LUISS Guido Carli
LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Department
of Marketing Analytics and Metrics

Chair of Machine Learning

"THE POWER OF GOOGLE: HOW TO PREDICT EUROPEAN ELECTIONS USING GOOGLE DATA"

Luigi Laura
SUPERVISOR

Paolo Spagnoletti
CO-SUPERVISOR

691761
STUDENT ID

Academic year 2018/2019

General Index

To my parents, Daniela e Renato, who are the most important persons of my life and who have been there all the steps of the way, I couldn't make it without you.

To my grandfather who taught me how to love unconditionally.

To Guglielmo, who makes me happy every day and whose love has encouraged me to never give up.

To Flaminia and Giulia, who are "a friend", and it is not so easy to be.

To Carolina who can forgive, all real friends should be able to.

To Camilla, whose friendship is unbreakable despite the distance.

To Martina, Mariapia e Almachiara who have lived with me for the last years and who stand by me every day. They are significantly more than just my friends, they are my "home".

“Torture data and they will confess everything”

-Ronald Coase, economist

CHAPTER 1

BIG DATA: PROS AND CONS

Today we are overwhelmed by information, which arise in number more and more rapidly. The exponential increase of information can be showed by the following data:

Google process more than 24petabyte of data per day, on Facebook are uploaded each day more than 10 million of photos and subscribers click “like” almost three milliard times per day, the number of messages on Twitter increases by 200 percent per year.

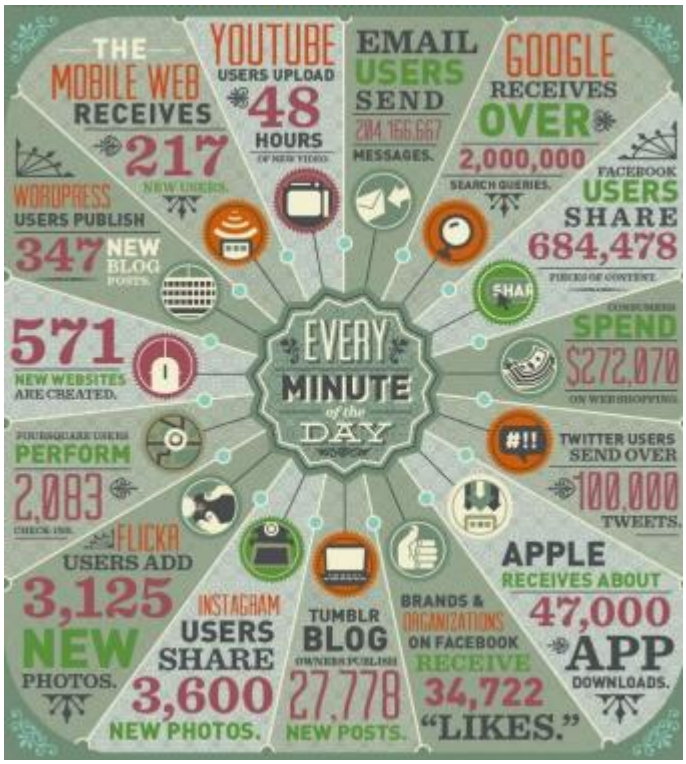


Figure 1. How many data can be generated in a minute?

According to Peter Norvig, expert of artificial intelligence of Google, the quantitative change will lead also to a qualitative change, changing quantity changes also essence.[1]

There is an increase of information availability, the improvement of the capability to elaborate and store data, and an economic convenience in making these processes. The information availability is given by the process of transforming each phenomenon in data and by the Internet of things, which allows to collect data on the status and functioning of objects, through sensors like the gps. Economic convenience is given by the reduction of costs thanks to artificial intelligence, which supports the storage and processing of data, with the marginal cost almost equal to zero because of the digital transmission of data. There is a passage from small data to big data, when the reference is not to a small subset, but to everyone. And the volume justifies the lack of precision of analyses of this huge amount of data.

Anyway, what is the correct definition of Big Data?

This is the starting point of Stephen Few's opinion. Few, IT expert, wrote the book "Big Data, Big Dupe. A little book about a big bunch of nonsense". In the first chapters of his book he underlines how there is not a single clear definition of Big Data. He lists six different definitions that we can encounter:

1. Data set that are extremely large
2. Data from various sources and various types

[1]Mayer Schönberger Viktor, Cuckier Kenneth, Big Data: a revolution that will transform how we live, work, and think. Milano: Garzanti, 2013

3. Data that is large in volume, derived from various sources, and produced and acquired at fast speed
4. Data that is extraordinary complex
5. Data that is processed using so-called advanced analytical methods
6. Any data at all that is associated with a current fad

1.1 What does Big Data mean?

In each of the precedent definitions it is possible to find problems. Let's start with the volume of data. It is clear that they are called Big Data because their volume is huge. But how large must a dataset be in order to qualify as Big Data rather than merely as data? It is not clear. And if we define Big Data as the data that cannot be processed using traditional database and software techniques, the problem is that we do not know what are "the traditional database and software techniques". Some other definitions focus on the complexity of data, but they lack a clear threshold. They do not explain what they mean by complexity. The same happens with the definition that focuses on the advanced analytical methods used to process Big Data. Are High-level programming skills really a parameter to distinguish current from past data analysis? They actually are not. Computers have evolved and skilled analysts have developed programming skills to do their work since computer became available for data analysis, it is not a prerogative of Big Data. "Now Big Data has become a buzzword to mean anything related to data analytics or visualization"[2] as said Ryan Swanstrom. Confused definitions can be manipulated to mean anything one wish.

[2] Swanstrom, Ryan (Data Science Blogger, "Data Science101" blog). URL: https://101.datascience.community/about/

Summing up, the first argument against Big Data is that nobody clearly knows what effectively means the term Big Data, at least not enough to be able to arrive at one single unique definition.

1.2 Big Data, Big opportunity?

One of the arguments in favor of Big Data underlies their potential. We just see a little part of data but there is an enormous amount of them that is hidden and just waits to be analyzed. The point is: can we find new uses for data that were originally generated for another purpose? In the past, it did not happen. As Few deduces in his book, if it is true that it is convenient for firms to collect and retain more data as possible in hope of finding unknown secondary uses which can be helpful in the future, so it has to be true that the organizations with most data today should be the most successful ones. But they are not. In its 2018 Big Data Maturity Survey, vendor At Scale –data warehouse virtualization platform- noted that while 78 percent of companies believe they are at a "medium" or "high" level of big data maturity, in reality, only 12 percent meet the criteria of a high level of maturity. And organizations at lower levels of maturity continue to struggle with multiple challenges in regards to big data analytics.[3]

Before embarking on any new analytics project, experts recommend that enterprises carefully weigh the pros and cons of big data to see whether the initiative will be worth the risks and costs.

[3] Harvey, Cynthia. Big Data Pro and Cons. Datamation: August 9,2018. Online article: https://www.datamation.com/big-data/big-data-pros-and-cons.html

Storing and collecting everything makes it harder to focus on the little, which has value. Moreover, even if the collection of data is made by hardware this does not mean that it is inexpensive. The only ones who really benefit from the collection and storage of all the data are the companies that make the hardware and software, companies that produce much of the data such as Facebook, and the companies that provide storage facilities.

Second argument against Big Data: their collection is not worth it.

1.3 Big Data: what really matters

Data in itself are just data and are not useful, the most important thing is the "sense making", as Few called the analytics, the one which permits to make sense of data. Breaking down data and dig into the details is just the starting point, training in statistical is fundamental, but it has to be supported also by logical thinking, critical thinking, ethical thinking, and other skills. The point seems to be the importance of the training of the skills needed for data sense making, not the data itself. Human makes data sense making, machines can just assist. Anyway, it is quite obvious that it is impossible to find a technology company who sustains this thesis; it would be against their business. In addition to this, we have to look back to the past. We want to collect more and more data but maybe we have to take a step back. Data cannot speak from themselves, they have to be approached using scientific methods and statistical models, it is not the quantity but the understanding of correlations: if we understand a subset, this one can substitute an entire dataset. We have to step back from quantity to quality, and in order to do this we need expertise.

Evgeny Morozov coined the term "technological solutionism" to describe our over-reliance on technologies to save the day. [4] But data cannot be left by themselves having the power to decide the day. An analysis which excludes any theoretical foundations and which is based fundamentally on correlations between data will result weak. If we leave data by themselves they will be able only to recognize correlation and not causation. But, as it will be explained in the third chapter, there can be a lot of spurious correlations. And we care about whom, how and why, not about casual correlations. We need models to represent and comprehend relationships and patterns, we don't build models for lack of data, contrarily we cannot think without models, they are part of our thought. Moreover, it can be affirmed that when we talk about data, size does not matter: if we conduct a poll on 10 million people to predict something, it might happen that we will not be able to predict a result that is as exact as a smaller poll's result.

Third argument against Big Data: they focus on quantity and not on quality leaving out the importance of expertise of the subject without which Data are not useful.

1.4 Big Data: what do they hide?

What about the idea that Big Data are being used for a second hidden purpose? This huge amount of data can be used in a potential harmful way. We do not even notice it and that is because people usually trust in technology and do not even imagine that technology can be malign. Instead, it is very important to limit technology and use it ethically. Governmental and non-governmental organizations use it to influence and

[4] Few Stephen. Big Data, Big Dupe. Analytics Press, 2018

harm us. There is a separation between technical models and people, an increasing gap of which we should be scary. Data are left by themselves and we see the data world as a world apart, distant from us. The problem behind data is that they favor efficiency and are only able to measure and quantify, but fairness is hard to quantify, because it is a concept. The result is a continuous restless production of unfairness. Human has virtue and they can learn, grow, evolve, change; automated systems are blocked until experts and engineers want them make a change. They are stuck in the past, and cannot imagine the future, imagination is a human prerogative.[5] This argument leads to another concern, the one about privacy. Data can be used to gain insights into our behavior, but that also implies a substantial loss of privacy. All information collected about each one of us is feed into a machine, which processes them. And we do not know anything about the process, we do not know how they will be used and who will use them. Mark Zuckerberg, one of the funders of Facebook has affirmed: "the age of privacy is over".[6] As S.S. Davidowitz explains in his book one of the Big Data risks is the one of potentiated corporation. It has resulted that one indicator of the probability that loans' money will be given back is the language used by who asks for the loan. Indicator that is important also to check others information on potential contractors. According to the study reported by S.S. Davidowitz the terms "without debts", "lower interest rate", "taxed", "least payment", "graduated" are the ones used by people who have more possibility to pay back the loan. "I will pay", "hospital", "God", "I promise", "thanks" are the ones

[5] O' Neil, Cathy. Weapons of Math Destruction. New York: Penguin, 2016.
[6] Zuckerberg, Mark. Quoted by Marshall Kirkpatrick. January 9, 2010. "Facebook's Zuckerberg Says The Age of Privacy is Over." ReadWrite.
<https://readwrite.com/2010/01/09facebooks_zuckerberg_says_the_age_of_privacy_is_ov/>

used by people who have higher probability to not pay back the loan. In particular, the researchers, Oded Netzer, Alain Lemaire and Michal Herzenstein, found out that who invokes God is 2,2 times more incline not to pay. This study can without any doubt be considered useful, but can also, without any doubt, be considered unethical. It means that in the immediate future a person should be worried not only about his economic situation, but also about his online activities, about his friends on Facebook - are they insolvent? If most of them are insolvent, I will be probably judged insolvent too-, about his language- a bad mood day can result in a loss of a loan!, a medical disease can be assed as a lie! Our IQ can be measured checking the "likes" on Facebook, just because we "like" one group on Facebook instead of others, we can be categorized in the "lower IQ group". This is scary. The use of Big Data can lead to discrimination. We are already judged for factors that are not related to our intelligence, like our physical appearance, the way we dress, or the way we move... we do not need other factors on the base of which we can be even more discriminated.

Another form of discrimination caused by Big Data is the price discrimination made by firms. Through the use of Big Data firms knows how high they can set the price in order to gain the maximum profit and squeeze out the consumers as much as possible. Firms have become able to know consumers' insights and so to take advantage of all their weakness, increasing the power they have on them.

On the other side, Big Data enables consumers to have more information and have the possibility to protect themselves and to compare all the offers that are on the market. Firms can use Big Data to know which consumers are better to target,

consumers can use Big Data to know which firms are better to choose. Government and institutions have the duty to guarantee to citizens that they will maintain unchanged their status and that there will not be illegal exploitations by the more powerful organizations both governmental or non.

Forth argument against Big Data: possible loss of privacy.

1.5 Potentiated government and ethical issues

Should Big Data be used by the Police to prevent crimes? If we know that a person has checked on Internet "how to rape a woman", should this information be used to arrest him and prevent that he actually does it in real life causing a victim? This is a huge and very difficult to solve ethical issue with implications in multiple, different areas. S.S. Davidowitz proposes a very simple idea: the use of these data on a geographical base to allocate funds. If in a specific area where there are more suicide correlated researches, funds will be allocated to increase the consciousness of suicide and of the ways in which is possible to ask for help. This can be possible. To take a person in a center of Mental Health just because he has searched on Google the word "suicide" is not. There is a huge difference in the attempt to predict the behavior of a city or of an individual. Data say that there are a lot of horrible researches but that rarely these are translated into horrible actions. We cannot act on the single person, it would be an unjustified and unethical lack of privacy, but we can act on the community, having the possibility to be aware of the problems and so to take action, make prevention and solve these problems.

Fifth argument about big data: ethical issue and difficulty in integrating legacy systems.

1.6 The other face of the coin: advantages of real-time Big Data Analytics

1.6.1 Smart Data

Even if after the lecture of Stephen Few's book it is difficult to still believe in the Big Data phenomenon and to be optimist about them, there is also another side of Big Data. Willfully or unintentionally we are in the era of Big Data. The term recalls to mind informatics complexities and makes us see it as something which does not concern us. But contrarily as what can be thought, we are Big Data, our choices and us. All the people on the planet who have an Internet connection are Big Data. The Internet of things -use of daily objects always connected to Internet- and social networks generate a total of 900 Exabyte, which is traduced in a number with thirty zeros. And it is not true that the shopping firms are the only who benefits from this, they obviously are the ones who have more immediate advantages and responses from Big Data analysis, and this is just because it is easier to ask to an unknown person what are his/her preferred brands than to ask what his intimate beliefs are. The process consists in exploiting Big Data created by online shopping, by gps, by sharing contents on social media and transforming them in "Smart Data", useful advices to redirect sales. As Daniel Keys Mora, an informatics, says, "you can have data without information, but not information without data".[7] Differently to the

[7] Corriere della Sera. Big Data: the new era of advantages for PMI is reality. Cured by Unicredit. Online article: https://www.corriere.it/native-adv/unicredit-longform04-big-data-la-nuova-era-dei-vantaggi-per-le-pmi.shtml

majority of technological trends Big Data are not a trend, they are a managerial necessity. In fact, before today's more streamlined big data analytics offerings, answering even seemingly simple questions such as "Who are my 10 best customers?" could take up to 60 days for business teams to analyze. Even after they figured out the right criteria, actually compiling and analyzing the data was a time-consuming process. The burden only grew as questions became more complex. With a powerful big data business intelligence platform, answering these questions becomes a relatively straightforward process. One of the most important benefits of big data is the ability to ask and answer questions more robustly. The whole process of answering complex questions can be shortened from months and weeks, to days and even hours or minutes.

1.6.2 Zoom in and Customization

It is thanks to contents that an user consults on websites, social networks, App and newsletters that it is possible to comprehend if he/she cares about natural textures, or about newest collections items or about what is worn by celebrities and customize the offer, directing specific images, article and videos according to his/her preferences. Website pages will be personalized and will differ from user to user, being modified in function of who is surfing on them. As for example happens in the Predictive Analysis, that is the adjustment of a marketing message to a probable action of the consumer -the ad of an hotel in Rome while he/she is on the way to Rome, made possible thanks to the gps localization. The zoom that Big Data consent to do is the future of the marketing. Through the analysis of Big Data we can know countries'

preferences, states' preferences, regions' preferences, specific areas' preferences, neighborhood preferences', until reaching each single consumer and internet user of the world. Segmentation provides the best way to sell something to someone, that is to know, even before he/she is aware of it, what he/she wants. And this is made possible by what S.S. Davidowitz calls the second power of Big Data: to provide honest data.

1.6.3 Google as a truth serum

The starting point of S.S. Davidowitz book is the assumption that everybody lies. We are used to lie to ourselves, much less to others. Most of people want to make a good impression even if polls are anonymous. This is called bias of social desirability and most of us are its victim. This problem is solved by the advent of the Internet. Even on the web, if we directly ask people what are their preferred brands or what is their favorite food through surveys, they can still lie, even if they are alone in front of their computer. But with the analysis of Big Data, of all the traces that we leave on the Internet, surveys are not required anymore. Why ask with the risk of an untruth answer when we can obtain it without even asking? Analyzing movements of web users, looking which websites they prefer, the bounce rate- how many users leave the page after the first page view-, their comments on social media, their likes on Instagram pictures, we can know what they want without any effort. Moreover, we can say that today is happening the opposite of what happened few years ago: in surveys customers had to answer to questions, today consumers are the one who make questions to Google. "Torture data and they will confess everything" said

Ronald Coase, an economist. Indeed, it is consumers who confess to Google what they would not confess to anyone else. S.S. Davidowitz shows it in his book with different examples, to take the most significant one: if we type something on Google he tries to complete our phrase automatically, and this gives us indicators of what people usually search. And if we type “is it normal to want ” the first Google’s suggestion is “to kill”. This sends back to another way in which Big Data can be used.

1.6.4 Smart cities

Figure 2. Words correlated with the concept of Smart Cities

Data can reveal which are the most dangerous areas of a city and give information that enables governments to make them a safer place. On the idea of “exploiting Big Data to make the world a better place” as Christian Rudder said, are being developed, and they are called Smart Cities. Smart Cities are an example of Big Data Management and Big Data Analysis. They are cities made with planning strategies with the aim of optimization and innovation of public services, all through the use of

Big Data. Thanks to lamppost with sensors it is possible to manage, in a better way, the rush hour traffic and to monitor pollution. Policy can reconstruct car routes that are suspicious through the closed circuit television which are increasingly present outside shops and bars. For the separate waste collection are used the tag RFID which enables bins to be interconnected. According to McKinsey's analysts, a managerial consulting multinational, in Europe the public administrations which have a good management of Big Data can save about 100 milliard of euros, increasing the operative efficiency. An amount that might increase if Big Data were used also to reduce frauds and errors, thus achieving fiscal transparency. Big Data can make governments optimize money and invest funds in what really became a problem to be solved. They can help to provide a sustainable environment with higher energy efficiency and less wastage of resources. Through predictive analysis is possible to analyze the growth of current infrastructure and plan for future needs.

Figure 3. Graphic representation of Smart Cities concepts

1.7 Slow Data Movement by Stephen Few: a meeting point

570 new websites are created each minute. The new generations are Millennials and Big Data are their future. Millennials are accustomed to technology that helps them find what they want, instantly. In the past, limited data sets meant businesses could only ask and answer a few questions. Now, with a powerful big data analytics platform, businesses cannot only answer more questions quickly, but also more questions about the questions themselves. This leads to cognitive improvements in the question-answering process, making the process itself more organic and fluid. By empowering analysts, and especially younger analysts, businesses can create an environment that encourages creativity, outside-the-box thinking, and increased critical and analytic analyses. “I keep on saying that the sexiest work of the next 10 years will be in statistical sciences, and I am not joking” affirmed Hal Varian, Google economists, and it seems to be that he was right!

But maybe, even if accepting to be in the era of Big Data, an advice of Stephen Few can be interesting and useful. S. Few in his book proposes the idea of a “Slow Data Movement”. He was inspired by the idea of Carlo Petrini, the one who founded the “Slow Food Movement”, a movement born to fight against the boom of fast-food restaurants like McDonald’s, based on the idea that the quality and beauty of food requires time: time in the process of producing, time in the process of preparing and time in the process of tasting and in the act of eating. S. Few extends this idea of the importance of slowing down and taking the time to appreciate to the data world. In a fast run world we are always in a rush and we do not realize how much we miss out

on during our race. Doug Laney was the first who proposed the three characteristics of Big Data, the 3Vs: volume, velocity, and variety. He made a good observation: the advent of machines and computers before and that of Big Data after (assuming that they can be considered as a new category which did not exist before), lead to a bigger volume, a faster velocity, and a wider variety. But are these the characteristics of Big Data that really matter? Are these the ones we have to focus on? Few suggests 3Ss in alternative to the 3Vs: small, slow, sure.

Small because, as already said, only a small part of data is useful, we have to be able to recognize what is meaningful and useful and to leave out the rest.

Slow because, how Daniel Kahneman underlines in his book "Thinking fast and slow", each of us is made up of two systems: System 1, the intuitive one; System 2, the reflexive one. And all those decisions which are not decisions of habits are made by System 2, thanks to conscious, deliberate, reflective and analytical reasoning. This can make us understand that going too fast without taking the time to reflect is never a good idea. We are surrounded by technology, by a digital world that speeds faster than us. We should not try to compete with it, but by contrast to slower it in order to be able to understand it.

Sure, because variety can be a synonymous of complexity and so can be everything except useful. Only when we recognize some data as relevant, because they can be exploited to make something, we can call that data sure, because they become reliable as soon as we identify them as valuable.

The advice of S. Few to focus on the 3Ss can be taken into consideration as a "meeting point", a way to use data effectively without harming anyone but instead with the purpose of enhancing our lives. The key to handle data is education. The data world moves forward every day and in order to be always update and able to understand it, organizations should not limit themselves to hire people with analytics skills, yet they have to form them in a constant process of training. If we leave the data world apart from human intelligence, without exploiting it by human intelligence, Big Data will probably just create risks, and the worrying fact is that they harm us without us even noticing, benefiting from the huge amount of information that everything we do, day-by-day, leaves behind itself.

CHAPTER 2

HOW I CAME UP WITH THE IDEA FOR THIS THESIS: 'EVERYBODY LIES' BOOK WRITTEN BY SETH STEPHENS DAVIDOWITZ

2.1 "The book "Everybody Lies"

The idea of this thesis came up from the book written by Seth Stephens Davidowitz "Everybody lies", in Italian "La macchina della verità". It has all started in 2008 with American presidential elections and the long debate about the importance racial prejudice in America. Barack Obama, the first Afro-American candidate of a big party, won. At the time, a lot of polls were made which suggested that the race was not a determinant factor, according to those polls, the majority of Americans did not care about the color of Obama's skin when they decided who they were going to vote. In that year, S. S. Davidowitz discovered Google Trends, an instrument that indicates users' search-frequency of words and phrases in different places and at different times. Using Google Trends he found out that the night of Obama's first election, when all the comments were about the praises on the new president, about one in one hundred researches on Google containing the word "Obama" included also terms such as "Ku Klux Klan" or "nigga". On the same night, researches and subscription to "Stormfront", a nationalist white site, have been more than 10 times higher than usual. In some American states "nigga president" researches exceeded those for "first black president".

These findings could explain the success gained eight years later Obama's first election by the current President of the United States of America Donald Trump, a

candidate that summed up the racist researches, the immigrants attacks, the anger and the resentment of people's worse inclinations.

At the beginning S. S. Davidowitz's study was rejected by five academic magazines, the latter sustained that it was impossible to believe that so many Americans hatched such racism. With the election of Donald Trump his findings have become more credible. At the beginning of the primaries, Nate Silver, an American statistician and writer, had said that there were no possibilities for Trump's victory. When it became evident that his first assumptions were wrong, he decided to analyze data to understand what was happening and noticed that the areas in which Trump had more success drew a very atypical map:

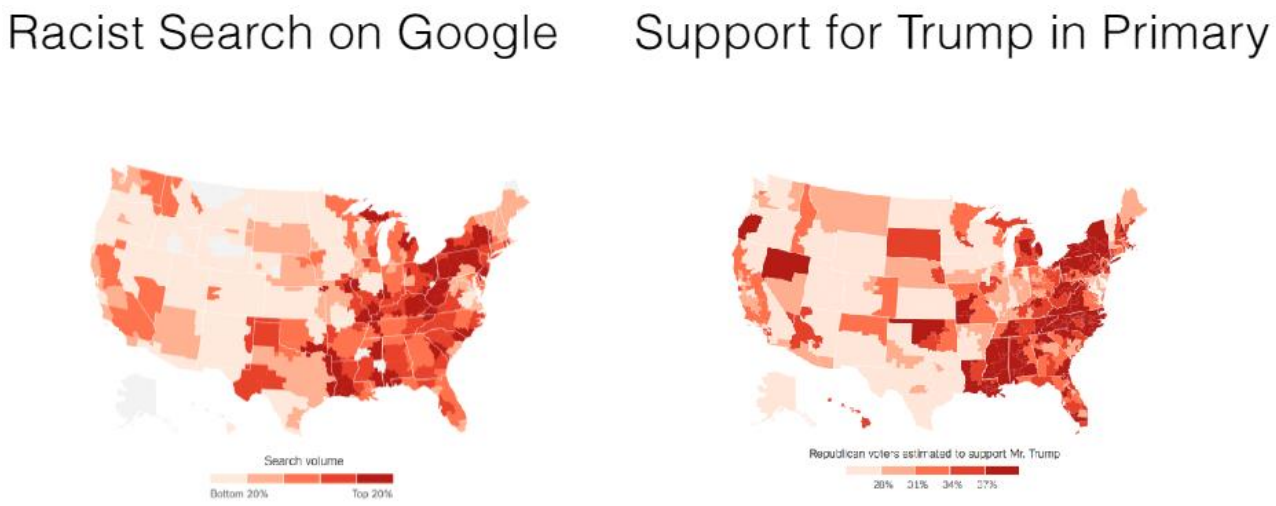


Figure 4. Nate Silver's findings- from "Everybody Lies" S.S. Davidowitz

He noted that the most correlated factor with the Trump support was the same one that S.S. Davidowitz had found previously: the areas which supported Trump were the ones in which people searched with an higher frequency the word "nigger" on Google.

2.2 The purpose of “Everybody lies” and what it teaches us

That is the beginning of my thesis: S.S. Davidowitz in his book shows how we can use Big Data to give new intuitions about the psychology and behavior of human beings, how Big Data can confirm and prove suspicious, but most of all, how they can reveal that the world does not go in the direction we imagine, but in the opposite direction. His book is a great and very interesting reading about using internet Big Data sources to study what people actually do and think vs. what they say they do and think. We tend to lie, especially about embarrassing or negative behaviors, when asked. But to Google, people confess the strangest things. In the blank space we do not type only words about which we want to know more, but we confess our secret passions and we ask for prying questions. Google database contains human beings’ worries, wishes, mental diseases, insecurities, political and religious choices and sexual preferences. In this way people’s search for information is, in itself, information. A big amount of information, which did not, exists just fifteen minutes ago and which increases hour by hour. In a day we produce on average 2,5 trillion of data’s byte. And each byte is information. Google search data is therefore a very powerful, and enormous source of information which to some extent is publicly available to anyone. From a tool to know the world it has become a tool to know human beings.

“La macchina della verità” teaches us how we can take advantage of Big Data, without losing ourselves in such an extraordinary amount of data but giving a sense to them and trying to gain conclusions analyzing them.

S.S. Davidowitz lists four main big power of Big Data:

1. Offering up new types of data is the first power of Big Data
2. Providing honest data is the second power of Big Data
3. Allowing us to zoom in on small subsets of people is the third power of Big Data
4. Allowing us to do many causal experiments is the fourth power of Big Data

2.2.1 First power of Big Data

The first power of big data is the redefinition of what is in data's category. Often the value of big data is not in their dimensions, but in the fact that they can give us new types of information to study, information that in the past we were not able to collect. S.S. Davidowitz tried to use Google searches to measure unemployment rate. Between 2004 and 2011 the main research strictly correlated with unemployment was not "employment office" but "Slutload", a pornographic website. Another correlated research was Spider Solitaire. We should not be surprised by these findings: it is presumed that unemployed people have a lot of time to waste. No one before had used researches on hobbies to predict unemployment rate, and maybe it is not the best way, but it should definitively be part of the best model to predict it.

2.2.2 Second power of Big Data

The second power of big data is to provide honest data. Big Data let us see what people really think and want, not what they say they do. This is the difference between data gained by a poll and data obtained by Google data. By analyzing the data of millions of people we can discover that we are not alone, we are not the only ones who have difficulties in life, love, sex, etc. Google shows us the reality, which is significantly different from what we see on social media. On Instagram and Facebook

everyone seems to have an amazing life and to be happier than us. The example that S.S. Davidowitz reports is the one about the most popular pornographic video: “Great body, great sex, great blowjob” has 80 million visualizations, but on social media has been shared only a dozen of times (most of all by porno stars). In the Facebook world it seems that every Saturday every adult human being spend the night taking part to unforgettable parties, in the real world most of the time we are alone at home.

2.2.3 Third power of Big Data

The third power is the possibility to zoom to small subsets of people. We can zoom on which dimension we want: age, cities, time, etc. We need millions of information in a dataset to be able to zoom clearly on small subsets of data. In his book S.S. Davidowitz reports the answer to the question “Is America the country of opportunities?”. Analyzing a representative subset of Americans and comparing it with similar data of other countries gives the traditional answer to this question. How many possibilities does a person whose parents are in the 20% of the lowest income bracket have to gain the 20% of the highest income bracket? Raj Chetty, an American economist and Harvard professor, zooming on the datasets, gave a more accurate response. First of all, he zoomed on geography and discovered that probabilities vary according to the place of the United States in which one is born. Then, his team was able to zoom even more, examining each small group of people who moved from one city to another. The definitive response to “Is America the country of opportunities?” is neither yes or no, but some areas are and others are not.

2.2.4 Fourth power of Big Data

The fourth power is the opportunity to make many causal experiences. They facilitate the execution of randomized experiments, being able to discover causal effects every time in every place, it is just necessary to be online. Google engineers were the first to understand that digital world experiments have an advantage compared to offline world experiments: offline experiments need a large amount of resources, cost thousands or hundreds of dollars, and take years to be completed; digital experiments are faster and cheaper. It is not necessary to recruit and pay participants; they can just be assigned randomly to a group. It is not essential that users answer to a survey; it is sufficient to measure and count the mouse moves and clicks. It is not needed to contact anyone; it is even not required to inform people that they are part of an experiment. Controlled randomized experiments have been named “test A/B”. In 2011, Google engineers made seven thousands tests A/B, and the number keeps increasing.

2.2.5 Limitations of Big Data

S.S. Davidowitz in his book does not show just the power, but also the limit of Big Data. He talks about “the dimensionality’s malediction”. It happens when there are a lot of variables and a low number of observations. The matter is that the probability of those dimensions will decrease if we lower the variables. Lawrence Summers, an American politician, academic and economist, asked S.S. Davidowitz if it was possible to use Big Data to make predictions about the stock market. The easy answer is “no”. In fact, as S.S. Davidowitz writes, it is extremely easy for one to fall victim

of the dimensionality problem and moreover is quite impossible to keep up with hedge funds. Numbers can be attractive; they can become an obsession and lead us to lose sight of more important considerations. The problem is that it is not always true that what we can measure is really what matters for us. For example, Facebook in its attempts to improve its website, has an enormous quantity of data on how the site is used. He track likes, clicks, sharing and comments, but none of these things can match perfectly with the most important question: “how is the experience of the website’s users?” To answer to this question, and not only to this question; to fill up the holes of its big amount of data, Facebook needs to use a traditional approach: ask people what they think. Even Facebook needs surveys. Big Data needs Small Data in order to have a complete view. Small Data are as important as Big ones.

CHAPTER 3

ATTEMPT TO PREDICT THE EUROPEAN ELECTIONS RESULTS

"La Macchina della Verità" has served as an excellent starting point. It has this unique capability of explaining statistical concepts in a clear way, without getting lost in technical jargon. The chapter before is only a brief sum up of what S.S. Davidowitz explains in his book, but it should be enough to understand what I tried to do.

What I tried to do, together with my team, is the same thing he did with American presidential election, but with the European elections in Italy. I tried to predict the results using the power of Big Data and following the advices of S.S. Davidowitz's book.

3.1 Tools

Two main tools were used in order to do this prediction: Google Correlate and Google Trends.

Data mining is a firm's process, which permits to analyze a big amount of data to find out eventual significant correlations between them. This process allows querying a database, through software of analysis, and going deeper in the data in real time. Google launched an instrument that enables us to do those things, and its name is Google Correlate. A tool that can discover trends more correlated to a general activity and so to the semantic universe of reference. It is possible to insert a trend, about any activity in the world, and see which search terms are connected to that trend.

Google Correlate is like Google Trends in reverse. With Google Trends, you type in a query and get back a series of its frequencies (over time, or in each U.S. state). With Google Correlate, you enter a data series (the target) and get back queries whose frequency follows a similar pattern.[8]

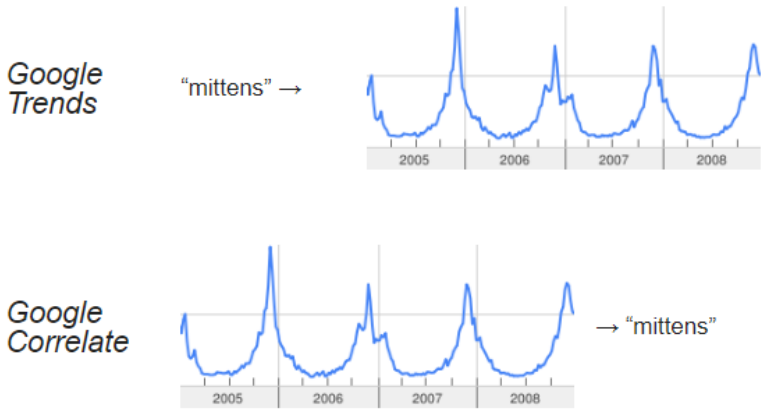


Figure 5. Difference between Google Trends and Google Correlate (https://www.google.com/trends/correlate/tutorial)

Correlated Queries: when you upload a data set (a time series, for instance), Google Correlate will compute the Pearson Correlation Coefficient (r) between your time series and the frequency time series for every query in our database. Correlation coefficients range from r=-1.0 to r=+1.0. The queries that Google Correlate shows you which are the ones with the highest correlation coefficient (i.e. closest to r=1.0).

Instead, Google Trends is a tool that allows us to know the frequency of web researches for a specific word or phrase. Research and visualization can be set by nation and by language and even by topic category. The trends are showed with a graphic which synthetize, by time, the trend and its popularity. In the section "researches' tendency" it is possible to visualize the "hot trends" of the moment,

[8] https://www.google.com/trends/correlate/tutorial

instead in the “classifies” section there are the “top charts”. In 2018 its graphic was restyled in order to make it understandable and usable to non-expert users. In fact it is very useful to whoever works in the digital world, but can also be interesting to less experts to be informed about the latest news and the latest tendency topics.

3.1.1 Examples

Google Correlate

As mentioned above, S.S. Davidowitz inserted on Google Correlate the unemployment rate in the Unites States from 2004 to 2011. He found out that the most correlated research in that period was not “employment office”, or something similar, it was in the principal voices but not the first. The first was “Slutload”, a pornographic website. This can sound as strange, but if we assume that unemployed have a lot of free time it is not such a surprise. Another research correlated is “Spider Solitaire”. This means that through a collective research on hobbies it is able to reveal the unemployment rate and should be part of the test to predict them.

Google Trends

S.S. Davidowitz used Google Trends to prove that Google is a truth serum. According to him, each of us sometimes type in the Google box something which reveals a thought or a behavior that we would hesitate to admit to the society. There is hard evidence of it. Americans, for example, search “porno” more than “weather”, but in polls only 8% of women and 25% of men confess to watch pornographic material.

- "porno"
- "weather"

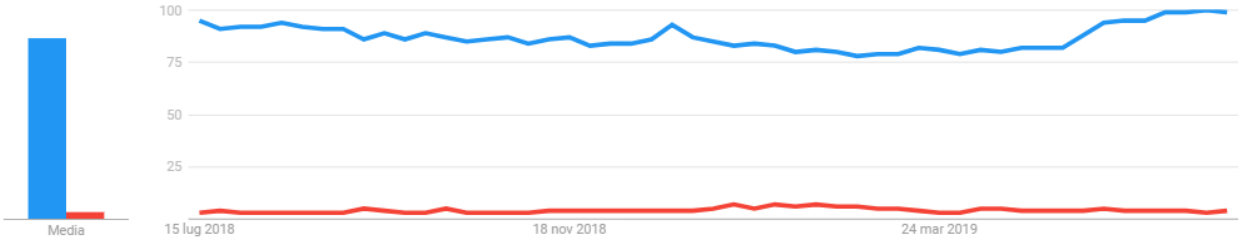


Figure 6. "porno" and "weather" researches' trends- Google Trends

3.2 Frame: actual political situation in Italy

The 26th of May 2019 there were the European political elections, and for the first time the result was not predictable. In fact, the decline of traditional parties and the rise of the radical right and of the other populist parties was critical for Socialist and Popular movements, the two blocks which have governed the Parliament for the last few years.

Especially, for Italy these have been very significant elections because they represent the first real test at a national level to the Government settled in June 2018. European elections were important to verify the total consent of the two main parties (Lega Nord and Movimento 5 Stelle) in respect to the opposition, but were also relevant to find out their relative strength: if the Lega Nord would have surpassed the Movimento 5 Stelle and which consequences it would have had for the Government.[9]

[9] Italian politics in 2019. The Post, January 2, 2019. Online article: https://www.ilpost.it/2019/01/02/eventi-notizie-2019/

In the Second Republic, after the passage from a proportional electoral system to a basically majoritarian electoral system, Italian politics was organized in two different coalition, two groups of political parties which competes for the majority both in Parliament and at the Government. To determine the advent of tripolarismo was the fast rise of the Movimento 5 Stelle at the 2013 political election. A party firmly contrary to any political alliance. In 2013, Italian politics had to face a tri-polar system and the impact was traumatic. Added to the economic crisis, after the elections, there was also a situation of uncertainty for the institutional equilibriums, with the Parliament stalled and expiration of the mandate for the President of the Republic, that is in the impossibility to dissolve Parliament to eventually hold new elections. A situation that was overcame with some innovations on both institutional and political level:

- It was named a commission of ten wise men
- It was reelected the resigning President of the Republic
- A Wide Agreements Government was formed (Letta's Government)
- Unusual role given to the Renzi's Government to strengthen the Parliament to adopt institutional reforms

These extraordinary measure lead Italian politics to the political election on the 4th of March in 2018, with a proportional electoral system. So between 2013 and 2018 Italy made some steps back, passing from a majoritarian electoral system to a proportional one. Electoral law affects the governability of Parliament. It is a determinant factor both for post-electoral consequences and pre-electoral choices of political parties. Only with the Rosatellum (26th October 2017) the homogeneity of electoral systems

between Camera and Senate was restored, but the political electoral system remains a proportional one.

There are two major problems connected to the tripolarismo. The first is the governability, tripolarismo increases the risk that after the election there will be obstacles to the Parliament majority and so that of the Government, therefore enhancing the danger of new institutional crisis. The second is about the survival of tripolarismo, a system that it is not consolidated, with a third pole that is formed only by one ever-changing party, and with the other two traditional poles which are evolving. By one side the main political parties kept on proposing the traditional platform of right and left wings, by the other side the electoral tendency and the Parliamentary majority formed between Movimento 5 Stelle and Lega Nord undermined the old contraposition between right-center and left-center.[10]

[10] Easy politics. Online article: https://politicasemplice.it/politica-italiana/situazione-politica-italiana-2013-2018

3.3 Starting point

The following graphic shows the trends of the three main parties and just how far Lega Nord was ahead of the other two political parties before the European elections.

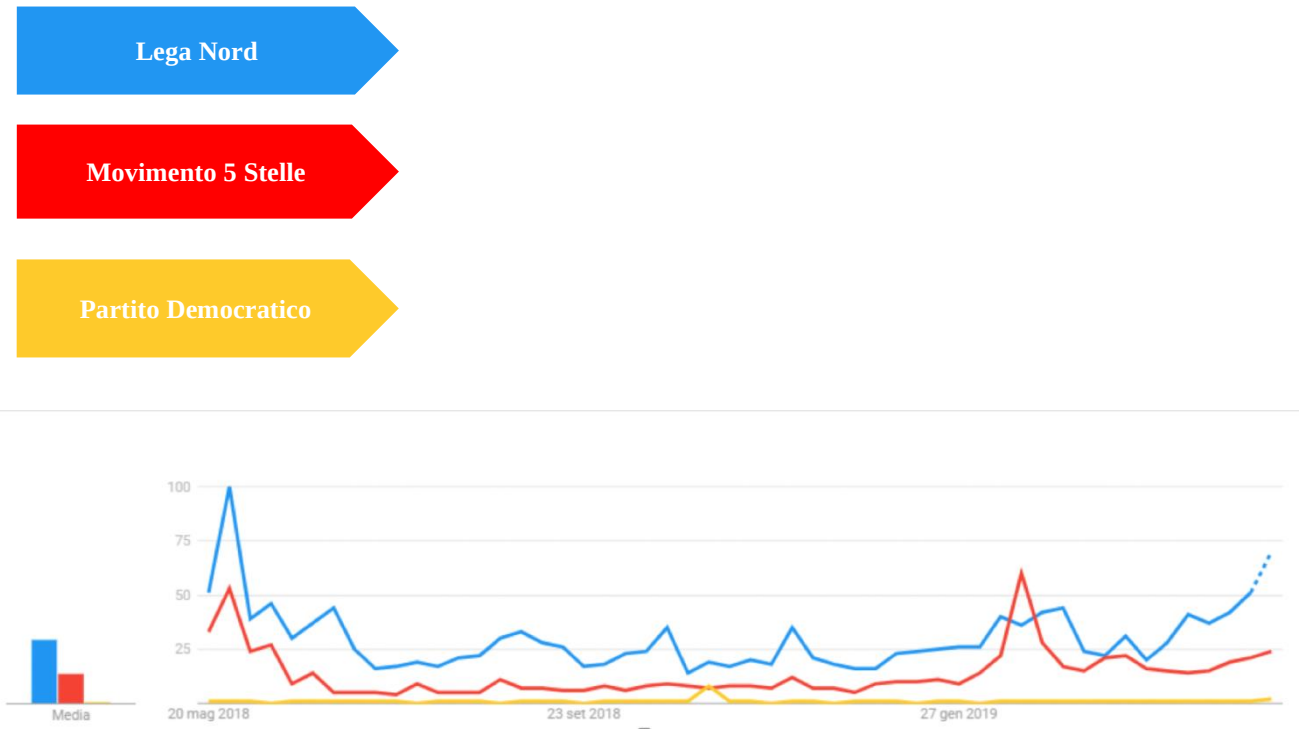


Figure 7. Political trend of the 3 main Italian parties- Google Trends

It is evident that the first party in Italy has been the Lega Nord for the last few months, followed by Movimento 5 Stelle and Partito Democratico. This reflects the fact that the actual government is a coalition government born from an agreement between Lega Nord and Movimento 5 Stelle. Anyway, the actual government was formed after the political election on 4th March 2018, election through which did not yield a majority able to vote on its own government. The ones who gained more votes, Movimento 5 Stelle with 32,7% and right-center, conjoined, with 37%, did not gained a real majority. These results reflected a dissatisfaction and disappointment of the Italian population. The government of coalition between Lega Nord and

Movimento 5 Stelle was an unusual and unexpected government between two parties with opposite political ideals, that is the reason why European political elections were considered so important: they would have given the possibility to know on which side, Lega Nord or Movimento 5 Stelle, is the Italian population, and consequently who would have gained more power within the coalition.[11]

3.4 Using Google Trends

Is it really possible to predict which candidates electors are going to vote only on the base of Google researches they make? S.S. Davidowitz affirms that we cannot limit ourselves to study only which candidates are most frequently searched, because a lot of people search for a candidate because they like him/her, but a same number of people search for a candidate because they hate him/her. Anyway, S.S. Davidowitz and Stuart Gabriel, financial professor at the California University, Los Angeles, found out a surprising indicator about how people are going to vote. A big amount of researches regarding elections contains the name of both the candidates. During American presidential election 12% of researches with "Trump" included also "Clinton", more of one quarter of researches with "Clinton" comprised also "Trump". These researches can give us some indicators on who is the candidate that a person sustains; it is sufficient to take care of the sequence in which names of candidates appear. The study of S.S. Davidowitz and S. Gabriel shows that the way people makes tech queries online is not random, in a research which contains the names of

[11] Finesi, Giorgio. From elections to Conte's government, almost 90 days of political crisis. Skytg24, June1, 2018. Online article: https://tg24.sky.it/politica/2018/06/01/governo-conte-storia.html

both candidates, a person is more willing to type first the one who he or she sustains. They proved that, confirming that in the three precedent American elections the candidate who won was the one who appeared as first in a major number of researches. Moreover, the order by which candidates were inserted anticipated the results of a specific state. This indicator can give us information that polls do not get because electors lie to themselves or are not willing to reveal their real preferences.

This is the first method used in the European elections prediction: comparing queries. We compared queries which included the name of the two main Parties' leaders: Salvini, leader of Lega Nord; Di Maio, leader of Movimento 5 Stelle.

- "Salvini Di Maio"
- "Di Maio Salvini"

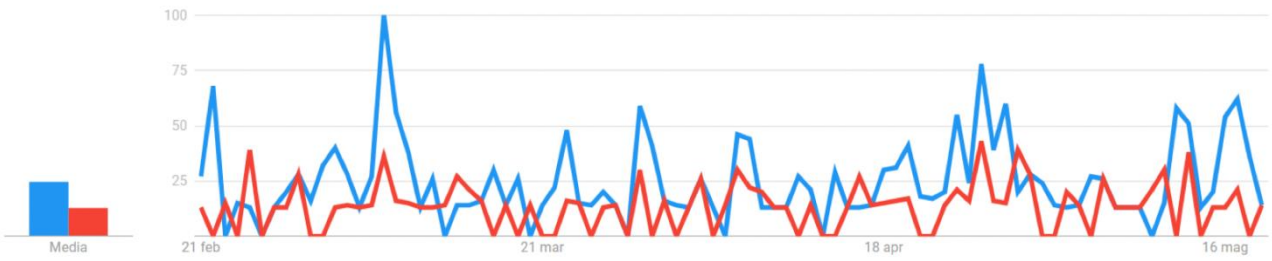


Figure 8. Comparing queries- Google Trends

According to this method there were no doubts that the winner of the European elections would have been Matteo Salvini, leader of the Lega Nord Party. The researches, which gave precedence to Salvini, exceeded the ones giving precedence to Di Maio.

3.5 Turning point

Anyway, we knew and proved that Salvini had not always been the most preferred one by the Italian population: Google Trends shows clearly that there was a "turning point" at which the trend change, shifting from Di Maio to Salvini.

- "Salvini"
- "Di Maio"

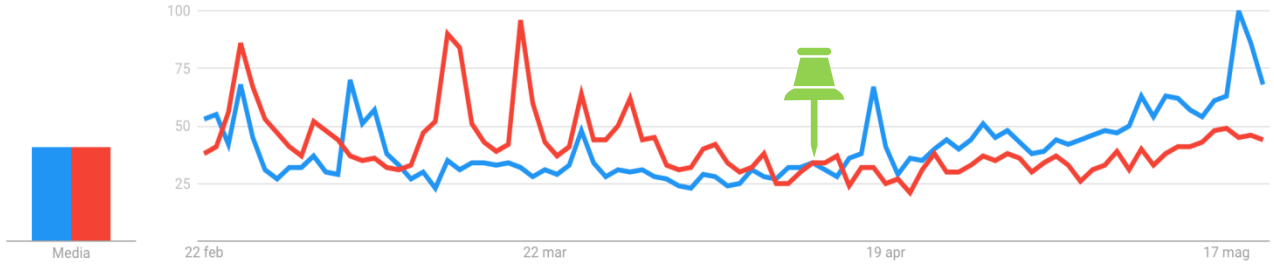


Figure 9. Turning point: trend shifts from Di Maio to Salvini- Google Trends

The following picture shows, region by region, the researches made during the period of the last governmental elections (03/04/2018). It is evident that the Preferred party was Movimento 5 Stelle, without any regional exception.

Figure 10. Research on Google during the period of the last Government Elections (03/04/2018)- Google Trends

This other image displays the researches on Google during the period of European election (05/26/2019). It is evident that the preferred Party was Lega Nord, without any regional exception.

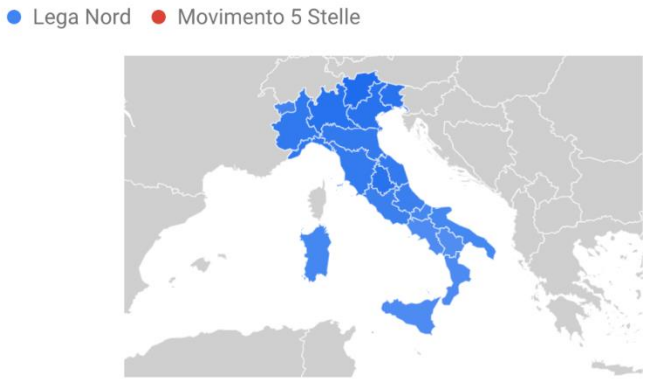


Figure 11. Research on Google during the period of the upcoming European Elections (05/26/2019)- Google Trends

3.6. Political situation right before the European election showed by Google Trends

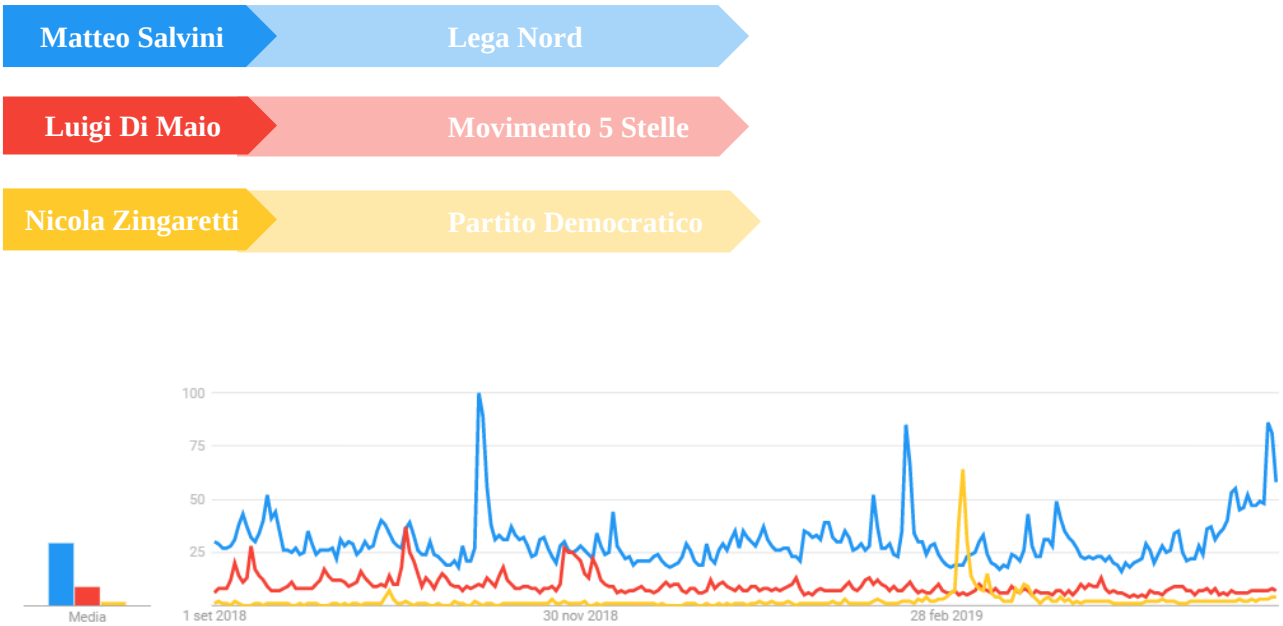


Figure 12. The trend of the leaders of the three main parties- Google Trends

Here we can see the trend of the leaders of the three main parties and observe that Matteo Salvini was far above the other two leaders, another indicator of his possible success at the European election. There is only one significant peak of Nicola Zingaretti, but is explained by the primaries of his party, the Partito Democratico.

3.7 Popularity of Matteo Salvini

The success of the figure of Matteo Salvini is reflected in his popularity on social media: " after the election on the 4th of March 2018, the Lega's leader gained 3 million followers on Facebook, surpassing Angela Merkel, Prime Minister of Germany (2.5 millions), Marine Le Pen, president of the Front National (1.5 millions) and all other European leaders. Even Trump, who has 22 million follower, is surpassed by Matteo Salvini in both engagement and involvement with his audience: 2,6 million in one week against 1 million and a half. And the America's population is 5 times that of Italy."[12] In Italy, in October 2018, Matteo Salvini was in the lead for numbers of followers on Facebook with 3.233.000 million, followed by Luigi di Maio (2.100.000 million), a really good result, taking into consideration that until June 2018 he was almost unknown to the public.

Usually during the political elections the leader substitutes the Party and after the elections Party's trend goes back to linearity. Instead, Matteo Salvini has substituted himself to the Party since he has become the leader of Lega Nord, as is evident in the following graph.

[12] From the interview to Luca Morisi, in charge for the communication strategy of Lega Nord- Panorama, November 12, 2018. Online article: https://www.panorama.it/news/politica/salvini-intervista-luca-morisi-web-facebook/

- Matteo Salvini
- Lega Nord

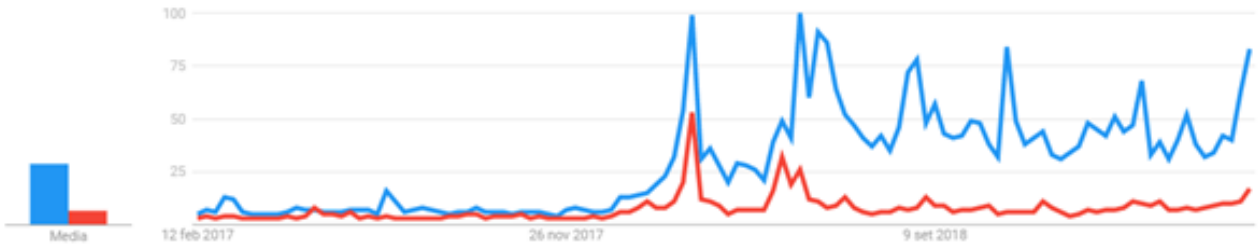


Figure 13. The trend of Matteo Salvini compared to the one of his party, Lega Nord- Google Trends

In order to prove such popularity of Matteo Salvini, we tried to compare the Google researches about him with the Google researches of some of the most influential people in Italy. Thank to Google Trends we were able to obtain the following graphic:

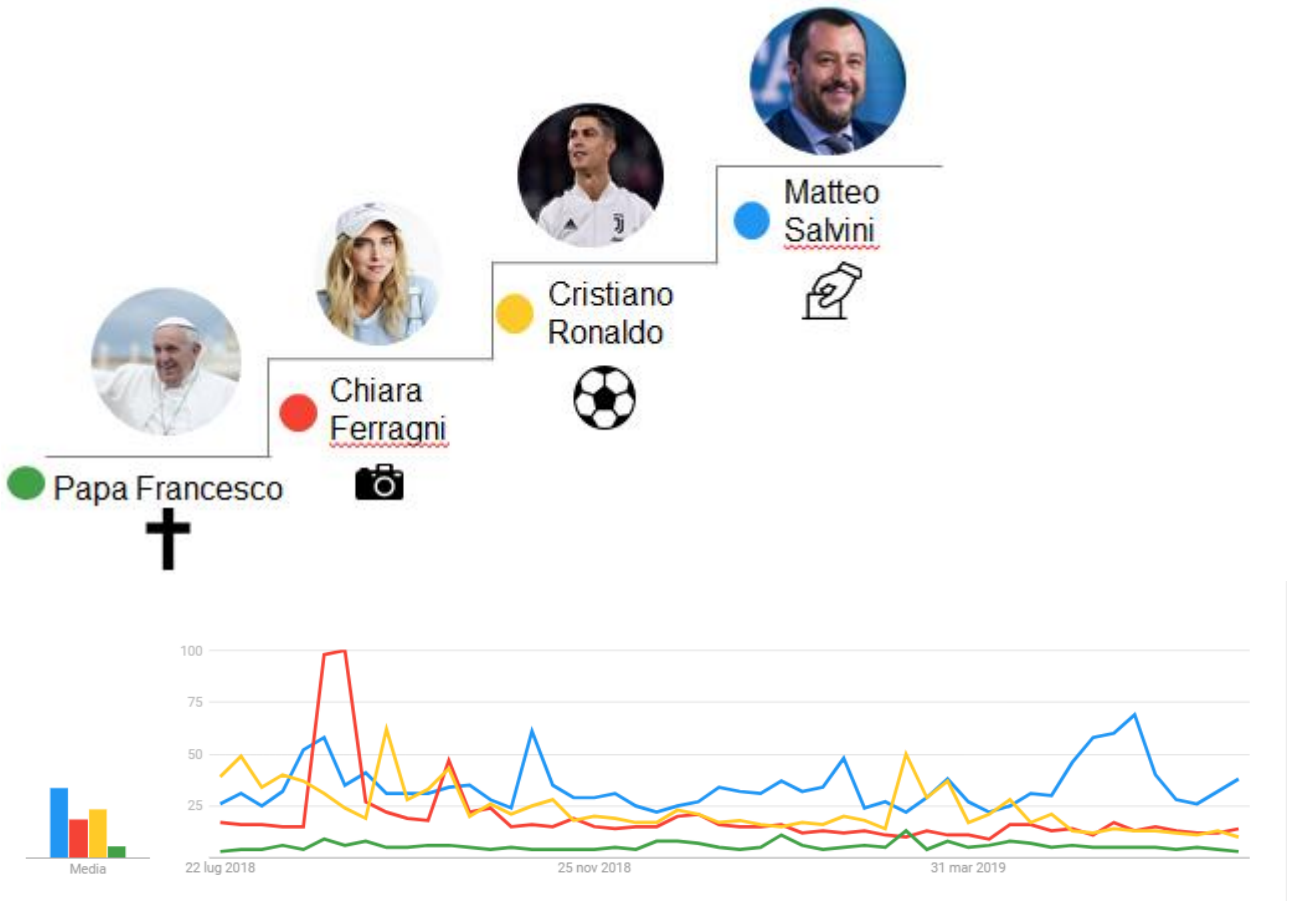


Figure 14. Popularity of Matteo Salvini compared to other famous Italian characters- Google Trends

We took three personalities who are influencers in different areas: The Pope, Francesco Jorge Mario Bergoglio, chief of the Catholic Church; Chiara Ferragni, entrepreneur and Italian blogger, in 2017 nominated by Forbes "the most important fashion influencer in the world"; and Cristiano Ronaldo, one of the most popular football players, at the moment engaged by the Juventus football team.

Except for a peak of researches on Chiara Ferragni during her wedding weekend at the end of August 2018- beginning of September 2018, even her, with 16,8 million followers on Instagram, was not able to beat Matteo Salvini in terms of Google researches' popularity. It seems that no other famous character in Italy can match Matteo Salvini's popularity.

3.7.1 Matteo Salvini and the Sentiment Analysis: "The Beast"

Matteo Salvini seems to be always instinctive and spontaneous, most repositioning are studied by a computer through an advanced system, which consents to Matteo Salvini to appear "evil", "despicable" or sometimes "nice", but always in harmony with the predominant humour of public opinion. This happens thanks to an informatics system that it is called "The Beast", for his cynic ferocity. Equipped with a capacity to understand what people want and to get inside the mind of people in an unnoticeable way. The centre of Salvini's system is Facebook. Is on Facebook's platform that he spreads his more effective messages through a system which analyses time by time in a scientific way thousands and thousands of tweets which obtain major results. And then, messages and keywords are prepared, ready to be

promoted by Salvini on Facebook. It is today evident that the Salvini's posts have been made to gain three principal objectives:

- the first is to occupy soon the media space, step in as first on the daily news, so that the others traditional media, opponents but also allies, are constrained to go after him
- the second is to polarize all the debates around the cross-road: for Salvini and against Salvini
- the third is to launch strong messages, taking out from the public opinion "negative feelings" such as anger, fear and aggressiveness

In fact, according to a study of the Michigan University (2005), the negative emotions are a way to gain attention, even if is the positive sentiment which takes back the voter to the electioneer. So true is it that at the end of each post, Salvini proposes a fragment of joy to who is listening to him, like "big kisses and good Friday to all"- by fading out the aggressive message launched right before.

Matteo Salvini has made a reverse engineering of a social campaign. It consists in downloading as much post as possible and starting to study how many likes those have obtained. Then a sentiment analysis is made, an analysis of the language used to understand the shadow and the emotions aroused.

Below there is an example of a sentiment analysis based on more than four thousands of Facebook posts published between 2011 and 2018, collected by API. [13]

[13] Piccinelli, Francesco. The secrets of Matteo Salvini's online strategy on social network. Wired.it. 15/02/2018. Online article: https://www.wired.it/attualita/politica/2018/02/15/matteo-salvini-strategia-social-network/

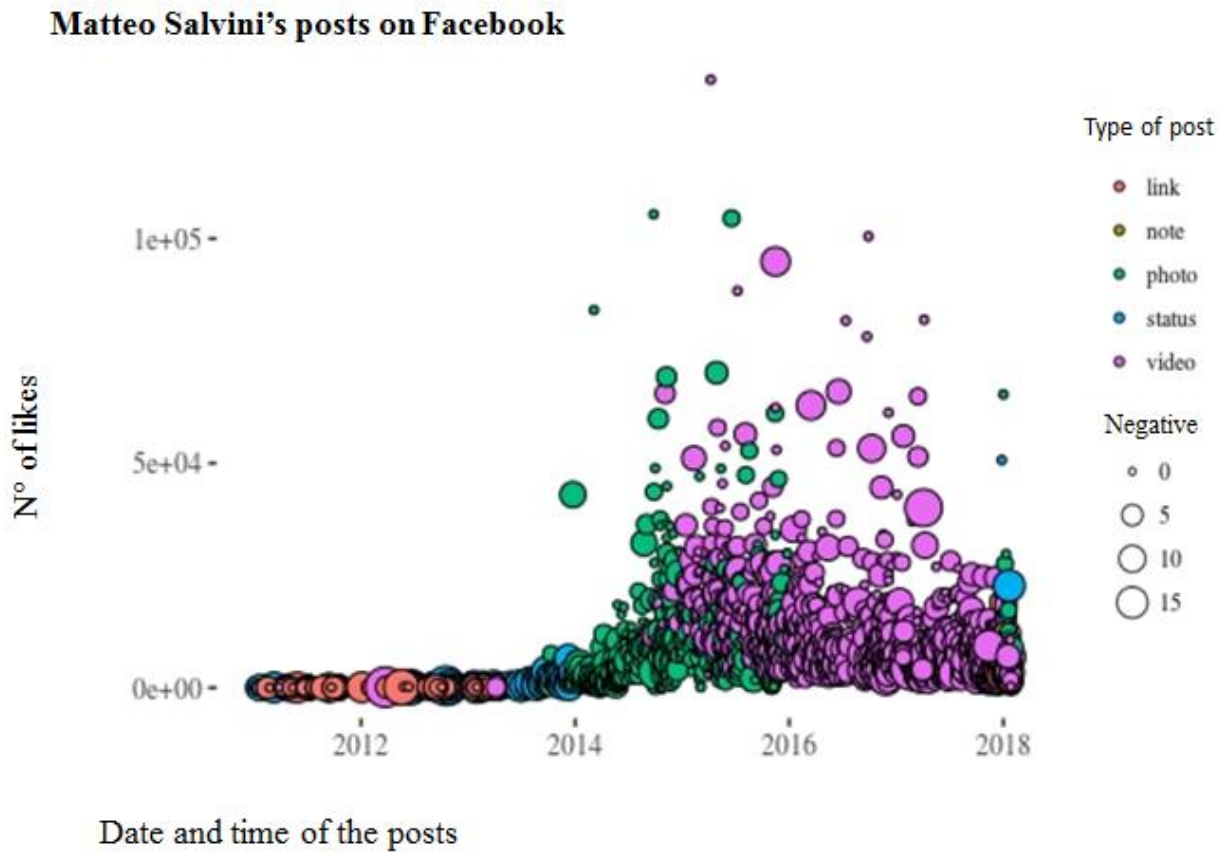


Figure 15. Sentiment analysis of Salvini's posts (2011-2018)

The sentiment analysis studies each phrase of a post through ten dimensions and assigns them a score. In this case the scores have been summed up to have an idea of how much high (or low) have been the tones of the texts in question.

How it is possible to see from the graphic, his strategy is mainly focused on video, but it remains diversified including photos, notes, links, status.

The evolution of the social strategy of Matteo Salvini has at least three explanations:

- from 2014 -the year in which the graphic shows a significant boost on social media- Salvini was getting close to a crucial test for his party and his leadership
- he wants to propose himself a credible leader
- the spin-doctor of Lega Nord have learned to refine the posts' language

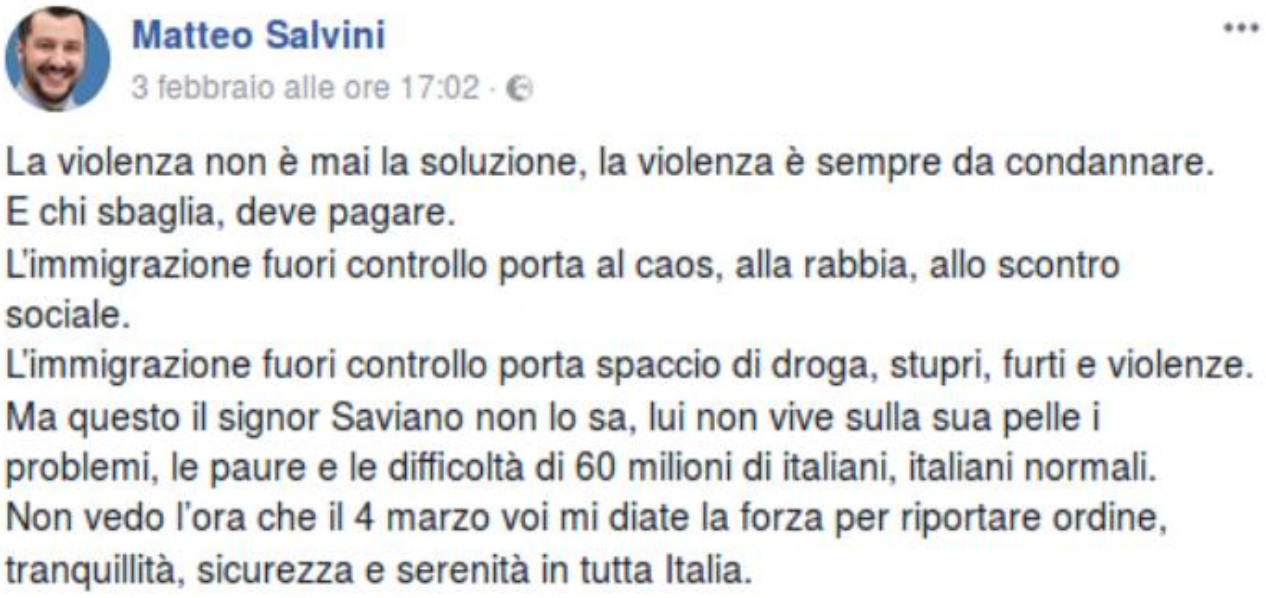


Figure 16. Matteo Salvini's post on Facebook on 3rd February 2018

How it has become accurate the way in which Salvini's staff write his posts on Facebook is evident from a post that Salvini shared on the 3rd February 2018 in which he commented the Macerata's shooting. First of all he started to use cognitive switches, such as "immigration out of control", "Saviano", "drug", "rapes". Secondly, from this post comes out a positive sentiment. A counter-intuitive fact: in the post Salvini comments the shooting made by an extreme right movement made to revenge the death of a girl. Nevertheless the algorithm notes a positive shade.

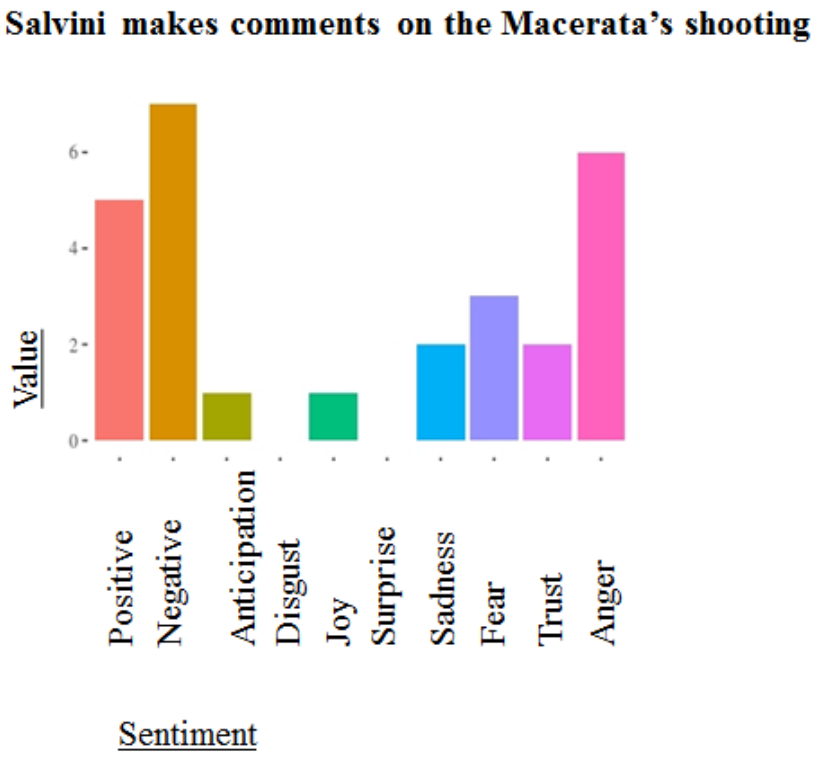


Figure 17. Sentiment analysis of Matteo Salvini's post on Facebook on 3rd February 2018

The rhetoric strategy is clear: to let the guard down leveraging on anger and fear, but suggesting that, giving trust to Lega Nord, things will get better. This is why, even if lower, there is also a positive side on this post. This can seems as an error, instead gives us an idea about the functioning of the sentiment analysis algorithms. Cross-referencing the words around ten dimensions it appears that the "hardest" comment on one of the more complex tragic events of the present has some intrinsic joy. But was not always like that. In the past Salvini's post was less complex. In the first posts of Salvini the predominant sentiments were anger or sadness. For example, even in the post in which he was celebrating the obtainment of the Secretary of Lega Nord in 2013 there is no joy at all.

Figure 18. Salvini's post on Facebook on 7th December 2013

This post reveals that the social network strategy was just at the beginning. It is not as well edited as the Macerata's one. The polarization of Matteo Salvini's posts has evolved in time. In the following graphic posts are aggregated by day. It shows how the polarization is a constant element in the Salvini's communication strategy, but only in the last year, 2018, has become a structural component in his posts.

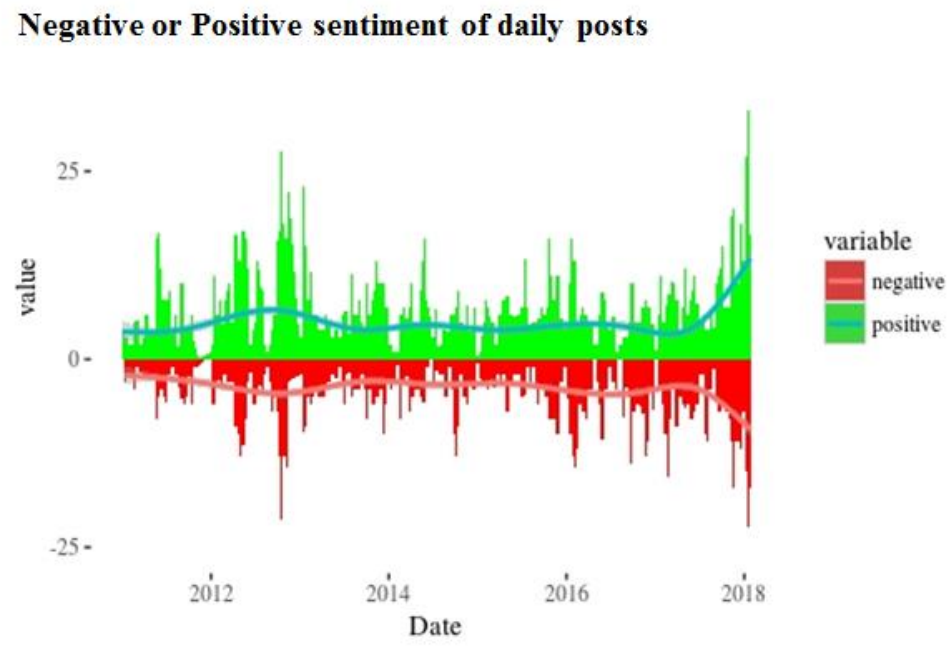


Figure 19. Negative or Positive sentiment of Matteo Salvini's daily posts

How much is effective this progressive refinement is evident in the likes curve. The daily likes in the last months of 2017 have increased significantly. There are a lot of peaks due to single episodes, but the trend from the end of 2017 is clearly remarked.

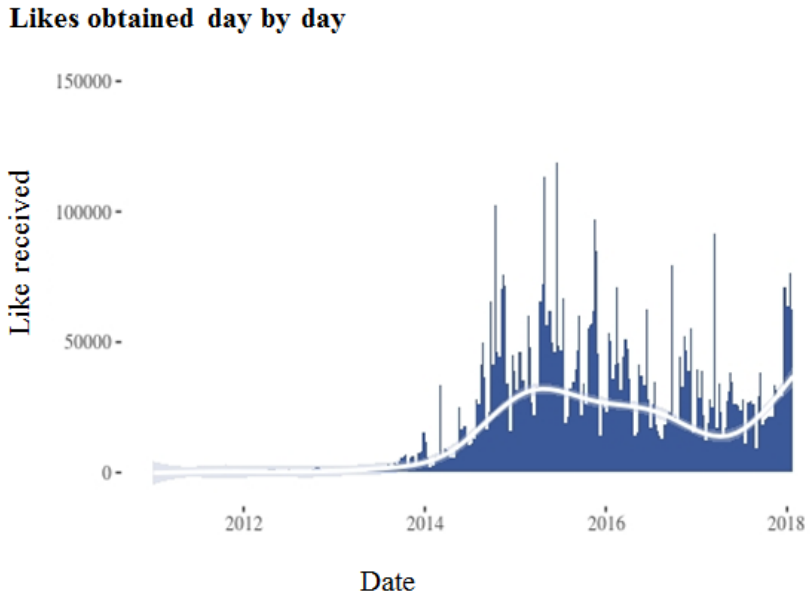


Figure 20. Likes obtained day by day by Salvini from 2011 to 2018

This ascending trend is the effect of two complementary factors: the first is about an increased competence in the sentiment management in the post writing; the second is the fact that- even if he almost always used video between 2005 and 2019- had different ways of communication (photos, links, status, notes).

3.7.2 How Matteo Salvini defeated Luigi di Maio through social network

It has proved to be successful the Salvini's decision to bet on the digital communication right before the European elections, deciding to sponsor Facebook posts and increasing the number of Twitter posts. On the other side the Movimento 5 Stelle decided to cut on the Facebook posts. A choice that sounds in contrast with the digital origins of the Movimento. It was Matteo Salvini the one who invested more on social media and in a more targeted way- setting age and gender of his audience. With a budget of 128.782 thousands of euros, three times of what the Movimento 5 Stelle allocated, amount spent from March to 25th of May and of which 41.000 thousands euros in the last decisive week. Budget spent to direct not only the promotion of public appointments, but also the comments on news and the key messages of the electoral campaign. This reminds us that we exist only if we are online. Otherwise we are irrelevant. Italians are connected on average 2,3 hours per day, in which 31 million of citizens (most of all voters) open Facebook – become for many the main source of information. [14]

[14]Lo Conte, Marco. The three social moves by which Salvini defeated Di Maio. Il sole 24 ore. 26/06/2019. Online article: https://www.ilsole24ore.com/art/le-tre-mosse-social-cui-salvini-ha-battuto-maio--ACPxLHJ

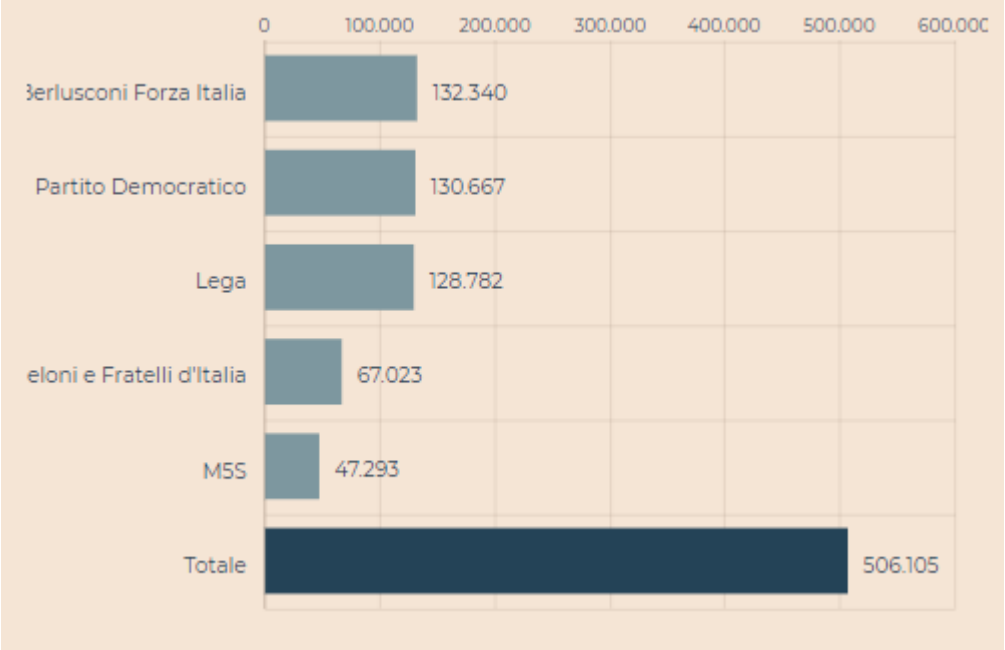


Figure 21. How much Italian political parties invested in social communication during European elections 2019

It is possible to gain votes of undecided electors at the very last moment thanks to the use of a coherent message with the capacity to intercept wishes and needs of the on the fence ones. On this side the campaign set by the Salvini's staff was winning because of two rhetorical artifacts: the question and the greeting. The custom is to say "who ask is the one who runs the show". The interrogative point has been transformed in a powerful tool of political engagement, which in the last months has constitutes the point of contact with large groups of "volatile" voters. The "big kisses" at the end of the posts has become a refrain who has rubbed off on supporters and detractors.

Contrary the Movimento 5 Stelle in the last weeks swerved in a moderate direction: moving on to moderate tones and praises to the institutions, moves which resulted in contrast with the historical approach of the Movimento, the "change". The fact that has distinguished the political communication of Salvini and the one of Di Maio is

most of all the “climax”: M5S limited himself to conquer the center, instead Salvini set up a strategy with a starting point and an objective.

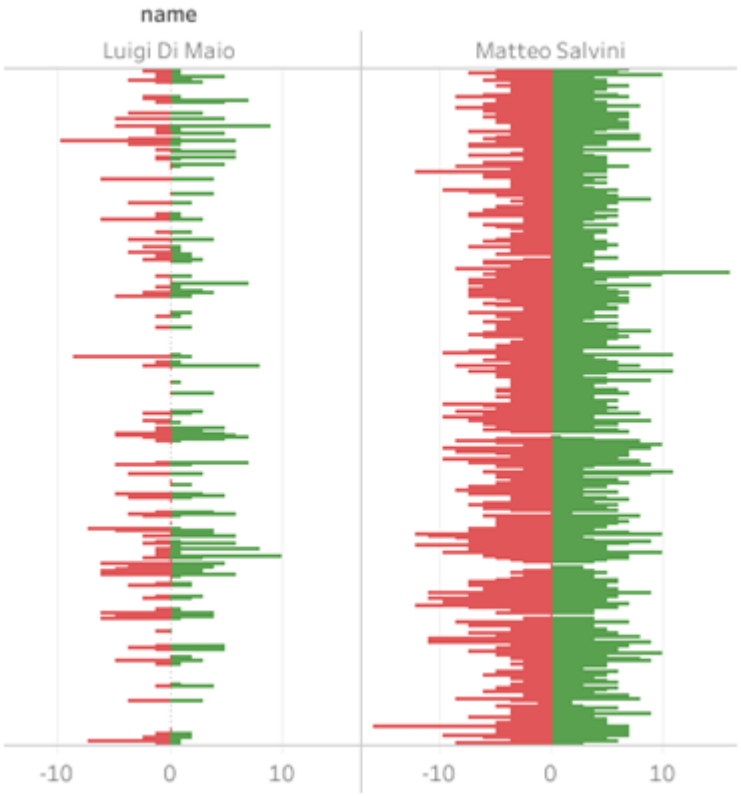


Figure 22. Analysis of the tweets on the positive/negative dimensions

The emotionally exasperated vocabulary of Salvini is opposite to the measured one of Di Maio, who tried to limit anger and fear. Salvini is not afraid to use a language full of contrasting emotions, differently to Di Maio he does not want to reassure anyone. He just want to maintain the center stage and to do that he needs to polarize the debate. This situation is clear in the analysis of the tweets on the positive/negative dimensions (figure 22).

The difference of tones arises also from the word clouds obtained through the sentiment analysis of the contents published on each other’s profiles of Facebook. The focus on “Italy” registered the strategic shift of the center of gravity of Lega

Nord: not anymore the fortress of Nord defense, but a point of defense of Italy against the bureaucracy, represented by the European Union, and the "danger of the stranger", represented by the immigrants. He founds a shared enemy to fight against. On the side of the Movimento 5 Stelle, words connected most of all to institutional activity of the Minister. Less slogans, less propaganda and more sober and less emotive vocabulary. The social communication of Di Maio is the attempt to preserve the original aggregative identity of the Movimento, at least in the language, but projecting it in the vocabulary of who is governing. This reflects the contradiction of the M5S, who is not anymore just a digital movements "from the bottom" , but neither a real political party. Di Maio's vocabulary embodies the incongruence of an anti-system non-party which today is in government and represents the institutions.

Figure 23. Word cloud of Salvini's posts

Figure 24. World cloud of Di Maio's posts

Furthermore it has to be noticed that in the last weeks before elections were several posts on Facebook and Twitter in which Salvini was busy in daily life activities, among which stood out the central topic of the food. These activities constituted a proximity between the Salvini's brand and the potential electors, to which followed the engagement made through the questions and the assertiveness in the political message of the last lines of electoral campaign. Besides, already in 1960 was the use

of television which consented to JFK to defeat Nixon at the presidential elections in USA. Today television has been replaced by Internet, and social media in particular.

3.8 Is the rate of interest a parameter to predict European elections results?

Here we have the trend of interest of the last three European elections and we can see also the percentage of participation corresponding to each election.

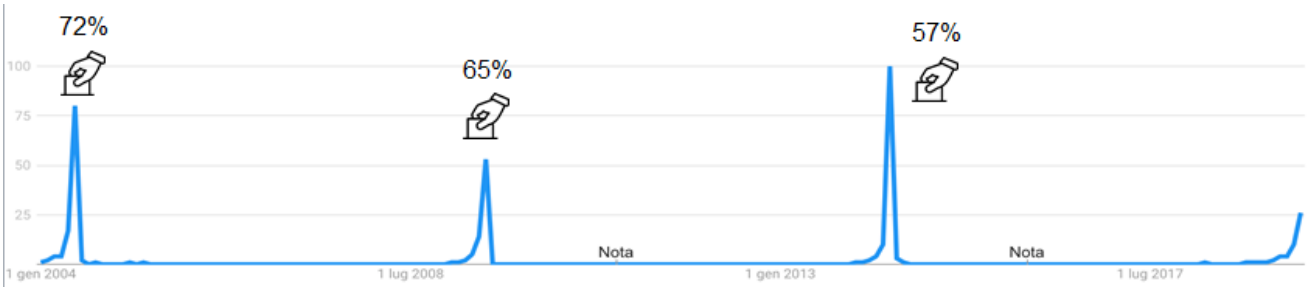


Figure 25. Trend of interest in the last European elections and percentage of participation- Google Trends

We can observe how interest in the European elections reached its peak at the last election (2014) even though the affluence was decreasing. It is evident that the rate of interest is not a parameter to predict European elections results. People search for information, but this does not mean that they are going to vote. On the other side, even if we ask people what they intend to do, vote or not vote, or who they are going to vote for, they are going to lie. Roger Tourangeau, professor at the University of Michigan, affirms that we lie in anonymous polls because of our weakness for "white lie". For one third of the time, people lie in everyday real life and this attitude is repeated also in surveys. Added to this, there is the use to lie to ourselves, we are reluctant to reveal to ourselves who we really are. We deceive ourselves and consequently we cannot be honest in polls. Another factor is the desire to make a

good impression on the interviewer. This is why the more the interviews are impersonal the better it is, as they will be more trustable. Internet polls are better than phone polls, which in turn are better than personal polls. People are more honest when they are alone and not in a room with other people. The problem, analyses Tourangeau, is that people do not have incentives to say the truth in polls. When there were not official sources, before the advent of Big Data, it was really hard to know what people really think and do. This is the second power of Big Data mentioned above: some online sources lead people to say things they would never admit in other circumstances. They become a digital truth serum. In research on Google people are online, alone and there is no interviewer. And most of all people have incentives: for example if they have a problem they have no interest to reveal it to a poll, indeed they would be ashamed, but they have reason to search for a possible solution on Google.[15] Even if we lie to ourselves Google can know the truth. Before elections we can lie to ourselves and think that we are going to vote, but if we had not looked for information about where is the polling place, data scientist can predict that affluence will be low. Anyway the power of Big Data is limited in the contest of elections because they cannot predict how many people will be lazy and will prefer to stay on the coach instead of going to vote, how many people will have a setback or how many people decide at the last moment not to vote. This is why the rate of interest can be a parameter but it is not one hundred per cent reliable. How the graphic above demonstrate: in the last election in 2014 the percentage of researches increases, but the affluence does not.

[15] From "Everybody lies" written by Set Stephens-Davidowitz- chapter 4 The serum of digital truth

3.8.1 Abstentionism

In Italy, this contrast between interest and affluence can be also explained by the dissatisfaction of Italian population, who search information on how to vote and who to vote, but reading everyday a new scandal in which one of the candidate is involved, decide at the end not to vote, because they would not choose any of the possible candidates and would not know who is the "least worst" choice. Today, we vote for the least worst, but a country should vote for its first choice, for the one in which it beliefs, the one who reflects its ideals. Sadly, this is not the case anymore.

Gianfranco Pasquino, political scientist, has individualized three causes of abstentionism[16]:

- Tendency to participate only to political elections which are considered "more important"
- The similarity of ideals and proposals of different candidates, with the consequences that the win of one or of the other will have the same impact on the Italian citizens' life
- Crisis of parties, which are not able to mobilize electors anymore

According to an enquiry of Cmr Intesa Sanpaolo for La Stampa, 2015, before regional elections, 52% of Italians, do not feel part of any party. The ones "without party" are such because they think that politicians do not care about people's problems and that voting is useless. Who has declared not to be sure to vote and who has said to be sure not to vote specified his/her motivations: 37,4% - politicians are

[16] Openpolis. June 23, 2016.Online article: https://blog.openpolis.it/2016/06/23/perche-le-persone-non-vanno-votare-cause-astensionismo/8870

not interested of "normal" people; 27,5% - to vote is useless, does not change anything; 15,2% - Parties suck.[17]

The following image shows as, according to the Demopolis Institute, almost 20 million Italians could have not vote on the 26th of May (27 million voting, 5 million in doubt, 18 million not voting)

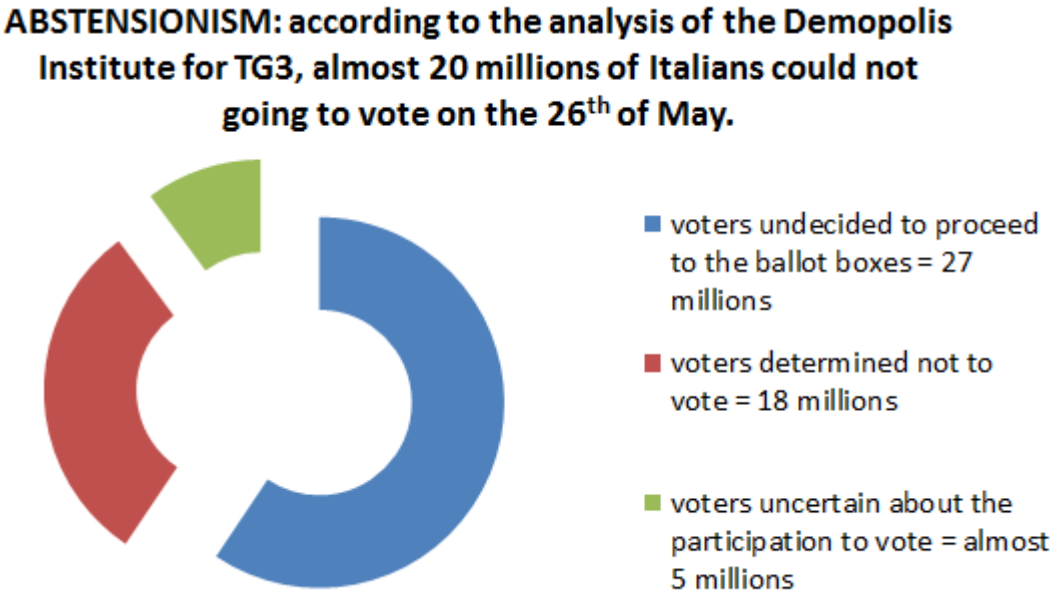


Figure 26. Abstentionism: signs of insecurity about European vote

It is possible to talk of political disaffection, in particular, between the youngest generations. The disaffection is due to a generational reason, in a society that promises but does not offer. Absence of perspectives, rate of unemployment, lack of future investments and life's instability lead to the diffusion of an anti-system attitude, that some political parties try to interpret making themselves promoters of a "wave of change", as Movimento 5 Stelle tried to do, and was able to, at least until it

[17] F.Q. Elections, The Italy of abstentionism. The 52% does not recognize himself in any party. Il fatto quotidiano, May 15, 2018. Online article: www.ilfattoquotidiano.it/2015/05/18/elezioni-litalia-dellastensionismo-il-52-non-si-riconosce-in-nessun-partito/1694292/

reveals itself as just one of the others, who, once obtained the consent, does not change anything, but it is ready to compromise to remain in power and make its own business. In addition to dissatisfaction there is a new tendency to be passive, not involved with political institution, a sort of presence-absence, a feeling of weak identification which can explain, at least in part, the electoral apathy.

According to the study "Electoral participation in Italy", carried out by Maurizio Cerruto, Sociology professor at Cagliari University, we can talk about two types of abstentionism: "abstentionism of apathy", which reflects the distance between elector and the political offer; "abstentionism of protest", as active expression of electors' dissatisfaction, which express distrust and often even hostility towards the political cast.

In particular, the European elections have always and everywhere registered a low rate of participation, probably because of a weak awareness of the weight of the European Parliament and of a distance condition between everyday life and European institutions. Euro-barometer reveals that only 48% of European citizens believe that their voice is relevant in the European Union, even if there are differences from country to country (in Sweden 90% of citizens believe that his/her voice has a weigh, differently from Italy, with just 24%).

AFFLUENCE TO EUROPEAN ELECTIONS

Peaks of participation:

Italy - 1979: 90,35%

Germany - 1980: 88%

France - 2007: 83%

United Kingdom - 1992: 77%

European elections average: 1979: 61,99% - 2014: 42,61%

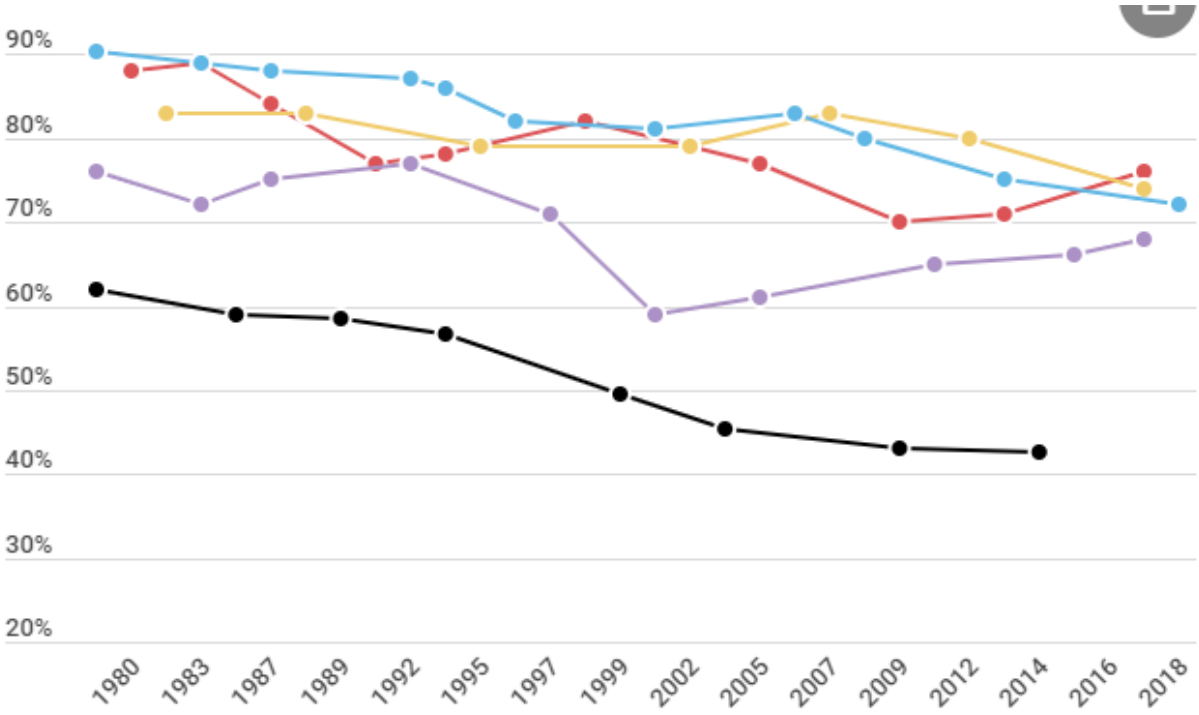


Figure 27. Affluence to European elections

Anyway, it is not about trust. Euro-barometer shows that European citizens use to trust more European Union than their own Parliament. So, why did just a minority vote for the European elections? This paradox can be due to two factors, as Alberto Alemanno, analyst of European politic, explains: “First of all, the European elections are still the sum of national elections, and not a transnational political event animated

by real European Parties; secondly, there is not a common public sphere which is able to talk about the European political system." Moreover European elections' affluence is difficult to be analyzed in time, because, as Jules Baley from SciencesPo University, Paris, notes, "How can we compare the affluence of 1979, when European Community was made of just nine countries of West Europe, with the one of 2014, when Union counted twenty-eight countries with different political cultures and different democratic traditions?"

Talking about European elections' affluence we also have to take into consideration that more and more citizens have difficulties to vote because they moved to another country. More than 10% of Rumen, Bulgarian, Croatian, Latvian and Portuguese citizens live in a member State that is not the one in which he/she was born. They have the right to vote in the cities where they live in, but it is difficult to identify yourself with parties and politicians who speak only to electors of their nationality. In 2014, 95% did not go to polling places because of language, bureaucratic and political obstacles.[18]

The creation of a one and only European electoral collegium, which permits to present transnational lists and parties, can be a solution to abstentionism. In this way parties will be boost to focus on topics of European interest, instead of focus, as now, on national issues. Maybe in that way the debate will be successful in deeply involving the electors.

[18] Ottaviani, Jacopo and Ferrari, Lorenzo. What happens with abstentionism in Europe. Internazionale, May 16, 2019. Online article: https://www.internazionale.it/notizie/jacopo-ottaviani/2019/05/16/elezioni-europee-astensionismo

3.8.2 Affluence of European election 2019 in Italy

Figure 28. Trend of interest in the last European elections and percentage of participation- Google Trends

The low participation has been confirmed also in these last elections. How the precedent graphic shows, even if the interest increased in the European election 2019 (it is evident in the last peak of the graphic), the affluence did not. Actually it decreases from 57% at last European elections (2014) to 56% at the actual European elections (2019). Only in Trentino Alto Adige people voted more than in the rest of the country. The interesting and curios fact that can be observed is that, contrarily to what could be deduced from Google Trends, the fall of participation was registered most of all in the South of Italy. Google Trends reveals a higher interest in a lot of Southern regions, as the following picture shows. The researches for "European elections" are higher in the South of Italy.

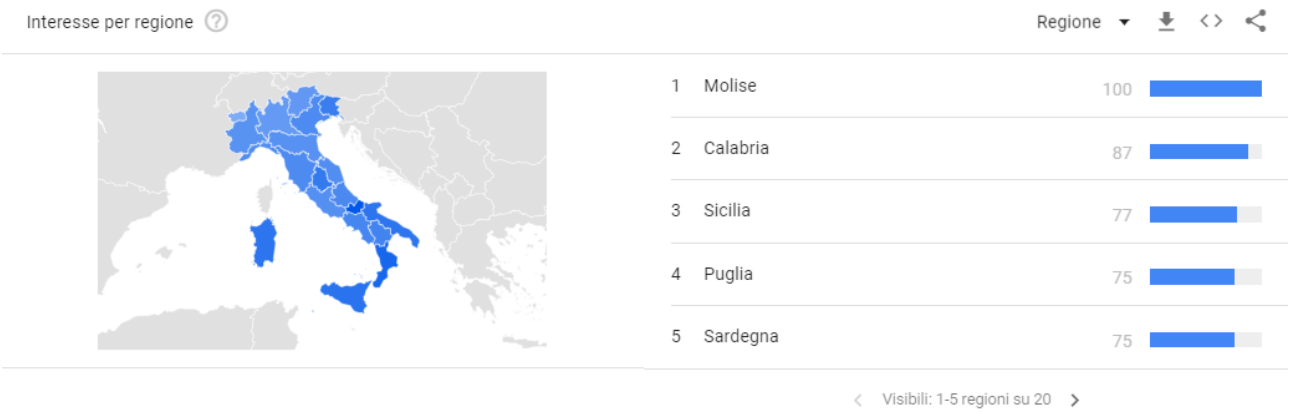


Figure 29. Researches for "European elections" region by region- Google Trends

Despite, the participation falls in the South: in Sardinia and Sicily it was under the 40% threshold, in Calabria it was just above 40%. In Campania and Basilicata the participation arrived to 47%, in Apulia it does not reach half of the voters, as instead happens in Molise (53%) and in Abruzzo, where however the participation shifts from 62% to 52%.

This data are in countertrend if compared with the data of the affluence in others European countries. There is an increase in participation in the principal European countries- in France 43,29% against the 35% of 2014, in Spain 34% against 24% of 2014.

3.9 Results of European elections 2019 in Italy

The low affluence to vote in the South of Italy is explanatory of the low success of the Movimento 5 Stelle -the first Party of the country until these elections- in fact, these regions have usually been sources of votes for the Movimento, which has always had high consensus in these areas.

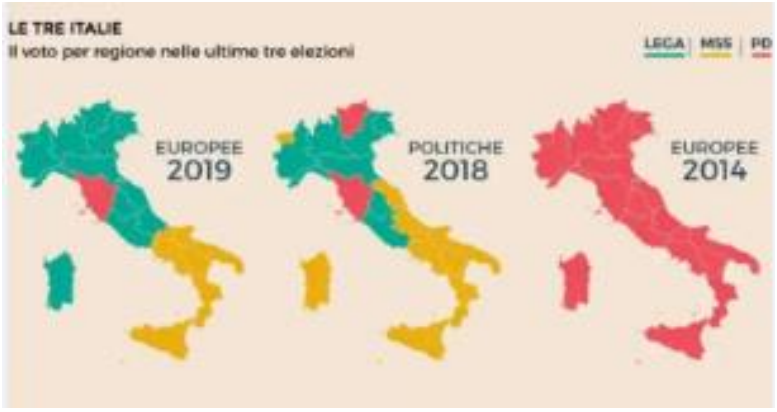


Figure 30. European elections 2019/ Political elections 2018/ European elections 2014- vote region by region

The results of European elections in Italy are the following:

- Lega Nord is the first Party with 34,3%
- Partito Democratico is the second Party with 22,7%
- Movimento 5 Stelle –which was the first Party- become the third one with 17,1%
- Affluence: 56%

According to these results, it is possible to say that the prediction made was right: Matteo Salvini, leader of Lega Nord confirmed his popularity, being the leader of the first Party. The analyses made using Google tools left no doubt. Anyway, as already said, we did not expect such a low participation in the South and we did not expect such a successful outcome for the Partito Democratico.

In the South, taking into consideration our analysis and seeing an increasing interest, it was expected a higher participation, also due to the fact that, compared to the past, in which more people still did not use internet and used more traditional ways to gain

information, today Google is supposed to be the main instrument in the whole country being used to search for everything.

In order to explain the 56% threshold of total affluence, and the even worse rate of participation in the South, it could be interesting to read the report of the linguist Tullio De Mauro. According to the former, in Italy there are 13 million of semi-illiterate, who just know how to sign, but do not understand what they read, and 13 million of illiterate, who have lost the fluidity in writing and reading. The total is 28 million on 52 million under the threshold of literacy's sufficiency. How are they going to vote? Watching television. This is why the high quality of television programs is so important. From 2000 the newspapers' copies sold have diminished by 32%: from 6 million to 4 million. Subtracting the copies of sport's newspapers, only 461 thousands of copies remain. The percentage of Italians who have a comprehension of political topics is under the 30% threshold, and the 33% is not able to completely understand a newspaper's article. And what about Europe? Europe is unknown to more than 57% of the Italian population. A research made by Ocse confirms De Mauro's opinion, according to the Organization for the Cooperation and Economic Development, 47% of the Italian population, that is one Italian out of two, informs himself, votes and works following just an elementary capacity of analysis. This costs to the world economy 1.2 trillion of dollars. Save the Children affirms that in Italy 48% of the children between six and seventeen years old have not read a single book, if not the scholar ones; 55% of them have never been to a museum.[19] The disinformation and the ignorance are determinant factors for participation's rate.

[19]Valentini, Carlo. We are a nation of illiterates. Italia Oggi. 25/05/2016. Online article: https://www.italiaoggi.it/news/vinco-l-analfabetismo-finanziario-2372226

It is not a case that in the Northern European countries the participation to vote has always been high.

Shifting to the second factor which was not expected, the success of Partito Democratico, it can be interpreted as a sign sent by the Italians: there is a part of the country who does not want to sustain anymore populism and nationalism of the actual government and of Lega Nord in particular. It is a request by them to the PD to make opposition and to become an alternative. This was not reflected by our analysis, maybe because the popularity of Partito Democratico has increased just in the last period and it was still not enough to be significantly recognizable on Google tools. Furthermore, as is evident in the figure 27 the PD after the European elections of 2014 had lost all the popularity. It is starting to regain it just today.

3.10 Fun Fact

In the attempt to try to quantify the success of Matteo Salvini and sustain the thesis that he would have won the European election, using Google Trends and verifying some correlations, came up that a strong correlation with Matteo Salvini researches and researches on today's weather.

- "Matteo Salvini"
- "Weather today"

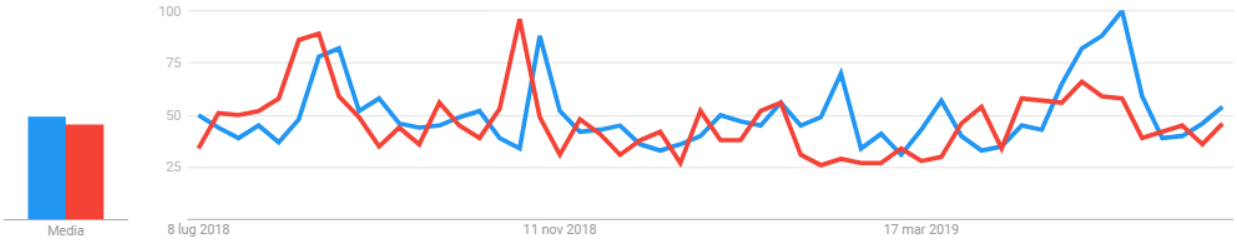


Figure 31. Correlation between "Matteo Salvini" and "weather today" Google researches- Google Trends

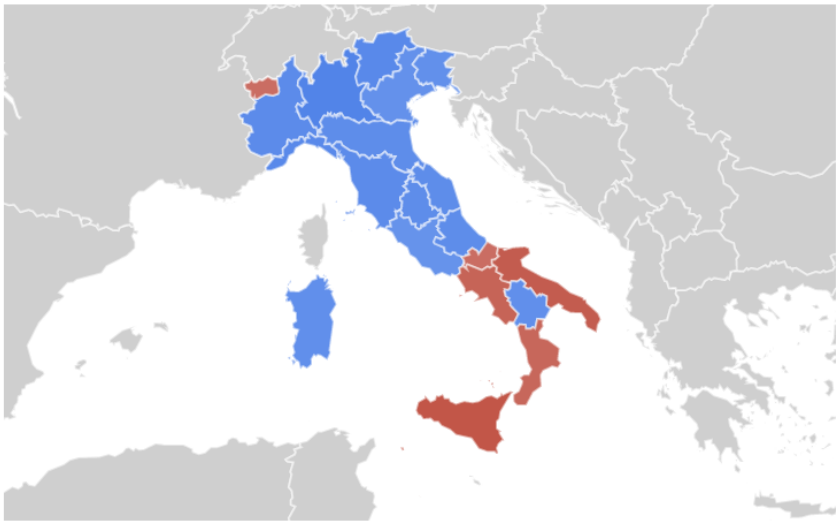
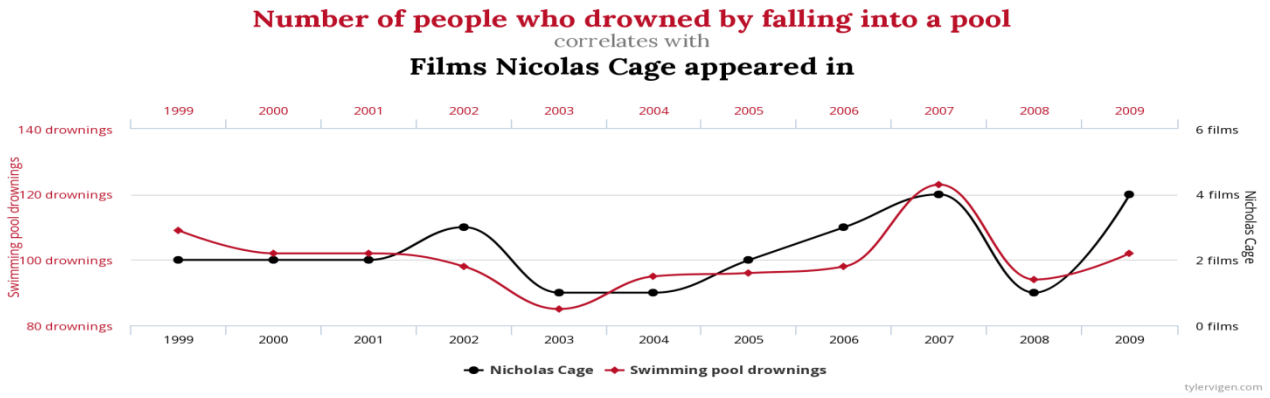


Figure 32. Correlation between Matteo Salvini and weather Google researches region by region- Google Trends

It seems that as Italians wake up they search for the weather and right after for Matteo Salvini. It is almost evident why people check the weather conditions: climate changes, in the last few years, led to an unusual weather, and we are always worried about possible rain. But why should they search for Matteo Salvini with almost the same frequency, if not more, by which they check the weather conditions?

With a shift of a week the correlation factor is 0.58. The conclusion can be that people search Salvini and within a week the sky becomes greyer!

Often Google Trends gives correlations that are not significant at all, or that are just casual and not causal, maybe this is one of these casual correlations. These are called "spurious correlations" -one of the most famous is the correlation between the number of films in which Nicolas Cage appeared in and the number of people who drowned by falling into a pool- a mathematical relationship in which two or more variables are associated but no causally related, due to either coincidence or the presence of a third "confounding factor", which is unseen. Anyway the weather-Salvini correlation can be considered interesting and "funny" in showing the popularity of Matteo Salvini, who has been able to become a fixed idea for Italians.



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

Figure 33 and 34. Spurious correlation between number of people who drowned by falling into a pool and films Nicolas Cage appeared in

CONCLUSION

This thesis could not have been made without Big Data.

As Vinton G. Cerf, vice president and Chief Internet Evangelist, Google, reminds "the terms artifacts means 'made by man' as opposed to 'found in nature'. It seems reasonable to view the progress of our species as the story of the artifacts we have created, applied, evolved and even abandoned in favor of new ones. There are almost always enabling tools that make us more efficient, more effective and more able to overcome human limitations. Artifacts often have economic consequences since they may reduce costs, increase productivity, free time for new endeavors, enable new ways of accomplishing old goals and way to accomplish new goals (…). Human progress is the story of our invention and adoption of artifacts. The Internet is simply another, rather large-scale artifact of human invention. What makes it so interesting is its scale and fabric of cooperation, collaboration and coordination that allows it to work. There is no central control."[20]

Given the words of Vinton G. Cerf, aside from the argument about Big Data seen as something new, which did not exist before, it can be affirmed that this huge amount of data, if used in the correct way, constitutes an enormous potential.

Big Data give to the firms the possibility to improve the operative efficiency, lower the costs, enhance productive performance, fasten the deliveries, and simplify the decisional process. All these innovativeness can happen only if there will be significant investments on technologies and most of all on human skills training. We have to focus on the fact that we make the difference, not the data by themselves -

[20] Cerf, Vinton G. Foreword of the book Internet Economics by Cellini Paolo. Roma, 2015.

computer security experts who work to prevent cybercrimes, are able to do that thanks to their skills and years of experience, because they can examine information and identify the one which can do harm and stop them. The merit goes to them, not to data. The hard work in the attempt to gain profit from Big Data is the achieving of the ability to define what we need to know from data and consequently to skim data and identify which can be those that will be able to help us to achieve the objective. In a day-by-day more competitive market, firms had to conform to the market and reorganize themselves, basing their strategies on the information obtained from the analysis of Big Data. Anyway, the maximization of profit through the use of Big Data is not possible without considerable initial investments.

Big Data can be helpful also in social and public contexts, as we have seen, in fact, they can be exploited to know citizens, cities and states problems and used to make prevention campaigns in order to reduce the number of victims, or murders, or incidents. The condition is that they have to be collected in the respect of the privacy without crossing the line of legality.

Privacy, loss of quality, ethical issues, hardware needs, need for talent, rapid change and “it is so much greater than reward” are just some of the accuses that can be moved against Big Data. As with any issue there are different opinions and this thesis is just an attempt to show the two faces of the coin, some possible pro and some possible cons.

The only things that, maybe!, can be subject of agreement, is the fact that we always have to keep in mind the importance to exploit technology, and so Big Data, but to

not be exploited by it, and so by them. Technologies should be considered as a support and as help, and they are probably the greater ones that we can have, without them we would not be in a such progressive and modern society; however, our enthusiasm for progress and innovations should not divert us from our human nature, who was the one who created technology and who is the one who has to detain the control of it.

Acknowledgements

This thesis is the result of a team work. I would like to mention and give thanks to Antonio Miloso, Francesca Celio and Stefano Ciccarelli. We worked under the supervision of professor Luigi Laura, who involved us in this beautiful project. We started to work on the European elections prediction two months before the date, 26th May 2019. And we had the great possibility to meet the author of the book – "Everybody lies"- which has inspired this work, Seth Stephens-Davidowitz, and to expose to him our results.

I would like to thank those who have been part of this project and who have made it happen.

Me, Francesca Celio, Antonio Miloso, Seth Stephen-Davidowitz and the professor Luigi Laura the day we had the opportunity to present our work to S.S. Davidowitz. (Stefano Ciccarelli is missing in this photo but was there with us)

Bibliography

Cellini, Paolo. Internet Economics. Rome: Luiss Press, 2015

Few, Stephen. Big Data, Big Dupe. A little book about a bunch of nonsense. Analytics press, 2018

Mayer Schönberger Viktor, Cuckier Kenneth, Big Data: a revolution that will transform how we live, work, and think. Milano: Garzanti, 2013

O' Neil, Cathy. Weapons of Math Destruction. New York: Penguin, 2016

Seth-Davidowitz, Stephen. Everybody lies. Rome: Luiss Press, 2018

Sitography

Finesi, Giorgio. “From elections to Conte’s government, almost 90 days of political crisis”. Skytg24, June1, 2018. Online article: https://tg24.sky.it/politica/2018/06/01/governo-conte-storia.html

Harvey, Cynthia. “Big Data Pro and Cons”. Datamation. August 9,2018. Online article: https://www.datamation.com/big-data/big-data-pros-and-cons.html

Lo Conte, Marco. “The three social moves by which Salvini defeated Di Maio”. Il sole 24 ore. 26/06/2019. Online Article: https://www.ilsole24ore.com/art/le-tre-mosse-social-cui-salvini-ha-battuto-maio--ACPxLHJ

Ottaviani, Jacopo and Ferrari, Lorenzo. “What happens with abstentionism in Europe”. Internazionale, May 16, 2019. Online article: https://www.internazionale.it/notizie/jacopo-ottaviani/2019/05/16/elezioni-europee-astensionismo

Piccinelli, Francesco. “The secrets of Matteo Salvini’s online strategy on social network”. Wired.it. 15/02/2018. Online article: https://www.wired.it/attualita/politica/2018/02/15/matteo-salvini-strategia-social-network/

Swanstrom, Ryan (Data Science Blogger, “Data Science101” blog). URL: https://101.datascience.community/about/

Valentini, Carlo. “We are a nation of illiterates”. Italia Oggi. 25/05/2016. Online article: https://www.italiaoggi.it/news/vinco-l-analfabetismo-finanziario-2372226

Zuckerberg, Mark. Quoted by Marshall Kirkpatrick. “Facebook’s Zuckerberg Says The Age of Privacy is Over.” January 9, 2010. ReadWrite. URL: https://readwrite.com/2010/01/09facebooks_zuckerberg_says_the_age_of_privacy_is_ov/

Corriere della Sera. Big Data: the new era of advantages for PMI is reality. Cured by Unicredit. Online article: https://www.corriere.it/native-adv/unicredit-longform04-big-data-la-nuova-era-dei-vantaggi-per-le-pmi.shtml

F.Q. Elections, The Italy of abstentionism. “The 52% does not recognize himself in any party”. Il fatto quotidiano, May 15, 2018. Online article: www.ilfattoquotidiano.it/2015/05/18/elezioni-litalia-dellastensionismo-il-52-non-si-riconosce-in-nessun-partito/1694292/

“Italian politics in 2019”. The Post, January 2, 2019. Online article: https://www.ilpost.it/2019/01/02/eventi-notizie-2019/

Openpolis. June 23, 2016.Online article: https://blog.openpolis.it/2016/06/23/perche-le-persone-non-vanno-votare-cause-astensionismo/8870

Panorama, November 12, 2018. Online article: https://www.panorama.it/news/politica/salvini-intervista-luca-morisi-web-facebook/

Politica semplice. Online article: https://politicasemplice.it/politica-italiana/situazione-politica-italiana-2013-2018

Abstract

This thesis is based on a team project which tried to predict the European elections results. This was possible through the use of Google tools and so thanks to the power of Big Data. This is why the first chapter is dedicated to Big Data. Today we are overwhelmed by information, which arises in number more and more rapidly.

There is a passage from Small Data to Big Data, when the reference is not to a small subset, but to everyone.

There are at least four issues which can be raised against Big Data.

The first argument against Big Data is that nobody clearly knows what effectively means the term Big Data, there is still not a clear single definition of Big Data, as Stephen Few, IT expert, sustains in his book "Big Data, Big Dupe. A little book about a big bunch of nonsense". He lists six different definitions that we can encounter: data set that are extremely large; data from various sources and various types; data that is large in volume, derived from various sources, and produced and acquired at fast speed; data that is extraordinary complex; data that is processed using so-called advanced analytical methods; any data at all that is associated with a current fad.

The confusion resulted from these different definitions can be reassumed by the opinion of Ryan Swanstrom: "Now Big Data has become a buzzword to mean anything related to data analytics or visualization"[21] .

The second argument against Big Data is that their collection is not worth it: if it is true that it is convenient for firms to collect and retain more data as possible in hope of finding unknown secondary uses which can be helpful in the future, so it has to be

[21] Swanstrom, Ryan (Data Science Blogger, "Data Science101" blog). URL: https://101.datascience.community/about/

true that the organizations with most data today should be the most successful ones. But they are not.

Third argument against Big Data: they focus on quantity and not on quality leaving out the importance of expertise of the subject without which Data are not useful. Forth argument against Big Data: possible loss of privacy. All information collected about each one of us is feed into a machine, which processes them. And we do not know anything about the process, we do not know how they will be used and who will use them. Mark Zuckerberg, one of the funders of Facebook has affirmed: "the age of privacy is over".[22]

Fifth argument about big data: ethical issue and difficulty in integrating legacy systems. Should Big Data be used by the Police to prevent crimes? If we know that a person has checked on Internet "how to rape a woman", should this information be used to arrest him and prevent that he actually does it in real life causing a victim?

There are several issues about Big Data, anyway, there are not only negative aspects, there is also the other face of the coin: advantages of Big Data.

Smart Data: a powerful big data business intelligence platform gives to us the ability to ask and answer questions more robustly, answering becomes a relatively straightforward process.

Zoom in and customization: the zoom that Big Data consent to do is the future of the marketing. Through the analysis of Big Data we can know countries' preferences, states' preferences, regions' preferences, specific areas' preferences, neighborhood preferences', until reaching each single consumer and internet user of the world.

[22] Zuckerberg, Mark. Quoted by Marshall Kirkpatrick. January 9, 2010. "Facebook's Zuckerberg Says The Age of Privacy is Over." ReadWrite. <https://readwrite.com/2010/01/09facebooks_zuckerberg_says_the_age_of_privacy_is_ov/>

Segmentation provides the best way to sell something to someone, that is to know, even before he/she is aware of it, what he/she wants.

Google as a truth serum: why ask with the risk of an untruth answer when we can obtain it without even asking? Analyzing movements of web users, looking which websites they prefer, the bounce rate - how many users leave the page after the first page view -, their comments on social media, their likes on Instagram pictures, we can know what they want without any effort.

Smart cities: cities which are being developed on the idea of "exploiting Big Data to make the world a better place" as Christian Rudder said. Smart Cities are an example of Big Data Management and Big Data Analysis. They are cities made with planning strategies with the aim of optimization and innovation of public services, all through the use of Big Data.

S. Few in his book proposes an interesting idea, the one of a "Slow Data Movement". Few suggests 3Ss in alternative to the 3Vs (volume, velocity, and variety): small, slow, sure.

Small because just a small part of data is useful, we have to be able to recognize what is meaningful and useful and to leave out the rest.

Slow because, how Daniel Kahneman underlines in his book "Thinking fast and slow", each of us is made up of two systems: System 1, the intuitive one; System 2, the reflexive one. And all those decisions which are not decisions of habits are made by System 2, thanks to conscious, deliberate, reflective and analytical reasoning. This can make us understand that going too fast without taking the time to reflect is never a good idea.

Sure, because variety can be a synonymous of complexity and so can be everything except useful. Only when we recognize some data as relevant, because they can be exploited to make something, we can call that data sure, because they become reliable as soon as we identify them as valuable.

As already said this thesis is based on the power of Big Data, but the idea of this thesis came up from the book written by Seth Stephens Davidowitz "Everybody lies", in Italian "La macchina della verità". For S.S. Davidowitz it has all started in 2008 with American presidential elections and the long debate about the importance racial prejudice in America. Barack Obama, the first Afro-American candidate of a big party, won. At the time, a lot of polls were made which suggested that the race was not a determinant factor, according to those polls, the majority of Americans did not care about the color of Obama's skin when they decided who they were going to vote. In that year, S. S. Davidowitz discovered Google Trends, an instrument that indicates users' search-frequency of words and phrases in different places and at different times. Using Google Trends he found out that the night of Obama's first election, when all the comments were about the praises on the new president, about one in one hundred researches on Google containing the word "Obama" included also terms such as "Ku Klux Klan" or "nigga". On the same night, researches and subscription to "Stormfront", a nationalist white site, have been more than 10 times higher than usual. In some American states "nigga president" researches exceeded those for "first black president".

These findings could explain the success gained eight years later Obama's first election by the current President of the United States of America Donald Trump, a

candidate that summed up the racist researches, the immigrants attacks, the anger and the resentment of people's worse inclinations. That is the beginning of my thesis: S.S. Davidowitz in his book shows how we can use Big Data to give new intuitions about the psychology and behavior of human beings, how Big Data can confirm and prove suspicious, but most of all, how they can reveal that the world does not go in the direction we imagine, but in the opposite direction. His book is a great and very interesting reading about using internet Big Data sources to study what people actually do and think vs. what they say they do and think. We tend to lie, especially about embarrassing or negative behaviors, when asked. But to Google, people confess the strangest things.

"La macchina della verità" teaches us how we can take advantage of Big Data, without losing ourselves in such an extraordinary amount of data but giving a sense to them and trying to gain conclusions analyzing them. It has served as an excellent starting point. It has this unique capability of explaining statistical concepts in a clear way, without getting lost in technical jargon. What I tried to do, together with my team, is the same thing he did with American presidential election, but with the European elections in Italy. I tried to predict the results using the power of Big Data and following the advices of S.S. Davidowitz's book.

Two main tools were used in order to do this prediction: Google Correlate and Google Trends. Google Correlate is a tool that can discover trends more correlated to a general activity and so to the semantic universe of reference. It is possible to insert a trend, about any activity in the world, and see which search terms are connected to that trend. Instead, Google Trends is a tool that allows us to know the frequency of

web researches for a specific word or phrase. Research and visualization can be set by nation and by language and even by topic category. The trends are showed with a graphic which synthetize, by time, the trend and its popularity.

The 26th of May 2019 there were the European political elections, and for the first time the result was not predictable. In fact, the decline of traditional parties and the rise of the radical right and of the other populist parties was critical for Socialist and Popular movements, the two blocks which have governed the Parliament for the last few years.

The following graphic shows the trends of the three main parties and just how far Lega Nord was ahead of the other two political parties before the European elections.

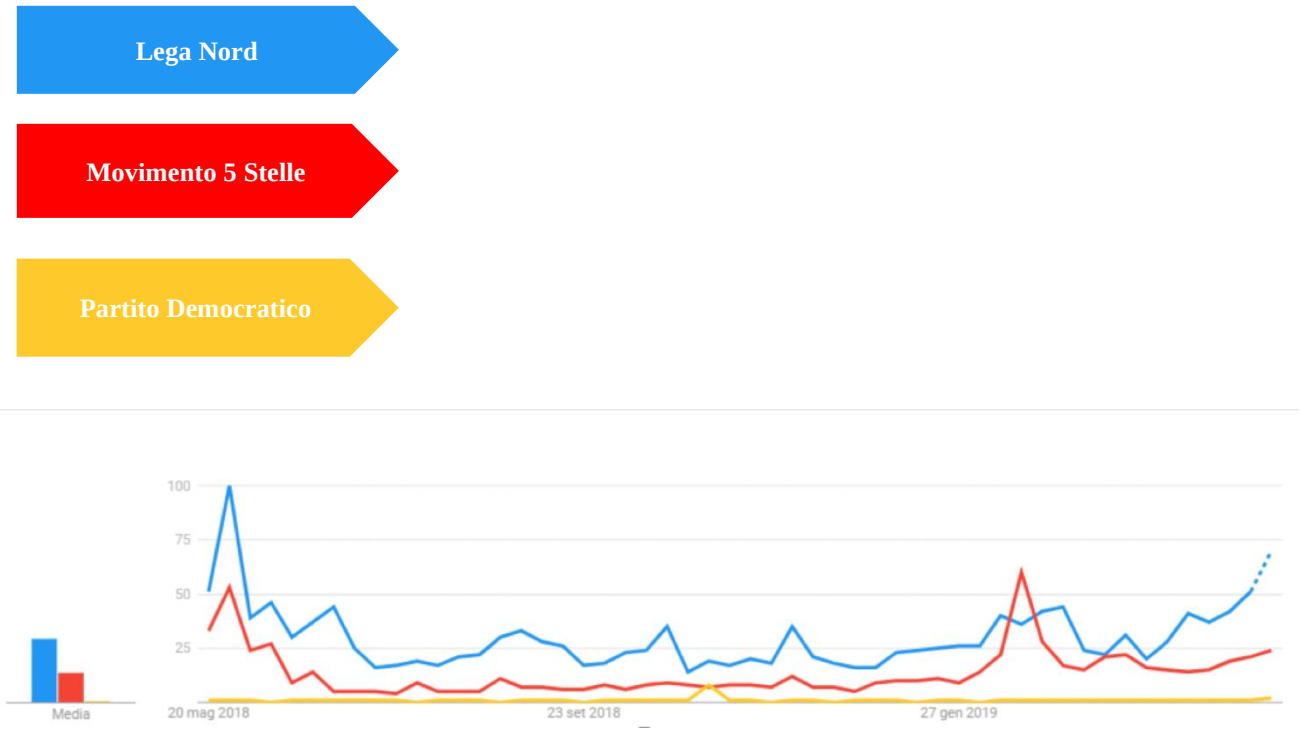


Figure 1. Political trend of the 3 main Italian parties- Google Trends

It is evident that the first party in Italy has been the Lega Nord for the last few months, followed by Movimento 5 Stelle and Partito Democratico. This reflects the

fact that the actual government is a coalition government born from an agreement between Lega Nord and Movimento 5 Stelle. Anyway, the actual government was formed after the political election on 4th March 2018, election through which did not yield a majority able to vote on its own government. The ones who gained more votes, Movimento 5 Stelle with 32,7% and right-center, conjoined, with 37%, did not gained a real majority. These results reflected a dissatisfaction and disappointment of the Italian population. The government of coalition between Lega Nord and Movimento 5 Stelle was an unusual and unexpected government between two parties with opposite political ideals, that is the reason why European political elections were considered so important: they would have given the possibility to know on which side, Lega Nord or Movimento 5 Stelle, is the Italian population, and consequently who would have gained more power within the coalition.[23]

S.S. Davidowitz and Stuart Gabriel, financial professor at the California University, Los Angeles, found out a surprising indicator about how people are going to vote. A big amount of researches regarding elections contains the name of both the candidates. These researches can give us some indicators on who is the candidate that a person sustains; it is sufficient to take care of the sequence in which names of candidates appear. The study of S.S. Davidowitz and S. Gabriel shows that the way people makes tech queries online is not random, in a research which contains the names of both candidates, a person is more willing to type first the one who he or she sustains. This is the first method used in the European elections prediction: comparing queries. We compared queries which included the name of the two main

[23] Finesi, Giorgio. From elections to Conte's government, almost 90 days of political crisis. Skytg24, June1, 2018. Online article: https://tg24.sky.it/politica/2018/06/01/governo-conte-storia.html

Parties' leaders: Salvini, leader of Lega Nord; Di Maio, leader of Movimento 5 Stelle.

- "Salvini Di Maio"
- "Di Maio Salvini"

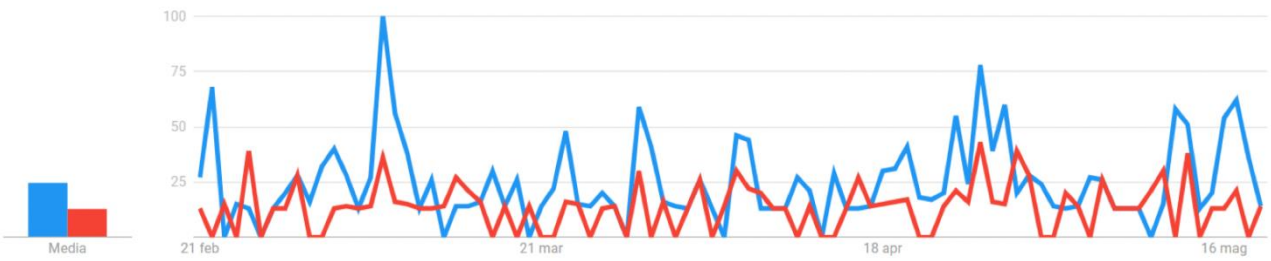


Figure 2. Comparing queries- Google Trends

According to this method there were no doubts that the winner of the European elections would have been Matteo Salvini, leader of the Lega Nord Party. The researches, which gave precedence to Salvini, exceeded the ones giving precedence to Di Maio. It was found out also that, usually during the political elections the leader substitutes the Party and after the elections Party's trend goes back to linearity, instead, Matteo Salvini has substituted himself to the Party since he has become the leader of Lega Nord, as is evident in the following graph.

- Matteo Salvini
- Lega Nord

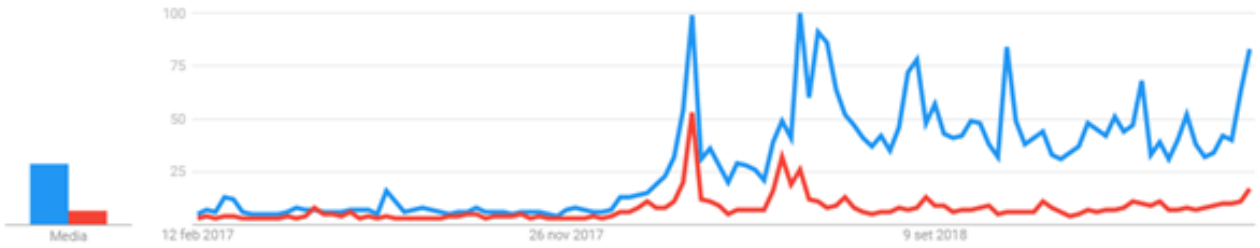


Figure 3. The trend of Matteo Salvini compared to the one of his party, Lega Nord- Google Trends

In order to prove such popularity of Matteo Salvini, we tried to compare the Google researches about him with the Google researches of some of the most influential people in Italy, through Google Trends. We took three personalities who are influencers in different areas: The Pope, Francesco Jorge Mario Bergoglio, chief of the Catholic Church; Chiara Ferragni, entrepreneur and Italian blogger, in 2017 nominated by Forbes "the most important fashion influencer in the world"; and Cristiano Ronaldo, one of the most popular football players, at the moment engaged by the Juventus football team. It seems that no other famous character in Italy can match Matteo Salvini's popularity.

Matteo Salvini seems to be always instinctive and spontaneous, most repositioning are studied by a computer through an advanced system, which consents to Matteo Salvini to appear "evil", "despicable" or sometimes "nice", but always in harmony with the predominant humour of public opinion. This happens thanks to an informatics system that it is called "The Beast". It is today evident that the Salvini's posts have been made to gain three principal objectives:

- the first is to occupy soon the media space, step in as first on the daily news, so that the others traditional media, opponents but also allies, are constrained to go after him

- the second is to polarize all the debates around the cross-road: for Salvini and against Salvini

- the third is to launch strong messages, taking out from the public opinion "negative feelings" such as anger, fear and aggressiveness

In fact, according to a study of the Michigan University (2005), the negative emotions are a way to gain attention, even if is the positive sentiment which takes back the voter to the electioneer. So true is it that at the end of each post, Salvini proposes a fragment of joy to who is listening to him, like "big kisses and good Friday to all"- by fading out the aggressive message launched right before.

In a post in which Salvini comments a shooting made by an extreme right movement made to revenge the death of a girl, comes out a positive sentiment. Nevertheless the algorithm notes a positive shade.

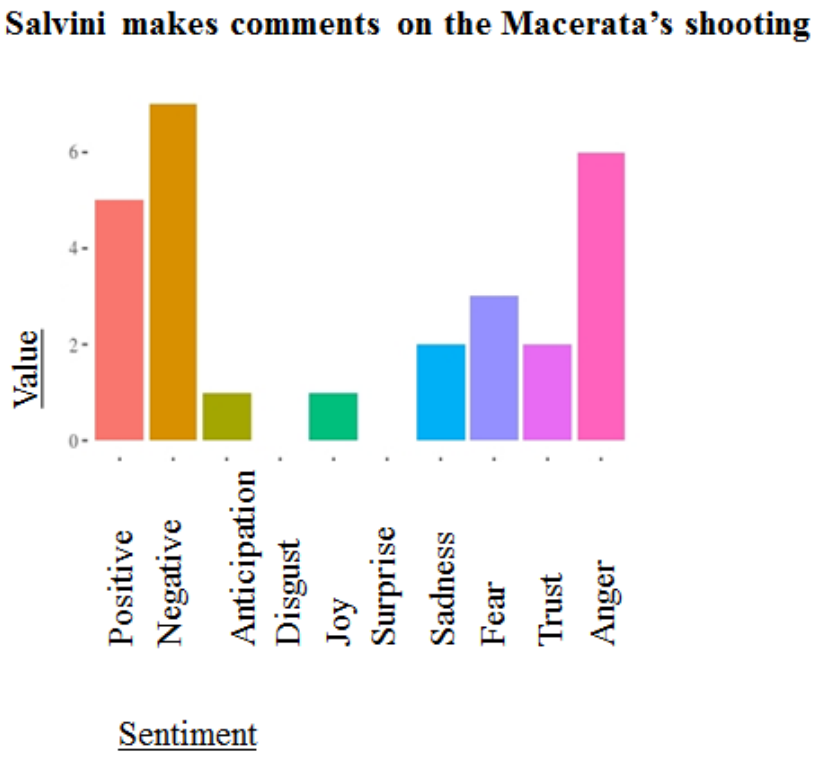


Figure 4. Sentiment analysis of Matteo Salvini's post on Facebook on 3rd February 2018

The rhetoric strategy is clear: to let the guard down leveraging on anger and fear, but suggesting that, giving trust to Lega Nord, things will get better. It has proved to be successful the Salvini's decision to bet on the digital communication right before the European elections, deciding to sponsor Facebook posts and increasing the number of Twitter posts. It was Matteo Salvini the one who invested more on social media and in a more targeted way- setting age and gender of his audience. With a budget of 128.782 thousands of euros, three times of what the Movimento 5 Stelle allocated, amount spent from March to 25th of May and of which 41.000 thousands euros in the last decisive week. Salvini set up a strategy with a starting point and an objective.

The emotionally exasperated vocabulary of Salvini is opposite to the measured one of Di Maio, who tried to limit anger and fear. Salvini is not afraid to use a language full

of contrasting emotions, differently to Di Maio he does not want to reassure anyone. He just want to maintain the center stage and to do that he needs to polarize the debate.

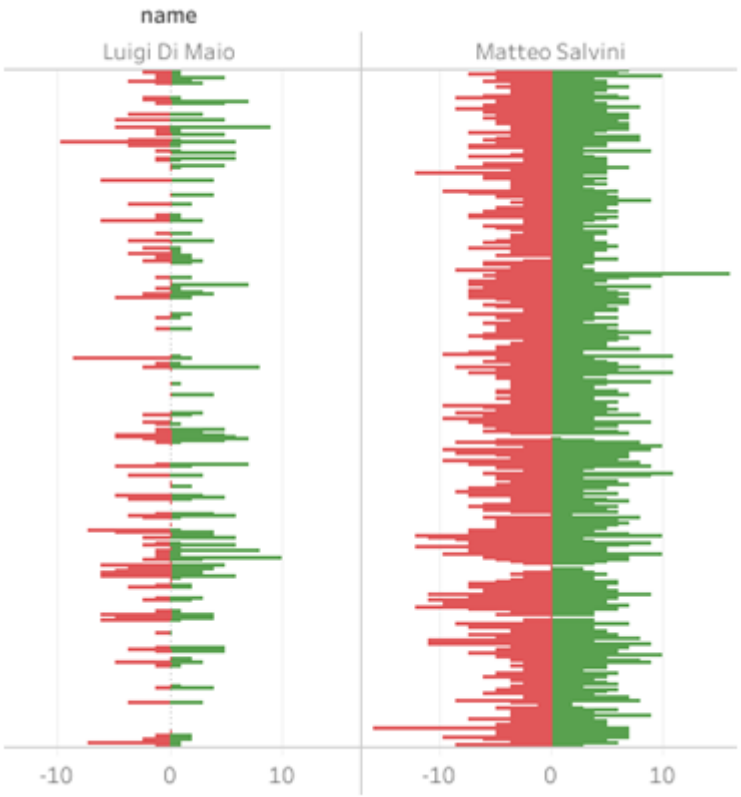


Figure 5. Analysis of the tweets on the positive/negative dimensions

Using Google Trends it was possible to have the trend of interest of the last three European elections and also to see the percentage of participation corresponding to each election.

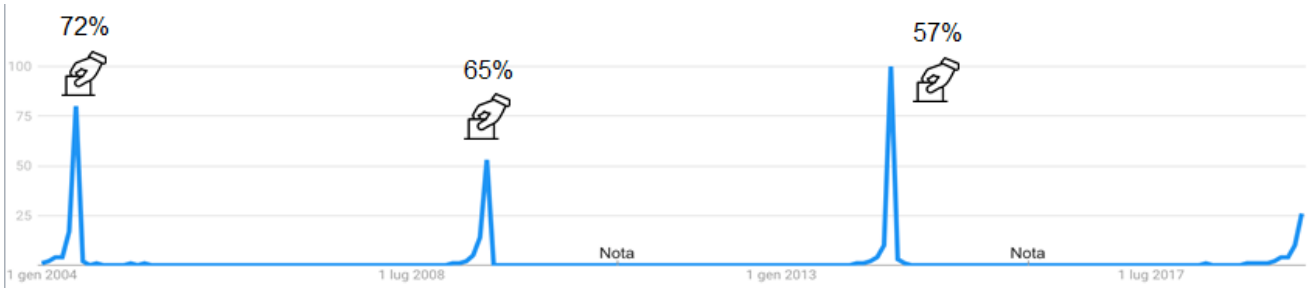


Figure 6. Trend of interest in the last European elections and percentage of participation- Google Trends

Looking at the rate of interest in the European elections it can be observed that it reached its peak at the last election (2014) even though the affluence was decreasing. It is evident that the rate of interest is not a parameter to predict European elections results. People search for information, but this does not mean that they are going to vote. Anyway, Euro-barometer shows that European citizens use to trust more European Union than their own Parliament. So, why did just a minority vote for the European elections? This paradox can be due to two factors, as Alberto Alemanno, analyst of European politic, explains: “First of all, the European elections are still the sum of national elections, and not a transnational political event animated by real European Parties; secondly, there is not a common public sphere which is able to talk about the European political system.” Moreover European elections’ affluence is difficult to be analyzed in time, because, as Jules Baley from SciencesPo University, Paris, notes, “How can we compare the affluence of 1979, when European Community was made of just nine countries of West Europe, with the one of 2014, when Union counted twenty-eight countries with different political cultures and different democratic traditions?” The low participation has been confirmed also in these last elections. Even if the interest increased in the European election 2019, the affluence did not. Actually it decreases from 57% at last European elections (2014) to 56% at the actual European elections (2019). Only in Trentino Alto Adige people voted more than in the rest of the country. The interesting and curios fact that can be observed is that, contrarily to what could be deduced from Google Trends, the fall of participation was registered most of all in the South of Italy. Google Trends reveals a

higher interest in a lot of Southern regions, the researches for “European elections” are higher in the South of Italy.

The results of European elections in Italy are the following:

- Lega Nord is the first Party with 34,3%
- Partito Democratico is the second Party with 22,7%
- Movimento 5 Stelle –which was the first Party- become the third one with 17,1%
- Affluence: 56%

According to these results, it is possible to say that the prediction made was right: Matteo Salvini, leader of Lega Nord confirmed his popularity, being the leader of the first Party. The analyses made using Google tools left no doubt. Anyway, as already said, we did not expect such a low participation in the South and we did not expect such a successful outcome for the Partito Democratico.

In order to explain the 56% threshold of total affluence, and the even worse rate of participation in the South, it could be interesting to read the report of the linguist Tullio De Mauro. According to the former, in Italy there are 13 million of semi-illiterate, who just know how to sign, but do not understand what they read, and 13 million of illiterate, who have lost the fluidity in writing and reading. The total is 28 million on 52 million under the threshold of literacy’s sufficiency. Shifting to the second factor which was not expected, the success of Partito Democratico, it can be interpreted as a sign sent by the Italians: there is a part of the country who does not want to sustain anymore populism and nationalism of the actual government and of Lega Nord in particular. It is a request by them to the PD to make opposition and to

become an alternative. This was not reflected by our analysis, maybe because the popularity of Partito Democratico has increased just in the last period and it was still not enough to be significantly recognizable on Google tools.