SAPIENZA
UNIVERSITÀ DI ROMA

Mapping news diet and polarization dynamics during electoral campaigns

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

Corso di Laurea Magistrale in Data Science

Candidate

Marco D'Ercole

ID number 1874366

Thesis Advisor

Prof. Walter Quattrociocchi

Academic Year 2023/2024

Thesis defended on 28 January 2025
in front of a Board of Examiners composed by:

Brutti Pierparolo (chairman)
Becchetti Luca
Petti Manuela
Quattrociocchi Walter
Silvestri Fabrizio
Spinelli Indro
Tieri Paolo

Mapping news diet and polarization dynamics during electoral campaigns
Master's thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: marcodercole1999@gmail.com

Abstract

Today, we all create content and are consumers of it. So, the truth of information tends to disappear and to be replaced by beliefs.

As pointed out by Professor Quattrociocchi, during the lectures of his course 'Data driven modelling for complex systems', tells of today's world and its complications. Are social networks and the content we read there actually a representation of reality, or are they simply a mirror of our beliefs that we surround ourselves with and that are always replayed to us because of the famous "algorithms"? Is democracy in danger today due to this self-manipulation that we inflict on ourselves by locking ourselves up in the famous "echo chambers"?

The questions about the world of social networks and how it can affect us are many, especially when we are campaigning. Last June (2024) there were European elections. I decided to pursue this project with Professor Quattrociocchi to analyze the Facebook posts of Italian political parties and major italian newspapers. The goal was to see if their Facebook pages were publishing polarized posts and if so what topics they were pushing their communication on.

My work was structured as follows:

In Chapter 1, I gave a review of the literature and state of the art on topics such as: news diet, echo chambers, polarizazion, the risks and the influences of the digital platforms, and the relationship between elections and computational social science.

Chapter 2 has as its theme the materials and methods that I used. Then I described the datasets available for my study, as well as the theoretical concepts I used to carry out my research.

In Chapter 3, I developed an in-depth analysis of data distribution, focusing especially on the numerical differences in posts per topic, the relationship between interactions and the number of posts per topic, and the comparison between the

number of posts and likes per topic.

In Chapter 4, I studied the correlation between newspapers and parties based on the number of posts per topic. The goal was to check for possible similar trends between them.

In Chapter 5, I did a clustering to see if there were similar trends among party groups and newspapers on the editorial choice to post more certain topics than others.

In Chapter 6, I studied entropy and KL divergence to check if polarization existed and if so, I determined the topics of preference from the social pages of parties and newspapers.

Finally, Chapter 7 is concerned with the due conclusions. I have summarized the work performed and the responses obtained. Although often when doing work, it is more the questions than the answers that arise.

Contents

Chapter 1

Literature and State of the Art

1.1 Introduction to the News Diet concept

The concept of **News diet** refers to the quantity, variety, and quality of information that an individual is exposed to daily through the media. Just as a diet affects physical health, a news diet affects a person's 'informational health' and their worldview.

A news diet can be diverse, including news sources from various fields and with different perspectives, or limited, characterized by a reliance on a few sources, often with a single political or ideological view. In recent years, with the proliferation of digital media and social networks, the concept of a news diet has become increasingly relevant. Today, people can choose where they get their information, with significant consequences on the formation of political opinions and perceptions.

A balanced and diverse news diet can foster a more comprehensive and critical understanding of events, whereas an unbalanced diet, with a prevalence of highly ideologized or polarized sources, can contribute to the effect of **echo chambers**, where people are only exposed to news that confirms their beliefs (There are different methods to extract the community structure of large networks [1]). The same capacity for interaction between different users will be subject to the rule of the

Bounded Confidence Model (BCM), by which people will only influence each other if the distance between their opinions is below a given threshold sigma (tolerance) [2].

1.2 The Echo chambers phenomenon

Figure 1.1. The phenomenon of echo chambers: we are driven to search for content in line with our preexisting ideas. (Geopop.it)

The phenomenon of echo chambers does not exclude the fact that there are contacts between users belonging to different factions, which unfortunately develop into manifestations of hatred. In fact, a single politically heterogeneous user cluster in which ideologically opposed individuals interact at a very high rate is also detectable. Politically motivated individuals can hypothesize that they provoke interaction by injecting content supporting their own ideas into information flows whose main audience consists of ideologically opposed users [3]. This clash, however, does not become an enrichment and a way to broaden one's view, but it is a war that only radicalizes support for one's own ideas even more.

This war between factions very often results in people coming to easily believe fake news in order to support their own thesis against the other. Alternative news consumers in particular, who are the ones trying to avoid the 'mass manipulation' of the mainstream media, are the most responsive to the injection of false claims. [4]

1.3 The role of the news diet in political polarization

In the study of political polarization, the news diet is a crucial element, as it shapes voters' opinions, their involvement in political discussions, and, ultimately, electoral behavior. It becomes of great interest to understand how newspapers and political parties position themselves on polarizing topics.

In fact, newspapers are supposed to devote space in their pages to many different topics concerning current affairs. The question that many scholars have asked over the years is whether newspapers instead focus the majority of their articles on specific topics. If this is true, and if newspapers specialize in different topics, the reason might not be market-related, i.e. writing what sells the most, otherwise all newspapers would focus on the same topics. Consequently, there could be two reasons: newspapers make precise and intentional editorial choices, or newspaper markets consist of different audiences that require different content offerings [5].

1.4 The influence of digital platforms on information

The spread of new information channels such as the Internet and social networks has therefore allowed both a wide dissemination of information and allowed readers to delve only into topics of interest.

On the one hand, therefore, this media revolution has made it possible for many citizens to be more aware of global and local issues, to take an active part in public debate, and to confront different ideas. Digital platforms also enable greater interactivity, allowing anyone to comment, share, and create content, transforming information from a one-way process to an open and continuous conversation. All this has led many scholars to want to analyze how much political support develops thanks to the work done by parties on social networks, but it is a very complex task in which political support for a candidate or a party is often confused with political attention towards them, and these two concepts are not necessarily related. [6]

However, there is evidence that content consumption on Facebook is strongly influenced by the tendency of users to limit their exposure to a few sites. Despite the wide availability of heterogeneous content and narratives, there is an important **segregation** and increasing **polarization** in online news consumption. [7]

There are scholars who suggest trying to avoid social fragmentation through mechanisms that provide a 'public sphere' that is used simultaneously by people with contrasting perspectives on facts and values. If a general public sphere is not available or is not feasible, it becomes even more important to ensure that in the course of deliberation people are exposed to a range of reasonably competing viewpoints. [8]

1.5 The risks of the digital information ecosystem

Figure 1.2. The amount of information we are subjected to (Isaac Lazaro)

Along with the extraordinary opportunity that the Internet world offers, a number of risks emerge that make the information ecosystem more fragile and vulnerable.

In particular, the very ease of access to and production of content has opened the door to negative phenomena, such as:

1) **Disinformation and fake news**:

The spread of false or misleading information has become easier than ever. The speed with which unverified news can go viral has made it difficult for readers to distinguish between reliable sources and manipulated content. This phenomenon undermines trust in institutions and the mainstream media, fueling the spread of conspiracy theories and misinformation. The spread of false information is amplified by people's desire to be consistent with their previous beliefs or a political party and it can lead people to firmly hold false beliefs [9].

2) **Echo chambers and information bubbles**:

Although access to different opinions is theoretically possible, many individuals tend to surround themselves with information that confirms their preexisting beliefs. Social media algorithms, designed to maximize engagement, often show users content that reinforces their opinions, creating real echo chambers. This ideological isolation leads to increased political polarization, where different factions do not only not communicate with each other, but perceive the other side as increasingly distant and threatening [10].

Selective exposure to content is the main driver of content dissemination and generates the formation of homogeneous clusters, or 'echo chambers'. In fact, homogeneity seems to be the main driver of content dissemination, and each echo chamber has its own cascading dynamics [11].

The difference in dynamics can be seen, for instance, between the group of polarized users of conspiracy news and the group of polarized users of science news. The way each approaches the other group is different. Conspiracists are more focused on their community's posts, and their attention is more directed towards conspiracy content. However, polarized users of science news are less engaged in dissemination and more inclined to comment on conspiracy pages [12a].

3) **Information overload**:

The vast amount of accessible information can lead to so-called information overload. Individuals exposed to too much news struggle to select the most relevant

ones and understand their context, often ending up relying on cognitive shortcuts such as sensationalist headlines or populist sources. This reduces the capacity for critical analysis and increases the risk of being influenced by simplistic or misleading narratives.

4) **Algorithmic manipulation**:

The content we see online is increasingly determined by algorithms that respond to commercial rather than informational criteria. These algorithms favor content that generates the most interactions, which are often the most emotionally polarizing or extreme. As a result, the flow of information is distorted, making it more difficult for users to access a balanced view of the facts.

1.6 Political polarization during election campaigns

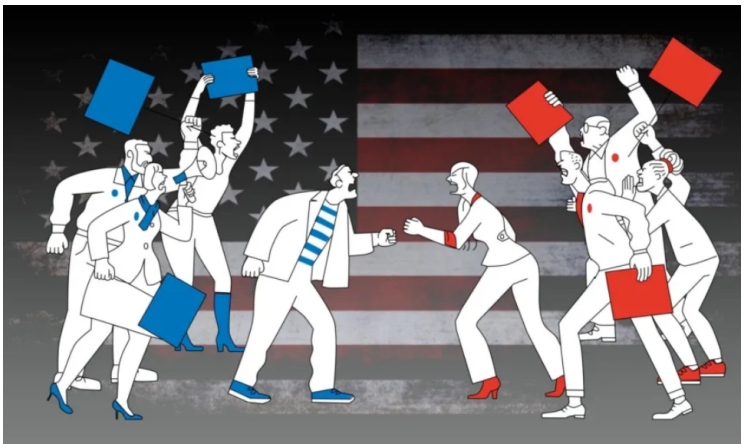

Figure 1.3. Polarization in the digital age exacerbates this already complex issue of societal and political division (Politics and rights review)

Political polarization is therefore a growing concern in many democracies, and this phenomenon becomes particularly evident during election campaigns. Polarization occurs when voters split into increasingly opposing groups with radically divergent political and ideological views and growing hostility toward the other side.

During elections, when the political debate intensifies, these divisions tend to

become more pronounced to such an extent that they even accept information that contains falsehoods in order to stick to their own views and reject dissenting information outright [12b]. Those, on the other hand, who have adopted a less systematic (more heuristic) approach to evaluate any information are more likely to end up with an account that is more consistent with their previous beliefs [13].

The media diet plays a key role in this process. Individuals who consume news from homogeneous sources, often ideologically aligned with their own beliefs, risk reinforcing their positions without ever being exposed to different points of view. This mechanism contributes to echo chambers, where people only receive confirmation of their pre-existing ideas, amplifying polarization. In addition, polarized media tend to present issues in a more divisive manner, accentuating differences rather than seeking points of common ground, further worsening the phenomenon during election campaigns.

Studies have shown that polarization can also be measured by engagement, with some theories claiming that the number of comments is positively correlated with their polarization [14]. In addition to this, some results show that the emotional behavior of communities is influenced by the participation of users within the echo chamber. In fact, greater involvement corresponds to a more negative approach. Furthermore, it was observed that, on average, more active users show a faster shift towards negativity than less active ones [15].

1.7 Elections and computational social sciences

Debates on elections within the **computational social sciences** (an interdisciplinary field that uses computational tools and methods to analyze, simulate, and understand social phenomena [16]) are becoming increasingly central as digital data analysis and the power of algorithms are redefining the way we understand electoral behavior and democratic processes. Elections, once studied primarily through traditional polling and qualitative research methods, are now examined

using advanced computational analysis techniques on vast datasets from social media, digital platforms, and other online sources. Studies are also done considering the political and cultural system that differs from nation to nation. In fact, not all voters behave the same way in the digital universe [17].

The thesis supports that studies on Facebook yield better results than analyses on X and that the sentiment analysis approach (volume/sentiment), although the most widely used, yields worse results than approaches such as regression or based on profile/post interactions [18]. However, many analyses have been developed in recent years that try to predict election results or study the causes of the results by developing different counting methods. Considering Bayrak and Kutlu's work [19] the detection of the location of the voters' accounts is used, and then four different methods were developed: the simple user counting method, the city-weighted counting method, the prediction method based on the nearest city and the method using the results of previous elections.

1.8 Threats to democracy

There are different issues in which computational social science can be used to serve democracy: using it to address the three main threats identified above: **hate speech**, **misinformation** and **coordinated influence campaigns from abroad** [20]. In order to better understand these phenomena, let us summarize the potential of digital technologies:

1) One of the main themes is to understand how **digital technologies influence electoral behavior**, from participation patterns to the dynamics of political persuasion. Through the analysis of posts, tweets, interactions and social networks, computational social science researchers are able to accurately map the evolution of political opinions, the flow of information, and the formation of coalitions or divisions between groups of voters. However, it is a study that must be done with great care, not taking for granted the fact that often neglected factors that are

independent of the political position of parties such as the political commitment of the authors, whether their party is in a government position or in opposition, the style, genre or date of the texts can influence the type of communication [21].

2) A second debate concerns **the impact of disinformation and online election manipulation strategies**. By studying phenomena such as fake news, coordinated disinformation campaigns, and the use of bots on social networks, researchers seek to understand how these factors influence voting and public debate. The use of computational techniques has made it possible to identify not only the dissemination of false information, but also the behavioral patterns of voters exposed to such content.

3) Another key theme is the growing capacity for **electoral micro-targeting through big data analysis**. Electoral campaigns today use sophisticated computational models to segment voters, send personalized messages, and optimize political communication strategies [22]. This raises ethical concerns regarding the privacy, manipulation, and fairness of the electoral process, fueling the debate on how to regulate the use of personal data in politics. In the 2016 US election, studies show that it was evident that Trump voters primarily consumed hyperpartisan news, many of which, such as Infowars and Breitbart, played a key role in amplifying subcultural messages and overwhelmingly supported Trump's candidacy [23].

4) Furthermore, the use of **computational methods to study political polarization** and its consequences during elections is a rapidly expanding field of research. Researchers examine how the media diet, online behavior and social networks influence the level of polarization among voters, and how these mechanisms lead to radicalization and fragmentation of the public sphere. Computational techniques make it possible to analyze huge amounts of data and map changes in political preferences with unprecedented precision.

Ultimately, debates in the computational social sciences concerning elections focus on the balance between opportunities to improve understanding of electoral processes and the risks associated with manipulation and distortion of the democratic

process. The ability to collect and analyze large volumes of digital data has opened up new frontiers for the study of elections, but it has also raised fundamental questions about how to maintain the integrity and transparency of the electoral process in the digital age.

However, this work only scratches the surface of the full potential of computational social science to advance theory development at multiple levels of analysis. Indeed, it has great potential, and in the coming years its potential will be seen even more when it is able to link the macro levels of theories on topics such as cultural change to the microlevel processes of decision-making [24].

Chapter 2

Materials and Methods

2.1 Working Material

2.1.1 Datasets and variables

My analysis was developed through the use of **two datasets** containing all Facebook posts from January 1, 2024 to June 9, 2024 from:

- The 20 most important **Italian newspapers** ("Domani", "Tg La7", "Il Fatto Quotidiano", "Il Dubbio", "MF-Milano Finanza", "Il Sole 24 ORE", "Corriere della Sera", "Sky tg24", "Libero", "ANSA.it", "Fanpage.it", "il manifesto", "La Verità", "Tgcom24", "Avvenire", "RaiNews", "Il Foglio", "Il Giornale", "la Repubblica", "La Stampa").

- The **Italian parties** in Parliament ("Partito Democratico", "Sinistra Italiana", "Italia Viva", "Forza Italia", "Fratelli d'Italia", "Movimento 5 Stelle", "Matteo Salvini", "Europa Verde - Verdi", "Azione"). The criterion by which the page for each party was chosen was the Facebook page indicated on the party website. In the specific case of Lega, the profile of Lega was not indicated but that of its secretary 'Matteo Salvini'.

The period considered was chosen in this way because the ends are: the beginning of the year and at the bottom the last voting day of the European elections. It is precisely in this vein that my studies developed, namely, to analyze what happened online in the pre-election period.

Both datasets contain the same variables and are therefore structured in the same way. Specifically, you have in each row the details of a post, told by the columns (variables), specifically:

1) **Url**: The website address of the post.

2) **Page name**: Name of the page (newspaper or party depending on the dataset).

3) **User name**: Simplified page name, i.e. all lowercase and no spaces.

4) **Followers at posting**: Number of followers who were following the page when the post was made.

5) **Post created**: The day the post was made.

6) **Type**: Type of post, e.g. whether there is a video, photo, etc.

7) **Total interactions**: The total number of interactions the post had.

8) **Likes**: The total number of likes the post had.

9) **Comments**: The total number of comments the post had.

10) **Shares**: The total number of shares the post had.

11) **Love**: The total number of 'love' reactions the post had.

12) **Wow**: The total number of 'wow' reactions the post had.

13) **Haha**: The total number of 'haha' reactions that the post had.

14) **Sad**: The total number of 'sad' reactions the post had.

15) **Angry**: The total number of 'angry' reactions that the post had.

16) **Care**: The total number of 'care' reactions that the post had.

17) **Clean post**: The content of the post.

Then, other characterizations were added to these variables. First, each post was assigned a topic based on the content expressed by the text of the post. This process was done through the **topic modeling** procedure and resulted in 17 topics: 'Ukraine', 'It Politics', 'Elections', 'Tech', 'Entertainment', 'Protest', 'Migration', 'Unclear', 'Rights', 'Climate', 'Mena', 'Economy', 'Health', 'Sport', 'Crime', 'Education', 'Religion'. I then added the variable **'Political tendency'** to indicate whether the party/newspaper has a tendency to left, centre or right-wing political thoughts. I then added another column to the dataset called **'New clean post'**. Here I inserted the cleaned post text and reduced it to the essential to be able to make appropriate sentiment analysis characterizations during my study. Finally, I added the column **'Engagement Rate'**, in which I entered the numerical value derived from the number of interactions received by the post given the number of followers on the page at the time the post was created.

2.1.2 Topic Modeling

In order to assign a topic to each post, the following procedure had to be developed in the following steps [25]:

Sampling of Posts

Given the high volume of data, the first step was to create a representative sample of all posts. This sample of posts made it possible to reduce the computational load and facilitated an in-depth analysis of the main content present, while still maintaining the thematic variety required for accurate classification.

Clustering via BERTopic

To analyze the textual content of the selected sample, the BERTopic model [26] was used, a clustering method based on sentence embedding and topic modeling techniques. BERTopic is based on BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model that generates robust semantic embeddings of texts. Thanks to these embeddings, BERTopic grouped similar posts into thematic clusters, representing sets of posts dealing with related topics. The main advantage of BERTopic is its ability to recognize latent topics with higher semantic granularity than traditional methods.

Assigning Names to Clusters

Once the clusters were generated, each group was examined to identify the main theme shared by the included posts. Then, each cluster was assigned a representative name that summarized the general content. Assigning a label to each cluster is a crucial step as it allows an intuitive understanding of the themes without the need to analyze each individual post within the cluster.

Propagation of Topics Using TF-IDF

After labeling the main clusters, these topics were extended to all other posts in the data set through the use of TF-IDF (Term Frequency-Inverse Document Frequency) [27]. TF-IDF was used to determine the most representative topic for each post in the sample by assigning the topic that best matched the key terms found in each text. In this way, it was possible to efficiently propagate the topics to the remaining posts, ensuring a consistent classification at the data set.

2.1.3 Text Cleaning Operations

It is useful to have in our dataset a column containing the posts with their clean, normalized version. Therefore, I performed the following cleaning operations to create the variable 'New clean post'.

1) **Removing URLs**:

URLs, which often appear in posts to link to external websites, were removed using a regular expression. This removes all links beginning with "http" or "www" to improve text clarity.

2) **Removing Hashtags**:

Hashtags were removed to reduce the noise in the text. The regular expression # finds and removes all terms that start with the symbol #, including single words or phrases without spaces.

3) **Removing Non-Alphabetic Characters (Including Emojis)**:

All non-alphabetic characters, such as numbers, punctuation, and emojis, were removed to retain only alphanumeric text. This simplifies the content and reduces unnecessary variation.

4) **Removing Page Names**:

The intended goal was to remove the names of political party pages from posts to avoid influencing the analysis with potentially recurring terms. The process involved creating a unique list of page names and removing them using regular expressions.

5) **Removing Italian Stopwords**:

Common words ("stopwords") such as "il" (the), "e" (and), and "un" (a) were removed, as they often don't add meaningful information to text analysis. A standard list of Italian stopwords was used for this step.

6) **Converting to Lowercase**:

All text was converted to lowercase to standardize word forms, reducing differences between uppercase and lowercase versions of the same word (e.g., "Italia" and "italia").

7) **Removing Numbers**:

All numbers were removed with a regular expression, as they might represent irrelevant information for topic analysis.

8) **Removing Punctuation**:

Punctuation was removed, leaving only words. This allows a focus on the meaning of the text without distracting us from commas, periods, or other punctuation marks.

9) **Removing Extra Whitespace**:

Extra spaces between words were removed to ensure that each post consists of continuous and well-structured text.

2.2 Working Methods

Before proceeding with the study of the data, it is necessary to clarify and define some of the indices I used in my study.

2.2.1 Shapiro Wilk Test

The Shapiro-Wilk Test is a statistical method used to assess the normality of a data distribution. In other words, this test checks whether a sample comes from a population that follows a normal (Gaussian) distribution. It is one of the most commonly used normality tests due to its effectiveness and power, especially for small sample sizes.

Principle of Operation

The test compares the sampled data with an ideal normal distribution. The test statistic is based on the correlation between the observed data and the expected data based on the normal distribution. More precisely, the test calculates a statistic called W, which is the ratio between the sum of the squared deviations of the data's ranks and the sum of the squared deviations of the expected ranks if the data were normally distributed.

Interpretation of Results

The W value ranges from 0 to 1.

W close to 1 indicates that the data are likely to be distributed normally.

W significantly less than 1 suggests that the data deviate from normality.

To determine whether the data are normal, the W value is compared with a critical value from specific tables, or a p-value is used.

If the **p-value** is less than a predetermined significance level (often 0.05), the null hypothesis of normality is rejected, indicating that the data do not follow a normal distribution.

2.2.2 The Silhouette Index

The Silhouette Index is a metric used in clustering analysis to assess the quality of clusters generated by an algorithm [28]. Specifically, it measures how well each data point is assigned to its own cluster, indicating if the clusters are distinct and cohesive. The Silhouette Index is commonly used to compare different clustering configurations and to select the optimal number of clusters for a dataset.

How it works

For each point in a dataset, the Silhouette Index calculates a value (called the "silhouette coefficient") that reflects the cohesion and separation of the point with respect to other clusters. The silhouette coefficient of a data point i is defined by:

- **Cohesion** $a(i)$: The average distance between point i and all other points within its own cluster. This value shows how close points are to each other within a cluster.

- **Separation** $b(i)$: The average distance between point i and all points in the closest cluster (the closest cluster other than the one to which the point belongs).

The silhouette coefficient for a point i is calculated as

$$s(i) = \frac{b(i) - a(i)}{max((a(i), b(i))}$$

This value ranges from -1 to 1:

- **Values close to 1** indicate that the point is well assigned to its cluster and is distant from neighboring clusters.

- **Values close to 0** indicate that the point lies on the border between two clusters.

- **Negative values** suggest that the point may be assigned to the wrong cluster, being closer to a different one.

Interpreting the Silhouette Index

The silhouette index for an entire dataset is the average of the silhouette coefficients for all data points. A higher average value indicates better separation between clusters and, thus, a more meaningful clustering:

- **Silhouette index close to 1**: clusters are well separated, and points within each cluster are tightly grouped.

- **Silhouette index close to 0**: clusters are poorly defined or overlap, suggesting weak separation.

- **Negative silhouette index**: some points may have been assigned to incorrect clusters.

2.2.3 Entropy

Entropy is a concept from information theory, initially developed by Claude Shannon, that quantifies the amount of uncertainty, disorder, or unpredictability within a system [29]. In simple terms, it measures how much "information" or "randomness" is contained in a dataset or distribution. Entropy is fundamental in many fields, including physics, machine learning, and data science, where it serves as a key measure to analyze the randomness or complexity of information.

How it works

In the context of information theory, the entropy represents the average amount of information produced by a source of data. It is often calculated for a random variable or a probability distribution and is defined mathematically as

$$H(X) = -\sum(P(x_i)log(p(x_i))$$

Where:

- $H(x)$ is the entropy of the random variable X.

- $p(x_i)$ is the probability of each outcome x_i.

- The logarithm is usually taken in base 2 (for information measured in bits) or

in natural logarithm form (for information measured in nats).

The value of entropy depends on the probabilities of the possible outcomes.

- **High entropy**: if the probabilities are distributed relatively evenly across many outcomes, the entropy is high, indicating high uncertainty or randomness.

- **Low entropy**: if the probabilities are skewed heavily towards one or few outcomes, the entropy is low, reflecting a more predictable or ordered system.

Interpretation of entropy

Entropy gives insight into the "information content" of a system. In data science and machine learning, it is commonly used to assess the purity or disorder in data:

- **High entropy**: suggests that the data is diverse or mixed, meaning it has a higher degree of uncertainty or unpredictability.

- **Low entropy**: implies that the data is more uniform or pure, with less uncertainty.

2.2.4 Kullback-Leibler (KL) Divergence

Kullback-Leibler divergence, also known as KL Divergence, is a measure of the dissimilarity between two probability distributions [30]. Introduced by Solomon Kullback and Richard Leibler, it quantifies how one probability distribution diverges from a reference distribution. Unlike other distance metrics, KL Divergence is not symmetric. It measures the "information loss" when an approximate distribution is used instead of the true distribution.

Formula and Interpretation

The KL divergence between two probability distributions P (the true distribution) and Q (the approximate or estimated distribution) is defined as:

$$D_{KL}(P||Q) = \sum P(i)log(\frac{P(i)}{Q(i)})$$

Where:

- $P(i)$ represents the probability of event i in the reference distribution P.

- $Q(i)$ represents the probability of event i in the approximate distribution Q.

KL divergence is always positive or zero. A value of 0 indicates that P and Q are identical (in terms of probabilities for each event), while a value greater than 0 indicates some degree of dissimilarity: the larger the value, the more the two distributions differ.

Meaning and Use of KL Divergence

KL divergence essentially measures the inefficiency of using the distribution Q instead of P. It can be interpreted as:

- **Information loss**: represents the amount of information "lost" when using an incorrect distribution Q to describe the true data behavior represented by P.

- **Relative distance**: provides a measure of relative "distance" between two distributions, though it is not a true distance metric since it is asymmetric and does not satisfy all the properties of metric distances.

Chapter 3

Data distribution

3.1 Number of post and topics

The question in the **case study** is: how does affective polarization emerge by observing the discussion of political parties and newspapers on the social media platform?

To answer this question, we examine the level of sparsity of clusters to see whether a cluster structure emerges.

Let us first analyze the distribution of data by topic.

Starting with the **dataset in newspapers** 3.1, it can be seen that the most discussed topics were, respectively, Economy (17751), Unclear - i.e. posts whose affiliation cannot be clearly defined (17207), Entertainment (16087) and Crime (13229). None of the most discussed topics therefore directly concern politics, but topics that are certainly related, such as the economy. The topics of least interest to newspapers were Rights (4829), Tech (5607) and Religion (5618).

However, changing to the **distribution of political parties by topic** 3.2, we have this much more polarized situation.

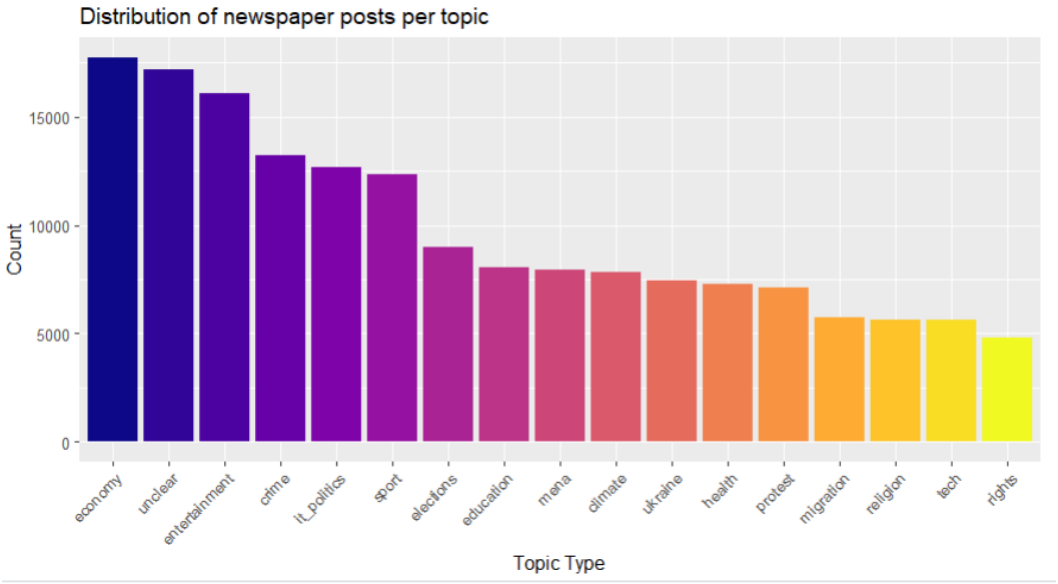

Figure 3.1. Distribution of newspaper by topic

The posts on Italian politics take the first place with 1490 posts. However, it was plausible that the parties would talk primarily about their policy. This is followed by Education (401) and Protest (368). The least discussed topics were Tech (88), Health (108), Crime (113).

We check how the number of posts per topic changes according to the political tendency of newspapers and parties to see whether this is an influential component or not.

Let us start with the newspaper dataset (Figure 3.3) and consider these divided in this way:

- **Center-wing newspapers**: "ANSA.it", "Corriere della Sera", "Il Dubbio", "Il Foglio", "La Stampa", "Sky tg24".

- **Right-wing newspapers**: "Il Giornale", "Il Sole 24 ORE", "Rai News", "La Verità, "Libero", "MF-Milano Finanza", "Tgcom24".

- **Left-wing newspapers**: "Avvenire", "Domani", "Fanpage.it", "Il Fatto Quotidiano", "Tg La7", il manifesto".

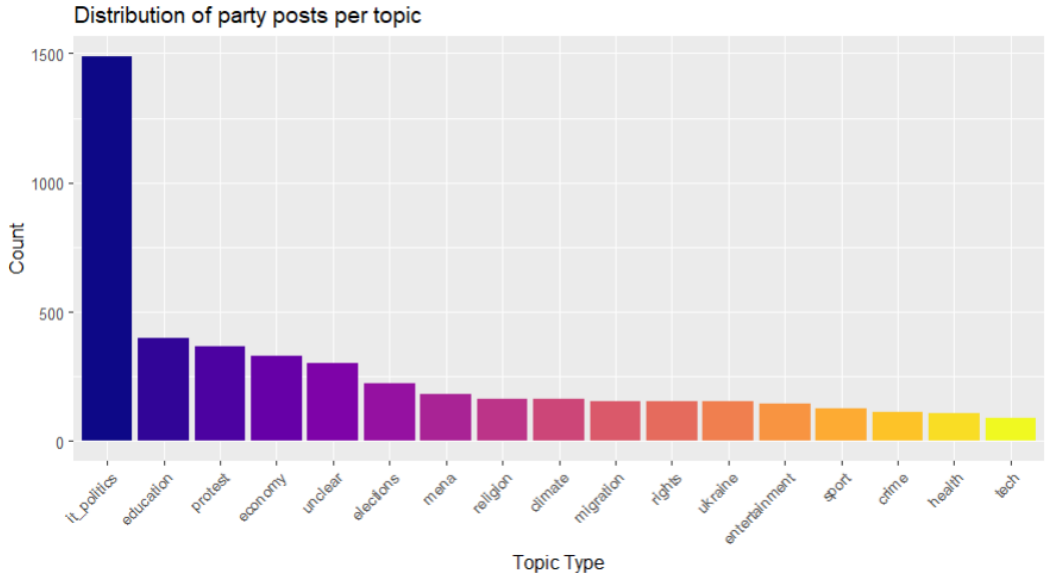

Figure 3.2. Distribution of political parties by topic

The barplot shows the biggest difference in the distribution for the topic 'Economy', it is influenced by the newspapers "Il Sole 24 Ore" and "MF - Milano Finanza", which deal with the economy as the main topic.

Let us see more in particular the discussions for the other topics:

- The topics that characterize the **center** more are: entertainment, It politics, crime, sport.

- The topics that characterize the **left** more are: crime, entertainment, sport, It politics.

- The topics that characterize more the **right** are: economy, entertainment, It politics, sport.

There are no big differences (with the exception of economy) between the different political belongings.

Let us see what happens instead in the party dataset thus divided by trend (Figure 3.4).

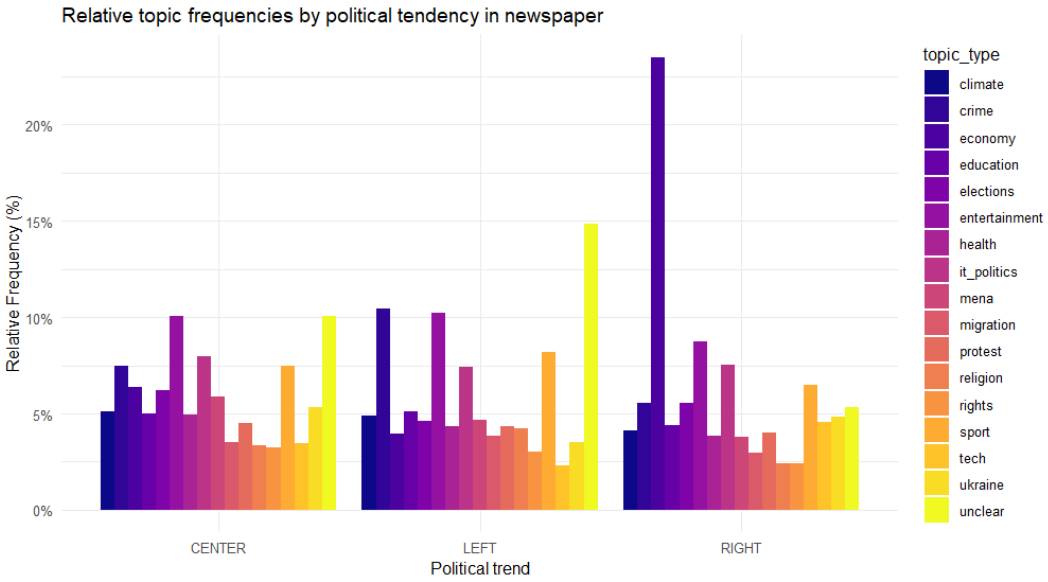

Figure 3.3. Relative topics distribution considering political tendency for newspaper

- **parties on the center wing**: Azione, Italia Viva

- **right-wing parties**: Fratelli d'Italia, Forza Italia, Matteo Salvini

- **left-wing parties**: Partito Democratico, Sinistra Italia, Europa Verde - Verdi, Movimento 5 Stelle

In the Italian parties dataset, this is the situation: It politics, education, protest, and economy are in the top rank for all the 3 groups.

- The topics that characterize more the **center** are: Ukraine, elections, entertainment.

- The topics that characterize the **left** more are: mena, protest, elections,and climate.

- The topics that characterize the **right** more are: climate, migration, religion.

After visualizing the distribution of the number of posts per topic considering political trends, it is now time to analyze the data for each page (party/newspaper depending on the dataset). We see in the figure 3.5 first of all the newspaper dataset:

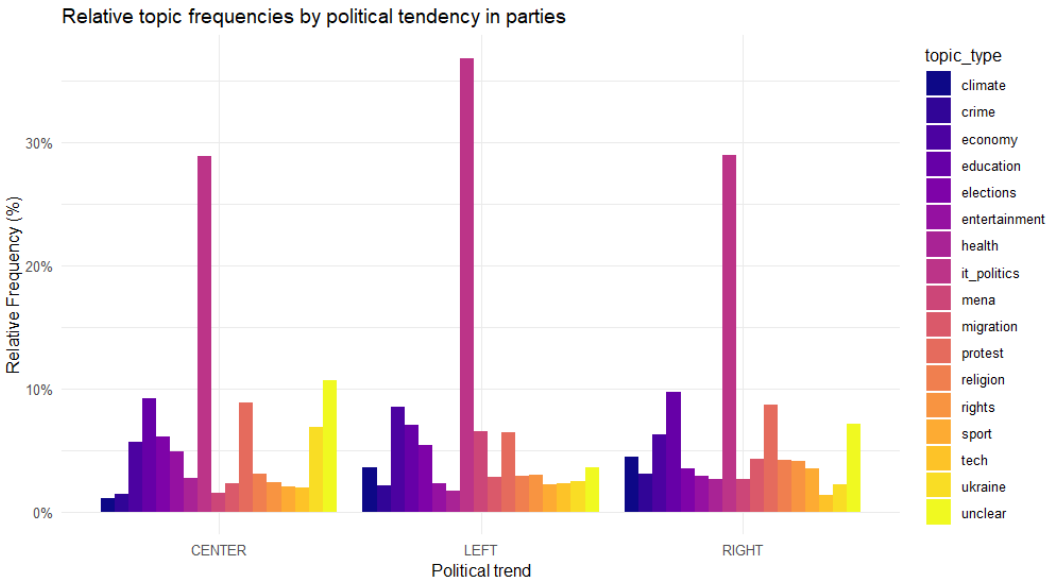

Figure 3.4. Relative topics distribution considering political tendency for parties

Looking to the table, I colored in green the three most discussed topics for each newspaper and in red the three lowest topics for each newspaper. The topics discussed more by each newspaper are visible in table 3.1.

We also do an analysis by horizontal, i.e. we see the topics that were particularly treated more by one newspaper than by the others (table 3.2).

Let us now check the situation for the party data set in figure 3.6.

Without considering It Politics the highest values are looking in table 3.3.

However, we should not only focus on the topics that are most discussed by each party but also note those topics that are highly discussed by certain parties and few by others. For example, it should be noted that Italian politics is treated a lot by all parties, but Azione and Matteo Salvini have less concentrated values in it.

Then, it is noticeable for each topic which party deals with it the most. The information is visible in table 3.4.

topic_type	ANSA.it	Avvenire	Corriere della Sera	Domani	Fanpage.it	Il Dubbio	Il Fatto Quotidiano	Il Foglio	Il Giornale	Il Sole 24 ORE	La Stampa	La Verità	Libero	MF-Milano Finanza	RaiNews	Sky tg24	Tg La7	Tgcom24	il manifesto	la Repubblica
climate	6.43	2.40	5.20	3.65	5.62	1.52	3.20	1.57	3.82	3.65	5.61	2.36	2.06	1.04	7.34	6.18	8.71	7.50	2.67	6.11
crime	7.45	2.97	8.69	3.33	19.50	10.14	6.40	1.71	8.33	1.51	9.95	3.96	6.21	0.41	6.43	7.22	8.18	12.14	2.67	12.96
economy	7.22	5.87	6.45	4.61	2.62	2.64	4.07	4.92	8.48	37.90	4.49	11.20	5.20	59.82	2.77	8.21	3.86	5.07	5.21	4.57
education	4.52	5.09	4.66	7.27	3.42	7.95	5.49	5.32	4.05	5.43	5.37	6.13	5.61	2.82	3.68	4.96	4.81	4.45	5.88	5.72
elections	6.08	3.39	5.18	7.07	2.28	11.08	4.98	6.70	4.86	5.47	5.21	11.20	6.79	3.86	4.51	6.94	7.53	5.75	6.55	4.43
entertainment	7.42	3.78	10.39	8.13	15.61	2.14	7.78	5.82	11.43	5.13	13.73	4.52	6.03	9.52	9.23	14.29	6.72	11.54	8.28	12.37
health	3.97	4.77	6.24	3.08	5.23	3.05	3.14	2.13	3.78	4.18	5.14	5.85	1.77	2.03	3.20	6.20	2.98	5.66	1.87	6.83
it_politics	8.55	4.49	4.76	19.78	1.91	22.04	10.06	19.54	12.28	4.65	5.73	15.16	28.52	2.17	3.31	3.26	8.23	3.69	11.82	5.04
mena	9.45	6.11	3.34	6.67	1.61	5.85	5.35	9.70	5.61	4.04	4.33	5.13	5.85	1.53	5.55	4.27	8.60	2.82	17.43	2.38
migration	3.16	3.57	3.45	4.19	3.35	4.78	3.46	1.95	3.92	2.17	4.40	2.99	4.17	0.65	3.47	4.09	5.02	4.61	5.14	4.18
protest	4.62	3.57	4.31	5.27	2.36	5.19	4.47	3.69	4.73	4.21	4.75	4.91	5.20	2.37	3.64	4.73	6.75	4.54	8.35	4.50
religion	2.93	33.04	3.01	2.39	3.06	2.72	2.45	5.68	2.80	1.78	3.16	5.21	1.68	1.02	2.97	3.15	2.71	3.11	2.81	3.50
rights	2.73	3.18	2.83	3.20	1.97	5.93	2.97	2.76	2.44	1.89	3.50	2.61	2.33	1.08	3.16	3.82	4.49	3.78	3.81	3.44
sport	11.26	2.72	7.64	5.20	15.14	1.73	7.24	6.78	10.54	3.38	7.32	2.33	5.96	1.82	20.19	5.28	6.31	7.10	2.27	5.55
tech	3.52	3.00	4.31	2.81	2.05	1.28	2.00	2.19	3.17	8.65	2.78	2.74	0.81	5.67	2.95	3.98	2.62	3.66	2.00	2.72
ukraine	7.58	3.46	3.51	4.95	1.27	5.32	4.05	9.13	5.14	3.80	3.59	6.98	7.11	2.16	13.81	4.32	6.87	2.80	7.75	2.25
unclear	3.11	8.59	16.05	8.42	13.00	6.63	22.87	10.38	4.61	2.18	10.92	6.74	4.69	2.04	3.78	9.10	5.60	11.79	5.48	13.44

Figure 3.5. Percentage of posts per topic given the newspaper

topic_type	Azione	Europa Verde - Verdi	Forza Italia	Fratelli d'Italia	Italia Viva	Matteo Salvini	MoVimento 5 Stelle	Partito Democratico	Sinistra Italiana
climate	1.18	7.12	0.80	1.90	1.03	6.76	3.18	2.82	1.89
crime	0.71	2.54	0.00	0.63	2.05	5.30	1.59	2.54	2.08
economy	7.08	9.67	15.66	8.07	4.52	3.20	9.94	5.37	8.49
education	6.37	6.87	3.21	6.33	11.70	13.15	8.75	6.21	6.04
elections	5.42	5.85	7.63	3.16	6.78	2.83	5.77	5.08	4.91
entertainment	6.13	1.53	1.20	0.95	3.90	4.38	2.78	3.67	1.70
health	3.07	2.04	2.41	1.58	2.46	3.29	0.60	1.69	2.64
it_politics	13.44	29.77	33.73	55.70	42.30	12.42	35.79	48.31	35.28
mena	1.42	6.87	2.41	2.37	1.64	2.92	1.99	3.11	12.83
migration	2.36	3.56	5.22	3.80	2.26	4.38	1.79	3.39	2.83
protest	14.62	7.12	5.22	5.06	3.90	11.60	4.37	4.80	9.06
religion	1.42	2.04	4.02	1.58	4.52	5.75	5.96	1.13	1.89
rights	3.54	3.56	4.82	1.58	1.44	5.39	1.99	3.11	3.40
sport	0.94	2.29	4.42	1.58	3.08	4.38	3.18	2.82	0.94
tech	3.54	2.54	1.61	1.27	0.62	1.46	4.17	0.56	1.70
ukraine	9.20	3.05	3.61	0.95	4.93	2.74	1.99	3.11	2.08
unclear	19.58	3.56	4.02	3.48	2.87	10.05	6.16	2.26	2.26

Figure 3.6. Percentage of posts per topic given the party

Ansa.it	Sport
Avvenire	Religion
Corriere della Sera	Entertainment
Domani	It Politics
Fanpage.it	Crime
Il Dubbio	It Politics
Il Fatto Quotidiano	It Politics
Il Foglio	It Politics
Il Giornale	It Politics
Il Sole 24 Ore	Economy
La Stampa	Entertainment
La Verità	It Politics
Libero	It Politics
MF Milano Finanza	Economy
RaiNews	Sport
Sky tg24	Entertainment
Tg La7	Climate
Tgcom24	Crime
Il Manifesto	Mena
La Repubblica	Crime

Table 3.1. Topics discussed more by each newspaper.

Climate	Tg La7
Crime	Fanpage. it
Economy	Mf - Milano Finanza
Education	Il Dubbio
Elections	La Verità
Entertainment	Fanpage.it
Health	La Repubblica
It Politics	Libero
Mena	l Manifesto
Migration	Il Manifesto
Protest	Il Manifesto
Religion	Avvenire
Rights	Il Dubbio
Sport	Rainews
Tech	Il Sole 24 Ore
Ukraine	RaiNews

Table 3.2. Which newspaper talked more about each topic.

Azione	Protest
Europa Verde - Verdi	Economy
Forza Italia	Economy
Fratelli d'Italia	Economy
Italia Viva	Education
Matteo Salvini	Education
Movimento 5 Stelle	Economy
Partito Democratico	Education
Sinistra Italiana	Mena

Table 3.3. The most discussed topic by each political party (with the exception of It Politics).

Climate	Europa Verde - Verdi
Crime	Matteo Salvini
Economy	Forza Italia
Education	Matteo Salvini
Elections	Forza Italia
Entertainment	Matteo Salvini
Health	Matteo Salvini
It Politics	Fratelli d'Italia
Mena	Sinistra Italiana
Migration	Forza Italia
Protest	Azione
Religion	Movimento 5 Stelle
Rights	Matteo Salvini
Sport	Forza Italia
Tech	Movimento 5 Stelle
Ukraine	Azione

Table 3.4. For each topic which party deals with it the most.

3.2 Interactions and topics

After all these analyses, the following questions arise: What does the different number of posts for different topics depend on? What leads newspapers or parties to discuss certain topics more than others? There could be several causes to investigate. It could be due to the amount of news and events in the given period, it could be due to the desire of the page owners to spread specific information and ideologies, or it could be due to the response that each topic generates in the followers of the pages. Therefore, let us proceed to **analyze the reaction that each topic elicits**.

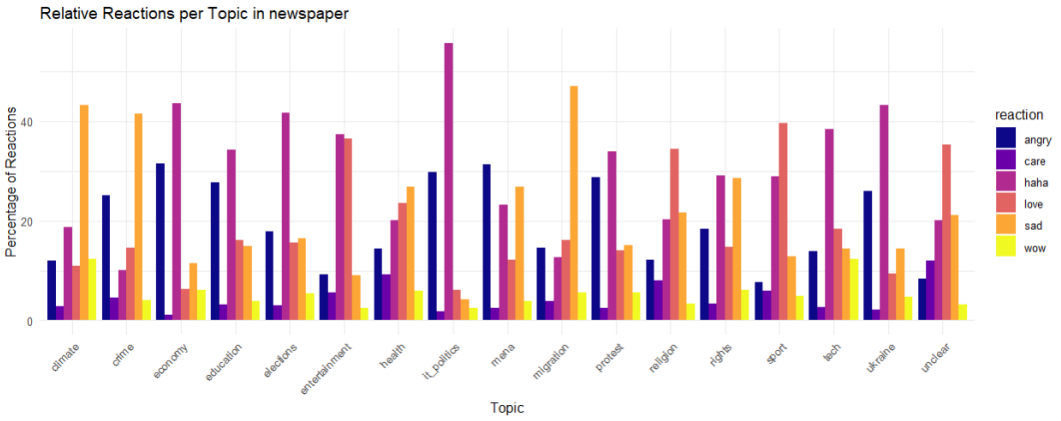

Figure 3.7. Relative reactions per topic in newspaper

In Figure 3.7 it is evident that for each of these reactions these are the topics that generate them the most:

- **Angry**: Mena (middle east and north africa).

- **Care**: -

- **Haha**: Economy, Education, Elections, Entertainment, It politics, Protest, Rights, Tech, Uckraine.

- **Love**: Entertainment, Religion, Sport.

- **Sad**: Climate, Crime, Health, Mena, Immigration, Rights.

- **Wow**: -

Let us instead see what happens in the posts of the political parties in figure 3.8, where we read from the graph which topic generates each reaction:

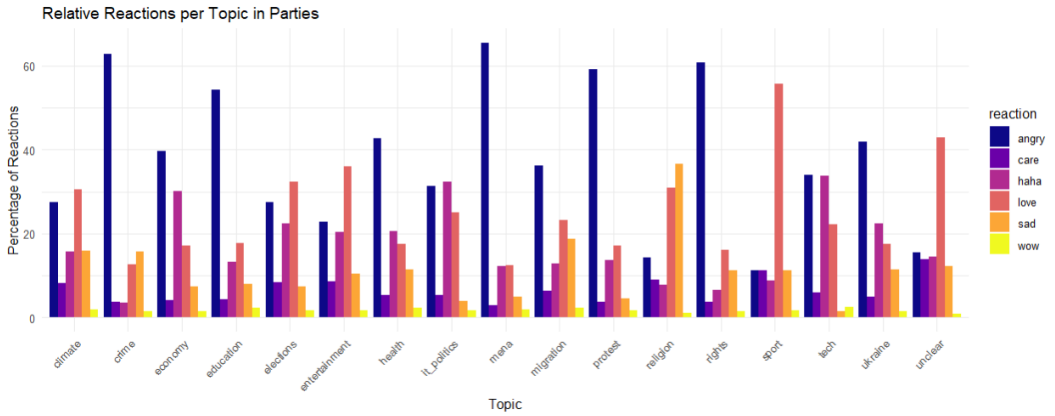

Figure 3.8. Relative reactions per topic in newspaper

- **Angry**: Mena, Crime, Rights, Protest, Education, Health, Ukraine, Economy, Migration.

- **Care**: -

- **Haha**: It politics, Tech.

- **Love**: Sport, Entertainment, Elections, Climate, Religion.

- **Sad**: Religion.

- **Wow**: -

Compared to the newspaper graph, it is much more evident in the party graph how many angry reactions there are, regardless of category (only religion and sport are exceptions).

The strong presence of the **angry reaction in political party** posts led me to want to visualize more clearly which topics generate this emotion. Here is therefore the following barplot 3.9.

The top 5 topics that generate the most negative emotions in political parties posts are: Crime, Rights, Education, Protest, and Health.

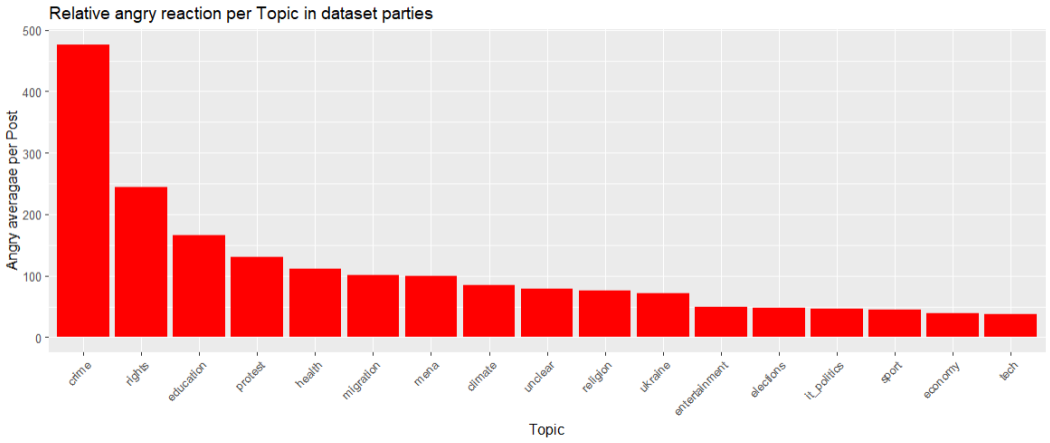

Figure 3.9. Reaction angry per topic in dataset parties

Continuing the reactions study generated by the posts, we check which reactions were generated by each party.

Looking to 3.10, it is possible to visualize that the highest reactions for each party are the following:

- **Love**: Forza Italia, Fratelli d'Italia, Italia Viva, Sinistra Italiana.

- **Wow**: -

- **Haha**: Azione, Europa Verde - Verdi, Partito Democratico.

- **Sad**: -

- **Angry**: Matteo Salvini, Movimento 5 Stelle

- **Care**: -

Let us do the same thing for the newspapers (Figure: 3.11)

- **Love**: Avvenire, Corriere della Sera, Fanpage.it, La Stampa, RaiNews, La Repubblica

- **Wow**: -

- **Haha**: Ansa.it, Il Dubbio, Il Fatto Quotidiano, Il Foglio, Il Giornale. Il Sole 24

	page_name	love_pct	wow_pct	haha_pct	sad_pct	angry_pct	care_pct
1	Azione	2.21%	0.44%	75.16%	1.71%	20.10%	0.38%
2	Europa Verde - Verdi	15.02%	1.29%	51.41%	5.50%	23.97%	2.82%
3	Forza Italia	38.88%	1.08%	31.90%	3.10%	17.01%	8.03%
4	Fratelli d'Italia	36.70%	1.17%	32.45%	1.69%	20.40%	7.58%
5	Italia Viva	50.30%	2.60%	18.31%	6.86%	10.31%	11.62%
6	Matteo Salvini	25.58%	1.76%	11.00%	14.03%	40.59%	7.03%
7	MoVimento 5 Stelle	18.71%	1.89%	22.64%	4.34%	46.42%	6.00%
8	Partito Democratico	23.48%	1.01%	36.40%	10.24%	23.02%	5.84%
9	Sinistra Italiana	30.11%	1.24%	16.54%	9.99%	35.24%	6.88%

Figure 3.10. Percentage of reactions for each political party

Ore, Libero, MF - Milano Finanza, Sky tg24, Tg La7

- **Sad**: Tgcom24

- **Angry**: Domani, La Verità, Il Manifesto

- **Care**: -

	page_name	love_pct	wow_pct	haha_pct	sad_pct	angry_pct	care_pct
1	ANSA.it	16.15%	5.31%	38.70%	19.44%	17.07%	3.33%
2	Avvenire	52.88%	2.17%	2.79%	22.85%	7.81%	11.50%
3	Corriere della Sera	28.20%	5.20%	27.27%	19.63%	13.07%	6.64%
4	Domani	9.74%	4.59%	33.43%	10.43%	39.77%	2.05%
5	Fanpage.it	31.66%	3.33%	14.17%	30.95%	10.93%	8.96%
6	Il Dubbio	7.69%	4.90%	35.36%	19.01%	29.48%	3.57%
7	Il Fatto Quotidiano	8.99%	3.74%	41.20%	11.43%	32.49%	2.15%
8	Il Foglio	9.30%	6.02%	60.44%	8.05%	14.34%	1.86%
9	Il Giornale	7.00%	3.73%	52.17%	6.66%	28.28%	2.16%
10	Il Sole 24 ORE	7.11%	6.64%	56.30%	11.16%	17.70%	1.09%
11	La Stampa	25.61%	4.67%	23.44%	22.75%	17.30%	6.23%
12	La Verità	4.34%	4.42%	36.14%	4.01%	50.26%	0.83%
13	Libero	4.87%	3.63%	59.48%	4.92%	25.84%	1.26%
14	MF-Milano Finanza	4.63%	9.09%	64.12%	5.87%	15.27%	1.02%
15	RaiNews	33.75%	6.61%	21.82%	19.76%	11.66%	6.40%
16	Sky tg24	18.49%	7.77%	33.50%	22.31%	14.46%	3.45%
17	Tg La7	12.40%	5.61%	36.99%	17.53%	23.33%	4.14%
18	Tgcom24	18.51%	5.32%	26.17%	29.04%	16.93%	4.03%
19	il manifesto	21.64%	1.73%	4.99%	21.78%	46.16%	3.71%
20	la Repubblica	27.42%	3.80%	26.10%	21.39%	15.88%	5.41%

Figure 3.11. Percentage of reactions for each newspaper

3.3 Number of post, interactions and topics

After studying the number of posts by topic and analyzing the interactions and reactions for posts by topic, we combine these two insights to see if there is a relationship between the number of posts and interactions.

Figure 3.12 highlights the five topics with the highest likes as a percentage of the total. Looking at Figure 3.1 again, it can be seen that the economy is treated a lot in the newspapers, yet receives few likes. With regard to the party dataset, on the other hand, sport is noted in the top 5, and is therefore a topic that attracts many likes, yet is treated little.

The table in Figure 3.13 is even more interesting.

In fact, in this table, I wanted to include two comparative pieces of information for each page. In the case of the party dataset, for example, there is for each party

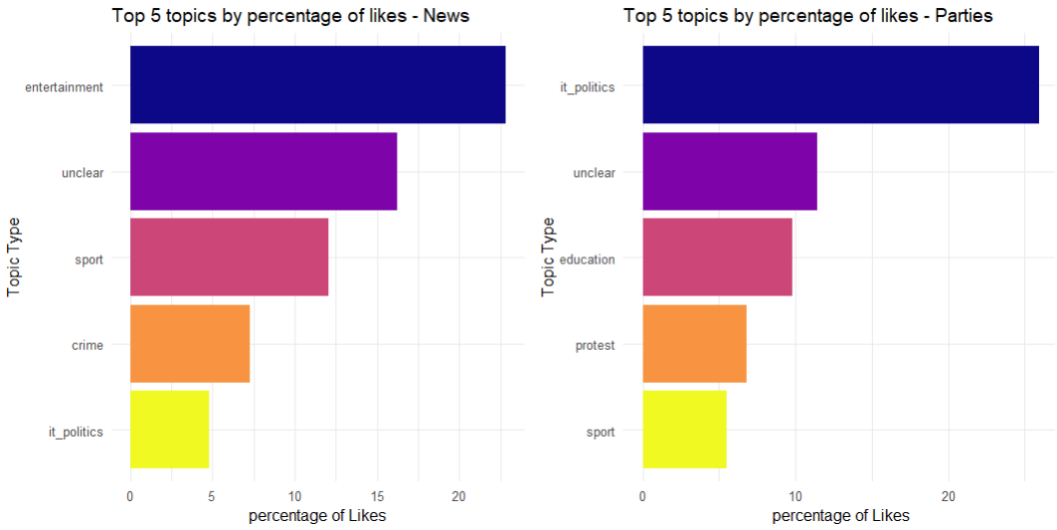

Figure 3.12. Top 5 topics by percentage of likes

topic_type	Azione	Rank Azione	Europa Verde - Verdi	Rank Europa Verde - Verdi	Forza Italia	Rank Forza Italia	Fratelli d'Italia	Rank Fratelli d'Italia	Italia Viva	Rank Italia Viva	Matteo Salvini	Rank Matteo Salvini	MoVimento 5 Stelle	Rank MoVimento 5 Stelle	Partito Democratico	Rank Partito Democratico	Sinistra Italiana	Rank Sinistra Italiana
climate	572.80	15	160.07	3	688.50	16	2128.50	9	367.00	16	4105.57	5	2121.94	9	1094.50	11	214.60	13
crime	87.33	17	120.00	12	NA	NA	1284.50	17	420.50	13	5565.53	8	1991.50	16	864.22	13	312.00	11
economy	230.20	5	101.89	2	711.41	2	1951.78	2	526.50	5	4139.94	13	1505.64	2	962.37	3	149.47	4
education	245.11	6	134.85	5	940.12	11	2314.57	3	385.09	2	5280.83	1	1327.45	3	1256.00	2	238.19	5
elections	494.43	8	127.87	7	1111.79	3	2319.25	7	661.18	3	7152.00	15	1234.66	6	1560.67	4	273.69	6
entertainment	71.35	7	1600.17	17	540.00	15	2075.00	15	548.53	7	4045.83	9	2456.29	11	1447.85	6	160.78	15
health	101.92	11	122.38	15	708.00	12	1975.90	10	527.92	11	5064.06	12	427.67	17	933.17	15	218.43	9
it_politics	135.37	3	145.20	1	1251.39	1	2469.68	1	495.65	1	4861.09	2	2556.82	1	1430.25	1	279.79	1
mena	109.83	13	124.41	6	324.83	13	1931.93	8	439.88	14	5450.22	14	2326.20	12	841.91	8	362.75	2
migration	375.00	12	127.00	8	884.38	4	1963.21	5	324.27	12	4122.06	10	2930.67	15	1117.83	7	272.80	8
protest	58.48	2	114.71	4	1103.15	5	2026.91	4	504.79	8	4169.65	3	1473.68	7	872.00	5	246.85	3
religion	43.33	14	130.12	16	1490.00	8	1561.20	11	863.00	6	6472.75	6	2147.37	5	1142.50	16	203.60	14
rights	87.33	9	128.21	9	843.58	6	2863.30	12	448.86	15	5917.34	7	2074.70	13	1166.00	9	198.00	7
sport	109.25	16	101.56	14	2983.18	7	2382.00	13	435.67	9	7227.19	11	2225.88	10	1514.70	12	131.20	17
tech	70.80	10	120.20	13	494.75	14	1628.12	14	385.67	17	4706.38	17	1948.76	8	457.00	17	171.00	16
ukraine	312.31	4	113.25	11	465.89	10	2335.17	16	462.75	4	6819.73	16	2145.60	14	1680.91	10	267.18	12
unclear	461.58	1	204.71	10	1077.20	9	3248.45	6	536.36	10	7899.77	4	2838.23	4	1305.88	14	328.33	10

Figure 3.13. Average of likes and position in the ranking of the most posted topics by the party

a column indicating the average number of likes to the posts per topic, and in the column next to it the position in the ranking of the most posted topics by the party.

Is there a relationship between likes and the number of posts related to a topic?

Here are the topics that seem to be most correlated with their ranking. All topics where the parties seem to have positive correlation between likes and number of posts, i.e. they have many likes on average in a topic and the ranking of the most popular topics per party is high compared to the other parties. In brackets are marked if the topic received many or few likes and the position in the ranking

position (RP) of the most discussed topics for that party.

Azione: Ukraine (many likes, RP: 4), **Europa verde – Verdi**: Climate (many likes, RP: 3), **Forza Italia**: Elections (many likes, RP: 3), Sport (many likes, RP: 7), **Italia Viva**: Religion (many likes, RP: 6), Entertainment (many likes,7), Climate (few likes, RP: 16), **Matteo Salvini**: Crime (many likes, RP: 8), **Movimento 5 Stelle**: Religion (many likes, RP:5), **Partito Democratico**: Entertainment (many likes, RP: 6), Migration (many likes, RP: 7), **Sinistra Italiana**: Health (many likes, RP: 9), Mena (many likes, RP: 2), Religion (few likes, RP: 14), Sport (few likes, RP: 17).

Here, they are the topics for each party where there seems to be low correlation between likes and rank position (RP):

Azione: Climate (many likes, RP: 15), Protest (few likes, RP: 2), **Europa verde – Verdi**: Economy (few likes, RP: 2), Protest (few likes, RP: 4) **Forza Italia**: Education (many likes, RP: 11) **Matteo Salvini**: Climate (few likes, RP: 5), Elections (many likes, RP: 15), Entertainment (few likes, RP: 9), Mena (many likes, RP: 14), Sport (many likes, RP: 11), Uckraine (many likes, RP: 16) **Movimento 5 Stelle**: Economy (few likes, RP: 3), Education (few likes, RP: 6), Mena (many likes, RP: 12), **Partito Democratico**: Mena (few likes, RP: 8), Protest (few likes, RP: 5), Religion (many likes, RP: 16), Sport (many likes, RP: 12), Uckraine (many likes, RP: 10) **Sinistra Italiana**: Crime (many likes, RP: 11), Economy (few likes, RP: 4).

With all these analyses, we noticed how much the topics for each newspaper lean towards certain topics rather than others, and we saw that there might be correlations. However, in order to be able to make these assertions, it is necessary to study these cases by means of precise indices that can validate these assertions.

Chapter 4

Correlation

4.1 Shapiro-Wilk Test

After studying the dataset of political parties and that of Italian newspapers, we try to understand the relationship between them. Let us analyze, that is, the correlation. In doing so, however, we consider only the topic frequencies for each page. In order to do this, we need to verify that the data are normally distributed. We therefore use the **Shapiro - Wilk Test** (*chapter 2.2.1*).

Figures 4.1 and 4.2 show two examples of test results for the two different data sets.

As can be seen from the results, the data do not follow a normal distribution (because $p<0.05$) so we cannot use Pearson's correlation but will use Spearman's correlation.

The difference is that while Pearson's correlation measures the linearity between the distribution of topics of a political group and a news outlet (thus a value close to +1 indicates that the two share a similar linear pattern, while a value close to -1 indicates an opposite pattern), Spearman's correlation (which is less influenced by outliers) reflects the similarity of patterns at the ranking level between topics.

Thus, if two entities assign similar importance to different topics (regardless of the numerical magnitude of frequencies), they will obtain a high Spearman value.

```
$Azione

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.84495, p-value = 0.00899
```

Figure 4.1. Shapiro - Wilk test for "Azione"

```
$`Corriere della Sera`

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.79217, p-value = 0.001588
```

Figure 4.2. Shapiro - Wilk test for "Corriere della Sera"

4.2 Spearman's correlation

Here, in Figure 4.3 there is the Spearman's correlation between parties and newspapers.

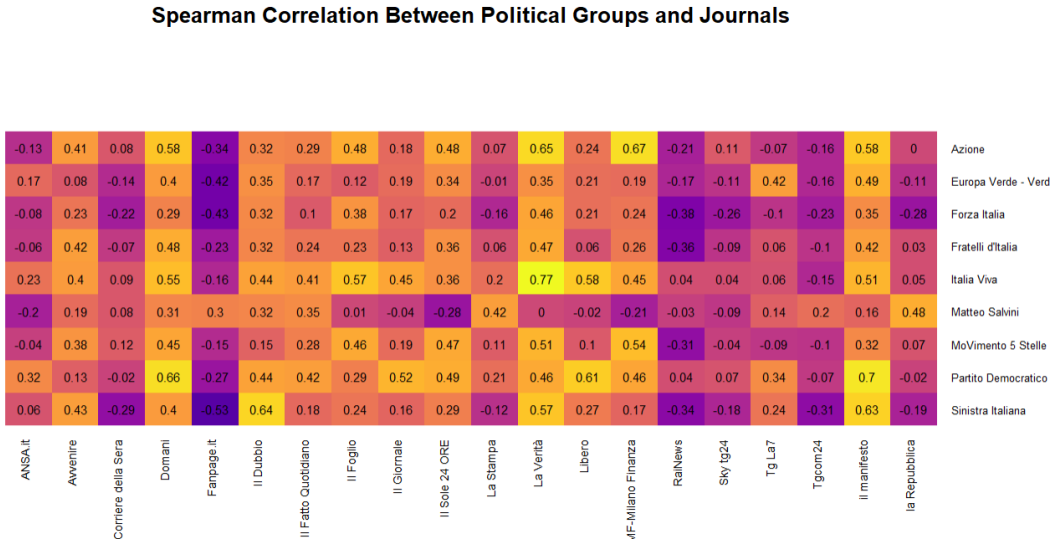

Figure 4.3. Spearman's correlation between parties and newspapers

There are not many high correlations; however, the strongest are the following:

- **Azione** with La Verità (0.65), MF - Milano Finanza (0.67).

- **Italia Viva** with La Verità (0.77).

- **Partito Democratico** with Domani (0.66), Libero (0.61), Il Manifesto (0.7).

- **Sinistra Italiana** with Il Dubbio (0.64), Il Manifesto (0.63).

Of course, the fact that there is a correlation does not mean that politically these parties and newspapers think alike, but that they talk in similar amounts about the same topics, and they are therefore similarly polarized.

Chapter 5

Clusterization

5.1 Dendograms

Clustering is a process that allows us to group together parties and newspapers that possess the same characteristics, in this case talking about similar topics.

Let us create clusters for the two datasets. To do so I used the Spearman method for the correlation between parties/newspapers, then I applied hierarchical clustering with the method "Average".

From this procedure, I obtained the dendograms in Figures 5.1 and 5.2.

A **dendrogram** is a type of tree graph used to represent the clustering structure between different elements. The dendrogram is constructed on the basis of similarities or distances between the data: similar elements (or close in terms of distance) are initially joined, and then, as they grow, they are combined into larger groups. Each bifurcation or node represents a union between groups or individual elements. The vertical axis indicates the level of similarity or distance at the time of joining; lower levels indicate greater similarity.

By cutting the dendrogram to a certain level, a number of groups or clusters can be chosen. The question arises. How can one tell at what level to cut the

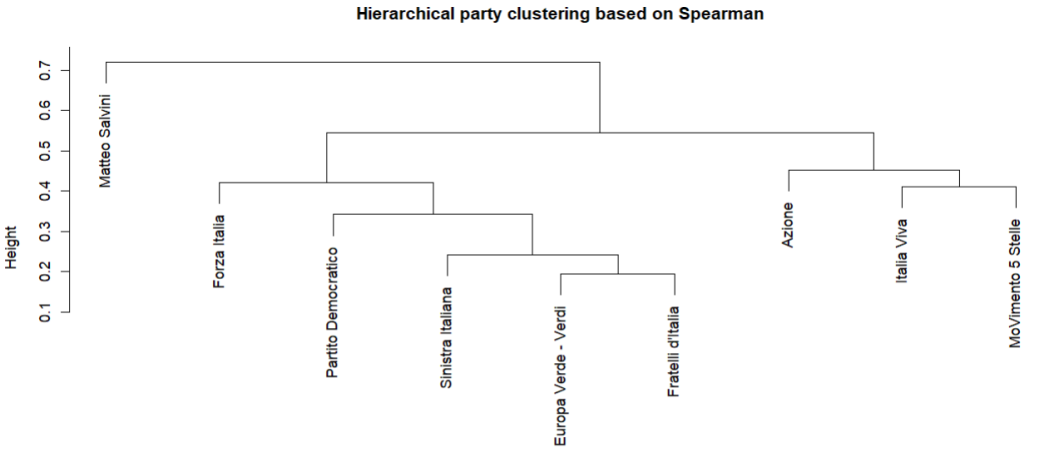

Figure 5.1. Party dendogram

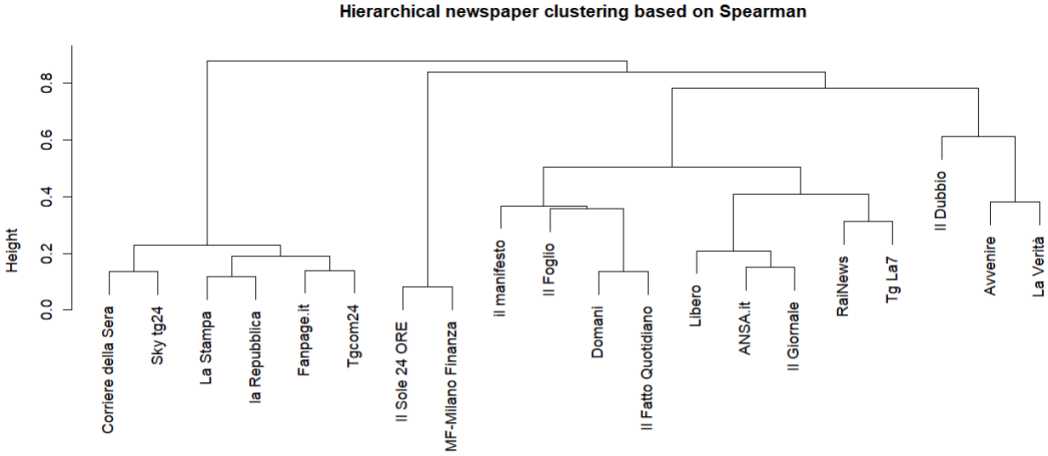

Figure 5.2. Newspaper dendogram

dendogram? We use the **silhouette index** (*chapter 2.2.2*).

5.2 The silhouette index

After various analyses, it turned out to be the best choice to divide the data into four clusters. In fact, in this case the silhouette index has the highest values, although these are not very high and satisfactory. We see the results in the Figures 5.3 and 5.4

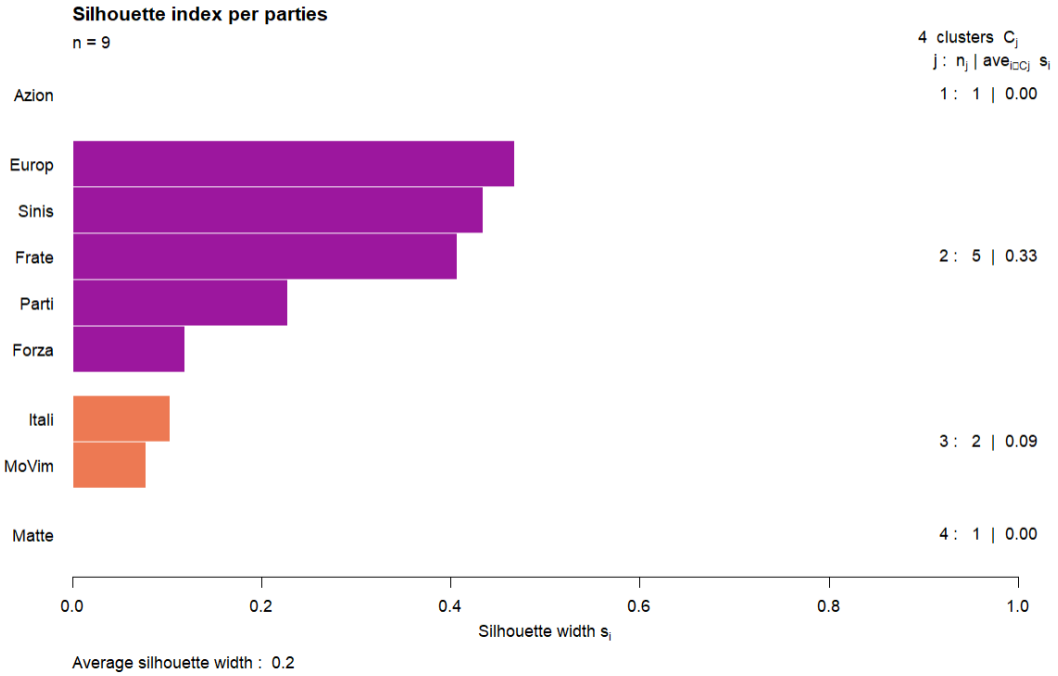

Figure 5.3. Silhouette index per parties

The silhouette index for the parties was averaged 0.2, so it was a very low characterization, and the groups are not so well divided. The data would be divided as follows:

Cluster 1 -> Azione (contains only 1 party).

Cluster 2 -> Europa Verde - Verdi, Sinistra Italiana, Fratelli d'Italia, Partito Democratico, Forza Italia (silhouette index equal to 0.33).

Cluster 3 -> Italia Viva, Movimento 5 Stelle (silhouette index equal to 0.09).

Cluster 4 -> Matteo Salvini (constituates only 1 party).

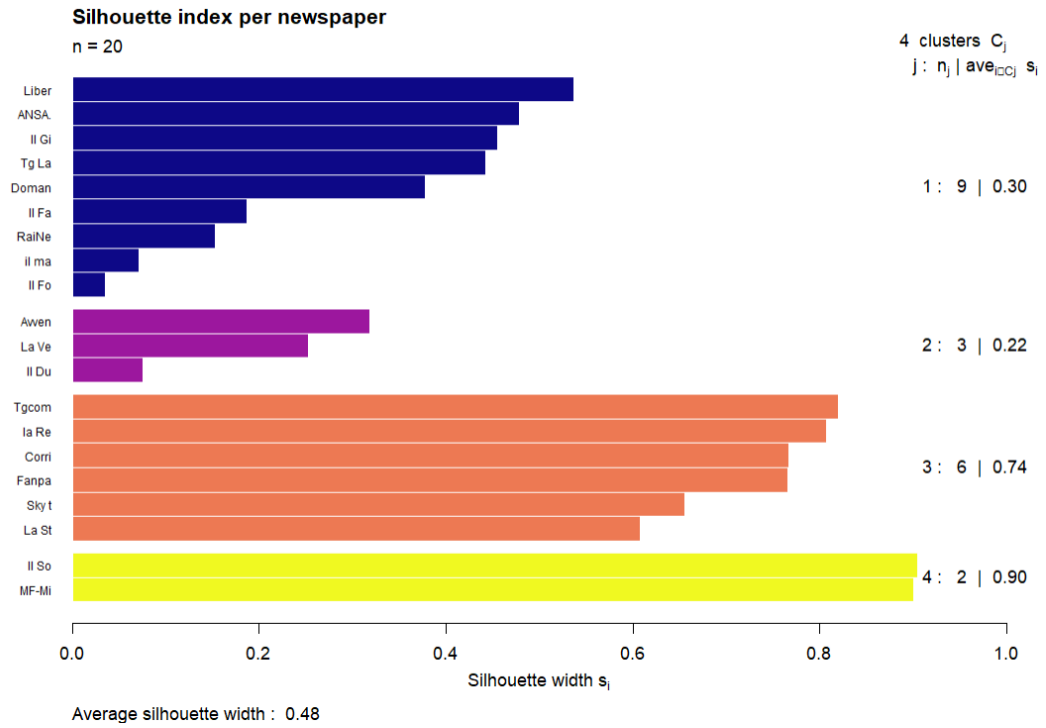

Figure 5.4. Silhouette index per newspaper

The situation is slightly better for the newspaper dataset. In fact, the average silhouette width is 0.48. (even if this is not so good).

The division of the newspapers into the clusters results in this way:

Cluster 1 -> Libero, Ansa.it, Il Giornale, Tg La7, Domani, Il Fatto Quotidiano, RaiNews, Il Manifesto, Il Foglio (Silhouette index equal to 0.30).

Cluster 2 -> Avvenire, La Verità, Il Dubbio (Silhouette index equal to 0.22).

Cluster 3 -> Tgcom24, La Repubblica, Corriere della Sera, Fanpage.it, Sky tg24, La Stampa (Silhouette index equal to 0.74).

Cluster 4 -> Il Sole 24 Ore, MF - Milano Finanza (Silhouette index equal to 0.90).

Figures 5.5 and 5.6 show the different clusters for newspapers and parties.

party	cluster
Azione	1
Europa Verde - Verdi	2
Forza Italia	2
Fratelli d'Italia	2
Partito Democratico	2
Sinistra Italiana	2
Italia Viva	3
MoVimento 5 Stelle	3
Matteo Salvini	4

Figure 5.5. Parties clusters

newspaper	cluster	newspaper	cluster
ANSA.it	1	Il Dubbio	2
Domani	1	La Verità	2
Il Fatto Quotidiano	1	Corriere della Sera	3
Il Foglio	1	Fanpage.it	3
Il Giornale	1	La Stampa	3
Libero	1	Sky tg24	3
RaiNews	1	Tgcom24	3
Tg La7	1	la Repubblica	3
il manifesto	1	Il Sole 24 ORE	4
Avvenire	2	MF-Milano Finanza	4

Figure 5.6. Newspapers clusters

5.3 The clusters

Having constructed the clusters, I wanted to analyze the topics that most characterize each cluster, to see if the party groups or newspapers talk about similar topics.

Figure 5.7 shows the clusters with the topics most discussed in each group.

In all clusters, the topic economics is present (fairly evenly distributed) and Italian politics (in some clusters massively in others more evenly), we then note

Cluster 1 -> shows in particular the higher presence of the topic Protests and a strong presence of Ukraine.

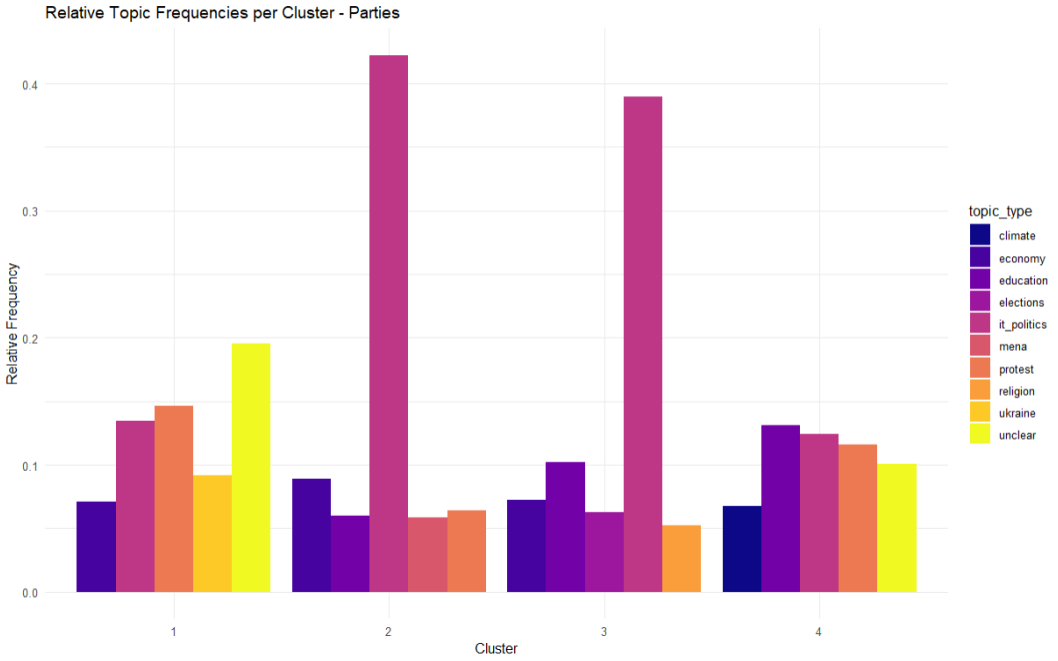

Figure 5.7. Relative Topic frequencies per parties clusters

Cluster 2 -> characterized by a strong presence of the Italian politics topic, with a smaller number of Protests, Education, and Mena.

Cluster 3 -> Here too there is a strong presence of the topic Italian Politics, followed by Education, Elections, and Religion.

Cluster 4 -> the most frequently discussed topic is Education, followed immediately by Italian Politics, Protests, and Climate.

What happens in the newspaper clusters instead? Let us see it in the figure 5.8

Cluster 1 -> The newspapers in this cluster talk about Italian Politics, Sports, Entertainment, and Mena.

Cluster 2 -> The newspapers in this cluster report on Italian Politics, Religion, Elections, and the Economy.

Cluster 3 -> The newspapers in this cluster report on Entertainment, Crime, Sport, and Climate.

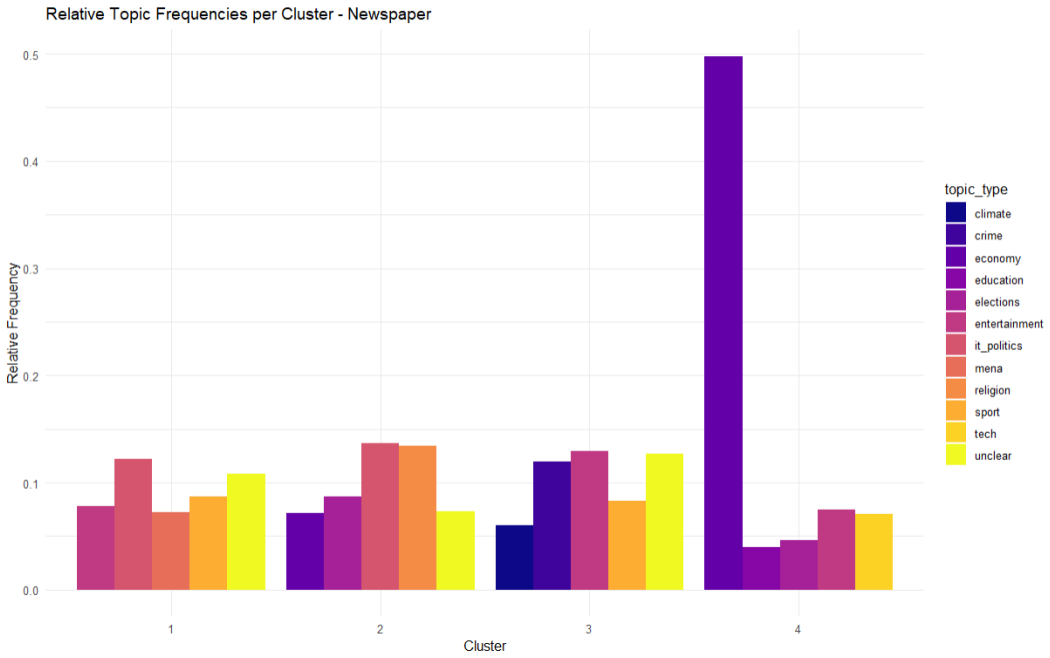

Figure 5.8. Relative topic frequencies per parties clusters

Cluster 4 -> The newspapers in this cluster talk a lot about the Economy, then to a lesser extent about Entertainment, Tech, Elections and Education.

As can be seen from both the low index results and the characterization of the clusters, the results are not the best. It is not really possible to divide newspapers and parties into clusters because clusters are not formed that are strongly polarized toward certain topics. The only case is for newspapers in cluster 4, in which Il Sole 24 Ore and MF - Milano Finanza are present. Here, there is a strong tendency towards the topic economy. However, in the remaining clusters, there is no specific characterization.

Consequently, we have to abandon the idea of clusters and focus on the study of each individual page (party and newspaper) in the individual, to see if they are biased towards a specific topic compared to the general distribution of the others. To do this, it is necessary to study the entropy.

Chapter 6

Entropy and KL divergence

6.1 Entropy

Entropy (*chapter 2.2.3*) is a measure of uncertainty or randomness in a dataset, quantifying the average level of information or surprise inherent in the possible outcomes. Then, it is strictly connected with polarization.

We then study the entropy graph of the two datasets (Figure: 6.1). Knowing that the lower the entropy, the less disorder there is and therefore a lot of polarization.

What immediately catches the eye is that for the most part it is the parties that are more polarized, while the newspapers tend to speak on a broader spectrum of topics. This can be guessed from the fact that the parties have very specific communication objectives, very much focused on Italian politics. In contrast, newspaper communication ranges from politics to crime, entertainment, sport, etc.

However, the one with the lowest entropy value is not a party but a newspaper: MF - Milano Finanza. In fact, this is well known for being a very specific newspaper that speaks strictly about economic issues. Il Sole 24 Ore, Fanpage.it, and Avvenire also obtain similarly low indices (and similar to those of the parties). These are also newspapers with a specific target.

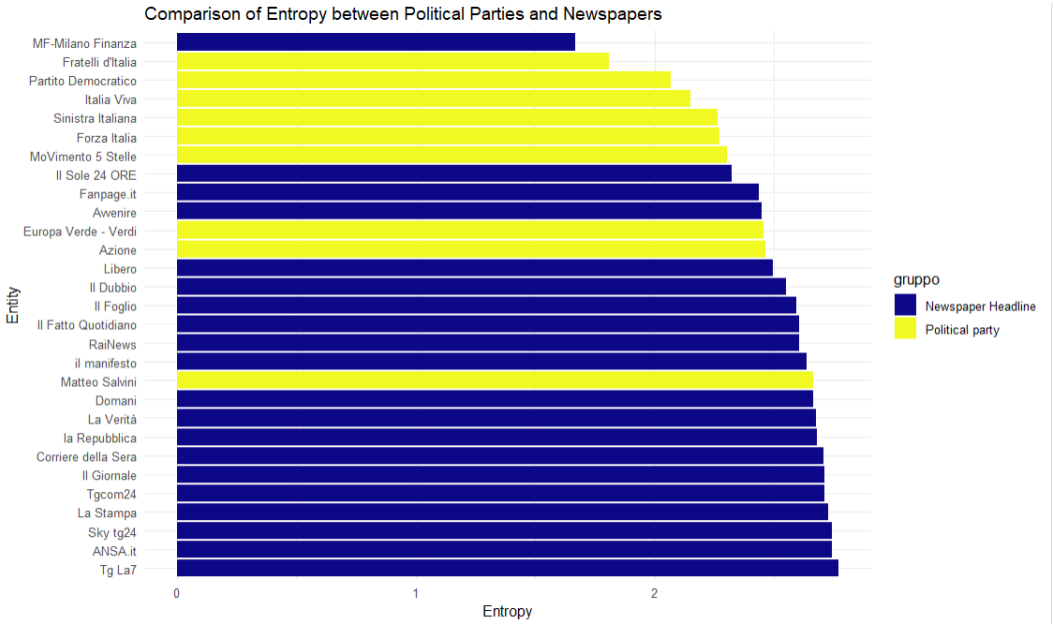

Figure 6.1. Entropy values

The less polarized newspapers, on the other hand, are newspapers that address a wide audience, primarily Tg La7, Ansa.it, and Skytg24.

Looking at the parties, the most polarized is Fratelli d'italia, the Partito Democratico, and then Italia Viva. Clearly detaching themselves from the others (and thus assuming a less polarized behavior) are Europa Verde - Verdi and Azione, but even more so Matteo Salvini, who talks about many different topics.

It's now necessary understand where the pages are polarized.

6.2 KL Divergence

With Kullback-Leibler divergence, it is possible to check where polarization exists. The KL divergence measures the difference between two probability distributions (*chapter 2.2.4*). For example, between the distribution of topics covered by one newspaper and the overall distribution of topics covered by all newspapers. This helps to understand whether a newspaper focuses on certain topics disproportionately to others, indicating a possible bias.

6.2.1 Political Parties

Let us start by looking at the polarization in political parties.

We check the Z score of the parties (Figure: 6.2) and exclude those that have |z|<3 because they are not polarized. Then we study the distribution of the others to check which topics deviate the most from the general distribution (Figure: 6.3).

Party	Z_score
Azione	-1.177908
Europa Verde - Verdi	-8.431953
Forza Italia	-5.554887
Fratelli d'Italia	-5.295193
Italia Viva	-6.991766
Matteo Salvini	-4.668333
MoVimento 5 Stelle	-8.039089
Partito Democratico	-7.246298
Sinistra Italiana	-6.447860

Figure 6.2. Z score for parties

Azione is the only party with a z score less than 3 and higher than -3. For this reason, it is not polarized and we can exclude it. Let us see the distribution of the topics for the other parties to check where it is possible to find polarization.

In Figure 6.3 it is possible to see the polarisations of the various parties. In particular, we have on the x-axis the parties and for each party the various topics, while on the y-axis the frequency differences are the following: That is, for each topic and party, the relative frequency is compared to the reference distribution. The 'frequency difference' is therefore a measure of how much this probability (frequency) of a topic for a party deviates from the reference distribution (the overall distribution).

The most striking results are related to Italian Politics. In fact, it seems that the range of the distribution is very wide and the weight of the differences is higher. One notices, for example, the great polarization towards Italian Politics of Fratelli d'Italia, followed by the Partito Democratico, then Italia Viva and the Sinistra Italiana. In contrast, Matteo Salvini is negatively polarized, i.e. he speaks very little

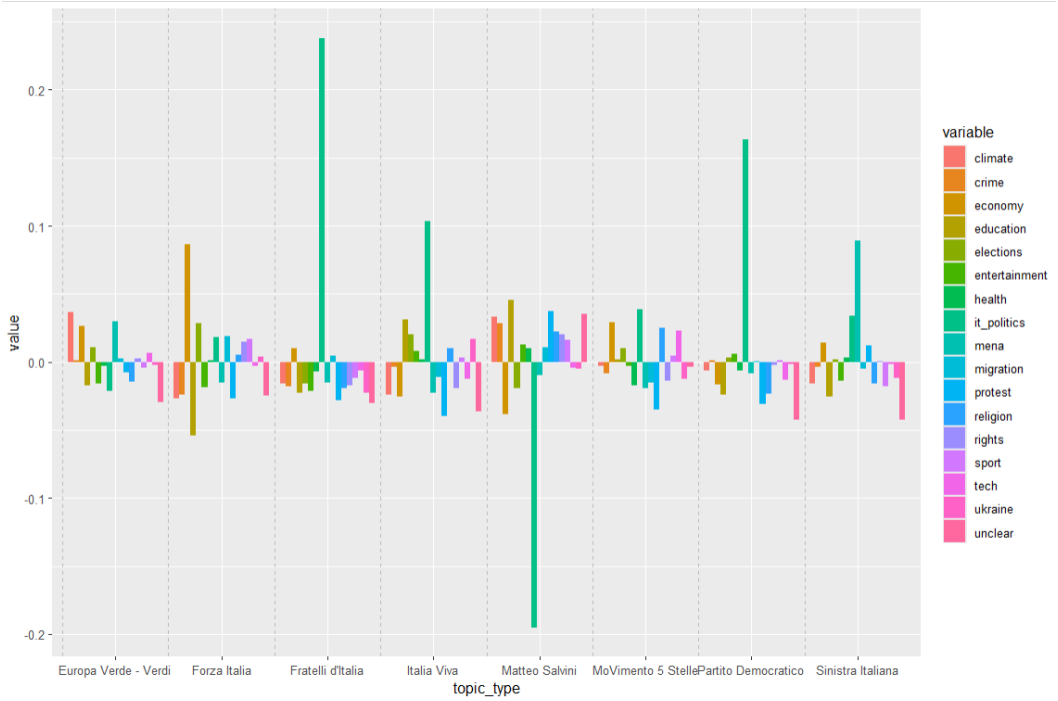

Figure 6.3. Polarization in parties

about Italian Politics.

In addition to this topic, there are also other polarizations, such as the Economy in Forza Italia (and the negative polarization of Forza Italia towards Education) or the polarization of Europa Verde - Verdi towards the Climate.

However, precisely because the breadth of the distributions is different for each topic, it is necessary to see individually for each topic the parties that are outliers with respect to the general trend (and therefore polarized). We therefore create a box plot (a statistical graph that represents the distribution of a dataset and provides an overview of the central values, dispersion, and potential outliers).

The box plot (Figure 6.4) makes **parties polarizations** much more evident:

- Azione: -

- Europa Verde - Verdi: This party is more polarized than the others in **Mena**.

- Forza Italia: -

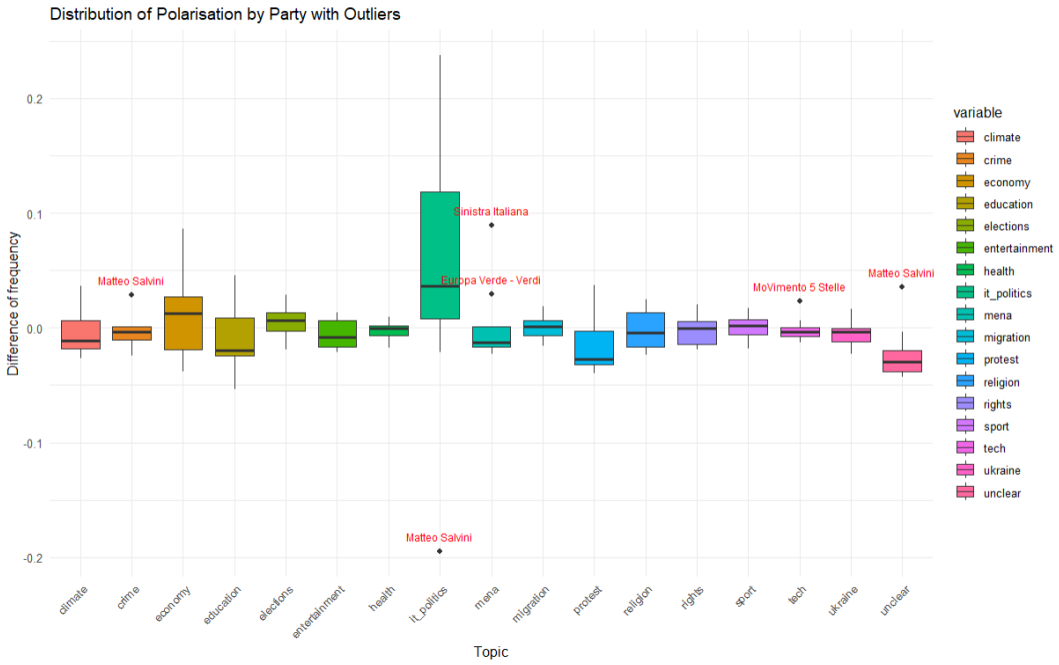

Figure 6.4. Boxplot for the parties polarization

- Fratelli d'Italia: -

- Italia Viva: -

- Matteo Salvini: develops **more Crime-oriented** communication than the average and develops much **less** discourse on **Italian Politics** than the average.

- Movimento 5 Stelle: is particularly polarized in **Tech**.

- Partito Democratico: -

- Sinistra Italiana: has a much more **Mena** (Middle East and North Africa) communication than average.

Four parties over eight are polarized.

6.2.2 Newspapers Headline

Let us check the Z score for the newspapers (Figure 6.5).

Newspaper	Z_score
ANSA.it	-0.274631
Avvenire	130.481977
Corriere della Sera	-12.794205
Domani	12.947274
Fanpage.it	42.449380
Il Dubbio	57.386020
Il Fatto Quotidiano	4.280485
Il Foglio	32.766023
Il Giornale	-10.575262
il manifesto	42.601840
Il Sole 24 ORE	78.478642
la Repubblica	-2.456391
La Stampa	-12.189619
La Verità	14.068762
Libero	50.135780
MF-Milano Finanza	201.906602
RaiNews	41.334376
Sky tg24	-11.494252
Tg La7	4.138948
Tgcom24	-5.988385

Figure 6.5. Z score for newspapers

I am not going to consider the newspapers that have a z score higher than -3 or lower than 3, because they are not polarized. For this reason, "Ansa.it" and "La Repubblica" will be excluded.

Let us now see the distribution of the topics for the other newspapers, to check where they are polarized (Figure 6.6).

Polarizations that cannot be overlooked because they are immediately obvious to the eye are as follows: MF - Milano Finanza and Il Sole 24 Ore are very polarized on economics, just as Avvenire is very polarized on Religion. Then one also notices the polarizations of Libero, Il Foglio, Il Dubbio, and Domani on Italian Politics, Il Manifesto on Mena, or Rainwes on Sport and Ukraine.

However, as done above, it is necessary, in order to better understand the most polarized newspapers, to analyze the graph of the distribution (the box plots) and thus clearly notice the outliers that deviate from the general distribution.

In Figure 6.7 the polarization of the newspaper is clear. Let us summarize them:

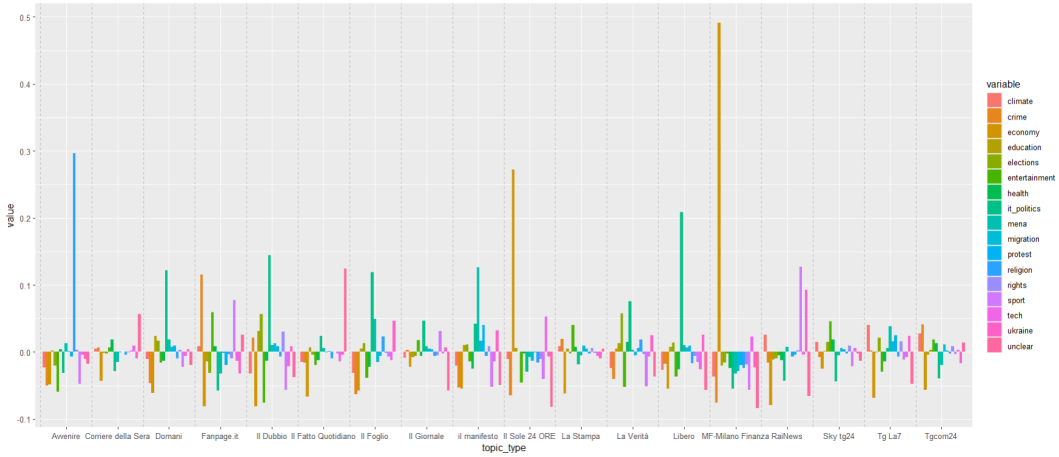

Figure 6.6. Polarization in newspapers

- Ansa.it: -

- Avvenire: very polarized in **Religion**.

- Corriere della Sera: -

- Domani: polarized in **Education**.

- Fanpage.it: polarized in **Crime** and **Sport**.

- Il Dubbio: polarized in **Education**, **Elections**, **Rights**.

- Il Fatto Quotidiano: -

- Il Foglio: Polarized in **Mena** (Middle East and North Africa) and **Religion**.

- Il Giornale: -

- Il Manifesto: Very polarized in **Mena** (Middle East and North Africa) and **Protest**.

- Il Sole 24 Ore: very polarized in **Economy** and **Tech**.

- La Repubblica: -

- La Stampa: -

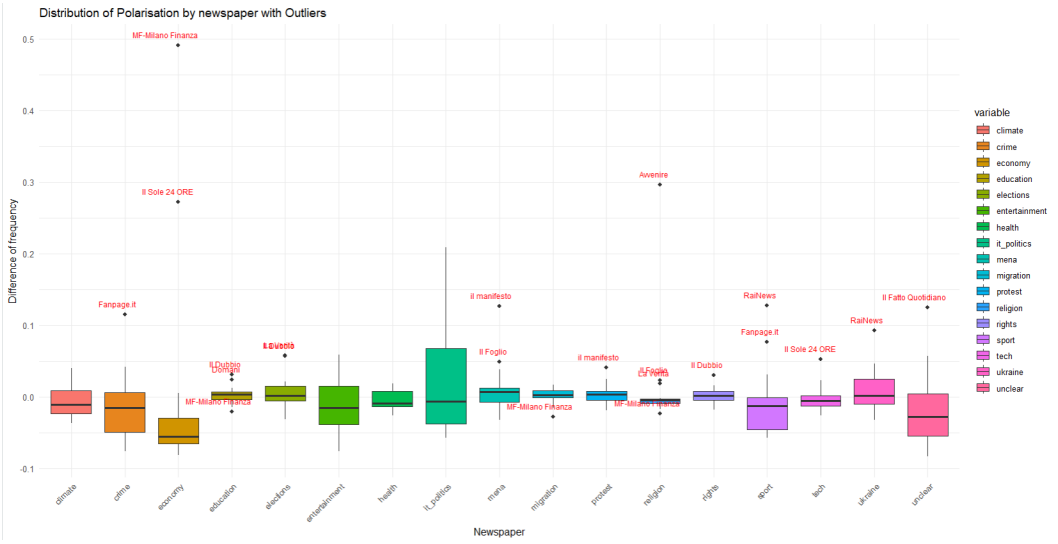

Figure 6.7. Boxplot for the newspapers polarization

- La Verità: polarized in **Elections** and **Religion**

- Libero: -

- MF - Milano Finanza: highly polarized in **Economy**. It is also **negatively** polarized (i.e. speaks little compared to the average) **on Education**, **Migration** and **Religion**.

- Rai News: polarized in **Sport** and **Ukraine**.

- Sky tg24: -

- Tg La7: -

- Tgcom24: -

So far half of the newspapers (10/20) are therefore polarized on certain topics. Exactly the same as with political parties (4/8). However, it is noticeable how the polarization in the newspapers is particularly stronger and more distributed among the different topics. In some cases the polarization was estimable because it was dictated by the field of interest of the newspapers (e.g. Il Sole 24 Ore and MF Milano Finanza with the Economy or Avvenire on Religion), in other cases it was

less estimable because it was the result of undeclared editorial choices and dictated more by an interest detected in the period under examination towards a topic).

Looking at the polarizations of parties and newspapers in comparison, one can see that both **Matteo Salvini** and **Fanpage.it** talk a lot about Crime. While **Sinistra Italiana** and **Europa Verde - Verdi** are very polarized on Mena (Middle East and North Africa) as are **Il Manifesto** and **Il Foglio** for newspapers. While **Movimento 5 Stelle** talks a lot about Tech as does **Il Sole 24 Ore**.

It should be remembered and kept in mind, of course, that if two parties or newspapers are polarized on the same topic, it does not mean that they are of the same mind, but that they talk about the same topic, for better or worse.

Chapter 7

Conclusions

The study carried out was a magnifying glass on the Italian newspaper and party communication landscape. The first half of 2024 was eventful and saw politicians in particular facing off in the run-up to the European elections, and journalists in turn reporting on the problems and challenges they saw in Europe. In doing so, through their Facebook pages, parties and newspapers chose to focus their communication on certain issues rather than others. Some of these have remained more general, others have chosen a targeted communication strategy in certain areas.

The study of possible biases towards certain topics initially led me to analyze the distribution of pages towards various topics. The newspapers were found to be less polarized, focusing on topics such as Economics, Entertainment, Crime, Italian Politics, and Sport. The parties, on the other hand, were much more polarized toward Italian Politics.

In order to understand more hidden communication strategies, I proceeded with analyzing the topic distribution by considering the center, the left, and the right political tendencies of newspapers and parties. This analysis revealed that in general for newspapers, the communication seems similar between different political tendencies. On the other hand, for the parties, after general topics addressed by all of them (It Politics, Education, Protest, and Economy), those from the center

have communication focused on Ukraine, Elections and entertainment; the left-wing newspapers more on Mena, Protest, Elections, and Climate. Those on the right on Climate, Migration, and Religion.

After the study based on political trends, I went into even more detail and analyzed the data of each individual page. Particularly surprising is the strong presence of the topic Crime for Fanpage.it, climate for TgLa7, and Mena for Il Manifesto. But also the wide space of RaiNews for Ukraine. For the parties, on the other hand, we note Protest for Azione, Education for Italia Viva, Partito Democratico, and Matteo Salvini; Mena for Sinistra Italiana; the others with Economy. But also noteworthy is the large focus on Matteo Salvini's other parties for Crime, such as Europa Verde - Verdi for Climate, and Ukraine for Azione.

Then, I analyzed how much the number of posts might have been influenced by the number of likes for that particular topic. Thus, I also studied the type of reaction each topic elicited. For example, I found that in both newspapers and parties, the topic 'Mena' arouses anger; Italian Politics, on the other hand, arouses laughter; Entertainment, Religion and Sport, on the other hand, arouse love. Angry is definitely the strongest emotion, which speaks to the belly of the population, and so I dwelt on this emotion, noting how topics such as Crime, Rights, Education, and Protest originate from it. The parties that generate the most anger in voters are Matteo Salvini (40.59% angry reaction) and Movimento 5 Stelle (46.42%). The causes could be the populist politics pursued by these two parties. Looking at newspapers instead, those with the highest percentage of angry are La Verità (50.26%), Il Manifesto (46.16%) and Domani (39.77%).

After the distribution study, I moved on to the correlation study. I had to apply the Shapiro-Wilk test, as the data were not distributed and I saw that the correlations between parties and newspapers are not high with a few exceptions.

The search for common polarization trends prompted me to look for clusters that could describe possible common phenomena between the pages. Yet the study of clustering did not yield any interesting results.

After analyzing distribution, correlation, and the possible presence of clusters, the research culminated with the study of entropy to determine whether Facebook pages are polarized or not, and then of KL divergence to determine what they are polarized into.

The entropy values showed a greater polarization of parties than newspapers with the exception of newspapers with a specific theme such as MF-Milano Finanza, Il Sole 24 Ore, Fanpage.it and Avvenire.

With the polarization analysis, the interest of Europa Verde - Verdi and Sinistra Italiana on the topic Mena was important, as was the concentration of Matteo Salvini on Crime and on speaking little instead compared to the average Italian politics. Finally, particular and specific is the Movimento 5 Stelle's media focus on Tech.

As for newspapers, on the other hand, Avvenire is polarized on Religion; Domani on Education; Fanpage.it on Crime and Sports; Il Dubbio on Education, Elections, and Rights; Il Foglio on Mena and Religion, Il Manifesto on Mena and Protest; Il Sole 24 Ore on Economy and Tech; La Verità on Elections and Religion; MF - Milano Finanza on Economy and treats little compared to the average Education, Migration, and Religion; while RaiNews talks particularly about Sports and Ukrainian.

Thus, if the entropy showed higher polarization in the parties because they were considered in absolute terms, the situation considering the comparison of the various distributions is that newspapers are highly polarized in different and specific topics.

The motivation and mode of the polarization is obviously not demonstrated and explained; however, we note the common polarization of Italian Left and Europa Verde - Verdi for Mena as for Il Manifesto and Il Foglio; as well as the topic Crime commonly dealt with by Fanpage.it and Matteo Salvini; and the topic Tech dealt with a lot by both the Movimento 5 Stelle and Il Sole 24 Ore.

Biases are a complex topic and difficult to understand because of the amount of motivations that can generate them and the difficulty of demonstrability. However,

their presence remains an essential factor to be studied for the resilience of democracy in the society today and tomorrow. A society that is changing due to new technologies and social algorithms that condition citizens' communication and information.

Bibliography

[1] Vincent D Blondel and Jean-Loup Guillaume and Renaud Lambiotte and Etienne Lefebvr, "*Fast unfolding of communities in large networks*, 2008, Journal of Statistical Mechanics: Theory and Experiment

[2] Walter Quattrociocchi ,Guido Caldarelli Antonio Scala, "*Opinion dynamics on interacting networks: media competition and social influence*", Nature, 27 maggio 2014

[3] Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2021). "*Political Polarization on Twitter. Proceedings of the International AAAI Conference on Web and Social Media*", 5(1), 89-96. https://doi.org/10.1609/icwsm.v5i1.14126

[4] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, Walter Quattrociocchi, "*Collective attention in the age of (mis)information*", Computers in Human Behavior, Volume 51, Part B, 2015, Pages 1198-1204, ISSN 0747-5632

[5] Limor Peer, Bobby J. Calder, Edward C. Malthouse, "*The daily diet of news: Variation in newspaper content*", Media Management Center, Northwestern University, January 2001

[6] Bovet, A., Morone, F. & Makse, H.A. "*Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump.*" Sci Rep 8, 8673 (2018). https://doi.org/10.1038/s41598-018-26951-y

[7] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, Walter Quattrociocchi, "*Anatomy of news consumption on Facebook*", Proceedings of the National Academy of Sciences, volume 114, number 12, pages 3035-3039, Year 2017, doi 10.1073/pnas.1617052114

[8] Sunstein, Cass R., "*The Law of Group Polarization*", december 1999, University of Chicago Law School, John M. Olin Law & Economics Working Paper n. 91

[9] Jennifer Jerit, and Yangzi Zhao, "*Political Misinformation*", Department of Political Science, Stony Brook University, Stony Brook, New York 11794, USA, Vol. 23:77-94 (Volume publication date May 2020)

[10] Eytan Bakshy , Solomon Messing, and Lada A. Adamic, "*Exposure to ideologically diverse news and opinion on Facebook*", Science, 7 May 2015, Vol 348, Issue 6239

[11] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, Walter Quattrociocchi, "*The spreading of misinformation online*", 2016 Jan 19.

[12] Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W. "*Science vs conspiracy: collective narratives in the age of misinformation.*" PLoS One. 2015 Feb 23;10(2):e0118093. doi: 10.1371/journal.pone.0118093. PMID: 25706981; PMCID: PMC4338055.

[13] Alessandro Bessi, Antonio Scala, Luca Rossi, Qian Zhang & Walter Quattrociocchi, "*The economy of attention in the age of (mis)information*", Volume 1, article number 12, (2014), 31 December 2014

[14] Ignacio-Jesús Serrano-Contreras, Javier García-Marín, Óscar G. Luengo, "*Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation?*", Vol 8, No 4 (2020): The Ongoing Transformation of the Digital Public Sphere

[15] Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W. "*Echo Chambers: Emotional Contagion and Group Polarization*

on Facebook." Sci Rep. 2016 Dec 1;6:37825. doi: 10.1038/srep37825. PMID: 27905402; PMCID: PMC5131349

[16] David Lazer , Alex Pentland et all., "*Computational Social Science*", Science, 6 Feb 2009, Vol 323, Issue 5915, pp. 721-723, DOI: 10.1126/science.1167742

[17] Alexander Hanna, Chris Wells, Peter Maurer, Lew Friedland, Dhavan Shah, Jörg MatthesAuthors Info & Claims, "*Partisan alignments and political polarization online: a computational approach to understanding the french and US presidential elections*", 28 October 2013.

[18] K. D. S. Brito, R. L. C. S. Filho and P. J. L. Adeodato, "*A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions*", in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 819-843, Aug. 2021, doi: 10.1109/TCSS.2021.3063660.

[19] C. Bayrak and M. Kutlu, "*Predicting Election Results Via Social Media: A Case Study for 2018 Turkish Presidential Election*", in IEEE Transactions on Computational Social Systems, vol. 10, no. 5, pp. 2362-2373, Oct. 2023, doi: 10.1109/TCSS.2022.3178052

[20] Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli, Michele Vespe, "*Handbook of Computational Social Science for Policy*", Springer, 2023

[21] Renáta Németh, "*A scoping review on the use of natural language processing in research on political polarization: trends and research prospects*", Volume 6, pages 289–313, (2023), 19 December 2022

[22] Blesik, Till, Murawski, Matthias, Vurucu, Murat and Bick, Markus. "1. Applying big data analytics to psychometric micro-targeting". "*Machine Learning for Big Data Analysis*", edited by Siddhartha Bhattacharyya, Hrishikesh Bhaumik, Anirban Mukherjee and Sourav De, Berlin, Boston: De Gruyter, 2019, pp. 1-30. `https://doi.org/10.1515/9783110551433-001`

[23] Alice Marwick and Rebecca Lewis, "*Media Manipulation and Disinformation Online*", Data & Society, 2017

[24] Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A. Bail, "*Computational Social Science and Sociology*", Vol. 46:61-81 (Volume publication date July 2020)

[25] Kherwa, Pooja and Bansal, Poonam, 2018, "*Topic Modeling: A Comprehensive Review*, volume 7, ICST Transactions on Scalable Information Systems, doi = 10.4108/eai.13-7-2018.159623

[26] Maarten Grootendorst, "*BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, 2022.

[27] P. Bafna, D. Pramod and A. Vaidya, "*Document clustering: TF-IDF approach*", 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016, pp. 61-66, doi: 10.1109/ICEEOT.2016.7754750.

[28] Starczewski, A., Krzyżak, A. (2015). "*Performance Evaluation of the Silhouette Index*. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L., Zurada, J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2015. Lecture Notes in Computer Science(), vol 9120. Springer, Cham.

[29] Berthold Bein, "*Entropy*, "*Best Practice & Research Clinical Anaesthesiology*, Volume 20, Issue 1, 2006, Pages 101-109, ISSN 1521-6896.

[30] Kosheleva O., Kreinovich V. "*Why deep learning methods use kl divergence instead of Least Squares: a possible pedagogical explanation.*

Ringraziamenti

Queste righe vogliono essere un caloroso abbraccio a chi in questi anni mi è stato accanto particolarmente. Ho sentito l'esigenza di scrivere queste righe non tanto perchè è cosa comune, ma quanto perchè se oggi sono questa persona è grazie sia agli studi intrapresi all'università, così come il risultato del contesto sociale che ho avuto intorno, e quindi in parole povere della presenza di tutti voi.

Questo è un lavoro che racconta uno degli interessi che mi appassiona di più, ovvero la politica. Ringrazio il Professor Quattrociocchi per avermi accompagnato in questo lavoro e la Professoressa Cuomo per avermi proposto questa opportunità.

La tesi chiude due anni di studio, ma soprattutto chiude un percorso che ho scelto e che mi definirà negli anni a venire. Voglio quindi tornare all'estate 2022 quando ho dovuto scegliere quale percorso magistrale iniziare, per questo ringrazio la famiglia Piazza. E' stato infatti a Los Angeles ospite dei miei amici, che ho iniziato a valutare l'idea di iscrivermi a questo percorso magistrale.

Dopodichè il mio più grande grazie va a mio papà e a mia mamma. La vostra presenza e il vostro consiglio non mi ha mai lasciato. Grazie alla vostra spintarella mi sono deciso nell'iniziare Data Science e in tutto questo percorso mi siete sempre stati accanto aiutandomi, consigliandomi, indirizzandomi, nelle scelte secondo voi migliori, dall'alto della vostra esperienza, ma allo stesso tempo fornendo soprattutto ascolto per comprendere ciò che vivevo e sentivo. Insieme a voi ringrazio anche tutto il resto della mia famiglia, dai miei fratelli, ai miei nonni, zii e cugini. Anche con la distanza sento il vostro amore sempre accanto a me.

E così sono iniziati i due anni di magistrale. In questo percorso ho conosciuto persone incredibili. Il gruppo classe si è rivelato unito, generoso, disponibile e ho fatto amicizie importanti che porterò avanti negli anni a venire. Per prima ringrazio Ludovica. Prima non a caso, ma perchè è sicuramente grazie a te in primis che oggi sono qui. Con te ho condiviso tutto dall'inzio. Esame dopo esame, video dopo video. Quando avevo un problema per te diventava un nostro problema e questo per me ha significato tantissimo. Iniziato il percorso universitario il mio gruppetto per un progetto si è spezzettato e tu mi hai accolto nel tuo gruppo, cosa non scontata, insieme a Simone.

Simone, secondo che voglio ringraziare, per la tua disponibilità e il tuo aiuto gratuito. Nonostante ci siamo frequentati poco questi due anni il tuo consiglio è stato per me sempre essenziale.

Voglio poi ringraziare Giacomo. Una persona in gamba, buona e geniale, che ho imparato a scoprire pian piano e a cui auguro ogni bene per il futuro, grazie per avermi aiutato nei momenti di difficoltà.

Ringrazio poi Elisa, Matteo, Chiara, Gianluca, Federico. Siete stati uno dei motivi per cui la mattina avevo grande voglia di svegliarmi prima dell'alba e fare il viaggione per venire in università.

Grazie poi ai miei compagni di studio in erasmus. Mihnea, Alex e Diana. La vostra è stata una generosità per me inaudita. Eravete un gruppo di tre amici già assestato e mi avete accolto, incluso e aiutato. Grazie! Siete la mia famiglia bruxelliana.

Gli amici che ho conosciuto in questi due anni si sommano ai tantissimi amici a cui voglio bene e che conosco da ormai tanti anni. Con molti di voi c'è stato grande confronto e condivisione sulle sfide che ho affrontato e quindi voglio ringraziare alcuni di voi particolarmente.

Grazie Alessandro per essere una certezza, per essere il mio grillo parlante, per essere sempre disponibile all'ascolto e fornirmi quella spalla di cui ho bisogno.

Grazie Andrea, il tempo con te sembra non passare mai. Abbiamo vissuto e condiviso tanto e spero che non diminueremo mai questa intensità. Sei un pezzo importante della mia vita e il fatto che il nostro rapporto abbia retto a tanti terremoti da molto valore al bene che ci vogliamo.

Grazie Fabiola, considero ogni tua parola dal valore d'oro. Penso sia veramente prezioso il tuo consiglio, e gli do un altissimo valore. Ti voglio tanto bene, e hai un posto prezioso nel mio cuore.

Grazie Lorenzo, la semplicità dei momenti vissuti insieme racconta il bene instancabile che ci vogliamo. Ti auguro il meglio nel tuo futuro e spero di esserci sempre anche io a viverlo con te.

Grazie Marta, sento il nostro come un bene silenzioso, che ci accompagna in ogni passo della nostra vita. Sei instancabilmente lì, e dico instancabilmente perchè a volte posso essere un pò pesante ma so che tu ci sei, e anche quando sono lontano, nei miei viaggi, ti sento costantemente con me, mai distante.

Grazie Nicolò, ci diciamo poche volte cose carine allora approfitto adesso. So che ci sei, e che ci sei per me, per il mio bene, e che ci sarai. Sento la nostra come una amicizia semplice, senza troppe pretese, ma grande, disponibile, ferrea. Grazie per essere nella mia vita.

Grazie Roberto, per tutto il bene che mi vuoi e per tutta la cura che hai verso di me, la sento e la vedo, in particolare oggi, visto che sono stato vestito completamente da te.

Grazie Serena, perchè con te ho vissuto le difficoltà del liceo, le sfide insieme, le preoccupazioni e le gioie, e continuo a sentire tutto questo ancora adesso dopo ormai, purtroppo, tanti anni da quando abbiamo lasciato il liceo Touschek.

A questi preziosi amici se ne aggiungono tanti altri, di gruppi diversi. Grazie ai Gen, Grazie agli Anni d'Oro di Marina di Massa, Grazie al Sistema Solare, a Stati Uniti, agli amici conosciuti in erasmus e a tutti quelli che hanno condiviso un pezzetto

di vita con me. Spero con tutti voi di avere modo di vivere e condividere molto di più in futuro e se così non sarà tranquilli perchè ormai è difficile dimenticarmi di voi che siete già parte di me.