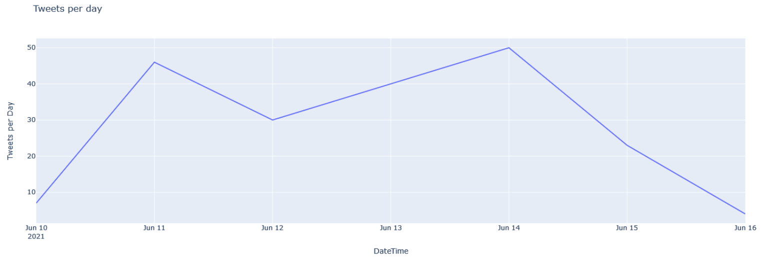


Figure 4.10: Time Analysis - User

From the attribute *TweetSource* it was possible to see from where the users usually tweet. The most used source are the following, as the Figure 4.11 shows:

- Twitter for Android (425061);
- Twitter for iPhone (288012);
- Twitter Web App (206770);
- Twitter for iPad (20902);
- TweetDeck (10822).

The Android source is the most used one, indeed according to some analysis [38] Android rules the Mobile Operating System in Italy with 70.04% of users, while IoS with 29.45%. Interesting enough, the five most used source is TweetDeck [39], a tool for managing Twitter publications and monitoring everything that happens around people business. Purchased by Twitter in 2011, the social media platform in question is divided into columns to manage the flows that characterize users online activity.

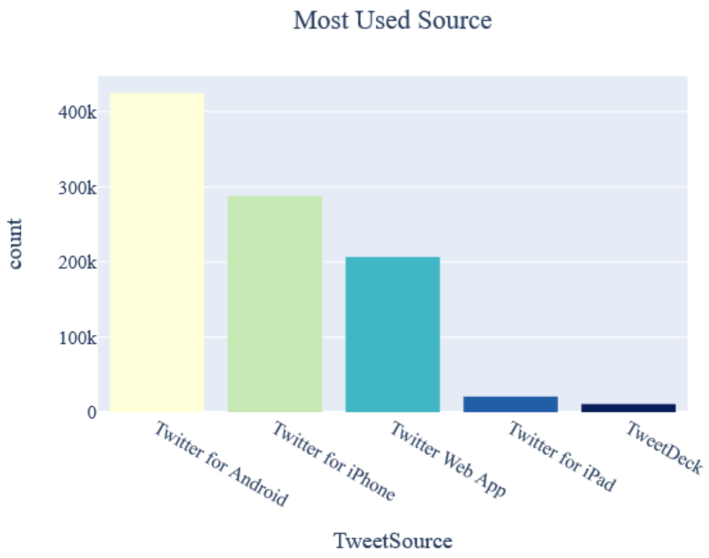


Figure 4.11: Most Used Sources

4.3.3 *Words Frequency and Trending Topics*

Creating a frequency-sorted word lists is one of the standard methodology for exploiting text data, since when studying a text, anyone is likely to need to know how often each different word form occurs in it [40].
A word list can be arranged in order of first occurrence, alphabetically or in

frequency order. Word frequency analysis is a technique for identifying the most common words in a text corpus. In technical terms, occurrences of each unique word in a document are collected in a term-document matrix which enables sorting by frequency. It can therefore provide interesting information about the words that appear (and do not appear) in a text, or in this case, in our tweets.

As we can see from the analysis of the most common words, both the selected user and his friends, apparently, use Twitter for discussing about political issues mainly. Words as "Salvini", an Italian right-wing politician, leader of the party "Lega" are frequent in both datasets.
In particular it seems that the selected user doesn't sympathize with this politician since there are words which are offensive like co****ne, m***a, and fascist. As confirmation of this, further on, we find the analysis of the hashtags in the next lines. We can't say the same about his friends, because even if Salvini and Meloni — an other Italian politician, leader of the national-conservative party "Fratelli d'Italia" since 2014 — are frequent words, we could not guess if they are supporters or not by looking only at this data.

Frequent Words	Count
salvini	886
italia	752
conte	560
roma	537
lavoro	494
via	492
governo	458
casa	441
meloni	439
vero	463

Table 4.7: Frequent Words - Friends

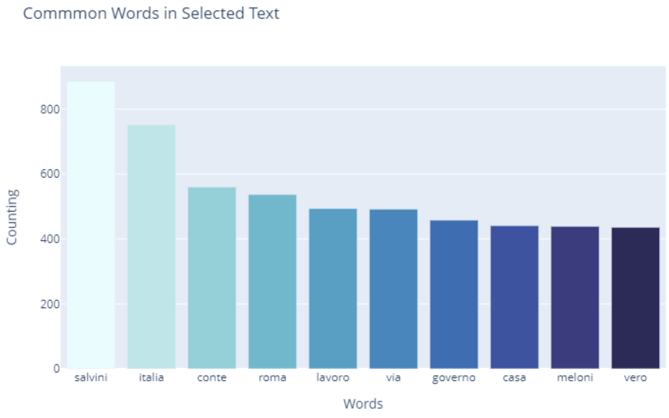


Figure 4.12: Frequent Words - Friends

Frequent Words	Count
salvini	19
fascista	10
roma	9
lega	8
meloni	8
co****ne	8
italia	6
me**a	6
partito	5
figliuolo	5

Table 4.8: Frequent Words - User

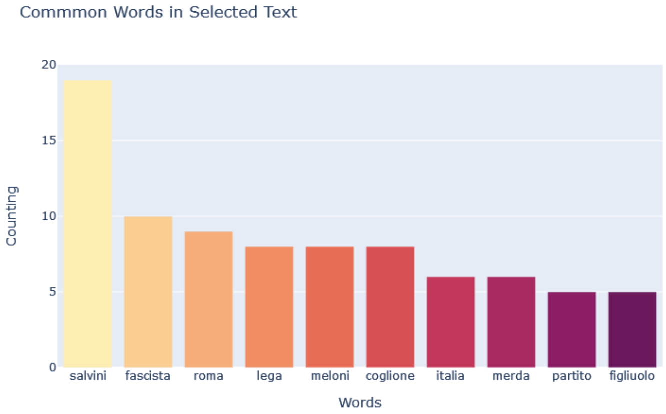


Figure 4.13: Frequent Words - User

Frequent Words	**Count**
lazio	9169
buongiorno	8851
roma	8249
italia	7427
lavoro	5533
repubblica	5025
storia	4839
governo	4798
presidente	4672
salvini	4571
covid19	4547
foto	4454
draghi	4445

Table 4.9: Frequent Words - Full DataSets

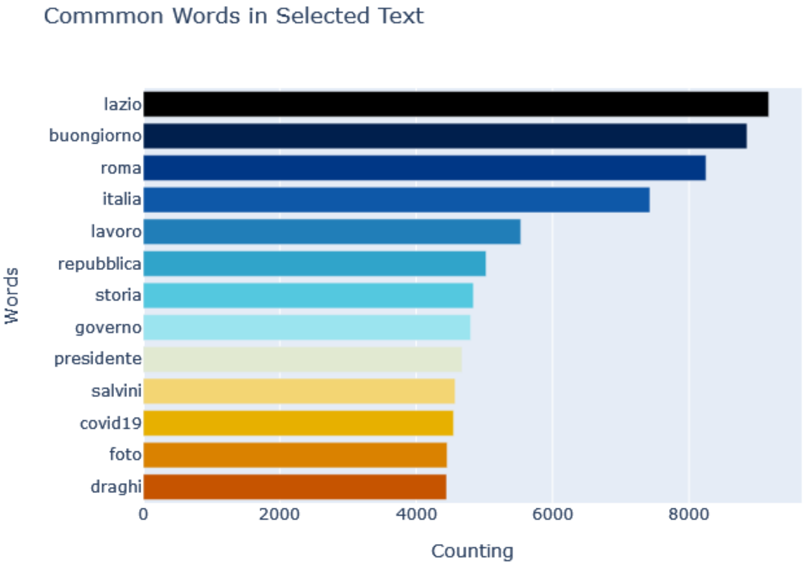


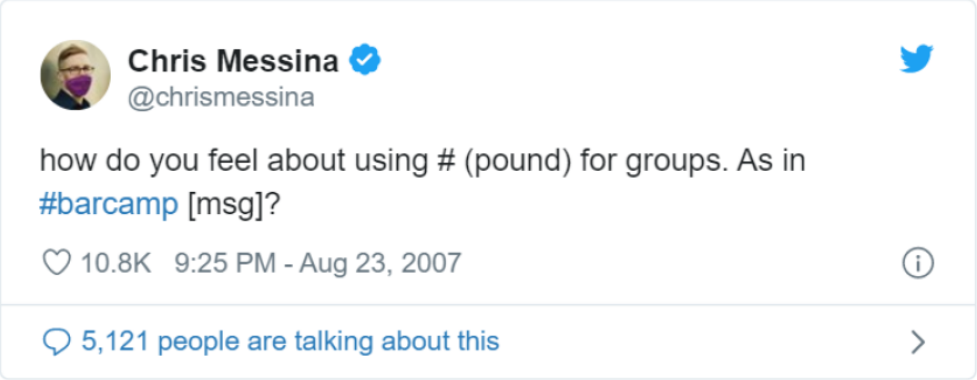
Figure 4.14: Frequent Words - Full DataSets

An analysis of all tweets was done, in order to understand the topics of the all entire downloaded tweets. By looking at the most frequent words in Figure 4.14 we can image three different main topics of discussions among users:

- football: lazio, roma;
- politics: repubblica, salvini, draghi, governo;
- daily news: lavoro, covid19, foto.

Further discussions about the topics of tweets and the separation in communities of the user will be done in Section 4.4.1.

Hashtags were introduced on the micro blogging platform as a way to classify tweets according to the topic. In this way users could search easily for a specific content and share information related to it. The very first Twitter hashtag was produced by social designer Chris Messina back in August 2007. The designer posted a tweet saying: "how do you feel about using # (pound) for groups. As in #barcamp [msg]?".



Chris Messina
@chrismessina

how do you feel about using # (pound) for groups. As in #barcamp [msg]?

♡ 10.8K 9:25 PM - Aug 23, 2007

5,121 people are talking about this

Figure 4.15: First Hashtag

A hashtag typically consists in a string of characters preceded by the pound symbol #, also called hash, and it could include numerical digits, creating a sort of label for the message itself, allowing the retrieval of all tweets dealing with the labeled topic. As a result, hashtags have become tools to find messages and take part in conversations, encouraging the creation of communities with the same interests, who wish to read and talk about the shared interest [41].

Nowadays, hashatgs are created by anyone who wants to summarize, or comment on a concept in few words. Specifically they are used in different ways like for example:

- promoting brands or events, like *#Euro2020*, *#Eurovision*;
- criticizing or praising ideas, *#vaccinoobbligatorio*;
- criticizing or praising people, *#conte*;
- spreading and providing updates on breaking news items, *#eruzione*, *#NotreDame*;

Besides, educators, social media experts and major companies from all around the world create new hashtags to bring in more followers.
Hashtags show up continuously on Twitter, becoming one of the main feature that characterizes the platform. Some of them have success and become very famous, used by people from anywhere in the world, while others die immediately after birth and are restricted to a few messages[42].

As it is possible to see from the two word clouds, even the most frequent hashtags reveal an interest for politics.
When analyzing the friends dataset, among the most used hashtags we find: *Salvini*, *Meloni*, *Draghi* who is the current Prime Minister of Italy since 13 February 2021, and took the place of *Conte*.
Other used hashatags are *Eurovision*, since during that period, Spring 2021, Italy was one of the country that participates to the 65th edition of the Eurovision Song Contest; *PatrickZaki*, a postgraduate student at the University of Bologna, who has been detained in Egypt since 7 February 2020.

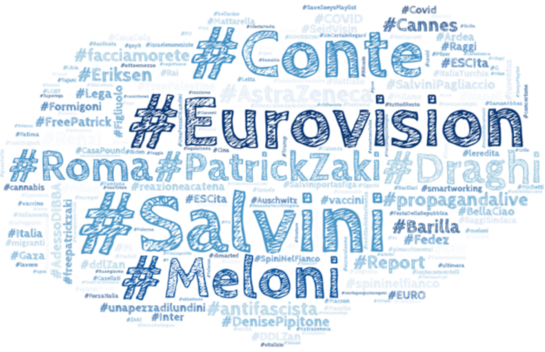


Figure 4.16: Hashtags Word Cloud - Friends

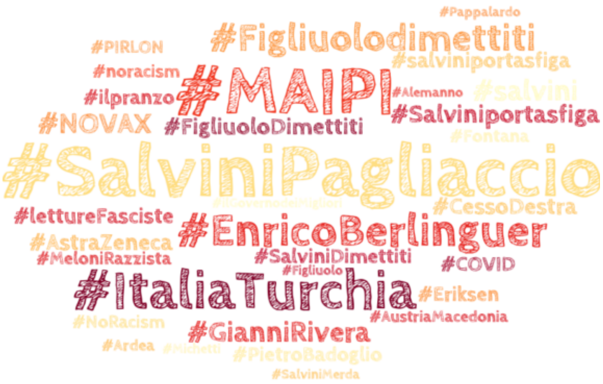


Figure 4.17: Hashtags Word Cloud - User

Thanks to frequents words analysis discussed before and the hashtags *SalviniPagliaccio*, *Salviniportasfiga*, *SalviniDimettiti* and *Salvinim***a* — Salviniclown, Salvini brings bad luck, Salvini resign, Salvinish**t — it's possible to perceive that the selected user doesn't agree with the politics of Salvini, and in general he doesn't sympathize with the actual Italian right-wing, since we can find also the hashtag *MeloniRazzista* — Meloni racist —.
Trought the hashtag *Figliuolodimettiti*, — Figliuolo resing — and *Covid*, we can imagine that the user expresses his feeling about the current Covid situation in Italy. Figliuolo is the new Extraordinary Commissioner for the Implementation of Health Measures to Contain the COVID-19 pandemic in Italy, appointed by Prime Minister Mario Draghi (since March 2021), and apparently the user doesn't like the way of organizing the implementation of the vaccination campaign against COVID-19 done by Figliuolo.

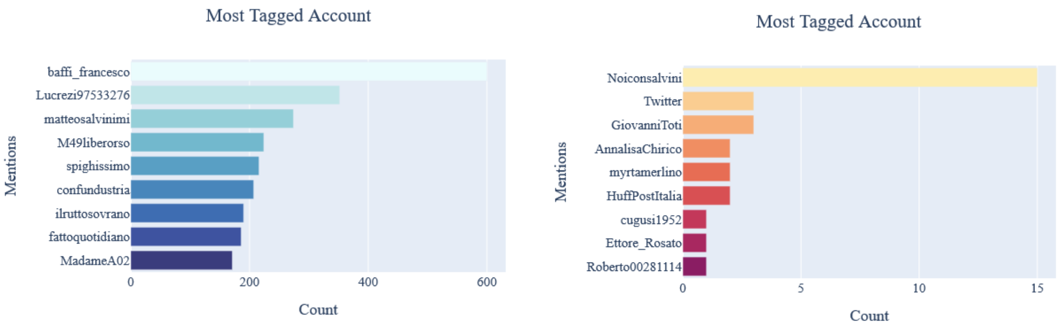


Figure 4.18: Mentions Count - Friends at left, User at right

In addition to analyzing the hashtags, the number of mentions received were also analyzed. The mentions were extracted from each tweet and a count was subsequently made.
The account *noiconsalvini* is the one who has gained the attention of the single user. "Noiconsalvini" is an official page that supports the Italian

politician. We can suppose that our user did tagged the account many times, because probably he replies to the contents published by the page that contain opinions which are not in line with his thinking. An other mentioned account is the one of *Giovanni Toti* President of the Liguria region.
The most tagged accounts in the other dataset are the ones of *baffifrancesco* and *Lucrezi97533276*, two people who declare in their Twitter profile as "Antifascista e Antirazzista" (anti-fascist and anti-racist). The account of the Salvini is, not surprisingly, the third most tagged account.

The accounts of *baffifrancesco* and *Lucrezi97533276* are among the top five most tagged accounts when analyzing the mentions of the dataset contained all the tweets. The three most tagged account are *OfficialSSLazio* (3225), the account of the Lazio football team, *repubblica* (1723) an Italian news report account, and we find once again the account of Salvini, *matteosalvinimi*(1486). Even his political party account, *LegaSalvini* (948) is one of the most mentioned one.

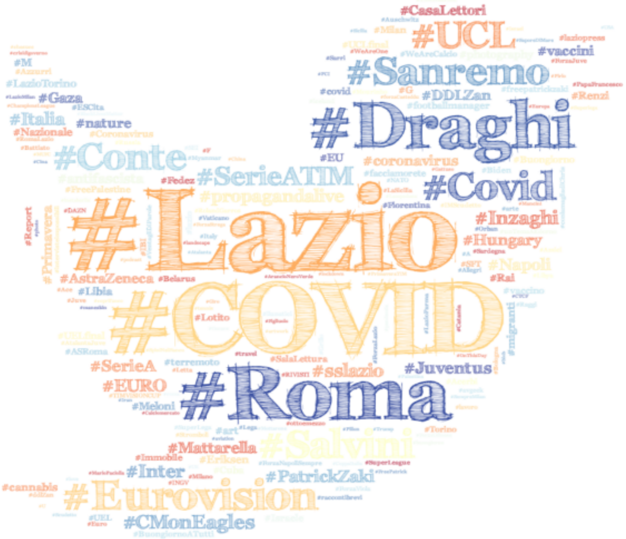


Figure 4.19: Hashtags Word Cloud - Entire Datasets

The word cloud displayed above, show us the most frequent hashatgs in all tweets. The hashtags *Lazio* is the most used one (2743), followed by *COVID* (1854) and *Roma* (1636). Others frequent hashtags are: *Draghi* (1590),

Eurovision (1385), *Salvini* (1333), *UCL*[3] (1271), *Conte* (1235).
The presence of these hashtags supports what was said when analyzing the most frequent words: users talk mainly of politics, soccer and daily news.

4.3.4 *Sentiment Analysis*

We dived the sentiment analysis task into two parts:

- discover emotion, trough emotion-oriented lexical resources, NCRLex, that should provide a list of words or expressions marked according to different emotion states;
- discover sentiment, trough a library called *feel-it* which is specific for Italian sentiment analysis.

4.3.4.1 *Emotion Detection with NRCL*

Emotion detection is an NLP task that has long been of interest to the field, and is usually conceived as a single - or multi - label classification in which zero (or more) emotion labels are assigned to variously defined semantic or syntactic subdivisions of the text.
The tweets were analyzed and processed through the indicator incorporated by NLTK, NRCL [43].

Emotion	**Description**
Positive	Comments contain supportive, cheerful and encouraging sentiment
Negative	Comments contain discouraging, dissatisfied and unhappy sentiment
Anticipation	Comment displays expectations towards the future. Expectations can be both optimistic and anxious
Anger	Comment contains annoyance, displeasure or hostility
Disgust	Comment displays strong disapproval aroused bu something unpleasant or offensive
Fear	Comment displays feeling of anxiety concerning future outcomes
Joy	Comment contains happiness and satisfaction
Sadness	Comment contains lower mood
Trust	Comment contains reliably and ability to believe in something

Table 4.10: Emotion Meaning

Thanks to NRCLex library, it was also possible to classify both emotion and sentiment of the tweets. The package contains approximately 27,000 words

[3] Uefa Champions League.

and is based on the National Research Council Canada affect lexicon and the NLTK library's WordNet synonym sets[44].
It was decided to translate this dictionary into Italian since the analyzed tweets are in Italian language. The detected emotions are anger, disgust, fear, sadness, trust, anticipation, joy and surprise, while the two sentiment are positive and negative, explained in Table 4.10, which came from the Plutchick wheel of emotions [45].
All these categories are not mutually exclusive, and hence, a word can be tagged according to multiple emotions or polarities (for example, the word applause is associated with various emotions such as surprise, trust and joy, as well as a positive feeling). Additionally, there are neutral words that are not associated with any emotion or polarity category.

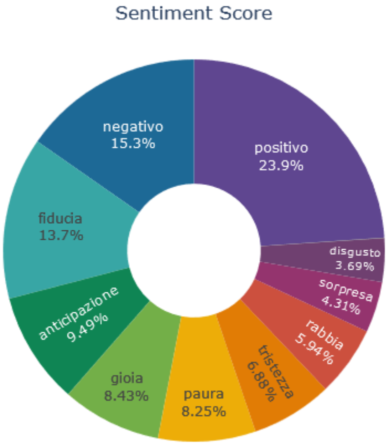


Figure 4.20: Sentiment Analysis - Entire Dataset

By looking at the Figure 4.20, it is possible to see that the most discovered feelings in the whole dataset are the *positive* (23,9%) and *negative* sentiments (15,3%), followed by *trust* (12,7%) and *anticipation* (9,49%) emotions. The least frequent emotion is the *disgust* one (3,69%).
When analyzing the single user tweets and his network of friends, the most frequent feelings are exactly the same ones, except that in the single user the emotion *anger* is more frequent if compared to the others.
Furthermore the emotion *disgust* both in the single user and his friends dataset is more frequent when comparing it to the whole dataset.
The complete distributions of sentiments and emotions are displayed in Figure 4.21.

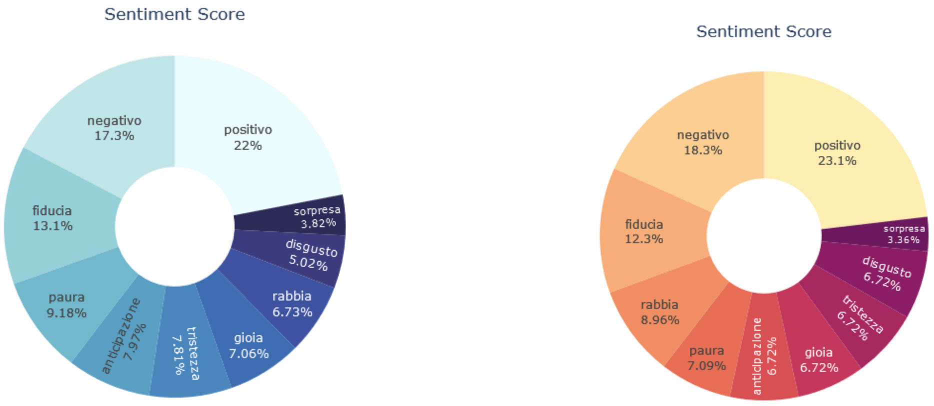


Figure 4.21: Sentiment Analysis - Friends and Single User

Since politics is one of the main topic in our datasets, it was decided to study the sentiment only for tweets that contain words related to politics, in order to understand how users feel when talking about the Italian political situation. At first it was created a list containing words connected to the italian right-wing like: *salvini*, *lega*, *meloni*, *governo*, *fratelli d'italia*, *partito*, *destra*. All the tweets were filtered and tweets which did not contained such words, were dropped.
Analyzing the dataset of the single user, from 200 tweets, 53 tweets contained the words in the created list. The situation is reversed: now there are more negative tweets (19,3%), and emotions like anger (14,7%), disgust (11%) and fear (10,1%) got a higher percentage.
We can notice that the percentage of positive feelings is still high: this happens because in the lexicon created by NRC there are plenty of words which are classified as positive anyway.

Moving on with the analysis, we do the same process for his network of friends. Considering tweets with only the selected words, we reduced the dimensionality of the dataset from about 50000 tweets into 7456 tweets done by 264 users. This numbers make us understand that the the 84% of the users have done at least one tweet containing those politic words. The Figure 4.22 at left, shows us that the majority of people are upset. If we compare the results with the analysis done in the original dataset, we notice that the negative (19.2%), fear (14%) and anger (6,94%) got a higher percentage once again. It is possible to draw the conclusion that both the user and most of his friends do not sympathize with the actual Italian right-wing.

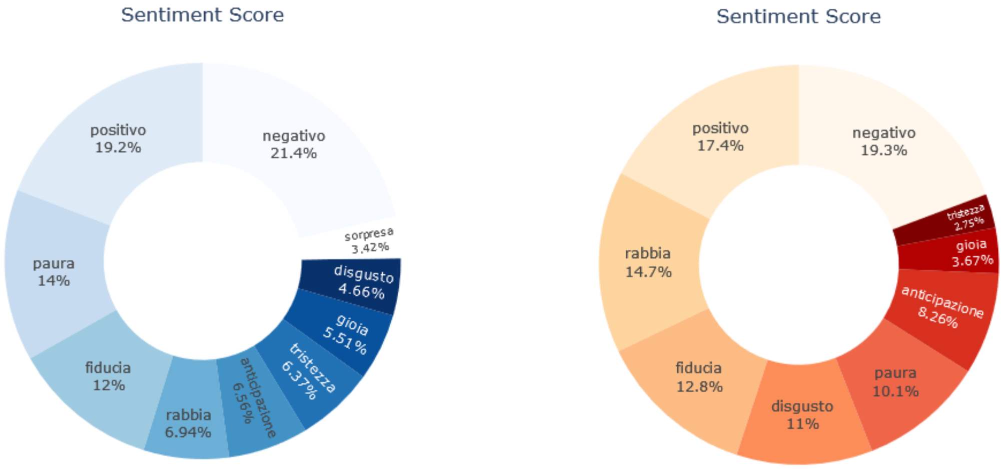


Figure 4.22: Sentiment Analysis Right-Wing - Friends and Single User

This process of selecting only some specific tweets was done with the whole dataset too. Among all the users, which were about 9000, about the 45% of them made at least one post containing the words contained in our list and the most detected sentiment was once again the negative one, overcoming the positive one found in the previous analysis.

4.3.4.2 *Sentiment Analysis with Feel-it*

Feel-it is a library recently created by Federico Bianchi, Debora Nozza and Dirk Hovy for Italian sentiment analysis [46]. The library wraps the HuggingFace internal APIs to provide a simple interface for emotion and sentiment prediction.
Feel-it provides an emotion classifiers, which detect 4 emotion: anger, fear, joy, sadness, and a sentiment classifier, used in our analysis that returns positive or negative sentiment.

The results obtained with Feel-it are the following: from the 200 tweets of the user, 156 were classified as negative, and only 44 are positive. When considering only the tweets with right-wing political references, the 53 tweets were classified exclusively with a negative sentiment as the Figure 4.23 shows. Examples of sentiment detection are shown in the Table 4.11.

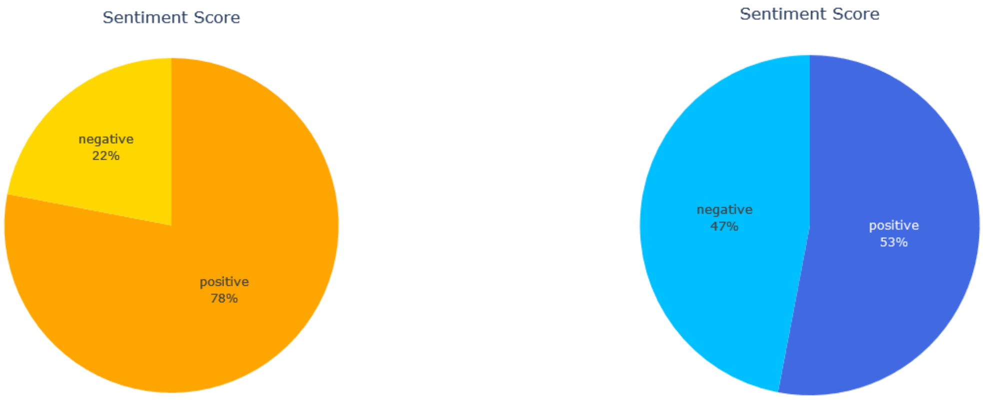


Figure 4.23: Sentiment Analysis Feel-it - Single User and Friends

Text	Emotion Detected
Qualcuno vuole rispondere a questo imbecille? @salvinimi Io non ce la faccio più!	negative
Siamo nel Far West, armi, fucili e pistole per tutti. Grazie Lega, grazie Salvini	negative
Trentasette anni fa...Nessuno più come lui. Pertini disse "lo porto via con me, come un figlio". #EnricoBerlinguer	positive
Posto un ricordo, indelebile	positive
La barzelletta del giorno. La Meloni che parla di illegalità, lei che ha nel suo partito delinquenti e mafiosi.	negative

Table 4.11: Sentiment Classification Example with Feel-it

The process was done with the dataset containing the tweets of all friends. At first, an analysis of the dataset was made without any tweet filtering. Feel-it rated 56% of tweets as positive, while the remaining 44% as negative. After applying the same kind of filtering to the tweets, as expected, the results change drastically: 82% of tweets are classified as negative while only 18% as positive.
Unfortunately it was not possible to perform an analysis of the complete dataset with Feel-it as the execution times would have been too long.

4.3.4.3 *Sentiment Analysis Conclusion*

From the analysis done with the two different libraries, NCRLex and Feel-it, we can say that both analysis gave the same results: when analysing the tweets without any changes, the most detected sentiment is the positive one, and the percentage of emotions like fear or anger is relatively low. But, if we considered only tweets that contain at least one word containing right-wing references, most tweets are classified as negative, and emotions like fear, anger and disgust obtained higher percentage.

4.4 NETWORK ANALYSIS

The network analysis aimed to create and characterize a social network, in order to see the connections among users. On Twitter, social networks are composed of users and the connections they form with other users when they mention and reply to one another [47].
In this framework, the link between nodes is created only if they have a specific kind of interaction, that is to say if a user u, which represents a node, has mentioned an other user u, i.e. the action of including a username in a tweet.
It was decided to study the network created by all the users. The graph was treated as a directed (if Alex mentioned John, a link from Alex to John is created), weighted graph, with 168339 nodes and 504806 links, and there are 2556 self-loops. The network characteristics are listed below:

Number of Nodes	168339
Number of Edges	504806
Weighted	Yes
Directed	Yes
Average Degree	2.985
Density	$1.78 \cdot 10^{-5}$
Number of self-loops	2556
Average Clustering Coefficient	0.025

Table 4.12: Characteristics of Twitter mention network

It was decided to compare the Twitter Network, with the Synthetic Network, Erdos-Renyi, Barabasi-Albert, Watts-Strogatz, and Configuration Model. To create the models, the predefined algorithms were used. With CM the parameters were the in and out degree, with ER the parameters were the number of nodes and edges, as well as the directed type, while for both WS and BA we set the number of nodes and some numeric values identified with a process of trial and error looking for the most similar number of edges as result. In Table 4.13 we find all the metrics of the Synthetic Networks.

	TN	ER	WS	BA	CM
Num of Nodes	168339	168339	168339	168339	168339
Num of Edges	504806	504806	505017	505008	501722

Table 4.13: Nodes and Edges of different networks

As regards the ER graph the numbers of nodes and edges are the same of our TN with a smaller CC value. The graphs created with the BA and WS models have higher number of edges, but both CC are lower respect to the TN. The CM model has the same number of nodes, but less number of edges. Comparing the degree distribution displayed in Figure 4.24 of our networks we noticed that the Configuration Model is the most similar one. This happens because CM is a random network model that completely relies on keeping the same degree as the RW network.

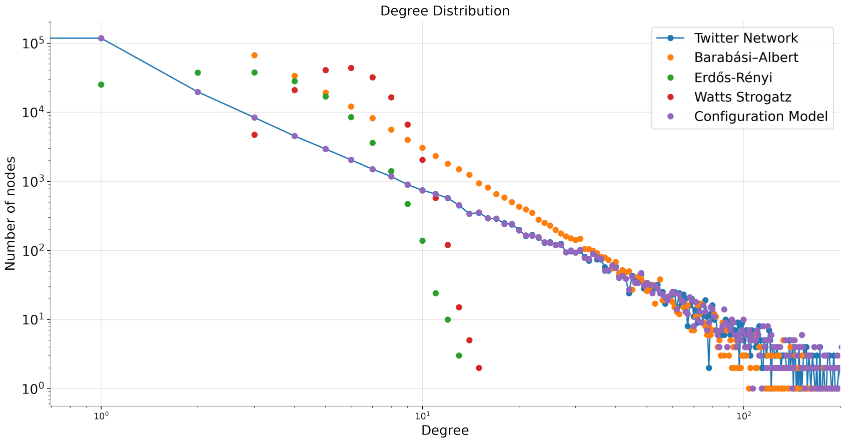


Figure 4.24: Degree Distribution

	TN	ER	WS	BA	CM
Cluster Coefficient	0.0250	0.0001	0.0780	0.0005	0.00017
Density	$1.78 \cdot 10^{-5}$	$1.78 \cdot 10^{-5}$	$3.56 \cdot 10^{-5}$	$3.56 \cdot 10^{-5}$	$1.56 \cdot 10^{-4}$

Table 4.14: Metrics of different networks

4.4.1 *Communities Detection in Twitter*

In order to estimate the importance of the filter bubbles phenomenon, it was decided to extract communities from the social graph with the Louvain method. To try to understand better what kind of communities these algorithms produce, we study the behavior of users regarding the community they belong to. First of all, in order to find communities and use the Louvain algorithm, which associate users to only a single community, it was necessary to treat the graph as an undirected one.

To explain the filter bubble effect, we try to understand the rationale for the formation of a community, so at first it was decided to label the communities according to their main features. To determine the labels of communities, it was decided to:

- see which are the most tagged accounts in each communities;
- analyze the words and hashtags present in the tweets of these users in order to catch the topic of debates, and analyze the sentiment of each communities.

It is in the nature of the Louvain algorithm to find a lower number of communities because it tends to avoid small communities. The algorithm manages to find 78 communities, with modularity $Q \simeq 0.53$, but it was decided to focus on and study only the most populated ones.

The biggest communities refers to some of the most debated themes and subjects during the analyzed period (late Spring 2021), in particular we found 5 communities displayed in the graph in Figure 4.25, made with Gephi, an open-source software for visualize and explore all kinds of graphs and networks.

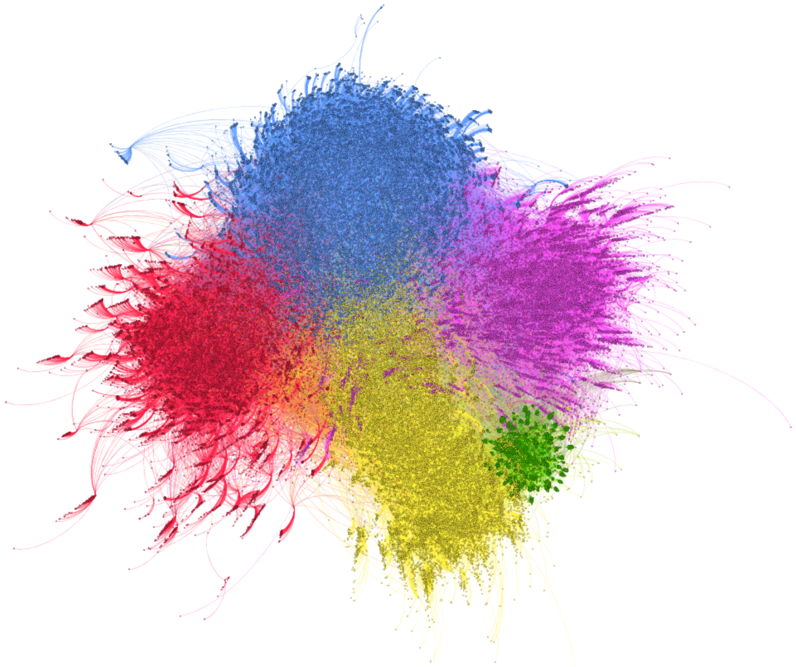


Figure 4.25: Community Detection with Gephi - Force Atlas 2 Layout. For better understanding the composition of the communities, the nodes that don't belong to those are not present in the Graph.

These communities are made of the:

- **Political Community**, in blue, 43651 nodes, 184159 links: topics of this community are political issues regarding the situation in Italy with no distinctions of political parties; some of the nodes that have higher degree values are: EnricoLetta (692), Matteosalvinimi (598), Matteorenzi (579), GiorgiaMeloni (551), CarloCalenda (522), pdnetwork (513), GiuseppeconteIT (451), RobertoBurioni (428), Lucrezi97533276 (437), Virginiaraggi (371), ItaliaViva (330), legasalvini (292), Luigidimaio (278), FratellidItalia (256). Beyond political figures and their parties, we find also newspaper headlines and press agencies like: Repubblica (1124), Corriere (692), LaStampa (485), ilfattoquotidiano (466), HuffPostItalia (382), La7Tv (376), RaiTre (315).
  Most used hashtags in this community are mainly hashtags that refer to Italian politics, like: *salvini*, *draghi*, *conte*, *mattarella*, *meloni*, but also others like: *propagandalive*, *rai*.

- **Football Communtity** in red, 29154 nodes, 79003 links: football is the main followed sport in Italy, and there's no surprise in finding a

community entirely focused on it. In particular it seems that a slice of the community is concentrated in the dispute between the two Italian teams, Rome and Lazio. In this community we find both team account, sports journalists and channels. Higher degree nodes are: OfficialSSLazio (442), RiccardoCucchi (404), Inter (346), FBiasin (334), SkySport (330), Gazzetta_it (316), SeriaA (279), OfficialASRoma (267), Juventusfcs (262).
The most used hashtags in this community are: *juventus*, *lazio*, *roma*, *ucl*, *laziotorino*.

- **Reading and entertainment Community** in yellow, 28605 nodes, 66091 links: in this community the main topic of conversations are reading and poetry. The YouTube (554) account was associated with this community. Other nodes are: Poesiaitaliana (284), Casalettori (275), Salalettura (173), LibriAmati (119), unTemaAlGiorno (89)[4].
Main hashtags in this community are: *untemaalgiorno*, *buongiornoatutti*, *casalettori*, *libridaleggere*.

- **American Community**, in pink, 23888 nodes, 31559 links: in this community we find mainly American users, indeed the detected language in this community is English. Nodes with higher degree values are: JoeBiden (221), POTUS (215), nytimes (214), AP (112), Washingtonpost (109), CNN (109).
Frequent hashtags are: *biden*, *news*, *travel*, *president*, *amazing*, *weareone*.

- **Music and show business Community** in green, 8529 nodes, 10834 links: the smallest community among the biggest one is composed mainly of tweets regarding music. During the download period of the tweets, Eurovision 2021 was held and one of the main topic in this community is indeed the exchange of opinions about singers. Higher degree nodes: RTL1025 (100), Sanremorai (71), vanityfairit (63), Fiorello (40), Eurovision (49), MarroneEmma (44), chiaraferragni (39), RadioItalia (37), EurovisionRai (35).
Within this community, there aren't many hashtags, but the ones used have a high frequency: *sanremo*, *eurovision*, *rai1*, *fantasenremo*, *musica*, *SerieATM*.

[4] In English: Italian poetry, reader house, room lecture, loved books, one topic at day.

An interesting fact is that the American community, even if it has more or less the same nodes of the Football and Reading communities, the number of links is way lower. This happens probably because connections are strictly limited within this community.
More details on the identity of the users mentioned above can be found in Appendix B.1.

5 PRESENTING THE RESULTS: THE DASHBOARD

In this section we will show the creation of a dashboard to view the results obtained from the analyzes carried out.
The created Dashboard is based on previously downloaded and locally saved data. Taking as input two datasets, the one of the single user, and the other related to his circle of friends, a series of analyzes are carried out and the results are shown through plots and cards created thanks to the Dash Core Components and Dash BootStrap Components. The creation of the Dashboard, was conducted on PyCharm, an IDE used in computer programming, specifically for the Python language.

The dashboard shows through simple plots – which displays for example the hours of use of twitter, the days in which the platform is used the most, the most used words and hashtags in his last 200 tweets, as well as the number of people tagged most – the behavior of the individual user, in relation to his friends (so the analyzes are made for both datasets). Finally, a sentiment analysis is made which shows the various emotions detected in the latest tweets.

As stated before, these analyses are made only with local saved dataset, but a future development of the dashboard could be to download data in real time: by entering the name of a user, and downloading the tweets of friends, it could be possible to understand if the selected user has a similar or completely different attitude respect to them, trying to answer questions that he might ask himself such as:

- Do I use Twitter in the same way as my friends?

- Do I usually tweet more in the evening or during the day? During the weekend or at the beginning of the week?
- Do I usually write long or short posts compared to those of my friends?
- Do my friends have more friends than me, or less?
- Are my friends and I interested in the same topics? Do we talk about same things? Do we share same interests?
- Do my tweets express positive or negative feelings?

At first, we will give an overview of what Dash is and how it works, then we will show in detail how the created Dashboard looks like, going deeper in the structure of the page.

5.1 WHAT IS DASH AND HOW IT WORKS

Dash, developed by Plotly Tecnologies Inc., is a user interface library for creating analytical web applications with HTML and CSS but completely written in Python. Behind the scenes, Dash uses Flask as the back-end server and React as the front-end JavaScript framework. Used by over 4K projects, it is one of the most popular library[1].
The documentation[2] is pretty expensive and covers various usage patterns in depth, allowing to create also complex applications.

Dash components are Python classes that encode the properties and values of a specific React components that serialize as JSON. Dash provided a toolset to easily package React components, which are written in JavaScript, as components that can be easily used in Dash. This toolset uses dynamic programming to automatically generate standard Python classes from annotated React propTypes. Each Dash app has two main parts:

- the layout, which determines the visual components displayed on the Dash app;

[1] https://github.com/plotly/dash
[2] https://dash.plotly.com/introduction

- the callback function, that connects the Dash components and defines their interactive features.

Here there are a few of the libraries in which Dash is broken down, used also for the realization of this project:

- Dash Core Components, responsible for high-level components like graphs, dropdowns, sliders, check-boxers, etc. ;
- Dash HTML Components, which contains almost every HTML tag like *div*, *button* and even *script*;
- Dash DataTable, a library for creating tables that are highly interactive and much like spreadsheet where it is possible to edit values, add or remove lines or columns, filter and more.
- Dash BootStrap Components, an independent community project that ports the BootStap project into Dash, giving access to components like modals, navbars, tabs, cards and more;
- Dash Core Components for Visualization, another external project that provides a few extra components like a component to run JavaScript, a network component and a data-table.

The layout has basically the structure of a tree of components. We use the keyword **layout** of the app to specify its layout. Then, using the two libraries, *dash_html_components* and *dash_core_components*, we can display the components on our dashboard.

5.2 DASHBOARD: TWITTER ANALYSIS

The dashboard created for this project starts with an H1 heading (html.H1) as the title of the dashboard and a H2 as subtitle.

In this framework there is no one single app, but a multi page app, composed of four different layouts:

- index page, where the user decides which analyzes to see;
- user analysis page, where it is possible to find the statics of a single user;

- friends analysis page where it is possible to find the metrics of his friends;
- complete analysis page, where both analysis are present in order to have a whole view of the metrics.

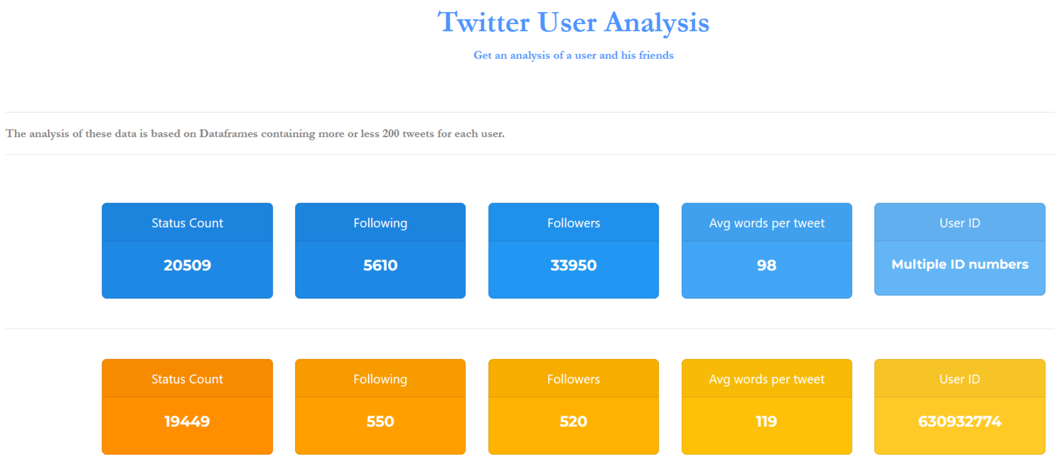
These layouts are divided into 3 different pyhton project, each one of it containing the dashboard of the single user, the one of his friends, and the other which shows both results, and another one, in which all the previuos layouts are imported.
The change of layout is possible due to the application of the callback functions. They are Python functions, but they get automatically called by Dash whenever its input changes. As a result, the function runs and updates its output as the code 2 shows.

```

@app.callback(Output(’page-content’, ’children’),
              Input(’url’, ’pathname’))

def display_page(pathname):
    if pathname == "/page-1":
        return layout_page_1
    if pathname == ’/page-2’:
        return layout_page_2
    elif pathname == "/page-3":
        return layout_page_3
    else:
        return layout_index
```

Listing 2: Python code for switching layout

The two main sections of the callback function are: the decorator which starts with *@app.callback*, and the function itself starts with def.

Within the decorator *@app.callback*, we specify the *Output* and the *Input* objects of the callback function, which are both the properties of Dash components.



Twitter User Analysis

Get an analysis of a user and his friends

The analysis of these data is based on Dataframes containing more or less 200 tweets for each user.

Status Count	Following	Followers	Avg words per tweet	User ID
20509	5610	33950	98	Multiple ID numbers

Status Count	Following	Followers	Avg words per tweet	User ID
19449	550	520	119	630932774

Figure 5.1: Cards in the Dashboard: Friends Metrics in blue, Single User Metrics in Orange

When choosing what layout to see, we find, at the top of the page, five different cards, created with the Bootstrap library. These five cards show respectively:

- the unique ID of a user;
- the status count: the exact number of status until the day the downloading of data;
- the number of followings;
- the number of followers;
- the average of the total words used in a tweet.

When we go to analysis of the user's friends network page, all the written metrics are to be understood as average values.

Scrolling through the dashboard, there are other information about the behaviour of the user behaviour on Twitter. The three different histograms tell us the most used source for tweeting, and when the user usually tweets, indicating the distribution of the days of the week and the top hours. When displaying the page of both user and friends analysis, the histograms are shown like the Figure 5.2.

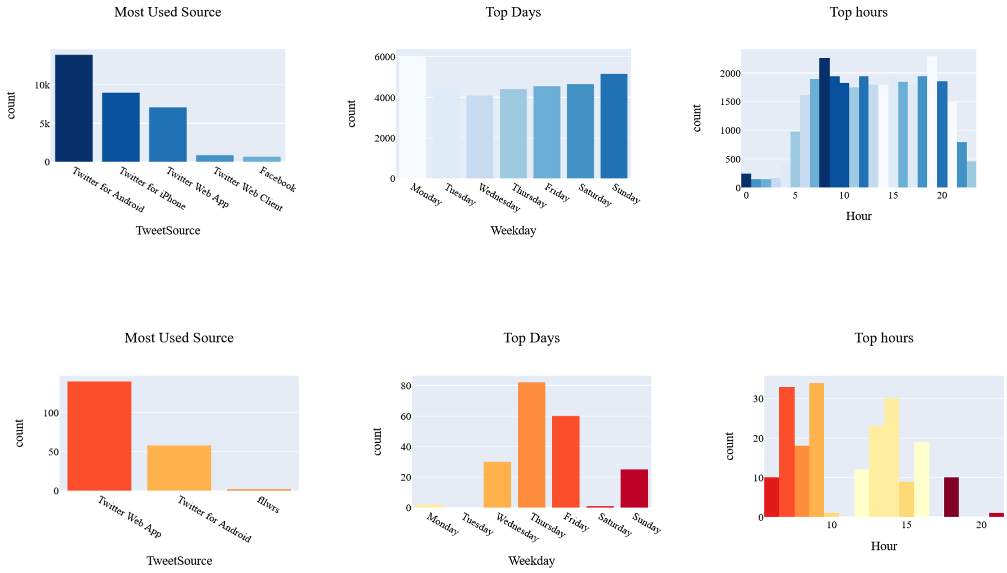


Figure 5.2: Behaviour Metrics: Friends Metrics in blue, Single User Metrics in Orange

Following these analyses, there are two count plots, which show the most frequent words and the most tagged accounts.

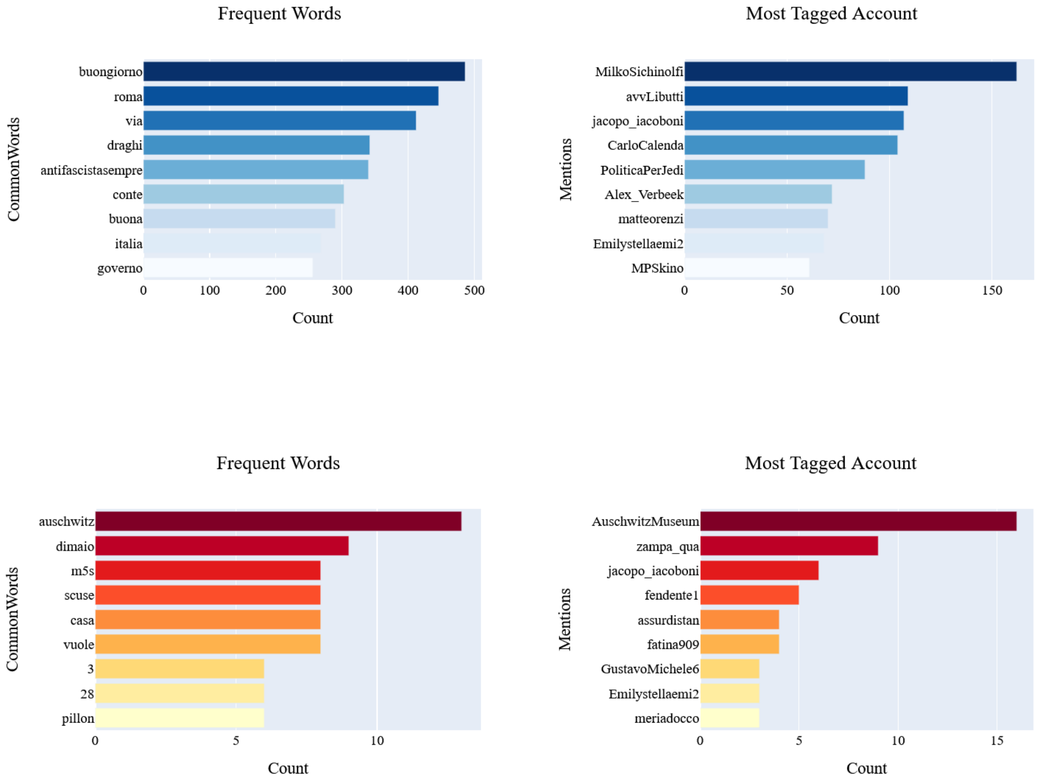


Figure 5.3: Counting Plots: Friends Metrics above, Single User Metrics below.

Moving on, we find two pie plots, one indicating the sentiment analysis made with NRCLex library (explained in Subsection 4.3.4.1), and the other that shows the polarity of the tweets.

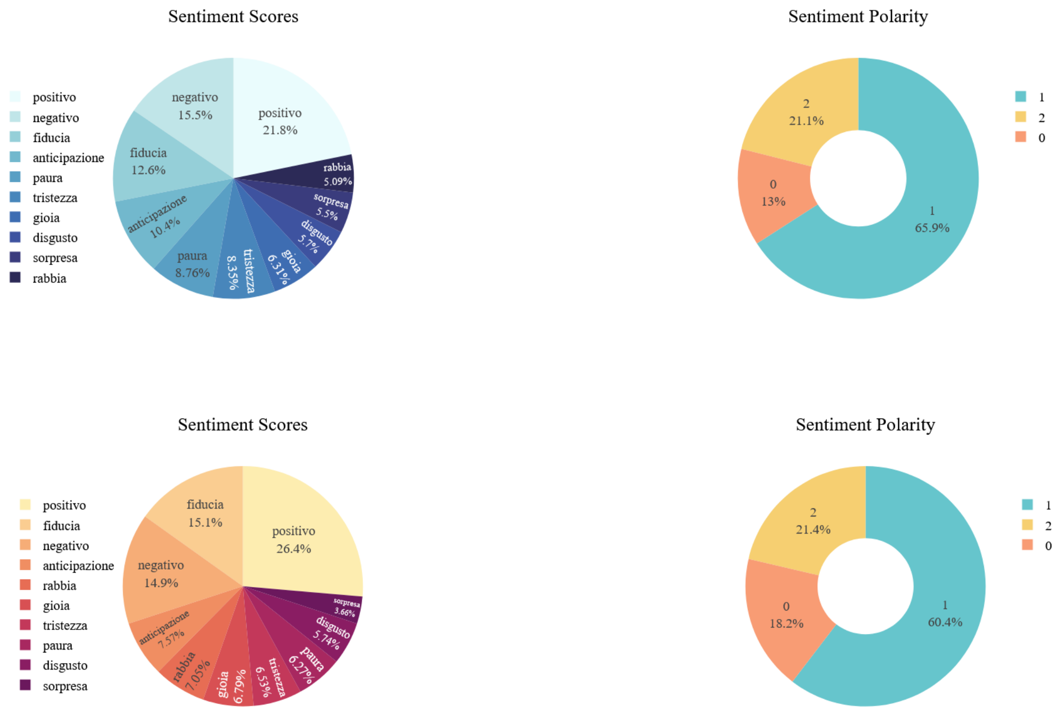


Figure 5.4: Sentiment Emotion at left, Polarity at right.

At the end of the dashboard, we find a word cloud of the most used hashtags wrapped in the shape of Twitter logo.

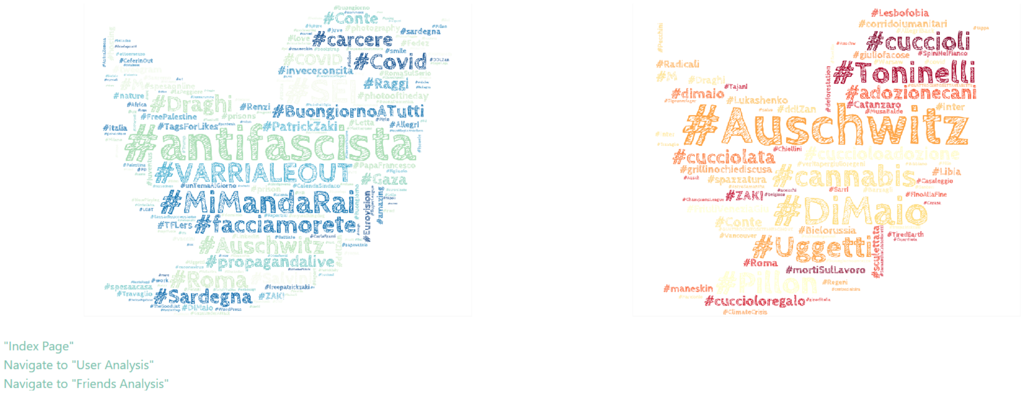


Figure 5.5: Hashtags Word Clouds: Friends Word cloud above, Single User Word cloud below.

As we can see from Figure 5.5, below the hashtags word clouds, there are 3 clickable links - Index Page, User Analysis and Friends Analysis - that redirect to the respective pages.

6 CONCLUSIONS

The aim of this thesis was to make an analysis on Twitter users for the identification and characterization of individual users, making a comparison with their list of friends, to verify if the attitude of the individual was uniform to the one of his friends, or not. The creation of a Dashbord was another main purpose of this project for giving a visual perspective of obtained results.

Several users were analyzed, focusing the study both on the use behaviour of the platform through statistical analyzes, and by analyzing the content of the posts. The statistical analyzes mainly concerned the discovery of attributes of individual users such as hours of use of the platform, which device was used to tweet, in which days Twitter was used most. Regarding the content analysis, we mainly focused on the count of unique words - which had a semantic meaning for the characterization of the topics in the tweets - as well as most used hashtags in the posts.
A sentiment analysis was also carried out using an indicator incorporated on NLTK, specifically with NCRL, and the use of a recently published library, Feel-it, created ad hoc for sentiment analysis on Italian texts, which returned respectively the emotions and sentiments of the single tweets. This analysis was essential for getting a general picture of the emotions that the tweets conveyed, analyzing also the emotions express regarding specific contents like Italian politics.

Having downloaded such an amount of data that could allow the creation of a social network, it was decided to create one, built through the mentions/user relationship. The basic metrics on the graph obtained were calculated, and they were compared with the synthetic networks. The main task of the SNA was aimed at finding the presence of communities within it. Through the Louvain algorithm, - which focuses its research on the optimization of modularity and binds a node to one and only one community, - 78 communities were found, concentrating the study only on the 5 most populated ones. It was seen how in these 5 communities were quite separate from each other, and the interactions between people belonging to two different communities were limited.

The statistics created were shown via a dashboard created with the help of Dash. The dashboard consists of 3 pages: one shows the analyzes relating to the individual user, the second one, those relating to the friends of the analyzed user, and finally a third page, which shows the analysis as a whole to have a more immediate comparison of the differences, or similar behaviors, that both parties may have.

There could be several future developments of this work: adapt the entire methodology proposed for the analysis of other platforms, and see if in other social networks, the presence of homophily among users is greater or less than in the Twitter platform. In addition, the Dashboard could be improved, allowing the download of data and real-time analysis of users, providing a tool for user to analyze his behavior on the SN and compare it with the one of his friends.

A

APPENDIX 1

A.1 LIST OF STOPWORDS

1 - a
2 - abbastanza
3 - abbia
4 - abbiamo
5 - abbiano
6 - abbiate
7 - accidenti
8 - ad
9 - adesso
10 - affinché
11 - agl
12 - agli
13 - ahime
14 - ahimè
15 - ai
16 - al
17 - alcuna
18 - alcuni
19 - alcuno
20 - all
21 - alla
22 - alle
23 - allo
24 - allora
25 - altre

25 - altri
26 - altrimenti
27 - altro
28 - altrove
29 - altrui
30 - anche
31 - ancora
32 - anni
33 - anno
34 - ansa
35 - anticipo
36 - assai
37 - attesa
38 - attraverso
39 - avanti
40 - avemmo
41 - avendo
42 - avente
43 - aver
44 - avere
45 - averlo
46 - avesse
47 - avessero
48 - avessi
49 - avessimo

26 - aveste
27 - avesti
28 - avete
29 - aveva
30 - avevamo
31 - avevano
32 - avevate
33 - avevi
34 - avevo
35 - avrai
36 - avranno
37 - avrebbe
38 - avrebbero
39 - avrei
40 - avremmo
41 - avremo
42 - avreste
43 - avresti
44 - avrete
45 - avrà
46 - avrò
47 - avuta
48 - avute
49 - avuti
50 - avuto

51 - basta
52 - ben
53 - bene
54 - benissimo
55 - brava
56 - bravo
57 - buono
58 - c
59 - caso
60 - cento
61 - certa
62 - certe
63 - certi
64 - certo
65 - che
66 - chi
67 - chicchessia
68 - chiunque
69 - ci
70 - ciascuna
71 - ciascuno
72 - cima
73 - cinque
74 - cio
75 - cioe

76 - cioè
77 - circa
78 - citta
79 - città
80 - ciò
81 - co
82 - codesta
83 - codesti
84 - codesto
85 - cogli
86 - coi
87 - col
88 - colei
89 - coll
90 - coloro
91 - colui
92 - come
93 - cominci
94 - comprare
95 - comunque
96 - con
97 - concernente
98 - conclusione
99 - consecutivi
100 - consecutivo

101 - consiglio
102 - contro
103 - cortesia
104 - cos
105 - cosa
106 - cosi
107 - così
108 - cui
109 - d
110 - da
111 - dagl
112 - dagli
113 - dai
114 - dal
115 - dall
116 - dalla
117 - dalle
118 - dallo
119 - dappertutto
120 - davanti
121 - degl
122 - degli
123 - dei
124 - del
125 - dell

126 - della
127 - delle
128 - dello
129 - dentro
130 - detto
131 - deve
132 - devo
133 - di
134 - dice
135 - dietro
136 - dire
137 - dirimpetto
138 - diventa
139 - diventare
140 - diventato
141 - dopo
142 - doppio
143 - dov
144 - dove
145 - dovra
146 - dovrà
147 - dovunque
148 - due
149 - dunque
150 - durante
151 - e
152 - ebbe
153 - ebbero
154 - ebbi
155 - ecc
156 - ecco
157 - ed
158 - effettivamente
159 - egli
160 - ella
161 - entrambi
162 - eppure
163 - era
164 - erano
165 - eravamo
166 - eravate
167 - eri
168 - ero
169 - esempio
170 - esse
171 - essendo
172 - esser
173 - essere
174 - essi
175 - ex
176 - fa
177 - faccia
178 - facciamo
179 - facciano
180 - facciate
181 - faccio
182 - facemmo
183 - facendo
184 - facesse
185 - facessero
186 - facessi
187 - facessimo
188 - faceste
189 - facesti
190 - faceva
191 - facevamo
192 - facevano
193 - facevate
194 - facevi
195 - facevo
196 - fai
197 - fanno
198 - farai
199 - faranno
200 - fare
201 - farebbe
202 - farebbero
203 - farei
204 - faremmo
205 - faremo
206 - fareste
207 - faresti
208 - farete
209 - farà
210 - farò
211 - fatto
212 - favore
213 - fece
214 - fecero
215 - feci
216 - fin
217 - finalmente
218 - finche
219 - fine
220 - fino
221 - forse
222 - forza
223 - fosse
224 - fossero
225 - fossi
226 - fossimo
227 - foste
228 - fosti
229 - fra
230 - frattempo
231 - fu
232 - fui
233 - fummo
234 - fuori
235 - furono
236 - futuro
237 - generale
238 - gente
239 - gia
240 - giacche
241 - giorni
242 - giorno
243 - giu
244 - già
245 - gli
246 - gliela
247 - gliele
248 - glieli
249 - glielo
250 - gliene
251 - grande
252 - grazie
253 - gruppo
254 - ha
255 - haha
256 - hai
257 - hanno
258 - ho
259 - i
260 - ie
261 - ieri
262 - il
263 - improvviso
264 - in
265 - inc
266 - indietro
267 - infatti
268 - inoltre
269 - insieme
270 - intanto
271 - intorno
272 - invece
273 - io
274 - l
275 - la
276 - lasciato
277 - lato
278 - le
279 - lei
280 - li
281 - lo
282 - lontano
283 - loro
284 - lui
285 - lungo
286 - luogo
287 - là
288 - ma
289 - macche
290 - magari
291 - maggior
292 - mai
293 - male
294 - malgrado
295 - malissimo
296 - me
297 - medesimo
298 - mediante
299 - meglio
300 - meno
301 - mentre
302 - mesi
303 - mezzo
304 - mi
305 - mia
306 - mie
307 - miei
308 - mila
309 - miliardi
310 - milioni
311 - minimi
312 - mio
313 - modo
314 - molta
315 - molti
316 - moltissimo
317 - molto
318 - momento
319 - mondo
320 - ne
321 - negl
322 - negli
323 - nei
324 - nel
325 - nell
326 - nella
327 - nelle
328 - nello
329 - nemmeno
330 - neppure
331 - nessun
332 - nessuna
333 - nessuno
334 - niente
335 - no
336 - noi
337 - nome
338 - non
339 - nondimeno
340 - nonostante
341 - nonsia
342 - nostra
343 - nostre
344 - nostri
345 - nostro
346 - novanta
347 - nove
348 - nulla
349 - nuovi
350 - nuovo
351 - o
352 - od
353 - oggi
354 - ogni
355 - ognuna
356 - ognuno
357 - oltre
358 - oppure
359 - ora
360 - ore
361 - osi
362 - ossia
363 - ottanta
364 - otto
365 - paese
366 - parecchi
367 - parecchie
368 - parecchio
369 - parte
370 - partendo
371 - peccato
372 - peggio
373 - per
374 - perche
375 - perchè
376 - perché
377 - percio
378 - perciò
379 - perfino
380 - pero
381 - persino
382 - persone
383 - però
384 - piedi
385 - pieno
386 - piglia
387 - piu
388 - piuttosto
389 - più
390 - po
391 - pochissimo
392 - poco
393 - poi
394 - poiche
395 - possa
396 - possedere
397 - posteriore
398 - posto
399 - potrebbe
400 - preferibilmente
401 - presa
402 - press
403 - prima
404 - primo
405 - principalmente
406 - probabilmente
407 - promesso
408 - proprio
409 - puo
410 - pure
411 - purtroppo
412 - può
413 - qua
414 - qualche
415 - qualcosa
416 - qualcuna
417 - qualcuno
418 - quale
419 - quali
420 - qualunque
421 - quando
422 - quanta
423 - quante
424 - quanti
425 - quanto

426 - quantunque
427 - quarto
428 - quasi
429 - quattro
430 - quel
431 - quella
432 - quelle
433 - quelli
434 - quello
435 - quest
436 - questa
437 - queste
438 - questi
439 - questo
440 - qui
441 - quindi
442 - quinto
443 - realmente
444 - recente
445 - recentemente
446 - registrazione
447 - relativo
448 - riecco
449 - rispetto
450 - salvo
451 - sara
452 - sarai
453 - saranno
454 - sarebbe
455 - sarebbero
456 - sarei
457 - saremmo
458 - saremo
459 - sareste
460 - saresti
461 - sarete
462 - sarà
463 - sarò
464 - scola
465 - scopo
466 - scorso
467 - se
468 - secondo
469 - seguente
470 - seguito
471 - sei
472 - sembra
473 - sembrare
474 - sembrato
475 - sembrava
476 - sembri
477 - sempre
478 - senza
479 - sette
480 - si
481 - sia
482 - siamo
483 - siano
484 - siate
485 - siete
486 - sig
487 - solito
488 - solo
489 - soltanto
490 - sono
491 - sopra
492 - soprattutto
493 - sotto
494 - spesso
495 - sta
496 - stai
497 - stando
498 - stanno
499 - starai
500 - staranno
501 - starebbe
502 - starebbero
503 - starei
504 - staremmo
505 - staremo
506 - stareste
507 - staresti
508 - starete
509 - starà
510 - starò
511 - stata
512 - state
513 - stati
514 - stato
515 - stava
516 - stavamo
517 - stavano
518 - stavate
519 - stavi
520 - stavo
521 - stemmo
522 - stessa
523 - stesse
524 - stessero
525 - stessi
526 - stessimo
527 - stesso
528 - steste
529 - stesti
530 - stette
531 - stettero
532 - stetti
533 - stia
534 - stiamo
535 - stiano
536 - stiate
537 - sto
538 - su
539 - sua
540 - subito
541 - successivamente
542 - successivo
543 - sue
544 - sugl
545 - sugli
546 - sui
547 - sul
548 - sull
549 - sulla
550 - sulle
551 - sullo
552 - suo
553 - suoi
554 - tale
555 - tali
556 - talvolta
557 - tanto
558 - te
559 - tempo
560 - terzo
561 - th
562 - ti
563 - titolo
564 - tra
565 - tranne
566 - tre
567 - trenta
568 - triplo
569 - troppo
570 - trovato
571 - tu
572 - tua
573 - tue
574 - tuo
575 - tuoi
576 - tutta
577 - tuttavia
578 - tutte
579 - tutti
580 - tutto
581 - uguali
582 - ulteriore
583 - ultimo
584 - un
585 - una
586 - uno
587 - uomo
588 - va
589 - vai
590 - vale
591 - vari
592 - varia
593 - varie
594 - vario
595 - verso
596 - vi
597 - vicino
598 - visto
599 - vita
600 - voi
601 - volta
602 - volte
603 - vostra
604 - vostro
605 - vostri

B APPENDIX 2

B.1 DESCRIPTION OF THE MENTIONED USERS

In Section 4.4.1 we explained the composition of the biggest communities found in the graph. Below we try to explain in detail the biggest nodes in each communities for better understating how they are characterize.

In the **Political Community** we find the different wings in the Italian Political Sphere. Below we find how these wings and parties are divided:

- right-wing parties, which are Lega and Fratelli d'Italia, guided respectively by Matteo Salvini and Giorgia Meloni;
- centre-left parties, which are Italian Democratic Party (PD), a social-democratic political party in Italy, whose secretary is Enrico Letta, elected by the national assembly in March 2021, after the resignation of the former leader Nicola Zingaretti; Italia Viva is a liberal political party in Italy founded by the former prime minister and former PD secretary Matteo Renzi;
- Virginia Raggi was the major of Rome until October 2021, and she is on of the member of the anti-establishment Five Star Movement (M5S), together with Luigi Di Maio, who he was the leader of the M5S, from September 2017 to January 2020, and since September 2019 he's serving as Minister of Foreign Affairs.

In this network we find also one of the main figure in the political discussion during the 2021 Spring, Giuseppe Conte, who served as Prime Minister of Italy from June 2018 to February 2021[1]. Furthermore the figure

[1] Conte has been also the president of the Five Star Movement since August 2021, but since the downloaded tweets were prior to this assignment, he was not included in the political division made above.

of Roberto Burioni is present too, an Italian virologist, physician and academic, who was particularly active on the popularization on subjects related to the pandemic situation in Italy.
Carlo Calenda is a member of the European Parliament since July 2019, and On 18 October 2020, he announced his intention to run as Mayor of Rome in the 2021 municipal election.
La Repubblica, Corriere della Sera, La Stampa, Il Fatto Quotidiano, Huff Post Italia are all online newspaper on politics, news, economics. La7Tv is an Italian free-to-air television channel which broadcasts several political television programs like "Non è l'Arena", "Tagadà" and "In Onda".

The **Reading and entertainment Community** is made up of accounts regarding poetry and in general of reading/writing accounts. Casalettori is a space for cultural sharing conceived and curated by Maria Anna Patti, as well as Salalettura (called "Il caffè Letterario") which its bio states that it is a literary meeting point.

The **Football Community** is composed mainly of football teams accounts like the one of Roma, Lazio, Inter and Juventus. Other main nodes are the on of FBiasin, the account of Fabrizio Biasin, a sports commentator and Riccardo Cucchi, a journalist and former sports commentator.

The **American Community** is composed mainly of American accounts: POTUS is the account of the current President of the USA, Joe Biden, who was inaugurated as the 46th president of the United States on January 20, 2021; Associated Press (AP), Washington Post, The New York Times (nytimes) and CNN are all US online newspapers.

The **Music and show business Community** contains node regarding mainly music. The account of Eurovision (an international songwriting competition organised annually by the European Broadcasting Union, featuring participants representing primarily European countries) and Sanremorai (the most popular Italian song contest and awards ceremony, held annually in the city of Sanremo, Liguria, whose winner partecipates to the Eurovision contest) belong to this community. In the 2021 edition, Sanremo was conducted by Fiorello, present in this commnunity too.
RTL 102.5 and Radio Italia are private Italian radio stations.

BIBLIOGRAPHY

[1] S. Hong and S. H. Kim, 'Political polarization on twitter: Implications for the use of social media in digital governments', *Government Information Quarterly*, vol. 33, no. 4, pp. 777–782, 2016.

[2] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer and A. Flammini, 'Political polarization on twitter', in *Fifth international AAAI conference on weblogs and social media*, 2011.

[3] A. Urman, 'Context matters: Political polarization on twitter from a comparative perspective', *Media, culture & society*, vol. 42, no. 6, pp. 857–879, 2020.

[4] A. Evette. 'Polarization in the twittersphere: What 86 million tweets reveal about the political makeup of american twitter users and how they engage with news'. (), [Online]. Available: `https://knightfoundation.org/articles/polarization-in-the-twittersphere-what-86-million-tweets-reveal-about-the-political-makeup-of-american-twitter-users-and-how-they-engage-with-news/` (visited on 08/12/2021).

[5] C. Farr. 'Jack dorsey: 'twitter does contribute to filter bubbles' and 'we need to fix it'. (), [Online]. Available: `https://www.cnbc.com/2018/10/15/twitter-ceo-jack-dorsey-twitter-does-contribute-to-filter-bubbles.html` (visited on 15/09/2021).

[6] S. Perez. 'Twitter's doubling of character count from 140 to 280 had little impact on length of tweets'. (), [Online]. Available: `shorturl.at/ehsKZ` (visited on 07/10/2021).

[7] E. Pariser, *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[8] P. M. Dahlgren, 'A critical review of filter bubbles and a comparison with selective exposure', *Nordicom Review*, vol. 42, no. 1, pp. 15–33, 2021.

[9] D. LR, A. Tamhane and N. Pervin, 'A clustering based social matrix factorization technique for personalized recommender systems', 2018.

[10] R. K. Garrett, 'The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation', *Journal of Applied Research in Memory and Cognition*, vol. 6, no. 4, pp. 370–376, 2017.

[11] J. Möller, D. Trilling, N. Helberger and B. van Es, 'Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity', *Information, Communication & Society*, vol. 21, no. 7, pp. 959–977, 2018.

[12] E. Bakshy, S. Messing and L. A. Adamic, 'Exposure to ideologically diverse news and opinion on facebook', *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[13] C. Cadwalladr and E. Graham-Harrison, 'Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach', *The guardian*, vol. 17, p. 22, 2018.

[14] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, 'The science of fake news', *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[15] A. Amrollahi *et al.*, 'A conceptual tool to eliminate filter bubbles in social networks', *Australasian Journal of Information Systems*, vol. 25, 2021.

[16] E. Dubois and G. Blank, 'The echo chamber is overstated: The moderating effect of political interest and diverse media', *Information, communication & society*, vol. 21, no. 5, pp. 729–745, 2018.

[17] H. Allcott and M. Gentzkow, 'Social media and fake news in the 2016 election', *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[18] E. Bozdag and J. Timmermans, 'Values in the filter bubble ethics of personalization algorithms in cloud computing', in *1st international workshop on values in design–Building bridges between RE, HCI and ethics*, vol. 296, 2011.

[19] M. Bastos, D. Mercea and A. Baronchelli, 'The geographic embedding of online echo chambers: Evidence from the brexit campaign', *PloS one*, vol. 13, no. 11, e0206841, 2018.

[20] J. An, D. Quercia and J. Crowcroft, 'Partisan sharing: Facebook evidence and societal consequences', in *Proceedings of the second ACM conference on Online social networks*, 2014, pp. 13–24.

[21] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi and M. Starnini, 'The echo chamber effect on social media', *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, 2021.

[22] R. D. Wimmer and J. R. Dominick, *Mass media research*. Cengage learning, 2013.

[23] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[24] H. Julien, 'Content analysis', *The SAGE encyclopedia of qualitative research methods*, vol. 1, pp. 120–121, 2008.

[25] D. Riffe, S. Lacy, B. R. Watson and F. Fico, *Analyzing media messages: Using quantitative content analysis in research*. Routledge, 2019.

[26] C. Grbich, *Qualitative data analysis: An introduction*. Sage, 2012.

[27] A. Luo. 'Content analysis | a step-by-step guide with examples'. (18 Jul. 2019), [Online]. Available: `https://www.scribbr.com/methodology/content-analysis/` (visited on 06/12/2021).

[28] B. D. Prasad, 'Content analysis', *Research methods for social work*, vol. 5, pp. 1–20, 2008.

[29] M. Bloor and F. Wood, *Keywords in qualitative methods: A vocabulary of research concepts*. Sage, 2006.

[30] L. Deng and Y. Liu, *Deep learning in natural language processing*. Springer, 2018.

[31] B. Liu, 'Sentiment analysis and opinion mining', *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[32] M. Annett and G. Kondrak, 'A comparison of sentiment analysis techniques: Polarizing movie blogs', in *Conference of the Canadian Society for Computational Studies of Intelligence*, Springer, 2008, pp. 25–35.

[33] J. Scott, 'Social network analysis', *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[34] A.-L. Barabási, 'Network science', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20 120 375, 2013.

[35] L. Luo and J. Fang, 'A study of how the watts-strogatz model relates to an economic system's utility', *Mathematical Problems in Engineering*, vol. 2014, 2014.

[36] M. Franceschet. 'Network science'. (), [Online]. Available: `https://www.sci.unich.it/~francesc/teaching/network/` (visited on 11/01/2022).

[37] M. E. Newman, 'The mathematics of networks', *The new palgrave encyclopedia of economics*, vol. 2, no. 2008, pp. 1–12, 2008.

[38] statcounter. 'Mobile operating system market share italy dec 2020 - dec 2021'. (), [Online]. Available: `https://gs.statcounter.com/os-market-share/mobile/italy`.

[39] (), [Online]. Available: `https://tweetdeck.twitter.com/`.

[40] J. Sinclair, 'A way with common words', *Language and Computers*, vol. 26, pp. 157–180, 1999.

[41] E. Kricfalusi. 'The twitter hashtag: What is it and how do you use it?' (2013), [Online]. Available: `http://www.techforluddites.com` (visited on 29/10/2021).

[42] P.-M. Caleffi, 'The'hashtag': A new word or a new rule?', *SKASE Journal of Theoretical Linguistics*, vol. 12, no. 2, 2015.

[43] 'Nrclex 3.0.0'. (), [Online]. Available: `https://pypi.org/project/NRCLex/` (visited on 03/08/2021).

[44] S. M. Mohammad and P. D. Turney, 'Crowdsourcing a word–emotion association lexicon', *Computational intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[45] R. Plutchik, 'The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice', *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[46] F. Bianchi, D. Nozza and D. Hovy, ‘"FEEL-IT: Emotion and Sentiment Classification for the Italian Language"’, in *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, 2021.

[47] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman and C. Espina, ‘Classifying twitter topic-networks using social network analysis’, *Social media+ society*, vol. 3, no. 1, p. 2 056 305 117 691 545, 2017.

[48] A. Bruns, ‘Filter bubble’, *Internet Policy Review*, vol. 8, no. 4, 2019.

[49] S. Nagulendra and J. Vassileva, ‘Providing awareness, explanation and control of personalized filtering in a social networking site’, *Information Systems Frontiers*, vol. 18, no. 1, pp. 145–158, 2016.

[50] F. Menczer, S. Fortunato and C. A. Davis, *A first course in network science*. Cambridge University Press, 2020.

[51] E. Bozdag and J. Van Den Hoven, ‘Breaking the filter bubble: Democracy and design’, *Ethics and information technology*, vol. 17, no. 4, pp. 249–265, 2015.

[52] B. Kitchens, S. L. Johnson and P. Gray, ‘Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption.’, *MIS Quarterly*, vol. 44, no. 4, 2020.

[53] A. Sarlan, C. Nadam and S. Basri, ‘Twitter sentiment analysis’, in *Proceedings of the 6th International conference on Information Technology and Multimedia*, IEEE, 2014, pp. 212–216.

[54] K. Verbert, E. Duval, J. Klerkx, S. Govaerts and J. L. Santos, ‘Learning analytics dashboard applications’, *American Behavioral Scientist*, vol. 57, no. 10, pp. 1500–1509, 2013.

[55] G. G. Chowdhury, ‘Natural language processing’, *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[56] A. Zubiaga, D. Spina, R. Martínez and V. Fresno, ‘Real-time classification of twitter trends’, *Journal of the Association for Information Science and Technology*, vol. 66, no. 3, pp. 462–473, 2015.

[57] B. Liu, ‘Opinion mining and sentiment analysis’, in *Web Data Mining*, Springer, 2011, pp. 459–526.

[58] S. Sun, C. Luo and J. Chen, ‘A review of natural language processing techniques for opinion mining systems’, *Information fusion*, vol. 36, pp. 10–25, 2017.

[59] S. Fortunato, ‘Community detection in graphs’, *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[60] M. Coscia, ‘Discovering communities of community discovery’, in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 1–8.

[61] I Feinerer, K Hornik and D Meyer, ‘Text mining infrastructure in r; journal of statistical software, 25 (2008), 5; 54 s.’,

[62] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

[63] S. Zirpe and B. Joglekar, 'Polarity shift detection approaches in sentiment analysis: A survey', in *2017 International Conference on Inventive Systems and Control (ICISC)*, IEEE, 2017, pp. 1–5.

[64] F. Calefato, F. Lanubile, F. Maiorano and N. Novielli, 'Sentiment polarity detection for software development', *Empirical Software Engineering*, vol. 23, no. 3, pp. 1352–1382, 2018.

[65] K. Krippendorff, *The content analysis reader*. Sage, 2009.

[66] V. Morini, L. Pollacci and G. Rossetti, 'Toward a standard approach for echo chamber detection: Reddit case study', *Applied Sciences*, vol. 11, no. 12, p. 5390, 2021.

[67] S. Campbell. 'Python RegEx: re.match(), re.search(), re.findall() with Example'. (), [Online]. Available: `https://www.guru99.com/python-regular-expressions-complete-tutorial.html` (visited on 30/07/2021).

[68] D. Zinoviev, *Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret*. Pragmatic Bookshelf, 2018.

[69] D. Easley, J. Kleinberg *et al.*, 'Networks, crowds, and markets', *Cambridge Books*, 2012.

[70] B. S. Khan and M. A. Niazi, 'Network community detection: A review and visual survey', *arXiv preprint arXiv:1708.00977*, 2017.

[71] G. Rossetti, L. Milli and R. Cazabet, 'Cdlib: A python library to extract, compare and evaluate communities from complex networks', *Applied Network Science*, vol. 4, no. 1, pp. 1–26, 2019.

[72] 'Cdlib'. (), [Online]. Available: `https://cdlib.readthedocs.io/en/latest/overview.html` (visited on 12/12/2021).

[73] 'Networkx, network analysis in python'. (), [Online]. Available: `https://networkx.org/` (visited on 12/12/2021).