Feminicide on Reddit: how is it perceived?

Social Media Analytics Project

Cervini Stella
Montalbano Daniel
Sabino Giuseppe

Contents

1 Introduction

Femicide is a term that encompasses the intentional killing of women or girls due to their gender. This tragic form of violence highlights the extreme consequences of gender-based discrimination and inequality. It goes beyond individual acts of violence, representing a systemic issue deeply rooted in societal norms, cultural biases, and power imbalances. The phenomena encompasses a range of scenarios, from domestic violence and intimate partner killings to honor killings, dowry-related murders, and other forms of gender-driven violence leading to the death of women or girls.

The number of women killed in Italy in 2023 whose death falls in this definition is **not unique and it changes based on the source**. This is due the fact that [1] the Penal Code doesn't specifically identify femicide as a distinct crime: it is classified as homicide (*Article 575*), but it doesn't exist as a "specific criminal offense", even if from 2013 a series of decrees have introduced harsher penalties for crimes involving women who had a qualified relationship (especially familial or emotional) with the perpetrator. In these cases, there exists the possibility of an aggravated offense.

Referring to the definitions provided by the United Nations Statistical Commission, ISTAT specifies that there are three types of "**gender-related killing**": **homicides of women by their partners**, those committed by another relative or person, whether known or unknown, but occurring through a modus operandi or in a **context linked to gender motivation**. Among these are included, for example, women who are victims of trafficking, forced labor, or prostitution exploitation. Women who have been **unlawfully deprived of their freedom**, who have been raped before their murder. Another criterion that allows us to speak of femicide is the hierarchical difference in position between the victim and the perpetrator; if the body has been abandoned in a public space; if the motivation for the murder constituted a gender-based hate crime (that is, if there was a specific prejudice against women on the part of the perpetrators).

With that said, the *Observatory for Femicides, Lesbicides, and Trans-cides (FLT) in Italy by Non Una Di Meno (NUDM)* affirms that **the total number of femicides, lesbicides, and trans*cides in 2023 was of 113** [2], while according to *FemminicidioItalia.info* **the number of feminicides is equal to 43**[5] [1].

2 Project's goals

The killings relating feminicides tend to have a large media exposure, both on traditional media (TVs, newspapers, ...) but also on social media, which is where people can actively exchange ideas and opinions.

Given this premise, the project has the objective to analyse data related to the subject. Our data source is Reddit.

In particular, after an introduction on the context, the project can be divided in *three main sections*:

[1] For the analysis we are interested in, we will consider this number as the true one.

1. **Network analysis**: it involves studying the connections and relationships between individuals or entities within a social network (in this case on social media). It's a method used to analyze how people or elements are interconnected, how information flows, and how communities form and interact on social media platforms. After we build the graph, here we will perform
   (a) Topological analysis on the network (e.g. find giants components and compute all the relevant metrics)
   (b) Community detection and relative metrics (e.g. assertativity, modularity, ...)
2. **Content analysis**: in social media analytic, it involves systematically evaluating and interpreting the content shared on social media platforms. In this scenario, we will perform:
   (a) Sentiment analysis on textual posts and comments
   (b) Identify the most significant events
3. **Miscellaneous**: topics analysis on the communities previously found. This involves the previous analysis on the network.

The goal is to verify whether the theme of feminicide in Italy is discussed online on this particular social media platform, what is the polarity on the topic and identify the most meaningful events that influences these ideas.

3 Data Collection

3.1 Data source: Reddit

Reddit is a social media platform and online community where users can engage in discussions, share content, and discover a wide range of topics. It's organized into smaller communities called *"subreddits"*, each dedicated to specific interests or relevant topics. Users can post links, text, images, and videos, and other users can upvote or downvote these posts and comments, affecting their visibility on the platform. Reddit is known for its diverse user base and the ability for people to anonymously interact and share content.

3.2 Data content

Feminicide is a broad theme and it encompass **a large range of interconnected issues**. For this reason, the data collection not only considers posts and comments tricly related to feminicide but we chose to take into account the subreddit *'Italy'* and search for different relevant words inside the posts. The main subject under analysis are then:

1. **Femicide** (discussed above)
2. **Violence**: violence is intricately related to feminicide due to its role as a primary factor leading to the intentional killing of women or girls because of their gender.

3. **Patriarchy**: patriarchy is deeply intertwined with the issue of feminicide due to its role in perpetuating societal norms, power imbalances, and gender-based discrimination. Patriarchy, a system where power and authority are primarily held by men, often leads to the marginalization, oppression, and devaluation of women. Within patriarchal structures, women might face limited rights, diminished autonomy, and systemic inequality. This power dynamic can manifest in various forms of gender-based violence, including femicide.

4. **Feminism**: feminism advocates for gender equality, challenges societal norms, and seeks to address the root causes of gender-based violence. Feminism aims to dismantle systems of inequality, including those that perpetuate violence against women. It highlights the importance of recognizing and rectifying power imbalances between genders. Feminist movements strive to change cultural attitudes, policies, and practices that devalue women and contribute to their vulnerability to violence, including femicide.

5. **25$^{\text{th}}$ November**: November 25$^{\text{th}}$ is observed as the International Day for the Elimination of Violence Against Women. It serves as a global call to action to raise awareness and combat all forms of violence against women, including femicide.

Besides these themes, we chose to specifically search for the keywords *Giulia Cecchettin* and *Filippo Turetta* because their casa has sparked considerable outrage and a public debate on the issue of femicide [3]. Given the fact that this was (unfortunately) the most significant case of feminicide in 2023, we used it as benchmark to check for post's peaks.
All the above concepts represent **keywords** that are used to retrieve data from Reddit.

3.3 Reddit's API: technical aspects

In order to collect data from Reddit, we make use of the library **PRAW**. Acronym for *"Python Reddit API Wrapper"*, it is a Python package that allows for simple access to Reddit's API. PRAW aims to be easy to use and internally follows all of Reddit's API rules [4]. Each keyword was retrieved singularly to grant the correctness of the data collected.

After the retrieving phase, the data was stored in a suitable data structure. In particular, for the analysis we chose to divide the data for *semantics*, that is two different dataframes were created:

1. the **first one** regarding the case of **Giulia and Filippo**;
2. the **second one** regarding **all the other subjects** (feminism, violence, ...) cited before.

By doing so, we was hoping to be able to grasp any correlation between news events and the discussion on all the topics that may be crucial in understanding such phenomena.

3.4 Other data sources

Data obtained from outside the Reddit platform was needed to provide some context to the analysis and provide anc objective reference point. In particular, **2 other data sources** were considered:

1. the list from all the victims of feminicide during 2023 retrieved from *FemminicidioItalia.info* [5]. FemminicidioItalia.info focuses on reporting news and in-depth analysis on recent and past cases of violence against women, stalking, and femicides that have occurred in Italy.

2. the list of all Italian towns and their region, to complement spatial and geographic analysis with temporal analysis. This data was retrieved from ISTAT [6].

4 Context of Analysis

Before continuing in the analysis, it is important to understand the context of the research. As mentioned above the number of feminicides that took place in Italy in 2023 is not well defined, but let us take the value sourced from *FemminicidioItalia.info* (43) as the true one. In order to put into context this value, we will provide a brief interpretation by plotting the total number of femicides per region and providing a timeline for them.

4.1 Geographical Distribution of Femicides in Italy

The below is a choropleth map of Italy, showing the number of femicides per region in 2023. The regions are colored according to the number of femicides, with darker colors indicating a higher number of femicides.

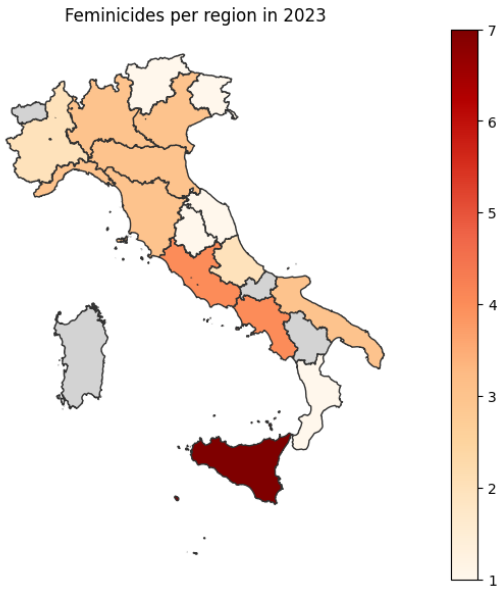

Figure 1: Feminicides choropleth map of Italy

The map shows that the highest number of femicides in 2023 occurred in the central and southern regions of Italy, with **Sicilia** having the highest number of femicides (7), followed by Campania (4) and Lazio (4). Overall, the lowest number of femicides occurred in the northern regions of Italy, with **Valle d'Aosta** having no femicides in 2023. The same is true also for **Sardegna**, **Basilicata** and **Molise**.

The map also shows that there is **not a clear north-south divide** in the overall total number of femicides in Italy, even if some of the southern and central regions have higher numbers of femicides than the northern regions.

There are a number of possible explanations for this phenomena. One possibility is that there are **cultural differences** between the north and south of Italy, with more traditional or conservative gender roles being more prevalent in the south. This could lead to a greater acceptance of violence against women in the south.

Another possibility is that there are **economic differences** between these two parts of Italy, with the south being more economically disadvantaged. This could lead to higher levels of poverty and inequality in the south, which could in turn lead to higher levels of violence against women.

4.2 Chronology of Femicides in Italy

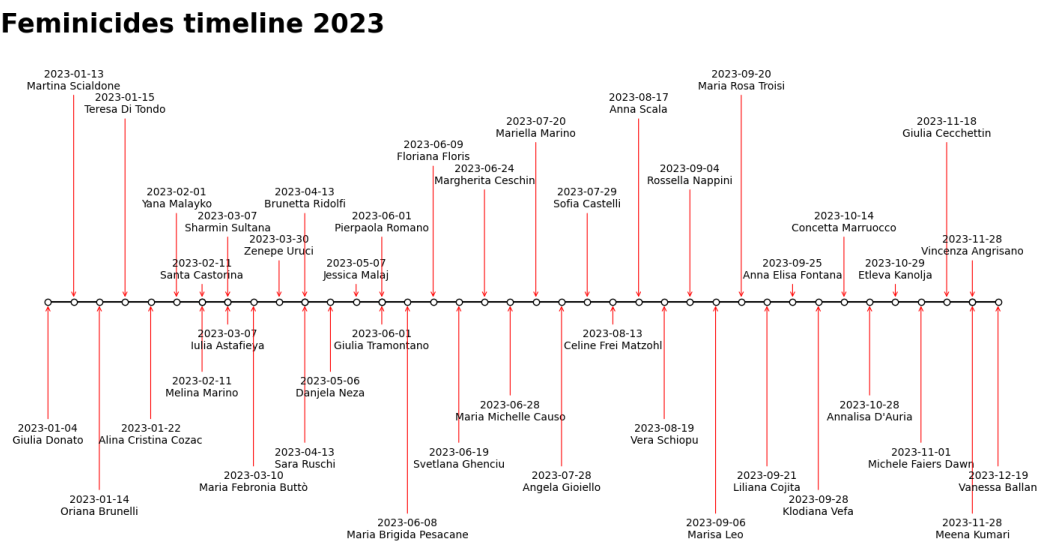

Figure 2: Timeline of Femicides in Italy

The timeline of femicides in Italy for 2023 paints a stark picture of the pervasive and persistent nature of gender-based violence in the country. It documents 43 reported cases of femicides, each representing the devastating loss of a life and the enduring impact on families, communities, and society as a whole. While the overall number of cases remained relatively consistent, the timeline reveals a troubling pattern of temporal variations, with spikes in femicides occurring in certain months. This suggests the potential influence of seasonal or cyclical factors.

5 Network analysis

Network analysis on social media involves studying the connections and relationships between individuals or entities within a social network. It's a method used to analyze how people or elements are interconnected, how information flows, and how communities form and interact on social media platforms.

5.1 Relationships definition

In our quest to understand the online community dynamics, we devised a **network architecture** aimed at analyzing the relationships between post authors and comment contributors. The primary aim was to systematically establish **connections between the creators of posts and every individual participating in discussions** through comments.

For each post in the dataset containing the Reddit's posts, we accessed the comments using PRAW, exluding anonymous comments where the author is identified as `None`. This allowed us to exclude anonymous posts/comments that influenced in a negative way the representation and to compile a comprehensive list of connections between post authors and comment contributors. This approach provides a solid foundation for in-depth analysis and exploration of the dynamics of online interactions, shedding light on the complex web of relationships that form within these digital communities.

Regarding the **nature of relationships** between authors and commenters within our graph, it's essential to consider concepts such as **multiplexity**: this refers to the presence of multiple types of connections between nodes. In our specific case, multiplexity manifests as users engaging in mutual interactions, where one user comments on another, and vice versa. This bidirectional interaction pattern forms a multiplex relationship, capturing the dynamic exchange between authors and commenters within the online community.

Analyzing these aspects within the graph can unveil the complexity and richness of relationships within the community, offering a more nuanced perspective on how authors and commenters interact across various dimensions.

Our network is sociocentric, characterized by the intricate relationships between the authors of posts and the individuals contributing comments. In this sociocentric context, the connections between nodes are based on the interactions initiated by post authors and the subsequent engagement of commenters. This sociocentric network is both unweighted and undirected. This implies that the presence or absence of connections between nodes is only determined by their interactions, without assigning weights to the relationships or indicating a specific directionality.

Following the removal of duplicates from the list of author-comment contributors, we proceeded to construct the graph using **NetworkX**. After that, we've cleaned up the graph by removing all isolated nodes. This step ensures that the analysis is centered on meaningful relationships and eliminates nodes with no direct connections in the context of the given dataset.

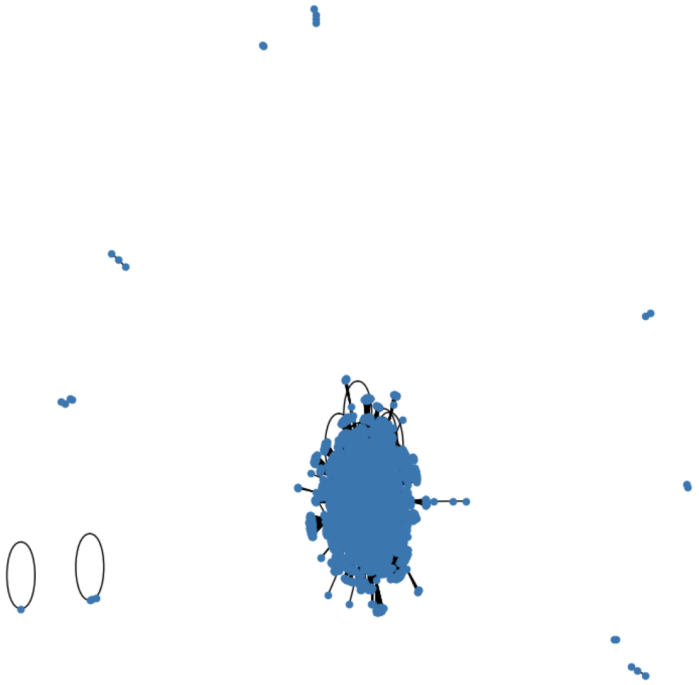

Figure 3: Initial graph

After having exluded certain nodes, our graph now comprises **4770 nodes connected by 8273 edges**.

5.2 Remove outliers

Based on our visualization, we've identified certain nodes that are relatively distant from the most highly connected nodes in the graph. To enhance the clarity of our analysis, we aim to **eliminate these outliers**. Our focus is on exploring and understanding the giant component of the graph, focusing our investigation on the most substantial and interconnected portion of the network.

5.2.1 Giant Component

We've utilized NetworkX to **highlight the giant component** within the graph, extracting the nodes that belong to this significant and interconnected part of the network.

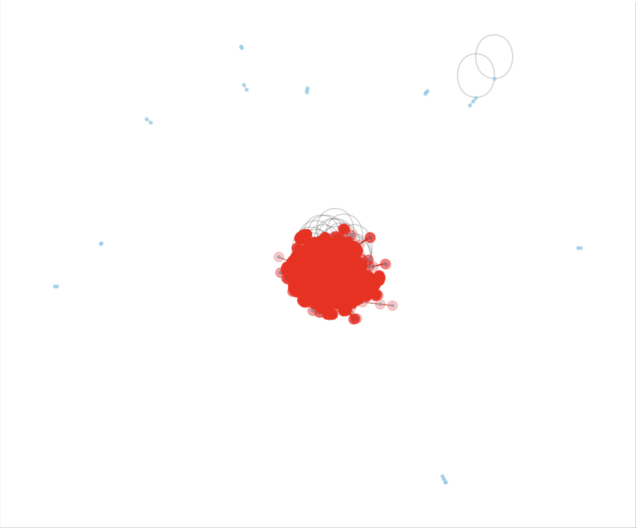

Figure 4: Giant Component Highlithed

Upon observation, the majority of nodes are concentrated within the giant component. Consequently, we have opted to exclude from the graph all nodes that do not belong to this substantial interconnected component:

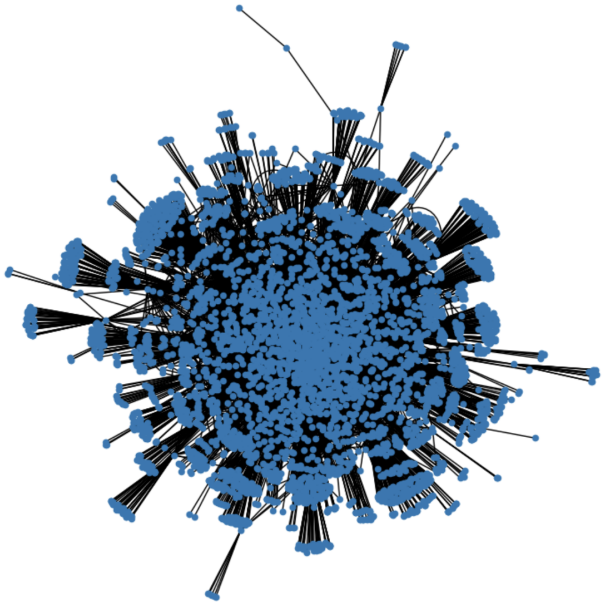

Figure 5: Final Graph

Upon examining this visualization, we've achieved a m**ore concentrated and denser graph**. This refinement sets the stage for a detailed analysis of the network. The updated statistics indicate that the graph now consists of **4744 nodes and 8256 edges**.

5.3 Metric analysis

- *Average Degree*: 3.48
  - This average degree value of approximately 3.04 provides insights into the connectivity of nodes within the graph. The degree of a node is a fundamental measure, indicating how many edges are connected to that particular node. In this context, the average degree serves as a central metric, revealing that, on average, each node in the graph is connected to ca. 3.48 other nodes.
- *Density*: 0.00073
  - The density of a graph is a measure of how many edges are present compared to the total number of possible edges. It ranges from 0 (sparse graph) to 1 (dense graph). In this case, the density is approximately 0.00073, indicating a sparse graph.
- *Average Clustering*: 0.018
  - The average clustering coefficient measures the tendency of nodes to form clusters or groups. It ranges from 0 to 1, where higher values indicate a higher tendency for nodes to form tightly-knit groups. In this case, the average clustering coefficient is approximately 0.018, suggesting a relatively low level of clustering in the graph.
- *Radius*: 5
  - The radius of a graph is the minimum eccentricity among all nodes. It represents the minimum distance from the farthest node to any other node in the graph. In this case, the radius is 5, indicating that the minimum distance from any node to its farthest neighbor is 5.
- *Diameter*: 9
  - The diameter of a graph is the maximum eccentricity among all nodes. It represents the longest shortest path between any two nodes in the graph. In this case, the diameter is 8, indicating that the longest shortest path between any two nodes in the graph is of length 9.

5.3.1 Degree Centrality

Degree centrality values offer crucial insights into the significance of nodes within a graph, emphasizing their connectivity. Nodes with higher degree centrality values are considered more central due to their increased number of connections, indicating a heightened influence or centrality in the overall network structure. The following list highlights the top 10 nodes by degree centrality in the graph, showcasing key influencers based on their substantial connections within the network.

Rank	Node (Username) - Degree Centrality
1	notsostrong134: 0.1206
2	WrongQuesti0n: 0.0428
3	MrShinzen: 0.0422
4	Single-Brain-1235: 0.0418
5	Pure-Contact7322: 0.0412
6	Visani_true_believer: 0.0388
7	Eugenio_Prigozzi: 0.0376
8	Cute-Warning7833: 0.0363
9	Millener89: 0.0354
10	LongRun00: 0.0305

Table 1: Top 10 Nodes by Degree Centrality

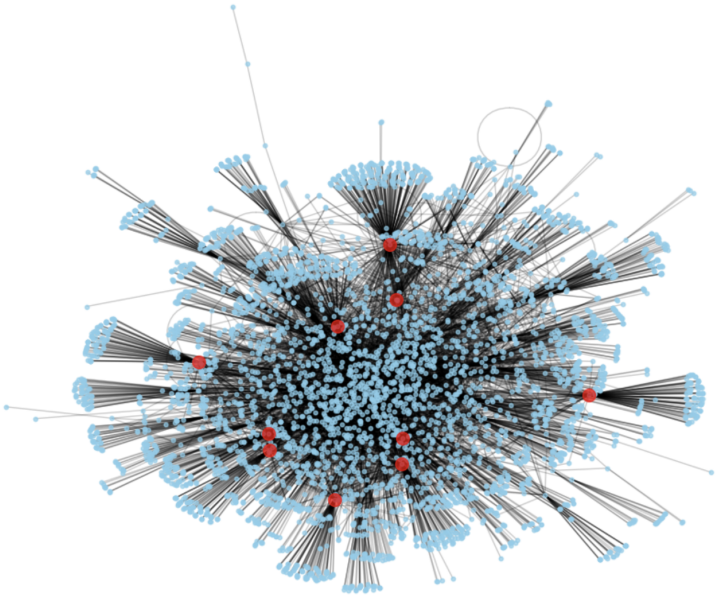

Figure 6: Graph with Top 10 (Degree Centrality) Nodes Highlighted

5.3.2 Eigenvector Centrality

Eigenvector centrality values capture the significance of nodes within a graph by considering not only their direct connections but also the connections of their neighbors. Elevated eigenvector centrality values highlight nodes with influential connections, suggesting a heightened overall influence or centrality within the network. The following list showcases the top 10 nodes possessing the highest eigenvector centrality in the graph, emphasizing their pivotal roles in terms of connectivity and influence.

Rank	Node (Username) - Eigenvector Centrality
1	notsostrong134: 0.6785
2	Pure-Contact7322: 0.1346
3	rango1801: 0.0691
4	Antistene: 0.0647
5	Kernel_Paniq: 0.0589
6	Visani_true_believer: 0.0560
7	Scatamarano89: 0.0554
8	Pineta1000m: 0.0534
9	Eugenio_Prigozzi: 0.0534
10	rotondof: 0.0470

Table 2: Top 10 Nodes by Eigenvector Centrality

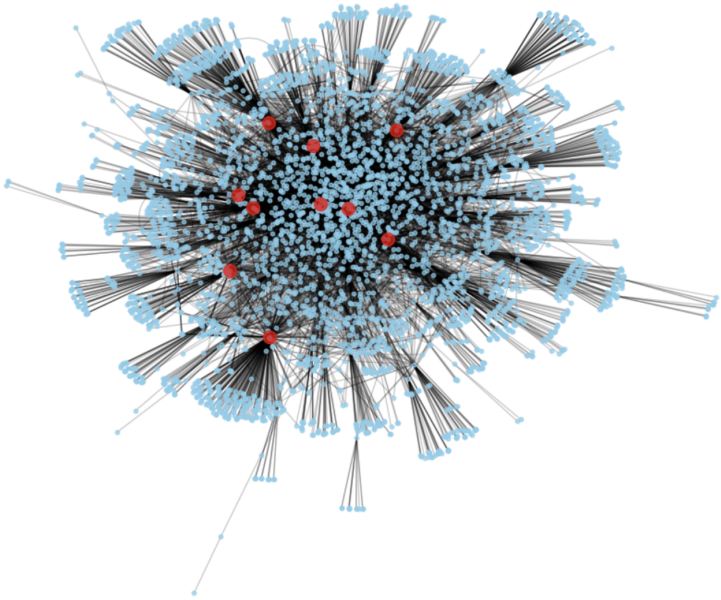

Figure 7: Graph with Top 10 (Eigenvector Centrality) Nodes Highlighted

5.3.3 Betweenness Centrality and Closeness Centrality

Betweenness Centrality identifies nodes crucial for bridging different graph components, facilitating communication and connectivity. Higher betweenness centrality values indicate nodes pivotal for maintaining efficient paths between other nodes.

Closeness Centrality, on the other hand, measures how rapidly a node can reach others in the graph. Elevated closeness centrality values highlight nodes centrally positioned, promoting efficient interactions.

In our visualization, top Betweenness nodes are plotted in red, while top Closeness nodes are marked in green.

Rank	Node (Username)	Betweenness Centrality	Closeness Centrality
1	notsostrong134	0.1361	0.3561
2	DurangoGango	0.0750	0.3698
3	Pure-Contact7322	0.0749	0.3754
4	SmellyFatCock	0.0748	0.3525
5	Syzygy82	0.0597	0.3411
6	ftrx	0.0595	0.3633
7	MrShinzen	0.0502	0.3569
8	ffioca	0.0450	0.3310
9	WrongQuesti0n	0.0438	0.3346
10	Single-Brain-1235	0.0421	0.3092

Table 3: Top 10 Nodes by Centrality

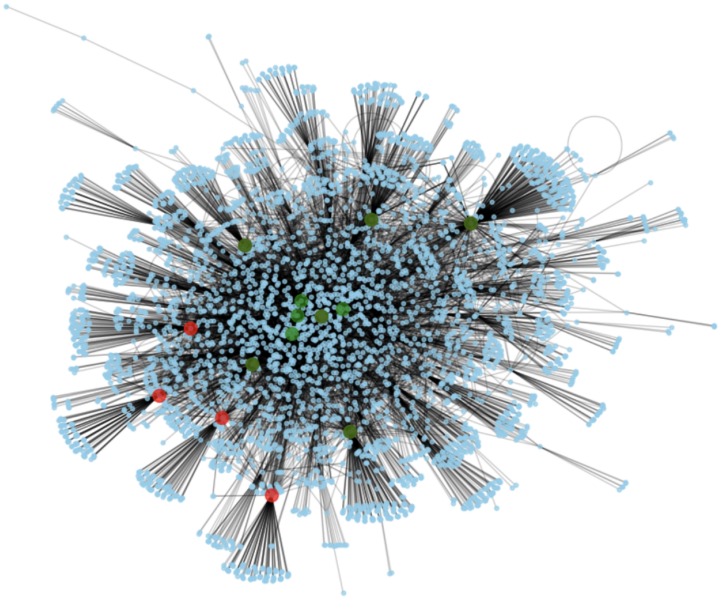

Figure 8: Graph with Top 10 Nodes for Betweenness and Closeness Centrality Highlighted

5.4 Community detection

Community detection is a crucial step in dissecting complex networks, and the Greedy Modularity algorithm emerges as an efficient tool for this purpose. With its ability to allocate nodes to communities, this algorithm provides valuable insights into the inherent structure of the graph, revealing clusters of nodes with similar connectivity patterns.

The assignment of a `community` attribute to each node facilitates in-depth analyses. This attribute allows us to explore the distinctive characteristics of each community and understand the dynamics of information flow within and across these groups. While we employed the Greedy Modularity algorithm instead of Louvain, the principles remain intact. The algorithm not only outlines the organizational structure of the network but also forms

the basis for exploring and interpreting nuanced interactions within the communities.

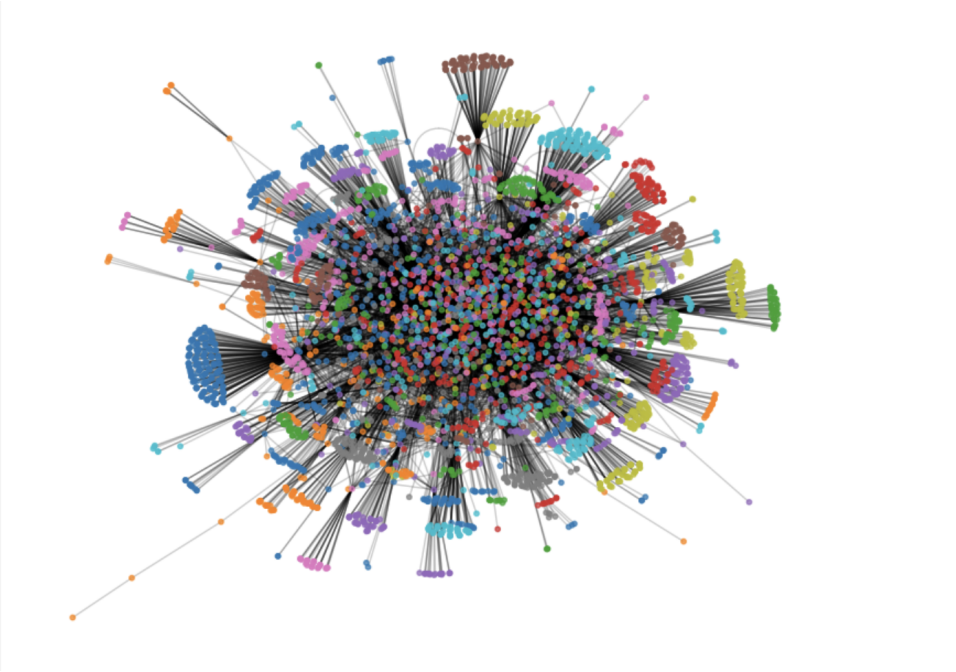

Figure 9: Community detection

The degree assortativity coefficient of the graph is a numerical representation of assortativity, which gauges the inclination of nodes with similar degrees to form connections. The obtained coefficient, in this instance, is -0.34. A negative value signifies disassortative mixing, implying a **higher likelihood of connections between nodes with dissimilar degrees**. In practical terms, nodes with elevated degrees exhibit a propensity to connect with nodes possessing lower degrees, and conversely, forming a distinctive pattern that characterizes the overall disassortative nature of the graph. This measure provides insights into the non-random organization of connections within the network, shedding light on the preferential attachment of nodes based on their degrees.

Community	Size
1	689
2	640
3	518
4	211
5	206
6	189
7	167
8	145
9	144
10	142
11	130
12	120
13	120
14	112
15	111
16	111
17	108
18	106
19	105
20	93
21	91
22	86
23	79
24	77
25	76
26	71
27	51
28	35
29	11

Table 4: Community Sizes

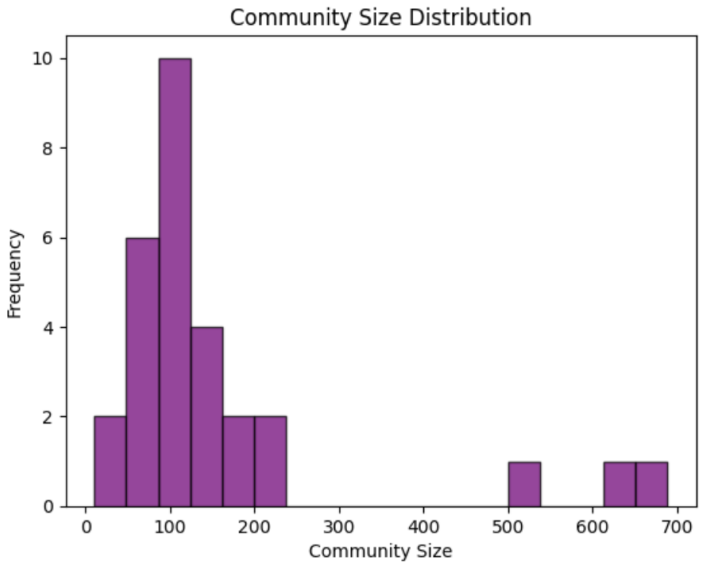

Figure 10: Communities sizes distribution

The **distribution of community sizes** within the network reveals a diverse landscape, ranging from smaller, tightly-knit groups to larger, more expansive communities. The variation in community sizes suggests a complex and multifaceted network structure, with some groups potentially forming around niche interests or tightly connected individuals, while others exhibit a broader, more inclusive membership.

To better analyze the communities we can compute **modularity**, which is a metric that quantifies the extent to which a network can be partitioned into distinct communities. A higher modularity value indicates a stronger community structure within the network. The modularity algorithm seeks to maximize the number of intra-community edges while minimizing the number of inter-community edges.

The modularity value is 0.6, that is indicative of **a well-defined community structure** within the network. This value falls within the typical range for networks with clear modular patterns. The communities identified by the Louvain algorithm contribute significantly to the overall modularity, indicating meaningful and distinct groupings of nodes.

5.4.1 Communities integration

To enhance our capacity for nuanced analysis of comment content within each community, we have enriched each comment entry with its respective community identifier. This augmentation empowers us to delve deeply into the themes and discussions unique to individual communities, offering valuable insights into the distinct topics and prevalent content within each community.

6 Content analysis

6.1 Sentiment analysis

Sentiment analysis applied to social media involves using natural language processing and machine learning techniques to automatically detect, extract, and analyze opinions, attitudes, or emotions expressed in social media content. It aims to understand the overall sentiment — positive, negative, or neutral — towards specific topics, products, events, or individuals within the posts, comments, or other user-generated content on social media platforms.

The goal of this section is to analyze users' opinions on a current and sensitive topic such as femicide. Taking into consideration the dataset related to keywords, the aim is to examine how sentiment has changed over the past year, which has been marked by numerous crime events targeting women.

6.1.1 Posts with the highest engagement

Most-interacted posts may be relevant in sentiment analysis because they can **provide insights into the overall sentiment of a particular topic**. The more interactions a post receives, the more likely it is to reflect the general opinion of the public. Some reasons why most-interacted posts are relevant in sentiment analysis could be identified as:

- they are a representative sample of **public opinion**, since the most-interacted posts are likely to be seen by a wider audience, which means that they are more likely to reflect the general sentiment of the public;
- they can be used to identify **influential voices**, because the posts with the most interactions are often written by individuals that have a large following. This can help us to identify influential voices that are shaping public opinion.

It is important to note that most-interacted posts may not always be representative of the entire population. For example, they may be more likely to reflect the views of people who are more active on social media. However, they can still provide valuable insights into public sentiment.

Below we report the most discussed posts on Giulia Cecchettin murder case and on topic related to feminicide.

Most discussed posts on the murder case

POST TITLE	NUM. COMMENTS	UPVOTE RATIO	SCORE
Uomini, non vi siete rotti le palle di tutti i post contro di voi?	801	0,62	252
Ho visto troppi post sulla Giulia Cecchettin	688	0,82	595
È un covo di incel	651	0,64	263
Filippo Turetta arrestato in Germania, il 22enne ha ucciso Giulia Cecchettin	383	0,94	187
Non c'è nulla di male ad istigare all'odio verso tutti gli uomini	337	0,73	166
Accoltellata dal marito per aver difeso la memoria di Giulia, a voi i commenti	307	0,66	118
"Gli uomini devono fare mea culpa, anche chi non ha mai fatto nulla […]"	287	0,75	153
Colpa del patriarcato o malattia mentale?	209	0,72	101
"Mi è scattato quacosa in testa"Filippo turetta	183	0,76	83
L'Italia e gli omicidi, qualche dat	172	0,95	580

Figure 11: Most discussed posts on Giulia Cecchettin case

Most discussed posts on topics related to Feminicide

POST TITLE	NUM. COMMENTS	UPVOTE RATIO	SCORE
39 FEMMINICIDI, NON 105	1.087	0,74	905
Imbecillità e legittima difesa	1.060	0,76	508
Ho visto troppi post sulla Giulia Cecchettin	688	0,82	587
Discoteche che fanno prezzi diversi e femminismo	683	0,59	47
È un covo di incel	651	0,64	255
Cosa ne pensate di Cecchettin E. e della sua nomina a Persona dell'Anno?	642	0,70	275
Essere uomini è difficile in italia.	618	0,53	8
(serio) Meloni pensaci tu! Lo stupro di Roma, il 22enne: "Io, trascinato e violentato nell'androne di un palazzo"	513	0,63	104
Ho visto Barbie, e non ho capito perchè molti uomini si sono offesi	508	0,85	353
Colpa del patriarcato e dello Stato	443	0,60	83

Figure 12: Most discussed posts on topics related to feminicide

By interpreting the meaning of most-interacted posts, we can gain a first understanding of public sentiment and use that information to grasp more knowledge about the context. In this case, we can notice that most-interacted post are the ones that contain a strongly polarised content, that are purposely made to generate a reaction in other users.

6.1.2 Text Preprocessing

1. **Tokenization**: the process of breaking down a piece of text into smaller units called tokens. In the context of natural language processing (NLP) and computational linguistics, these tokens are usually words or subwords. The primary goal of tokenization is to simplify the complexity of the text, making it more manageable for analysis.

2. **Removal of stop words**: stopwords are common words that are often filtered out during the preprocessing of text data because they are considered to be of little value in terms of conveying meaningful information. These words are extremely frequent in a language but typically do not contribute much to the understanding of the context or the sentiment of a text

3. **Normalization**: in the initial phase, using pattern, URLs and special characters were removed. This step is crucial to eliminate non-textual information and characters that might not significantly contribute to the subsequent analysis.The removal of URLs helps keep the text focused on the actual content, avoiding potential disruptions caused by external links. Cleaning special characters contributes to a cleaner representation of the text, reducing the risk of ambiguity in the results of the analysis.In a subsequent step, all tokens were converted to lowercase to ensure uniformity in their format. This process is known as lowercase normalization and is crucial for achieving consistency in the textual data. By converting all tokens to lowercase, variations in capitalization are eliminated, preventing duplication of words and ensuring that words with different cases are treated as identical.

6.1.3 Bert Model

The model employed for sentiment analysis is **BERT** (Bidirectional Encoder Representations from Transformers), specifically the bert-base-multilingual-uncased version [7]. BERT is a powerful natural language processing (NLP) model that employs a transformer architecture to understand the context of words in a sentence by considering the surrounding words.

Fine-tuning is a process where a pre-trained model, is adapted to a specific task or domain. In this case, the model has been fine-tuned for sentiment analysis on product reviews in six languages, making it adaptable to diverse linguistic contexts such as English, Dutch, German, French, Spanish, and Italian.

The model's task is to predict the sentiment of a product review by assigning it a number of *stars on a scale from 1 to 5*. This provides a quantitative measure of the sentiment expressed in the review, making it convenient for users to quickly gauge the overall sentiment.

The **versatility** of this model extends to its direct applicability for sentiment analysis in any of the specified languages, or it can be further fine-tuned for related sentiment analysis tasks. Specifically, in the case of Italian, the model was trained on a dataset consisting of 72,000 comments. The reported accuracy metrics include a **59% accuracy** for exact predictions and a **95% accuracy (off-by-1)**, indicating the model's ability to closely match human reviewer ratings with a maximum difference of 1 star

6.1.4 Results

At this point, the model that took the preprocessed text as input provides sentiment results summarized by the following Funnel Chart:

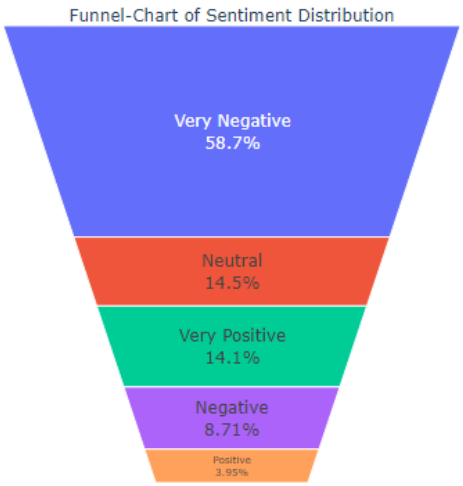

Figure 13: Sentiment analysis

The majority of comments, accounting for **58,7%**, have been categorized as **"Very negative"**, corresponding to a 1-star rating. Following that, **neutral comments**, without a particular positive or negative sentiment, **make up 35% of the total**. The positive feedback, represented by **"Very positive"** comments, comprises **approximately 15%** of the overall feedback. In conclusion, **negative and positive** comments make up **8.757%** and nearly **4%**, respectively, relative to the total comments. So, what immediately stands out is that **more than 65% of the comments express a negative sentiment**.

From this point onward, the "Very negative" comments will be considered as negative comments, and the "Very positive" comments will be treated as positive comments.

The **temporal distribution** of positive and negative comments is analyzed. The DataFrame containing comments related to the keywords is processed to include only comments post-2022, in order to analyze the current year (2023). Subsequently, comments are

grouped by date and sentiment. The graph represents the total number of positive and negative comments over time.

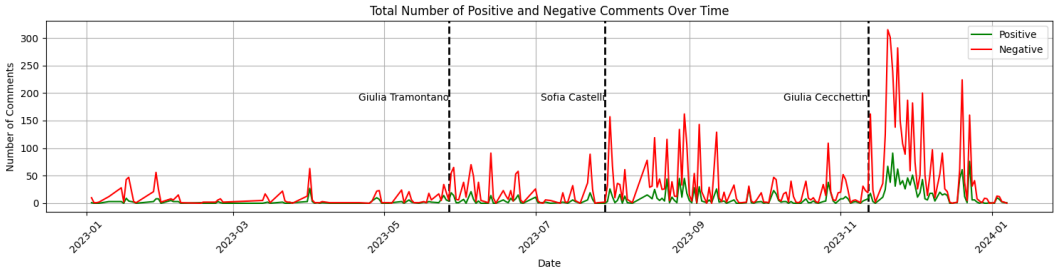

Figure 14: Number of Negative and positive comment over the 2023

As evident from the graph, the occurrence of **a woman's death aligns with specific peaks**, and this unfortunate event garnered television coverage as well. Upon initial observation, it appears that users are responding emphatically, expressing a collective stance against this incident. The visible spikes in activity suggest a heightened level of engagement and reaction within the community, possibly indicative of a broader sentiment or call to action in response to the reported tragedy. Let's consider **the latest tragic event**, the murder of Giulia Cecchettin, as it has elicited a more significant and notable response, being the most recent chronologically. Therefore, a comprehensive analysis is undertaken to **discern the shifts following this particular event**. The upper section attention is drawn to the nuanced distribution of comments, showcasing the percentages of negative, neutral, and positive sentiments for both time frames. Meanwhile, in the lower section, provides a breakdown of the total number of comments, distinguishing between the pre and post-event periods. This multifaceted examination aims to capture the evolving dynamics and sentiment variations within the community in the aftermath of the tragic incident involving Giulia Cecchettin.

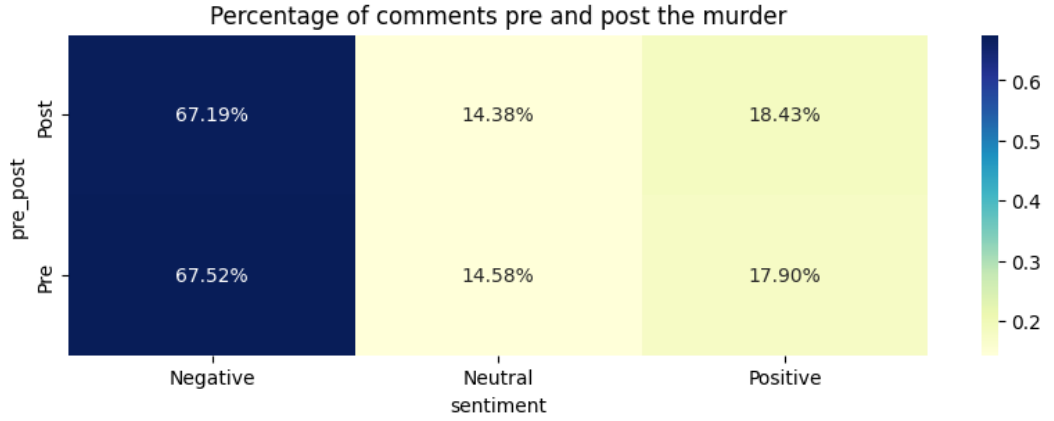

Figure 15: Percentage Pre and Post the murder of Giulia Cecchettin

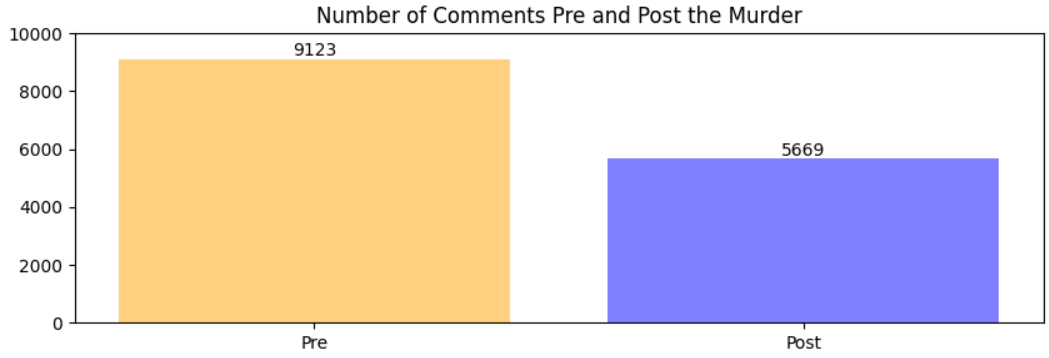

Figure 16: N° of comments Pre and Post the murder of Giulia Cecchettin

In contrast to earlier predictions, the percentage of sentiments has shown a **surprising stability**, with minimal fluctuations. Notably, the overall sentiment landscape has experienced only a marginal shift, specifically manifesting as a slight uptick in positive comments, amounting to approximately 0.5%.

The **significant rise in comments** following Giulia's death raises questions about the reasons behind this sudden surge. What stands out is that the number of comments starting from November 12^{th} makes up almost two-thirds of the total comments received in the preceding 10 months. While it might be tempting to think that there has been a complete shift in people's perspectives, a closer look suggests that it's probably not a total change in mindset. Rather, it seems like a large number of people are now actively taking a stance on this important and sensitive topic. This increased engagement could be due to various factors, such as heightened media coverage or discussions on social media platforms.

At this point, the reference dataset for the homicide has been used, *containing only comments with the words* 'Giulia Cecchettin' and 'Filippo Turetta'. With the utilization of this dataset, we have r**e-implemented the before mentioned procedure** to conduct sentiment analysis. This approach offers a more direct perspective on the specific case at hand. By focusing on comments containing the keywords 'Giulia Cecchettin' and 'Filippo Turetta we aim to gain deeper insights into the sentiments expressed within the context of this homicide investigation.

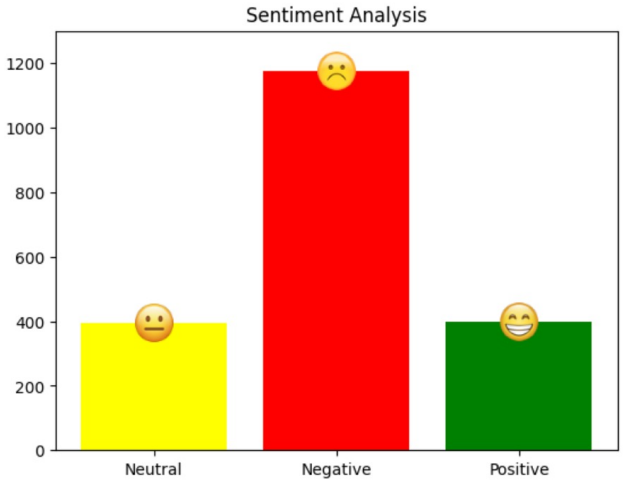

Figure 17: Sentiment analysis on the murder

In this case as well, the **number of negative comments is significantly higher** compared to the other two; however, the percentage has slightly decreased. It is now at 59% of the total. When we talk about understanding a negative comment, it goes beyond just identifying individual words. We aim to extract meaningful patterns and context from the comments. By focusing on trigrams, which are sequences of three words (in this case tokens), we gain a more nuanced understanding of the language used in negative sentiments.

Trigrams provide a way to capture the relationships between adjacent words, allowing us to identify recurring patterns or phrases that might be indicative of negative sentiment. This level of analysis helps in uncovering not only the presence of negative words but also the specific combinations and context in which they appear

	Trigram	Count
0	("'qualche',", "'anno',", "'fa',")	19
1	("'fino',", "'prova',", "'contraria',")	16
2	("'maschio',", "'bianco',", "'etero',")	15
3	("'maggior',", "'parte',", "'persone',")	14
4	("'maggior',", "'parte',", "'casi',")	14
5	("'cosa',", "'vuol',", "'dire',")	14
6	("'maggior',", "'parte',", "'donne',")	14
7	("'qualche',", "'giorno',", "'fa',")	14
8	("'fare',", "'mea',", "'culpa',")	14
9	("'frega',", "'cazzo',", "'nessuno',")	12

Figure 18: Top 10 trigram in Dataset with keywords

The presented trigram table provides insights into **recurring patterns and themes** within the analyzed text, and these patterns appear to align with topics commonly associated with discussions on femicide and gender-related violence. Trigrams referencing time,

such as events from the past or recent days, might indicate discussions about the persistence and evolution of gender-based violence over time. The presence of trigrams discussing demographic elements suggests a potential exploration of specific characteristics related to gender-based violence, such as the identity of perpetrators or victims. Trigrams acknowledging fault or mistakes could suggest a self-reflective tone, potentially relating to societal shortcomings or systemic issues contributing to gender-based violence. Trigrams expressing indifference or disregard might highlight instances where the text conveys a lack of empathy or concern, potentially reflecting societal attitudes that contribute to gender-based violence. The trigram table reflects a variety of themes that can be associated with considerations of personal responsibility, negative perceptions, exploration of possibilities, and discussions surrounding gender-based violence. Some trigrams suggest a critical reflection on the male essence, while others indicate the possibility of a connection between an individual's perception and the commission of a violent act

7 Miscellaneous: topics in communities

We're blending two powerful methods, network analysis and content analysis, for a comprehensive exploration. Our first step involves connecting comments to community IDs after detecting these communities. By gathering comments based on their community, we create **bundles of conversations unique to each group**.

Now, we're diving into the words used most frequently in each community, aiming to compare them. This helps us uncover the distinct language styles and hot topics that shape these groups. Our approach goes beyond just understanding the structure of the network; it delves into the actual discussions, offering a holistic view of how online communities connect and communicate

7.1 Token extraction

We've employed **Natural Language Processing (NLP)** tools from SpaCy to extract meaningful tokens from the comments. This step enhances our ability to discern key words and phrases within each community. By leveraging NLP, we're not only analyzing the network structure but also delving into the semantic richness of the conversations, providing a more nuanced understanding of the distinct linguistic patterns in each online community.

After this first phase, we've utilized the `Counter` module from Python's `collections` library to efficiently tally the occurrences of each token within the comments of each community. This approach allows us to identify and prioritize the most frequently used words, providing valuable insights into the prevalent topics and discussions within individual online communities.

7.2 PCA and Visualization

In the final phase of our analysis, we employed machine learning techniques to gain a comprehensive understanding of the distinctive characteristics of each online community.

Leveraging the `CountVectorizer` from the `scikit-learn` library, we transformed our tokenized data into a **Document-Term Matrix (DTM)**. Subsequently, we applied **Principal Component Analysis (PCA)** for dimensionality reduction, facilitating the visualization of community distribution in a two-dimensional space.

The resulting **scatter plot** provides a visual representation of the communities, with each point annotated by its corresponding community ID. Furthermore, to enrich our analysis, we included labels displaying the top tokens for each community. This dual representation enables us to not only observe the spatial arrangement of communities but also identify key contributing terms within each cluster. The inclusion of top tokens offers a nuanced perspective on the prevalent themes within distinct online groups.

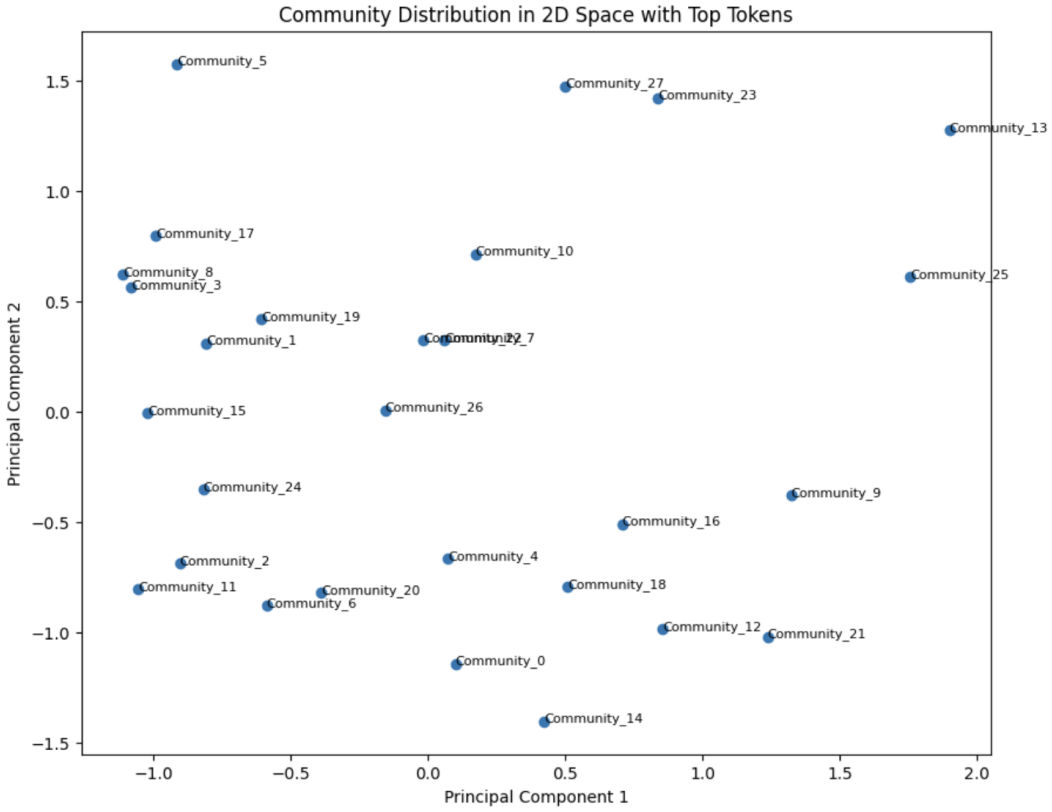

Figure 19: Community Distribution

To complement our visual exploration, we also generated a table presenting the **community numbers alongside their respective top tokens**. This tabular format enhances the interpretability of our findings, providing a detailed overview of the most significant terms characterizing each community. Together, the visual and tabular representations offer a comprehensive snapshot of the intricate interplay between community dynamics and content, highlighting on the diverse topics and discussions prevalent within each online group.

Community Number	Top Tokens
Community_0	potere, volere, dovere, problema, donna, vedere, sapere, pensare, venire, andare
Community_1	donna, uomo, problema, dovere, potere, volere, violenza, vedere, parlare, sapere
Community_2	donna, uomo, volere, dovere, problema, vedere, pensare, potere, femminismo, venire
Community_3	donna, dovere, uomo, violenza, parlare, Italia, problema, potere, venire, volere
Community_4	donna, dovere, volere, uomo, potere, venire, pagare, andare, Woke, sapere
Community_5	donna, violenza, problema, uomo, pensare, genere, parlare, post, mettere, vittima
Community_6	donna, ragazzo, volere, andare, potere, vedere, problema, film, uomo, pensare
Community_7	donna, dovere, potere, violenza, volere, sapere, situazione, perché, pensare, problema
Community_8	donna, problema, uomo, femminicidio, post, leggere, genere, potere, volere, dire
Community_9	potere, dovere, omicidio, andare, situazione, volere, pena, giustizia, sapere, punire
Community_10	dovere, potere, ladro, volere, violenza, problema, difesa, sapere, sparare, parlare
Community_11	donna, uomo, ragazzo, problema, dovere, vedere, volere, venire, amico, uccidere
Community_12	potere, amico, piacere, volere, andare, trovare, cercare, vedere, cosa, pensare
Community_13	madre, padre, violenza, genitore, figlio, schiaffo, prendere, pensare, mano, andare
Community_14	potere, dovere, volere, andare, sapere, ragazzo, pensare, problema, vedere, venire
Community_15	donna, uomo, dovere, potere, violenza, venire, problema, vedere, volere, cazzo
Community_16	problema, potere, andare, venire, dovere, genitore, colpa, bestemmia, situazione, volere
Community_17	donna, femminicidio, problema, post, dovere, uomo, volere, omicidio, potere, violenza
Community_18	volere, ragazzo, sapere, andare, amico, dovere, donna, potere, altro, dire
Community_19	volere, perché, potere, donna, venire, dare, problema, violenza, gente, Italia
Community_20	donna, uomo, andare, ragazzo, volere, potere, pagare, vedere, prezzo, perché
Community_21	dovere, potere, vedere, sapere, pensare, società, venire, dare, andare, credere
Community_22	uomo, donna, volere, RAL, canale, é, pensare, social, medio, sapere
Community_23	padre, potere, cazzo, madre, figlio, violenza, situazione, dovere, uomo, capire
Community_24	donna, uomo, volere, provare, potere, parlare, vedere, femminismo, ragazzo, dovere
Community_25	sociale, servizio, potere, figlio, dovere, sentire, chiamare, genitore, bambino, sapere
Community_26	pubblicità, livello, vedere, marito, volere, il, senso, colpevole, uomo, cambiare
Community_27	medio, giusto, donna, parlare, violenza, sapere, rattristare, arrivare, potere, mano

Table 5: Top 10 Tokens in Each Community

8 Conclusions

Our analysis has yielded valuable insights into the online discourse surrounding the critical issue of feminicide. The application of Network Analysis allowed us to unveil the intricate connections between users, identify influential contributors, and understand the flow of information within the Reddit communities discussing this topic.

Simultaneously, Content Analysis, facilitated by natural language processing (NLP) tools, provided a deeper understanding of the semantic richness of the conversations. By extracting meaningful tokens from comments, we discerned key words and phrases, shedding light on prevalent topics, concerns, and discussions within each online community.

Additionally, temporal variations and community-specific linguistic patterns were revealed, enriching our understanding of the multifaceted nature of the discourse.

It is crucial to acknowledge the limitations of our analysis, such as potential biases in online discussions and the need for context-aware interpretation of the data. Nonetheless, the insights gained from this research contribute to a broader comprehension of how online communities engage with and discuss femicide in Italy.

Moving forward, our approach can be extended to encompass a more extensive dataset or incorporate additional social media platforms for a comprehensive analysis.

References

[1] Numero femminicidi 2023: perché circolano dati così diversi?

[2] Tutti i dati del 2023

[3] Omicidio di Giulia Cecchettin

[4] PRAW: The Python Reddit API Wrapper

[5] Lista dei femminicidi in Italia nel 2023

[6] CODICI STATISTICI DELLE UNITÀ AMMINISTRATIVE TERRITORIALI: COMUNI, CITTÀ METROPOLITANE, PROVINCE E REGIONI

[7] bert-base-multilingual-uncased-sentiment